UNIVERSIDAD REY JUAN CARLOS

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

Temporal and behavioral patterns in the use of Wikipedia

Doctoral Thesis

Antonio José Reinoso Peinado

Ingeniero en Informática

Madrid, 2011

Thesis submitted to the Departamento de Sistemas Telemáticos y Computación in partial fulfillment of the requirements for the degree of
Doctor

Escuela Técnica Superior de Ingeniería de Telecomunicación
Universidad Rey Juan Carlos
Madrid, Spain

DOCTORAL THESIS

Temporal and behavioral patterns in the use of Wikipedia

Author:
Antonio José Reinoso Peinado
Ingeniero en Informática

Director:
Jesús M. González Barahona
Doctor Ingeniero de Telecomunicación

Madrid, Spain, 2011

June , 2011

WE HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER OUR SUPERVISION BY *Antonio José Reinoso Peinado* ENTITLED *Temporal and behavioral patterns in the use of Wikipedia* BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF *Doctor of Philosophy in Computer Science*.

Jesús M. González Barahona, Ph.D.
Thesis Director

The committee named to evaluate the Thesis above indicated, made up of the following doctors

Carlos Delgado Kloos, Ph.D.
Universidad Carlos III de Madrid
President

Baltasar Fernández Manjón, Ph.D.
Universidad Complutense de Madrid
Member

Israel Herraiz, Ph.D.
Universidad Politécnica de Madrid
Member

Eloisa Vargiu, Ph.D.
University of Cagliary
Member

Gregorio Robles Martínez, Ph.D.
Universidad Rey Juan Carlos
Secretary

has decided to grant the qualification of

Móstoles, Madrid (Spain), July , 2011.

The secretary of the committee.

I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be.

William Thomson, 1st Baron Kelvin, often referred as Lord Kelvin
Mathematical, physicist and engineer(1824-1907)

We often discover what will do, by finding out what will not do; and probably he who never made a mistake never made a discovery.

Samuel Smiles
Scottish author and reformer(1812-1904)

Acknowledgements

Writing this thesis is probably one the most difficult ventures I have ever undertaken and, of course, I am feeling as I have achieved a really important goal. However, on the other hand, I positively know that this is just another milestone, though a very important one. I can remember now, when I was sitting next to my university colleagues listening to the invited speaker in our graduation dissertation. The truth is that I cannot remember his name or position but I can perfectly remember some parts of his talk. I remember the way in which he insisted on the fact that the consecution of our degree had not have to be a goal for us at all but just another step and, most important, a new beginning. Because he thought that knowledge did not have any borders or limitations but just the ones we wanted to put over it. Education and learning did not have defined goals or achievements, according to him, but just particular milestones we had to pass through and, even, to put aside in the departure for new ones. He encouraged us to be always willing to undertake new challenges and to take part in new projects developing different approaches or ideas. Moreover, he persuaded us to conduct in the search for the new possibilities brought by the newest improvements and advances. As I have said, I cannot remember nothing at all about him but just these few ideas. That is the power and the awesome capability of knowledge: to survive to its creators and to persist through the time much longer than them.

Perhaps this thesis started just after listening to these words and it has been a really long trip. First I worked in a completely different research field in which I was partially successful because I gave my first steps in the research universe and I could taste the flavor of the acknowledgment given by any type of accepted publication. At a given moment, my research interests started to decrease and fall apart and I practically gave up from my research work. Then I met the Wikipedia phenomenon as a part of interesting discussions with my colleague, and good friend, Felipe Ortega who put me on the trial of Gsyc/Libresoft activities and introduced me to Jesus Barahona in a meeting that is probably the nearest beginning in time of the work your are just about to read.

Gathering of knowledge and its preservation for future generations has been an absolute concern for the humanity from immemorial times and information has always been a very appreciated and powerful good. One of the most special features of knowledge is its uncompleted character and the various ways in which it may be expanded. Moreover, knowledge can receive the contributions of individuals of any particular condition, disregarding their membership to particular cultural or academic groups. In my opinion, information and knowledge make people, not only more prepared, but also more free. Perhaps because of this, manipulation and domination have usually involved some kind of barriers for granting individuals' access to the information sources. Knowledge is an absolutely common good and, thus, it is everyone's responsibility to take care of it and also to contribute to its generation. It is because of this, I want to express, from this thesis, my total gratitude to the individuals who voluntarily and altruistically contribute to Wikipedia because they are contributing to the spread of knowledge and, thus, to a more free society and a more free world. In the same line, I would also like to thank everyone enrolled in any FLOSS project because, again, their contributions are pointing to the dissemination of a branch of knowledge consisting on software development. In general, I think we should be grateful to this kind of initiatives and if I just mention these two ones it is because their particular closeness to the topic addressed in this thesis.

As I have said before, this is (not this has been) a long and hard trip and it would be very difficult to thank all the people involved in it in any way. As I would regret to forget someone, I will do it briefly and in a general way. More than any one, I wish to thank Laura and my parents for supporting me

during all this time. Little after having met Laura I also met Libresoft, enrolled in a master on Libre Software and started my research activities. It has been a hard-to-get-on-well triangle relationship. Laura, more than any one, has had to share all these moments with me and knows about lots of hard days when frustation arose but also about moments plenty of satisfaction after any sort of achievement. In acknowledgment recognitions like this one, it is usual to read parts asking forgiveness to the couple or to the family for all the time invested to accomplish a certain work and, consequently, subtracted to them. I have to ask Laura for her forgiveness because there have been so many nights sleeping alone, so many evenings watching television alone, so many lates because of work. I can really think whether the result of all this work could serve as a little compensation and, in a certain way, I hope so. Once, I told you ”Do not love me although but love me because“ and think that if I have put so much effort in this work it is, in part, because of you, because you deserve the best of me. I can hardly thank to my parents, Luis and Maria, for their unconditional support and for having always encouraged me to undertake this particular project of having a PhD degree. I wish also thank, not my family-in-law, but in fact my family, Beatriz, Rafa, Eugenio and Olalla or Marga (in memoriam) for all their comprehension, love, and their always kind attention. I am really glad of offering this to all my beloved beings because the possibility of sharing this with them is perhaps one of the things that gives more value to this work.

Of course, I want to express my most sincere gratitude to all the people that have assisted me during the development of this thesis. Specially, I wish to thank Jesus Barahona for his guidance and assessment. I also want to extend my gratitude to all the Gsyc/Libresoft team for its support during this project. I have needed a great amount of help with lots of topics and I have always found a really cooperative and collaborative attitude from all the Gsyc/Libresoft members. I am grateful to them because, apart from their support, their treatment to me has always made me feeling part of the group even though I do not have any contractual relationship with it. As well, I wish to thank all my colleagues and directives at Universidad Alfonso X, specially Jesús Sánchez, Tomas García and Rafael Magro, for all their good advices and for their sane interest and encouragement to finish this work. Let me express a particular acknowledgment to Israel Herraiz and Felipe Ortega. Since Isra arrived at UAX and became my office mate, he has offered to me a really valuable advice and orientation. In addition, he has always been willing to collaborate in the resolution of several technical issues as well as to actively participate in the research activities developed as a part of this thesis. But more than this, I have learn from him how a good research has to be conducted, the passion for being always ready to face new challenges, the philosophy of freedom and collaborative efforts that we have to imbue in our work and the spirit for excellency and a high-quality job. Felipe introduced me to Libresoft and has contributed to this project since its first inception. Apart from technical issues, I was lucky to be near him at my my first conference attendance in Oporto where I could learn how researchers have to manage when divulgating and defending their principles and ideas. I also received a very special lesson from him at his PhD thesis dissertation and I hope that mine be at the same brilliant level. I want also to express my gratitude to Teresa Bravo for all her support and encouragement and, of course, her patience and sympathy expressed in all the email we have interchanged since a web page fortunately put us in contact again. In the same way, I want to thank Rocío Muñoz for her collaborative efforts and for our ”paper-therapy“ that has been really important and valuable for me. Lots of good friends have also been supporting me during this period, so thank you all and, please, let me express a special gratitude to Gabriel Pastor and to Ricardo Sosa.

Abstract

Reinoso Peinado, Antonio José. M.Sc. in Computers Science, Departamento de Sistemas Telemáticos y Computación, Universidad Rey Juan Carlos, Móstoles, Madrid, 2010. *Temporal and Behavioral patterns in the use of Wikipedia*.

Wikipedia stands as the most important wiki-based platform and continues providing the overall society with a vast set of contents and media resources related to all the branches of knowledge. Undoubtedly, Wikipedia constitutes one of the most remarkable facts in the evolution of encyclopedias and, also, a complete revolution in the area of knowledge management. Perhaps, its most innovative aspect is the underlying approach that promotes the collaboration and cooperation of users in the building of contents in a voluntary and altruistic manner.

The growth of Wikipedia has never stopped since its beginning as well as its popularity. In fact, the number of visits to its different editions has placed its web site within the top-six most visited pages all over the Internet. Such kind of success has spread the use of Wikipedia beyond typical academic environments and has made it become a complete mass phenomenon.

Due to this significant relevance, Wikipedia has revealed as a topic of increasing interest for the research community. However, most of the developed research is concerned with the quality and reliability of the offered contents. This previous research focuses on subjects such as reputation and trust, or addresses topics related to the evolution of Wikipedia and its growth tendencies. By contrast, this thesis is aimed to provide and empirical study and an in-depth analysis about the manner in which the different editions Wikipedia are being used by their corresponding communities of users. In this way, our main objective is the finding of temporal and behavioral patterns describing the different kinds of contents and interactions requested by Wikipedia users. Users' requests are expressed in the form of URLs submitted to Wikipedia as a part of the traffic directed to its supporting servers. The analysis presented here, basically, consists in the characterization of this traffic and has been developed by parsing and filtering the information elements extracted from the URLs contained in it. As we, necessarily, have had to work with a sample of all the requests to Wikipedia due to their incommensurable volume, we have, first, validated our results comparing them with trusted sources.

After having analyzed the traffic to Wikipedia during a whole year, this study presents a complete characterization of the different types of requests that make part of it. Furthermore, we have found several patterns related to the temporal distributions of such kind of requests as well as to the actions and contents involved in them. The influence of the most frequently searched topics and other contents positively considered by the community, as the featured articles, in the attention that articles get is also considered as a matter of interest. Finally, we have also analyzed the different categories of articles that attract more visits and search operations in the considered editions of Wikipedia.

Most of the objectives accomplished here are based on the results provided by the application developed ad-hoc to feed this study. The software engineering of this tool has been undertaken under the WikiSquilter project. We expect that this application can serve as a useful tool to characterize the traffic directed to wiki-based sites, particularly to any project supported by the Wikimedia Foundation.

Up to this work, no other analysis had been undertaken to study the use of Wikepedia in such a wide and thoroughgoing way. We hope that our efforts and results can serve as a significant contribution in the examination of the dynamics of use when interacting with knowledge management platforms like Wikipedia.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

" There is only one good, knowledge, and one evil, ignorance" *Socrates*

1.1 Introduction

The Wikipedia has successfully grown into a massive collaboration tool based on the wiki paradigm as the new approach to produce and access intellectual works. Its impressive figures about both articles and users have propitiated that the Wikipedia can be considered one of the largest compilations of knowledge that have ever existed. The number of articles in its different editions has never stopped growing [1] as well as its popularity, that situates the Wikipedia web page among the six most visited sites all over the Internet [2].

Undoubtedly, the Wikipedia initiative has evolved to a solid and stable project used as a valuable reference tool by million users. Its impact and degree of penetration in the so-called information society can be measured in terms of the vast number of visits that it receives every day. According to the statistics provided in dedicated web pages [3] by the institution that funds the project, the Wikimedia Foundation, the whole set of editions of Wikipedia were receiving more than 345 million visits per day by the end of May 2010.

With such an impressive portfolio, it is not rare at all that the scientific community decided to put its examining eye on subjects related to Wikipedia, mainly to determine whether the information it offers has and adequate level of quality and is reliable enough to be trusted. In this way, the academic works covering topics involving Wikipedia rapidly increased [4] and the Encyclopedia became a usual topic for discussion in several forums.

The relevance of Wikipedia can be considered from different perspectives, even from the adherence and the criticism that it has aroused. Its model for content generation may be thought as the result of the application of the paradigm based on the collaboration of individuals for the production of knowledge. This new approach has supposed a real collapse of the precedent centralized conception of how to create and disseminate knowledge in favor of a completely distributed, or at least de-centralized, model that pursues that anyone can get involved in the genesis of any kind of

[1] `http://stats.wikimedia.org` (Retrieved on 22 June 2010)

[2] `http://www.alexa.com/siteinfo/wikipedia.org` (Retrieved on 22 June 2010)

[3] `http://stats.wikimedia.org/EN/Sitemap.htm` (Retrieved on 22 June 2010)

[4] `http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_academic_studies` (Retrieved on 22 June 2010)

wisdom. Because of the consideration of Wikipedia as a successful implementation of this new model, the Encyclopedia and its supporting philosophy deserve the attention of researchers. The presence and relevance of Wikipedia in the current society and its contribution to the tie between knowledge and information technologies made of it a unique entity whose main features demand a deeply examination.

In this thesis, we are examining Wikipedia from a different, and not so explored point of view, because we are focusing on the use given to the Encyclopedia by its users. Regarding the massive dimension of this project and its absolute relevance in the propagation, transmission and generation of knowledge, we considered that the examination of users' behavior and attention to such kind of initiatives deserved our best efforts. Moreover, the approach used to analyze this subject is also quite inexperienced as we are basing our analysis in the characterization of the traffic containing the requests that users sent to Wikipedia. Of course, there have been other approaches to analyze some of the questions addressed here, but up to our knowledge, none of them has been used to cover so many topics as considered in this work.

Following sections present the main goals and objectives that have motivated this thesis and the main features of the Wikipedia and the *Wiki* paradigm, including its implications on knowledge management. Moreover, I also introduce the hardware architectures and software mechanisms deployed by the Wikimedia Foundation to support all its projects together with the most important topics about the interaction with Wikipedia. Finally, the organization and structure of the rest of this thesis is also presented.

1.2 Motivation

Most of the previous research on Wikipedia has focused on predictions about its evolution, different models for sustainable growth and, above all, on mechanisms for quality control to assess the reliability of its contents. Surprisingly, very few studies have been devoted to analyze the use of Wikipedia and the type of interactions requested by its users, even regarding the most easy-to-obtain measures. As an example, at the moment of writing this thesis there is no way of getting a list with the most visited articles for a given edition nor the topics most often submitted to the its search engine. There has been some initiatives of this kind in the past but, presently, they are out of service or they are not being conveniently updated and most of them have not been undertaken from an academic perspective. The most reliable information currently stems from the statistics offered by the Wikimedia Foundation itself which includes valuable data such as the number of articles, the number of registered users, the pageviews and also information about the contributions made by the Wikipedia editors. However, it does not include information about other kind of actions, such as edit requests or previews, that users solicit to be performed on articles.

As a result, I decided to carry out the study presented here in an attempt to determine the main characteristics of the use of Wikipedia by means of the analysis and characterization of its traffic. The challenge of finding both behavioral and temporal patterns, which could be useful to provide a better understanding of the use and the different kind of interactions between Wikipedia and its users, could hardly be more attractive and interesting to undertake.

As our analysis is completely based on the characterization of traffic, it will provide both well known metrics and new results. This interesting particularity will allow us to examine the trustworthiness of such kind of analysis by establishing different comparisons between our results and the ones derived from previous studies involving analyses of database dumps or statistics obtained from several types of surveys. In addition, the results of this type of analysis can lead to a great variety

of important benefits that include the availability of a detailed characterization of the Wikipedia traffic and the possibility of improvements to be performed on the supporting server systems to satisfy particular situations of overload and machine-stress. Fortunately, the necessary data related to the users' activity on Wikipedia have been easy to obtain thanks to the courtesy of the Wikimedia Foundation.

The Wikipedia philosophy completely adheres to the so-called *open movement* although this movement was conceived in a radically different environment related to the software production. This similar attitude towards the openness principles allows that everyone can get involved in the process of building knowledge and that this generated knowledge remains available to the whole community. Moreover, the Wikimedia Foundation offers dump files corresponding to the database records holding all the contributed contents and, even more, the access to the log information related to some internal operations is also granted for researchers and, in general, anyone interested. In this way, it is possible to obtain log files containing the requests submitted to the different editions of Wikipedia by their corresponding users. People's fundamental rights to privacy and confidentiality are not infringed or violated in any way. This is guaranteed because all the data susceptible of being used to perform any sort of identification, such as users' login names or Internet addresses, are completely removed in the Wikimedia systems prior to the sending of whatever information related to the requests made by the users.

This availability of information about the requests submitted by users to Wikipedia is unparalleled from a research point of view. All of the data are being obtained from the systems involved in the delivery of Wikipedia contents to the users asking for them. The Wikimedia Foundation maintains other projects besides Wikipedia and some parts of its system architecture are shared among all of them. Because of this, requests to Wikipedia are provided within the overall traffic to all the Wikimedia Foundation projects and resources. Considering the current scalability of the system, it is impossible to handle such traffic in a centralized system. Thus, we receive a sample consisting of the 1% of all these requests. Although it may seem that it is a not too large sample, we are receiving about 38 million log lines, corresponding to the same number of requests, a day. In general terms, this means that a whole year involves about 15,000 million log lines. This is an absolute challenge in terms of the necessary infrastructure to store relevant information but, specially, in terms of their processing.

Another special characteristic of this study is the reproducibility of the analysis undertaken as a part of it. This analysis has been performed on a feed consisting in log information from the Wikimedia Foundation systems that remains available in our systems properly stored. In addition, the most important tool used in this work has been the tailored application designed and developed to accomplish the fundamental tasks of parsing and filtering the data sources. This tool is libre software and it is offered under the suitable licenses to the research community. In this way, everyone interested in reproducing our empirical developments can get all the data elements as well as the adequate tools to do so.

On the other hand, we consider that there are important benefit derived of the study of the requests submitted to Wikipedia by its users. As an example, it will be possible to obtain a characterization of the overall traffic directed to the whole Wikipedia project as well as to particular editions. Such kind of characterization would allow, in addition, to determine the composition of the traffic in terms of the different kinds of requests that make part of it and their corresponding ratios. Moreover, it is possible to compare the measurements obtained from several editions of Wikipedia in order to assess if there are important differences among the way of conducting exhibited by users from different communities.

Although we could establish communities of users, at an early stage, according to their

users' native language, this would be only true for some particular languages due to their special characteristics. Nevertheless, the present globalized scenario allows that a great number of visitors of a given Wikipedia language edition correspond to countries not having this language as native. Despite of the fact that we are not yet able to geolocate the origin of the users' requests, it would be, undoubtedly, interesting to compare measures obtained from editions of Wikipedia corresponding to more generalized languages such as English or Spanish with the ones related to more restrictive communities of users such as the Polish or the Japanese ones.

URLs acting as users' requests are exceptionally rich in information elements, so they allow to study important aspects of the use of Wikipedia such as the kind of contents that attract more attention or the most searched topics. These metrics permit to perform comparisons related to the most popular subjects in the different editions of Wikipedia. Moreover, the distribution of the number of visits over the different types of articles deserves, in our opinion, an special interest because it may help to establish relationships between the different kinds of contents and the amount of traffic that they attract. Considering interactions consisting in requests for actions, their analysis can serve to model the way in which users are contributing to Wikipedia and some other aspects of their behavior when they make use of the services offered by the Encyclopedia. Besides this, the study of the submitted actions can lead to correlations between the number of visits to certain articles and the number of requests involving other types of actions submitted over them. These correlations can even be used as an indicator of the degree of participation and contribution exhibited by the community of users corresponding to a given edition of Wikipedia.

The influence of contents positively considered by the community, such as the featured articles, on the number of visits, and thus on the generated traffic, is also addressed. The Wikipedia community distinguishes the best articles giving them the special mention of featured articles. This work measures the impact of the consideration of an article as featured in its subsequent number of visits and editions and, furthermore, this kind of influence is analyzed for different editions of Wikipedia.

As far as we know, this thesis constitutes the most exhaustive examination performed on data reflecting the interaction and the information exchanges between the Wikipedia platform and its users. The thoroughness of this analysis can be regarded in terms of the coverage period (a whole year), the Wikipedia editions that have been considered, which are the largest in both traffic and number of articles, and, also, the set of information elements taken as object of study.

1.3 Research objectives

The main goal of this thesis is the finding of temporal and behavioral patterns related to the use of Wikipedia. As a result, this work aims to describe different aspects related to the way in which users are interacting with Wikipedia and making use of it. As the analysis of the traffic to Wikipedia is the basis or our study, obtaining a complete characterization of it is one of our most important concerns. In this we are specially interested in determining the different types of actions users submit as well as their corresponding frequencies. The temporal distributions of these requests, even regarding different units of time measurement, and their differences when considering several language editions constitute another important subject of interest for this work.

In the following, we will describe in detail the main objectives leading this work and the research questions in which they have materialized.

First, we will analyze the traffic to Wikipedia from a macroscopic perspective in the aim of classifying and quantifying, i.e. characterizing, the requests that make part of it. Our main objective related to traffic is twofold: First, we want to validate the results obtained from an analysis whose main

feed solely consists in requests sampled from the log information registered by the corresponding servers. On the other hand, we are aimed to study the composition of the traffic and the way in which it evolves. There are, of course, several aspects that may have some influence in the traffic directed to a specific edition of Wikipedia. These factors range from the degree of penetration of the Internet in a given society to the number of speakers of a certain language. In our case, we assess the influence of editions' size in the traffic they attract because of the immediate availability of the two informations. The following questions present our main aims concerning this topic:

1. **Can we trust the results obtained from the analysis of requests sampled from the Wikimedia Foundation Squid servers?** As the analysis performed as a part of this thesis constitutes a considerably innovative approach to the study of Wikipedia, a thorough validation of its comparable results is absolutely required to ensure the reliability of the rest of them. The verification we realize entails the validation of both the sample of data that our feed consists of and the process consisting in the parsing and filtering of the sampled requests that our application performs. Validation is possible because of the availability of trusted information sources emanating from the Wikimedia Foundation itself as well as from other previous analyses. In this way, to properly solve this question we will compare some of our results with reliable information always taking into account the sampling factor used to build our sample.

2. **Can we obtain a characterization and quantification of the types of requests composing the traffic to the different editions of Wikipedia?** To deal with this question, we will analyze the traffic directed to each considered edition of Wikipedia using regular expressions. In this way, we will be in position of obtaining a characterization of the overall traffic and we will be able to determine the number of requests consisting in visits to articles or in edits on them. Moreover, we will also quantify the number of requests asking for any kind of action and, also, for particular ones such as search operations. Finally, requests specifying css skins and other kind of visualization choices will be also computed.

3. **Is there a proportional relationship between the size of the Wikipedia editions and the amount of traffic they attract?** To answer this question we will compare the size, in number of articles, of the largest editions of Wikipedia with the amount of traffic they attract. Furthermore, we will compare the evolution of the measures, size and traffic, during the whole year 2009.

   Next, we are going to basis our examination in the traffic filtered by our application. Requests composing this traffic are referred to specific information elements (fundamentally certain namespaces) and actions in whose quantification and temporal distribution we are interested. Our analysis, here, focuses on temporal and behavioral aspects obtained from the traffic that can be helpful in the description of the interaction between Wikipedia and its users. The proposed questions are:

4. **Can we identify patterns temporarily repeated which involve specific types of requests to Wikipedia?** In order to provide a suitable answer to this question, we analyze the requests submitted to Wikipedia during different time units. This allows to obtain different perspectives corresponding to particular periods of examination. To achieve even more accuracy, we analyze each type of requests separately in order to avoid side-effects due to the influence of scale considerations. For the same reason, requests corresponding to different editions of Wikipedia are considered apart.

5. **Are visits to the Wikipedia contents related with edits and the other type of actions in any way?** To deal with this question I will put in relation the figures about the different types of requests issued in the same periods of time looking for positive correlations among them. Relationships between different types of requests may suggests specific ways of conducting from users when they interact with the Encyclopedia. Moreover, this kind of comparisons can help to map the contributions submitted to the different editions among their respective users and can also lead to establish the degree of participation of specific communities.

   Finally, we focus on the traffic directed to particular contents. Wikipedia establishes several mechanisms to promote and present contents considered of high quality and we tackle the evaluation of their effectiveness in terms of amount of traffic attracted. In addition, we are interested in the topics corresponding to the articles that attract the highest numbers of visits and in the comparison of these topics among the different editions of Wikipedia. Moreover, Wikipedia also offers a built-in search engine and we are interested in studying the kind of topics submitted to these engines by users. The following questions summarize this two research initiatives:

6. **Does the promotion of articles to the featured status affect to the number of visits that they receive?**
   We consider this question from a twofold perspective. To begin with, we analyze the impact that featured articles presented in the main pages of several Wikipedia editions as quality content attract in terms of number of visits. Furthermore, we also analyze the number of visits attracted by articles involved in promotion process as a reflect of the different dynamics exhibited by particular communities of users when looking for a consensus about the consideration of articles as featured. A great amount of visits to featured articles can be interpreted as the incipient interest of a given community for high quality articles and, probably, a use of the Encyclopedia not directly related with the search for specific information. In the case of featured articles presented in the main page, users have to visit this page previous to these featured articles. That means that the visits to this kind of articles are not the result of search operations, whatever they have been issued from external web searching engines or are coming from the own Wikipedia's search system. On the contrary, the origin of those visits is the corresponding main page where users' attention has been derived to the featured content. Of course, it will be of great interest to determine whether the promotion of articles to the featured status has the same repercussions and effects in different editions of Wikipedia.

7. **What are the topics to which correspond the articles that receive the highest numbers of visits and edits?**
   This question has a qualitative nature and it is aimed to determine what specific kind of articles maintained by each Wikipedia edition attract more attention from its community of users. In the same way, we will also determine the types of articles that receive more contributions in the form of edit operations. Both results can serve as good indicators of the type of use that the different communities of users made of Wikipedia. To properly solve this question, we have used a content characterization based on the categories presented in a previous work ([Spo07b]).

8. **Do search requests involving particular subjects have an impact on visits to articles related to same topics ?**
   This question has, again, a qualitative nature and it is, firstly, aimed to determine and categorize the subjects most repeatedly searched using the Wikipedia built-in search engine. We will

apply the same categorization used to determine the most visited and edited articles. In order to determine the influence of search operations in visits to articles, we will correlate both types of requests.

1.4 The Wikipedia project

Although Wikipedia is currently a consistent and enough well known initiative, we consider appropriate to introduce here some of its aspects and features, specially those more closely related to the work presented in this thesis. Thus, the main objective of this section is to to provide the readers with an adequate context and to properly present the Wikipedia scenario. Furthermore, we will go behind the stage and we will present the underlying supporting systems that are implementing Wikipedia and the rest of the other Wikimedia Foundation projects.

In this way, after a brief general presentation, following sections will focus on describing the main terms of the interaction between Wikipedia and its users as well as on the software and hardware infrastructure deployed by the Wikimedia Foundation to support all of its project and, of course, Wikipedia. Therefore, the presentation of the way in which Wikipedia organizes the information and the possibilities of interaction it offers will permit to obtain a better comprehension of the different types of requests that users may issue asking for particular contents or for certain actions or services. To provide a more detailed idea of these interaction elements, we will present them associated to the corresponding items of the web interface. In this way, the important differences between several concepts will be conveniently highlighted.

On the other hand, having a precise picture of the different kind of systems making part of the Wikipedia supporting architecture will serve to figure out how the different contents are stored and delivered to users. In addition, it will be possible to identify the systems specially arranged to improve or ameliorate the overall functionality in any way. Finally, and in what our research is concerned, this part acts as a valuable preamble to the kind of data that will be part of our information source and main feed.

1.4.1 Introducing Wikipedia

The Wikipedia phenomenon is built upon the *Wiki* paradigm, firstly developed in 1994 by Ward Cunningham in his *WikiWikiWeb* [5] site. The main principles of this new approach can be summarized in a few points:

- Every user who is able to visit a *Wiki* site is able to contribute to it just using his, or her, web browser.
- Articles having related contents can be associated using inter-article special links that can be considered as equivalent to the HTML hyperlinks commonly found inside web pages.
- A *Wiki* site aims to involve visitors in its creation process so they can contribute and collaborate in the production of knowledge.

At the moment of writing this thesis (May, 2011), the figures about Wikipedia are really impressive and stunning. In fact, it has more than 270 editions corresponding each to a different language which

[5] `http://en.wikipedia.org/wiki/Wiki` (Retrieved on 22 July 2010)

group, in total, more than 15 million articles. Finally, Wikipedia has attracted the attention of more than 15 million users who have completed the registration process in, at least, one of its editions.

This situation results particularly relevant due to the fact that all the Wikipedia contents are contributed in a completely voluntary manner by its community of users. These users are individuals, even not registered in the platform, which do not necessarily belong to any academic or scholar sphere and who are not usually qualified experts in the area they are contributing. This fact, which can be regarded as the most characteristic feature of Wikipedia, is, at the same time, its most controversial topic and it is often wielded by its detractors as the most important and serious drawback because it can compromise the quality and reliability of Wikipedia contents.

According to the own Wikipedia history [6], it had a former predecessor known as the *Nupedia* project [7] which consisted in a web encyclopedia holding free licensed articles from a reputed group of experts. At this early stage, Wikipedia was intended as an incubator of ideas to be developed by the Nupedia experts in the corresponding articles. Surprisingly, the growing of Wikipedia rapidly caught up the pace of the Nupedia and, actually, overtook it.

The first edition of Wikipedia, corresponding to its English version, came to the light in January, 2001. Its diversification on several language editions rapidly contributed to its growing boom. In fact, and according to the information offered by the own Wikipedia pages [8], new Wikipedia articles have been growing at an exponential rate until 2006.

1.4.2 The model of interaction of Wikipedia

As this thesis is devoted to collect and analyze information related to the use of Wikipedia, this section briefly describes the main features of its articles and presents the different choices and options available for users when they are visiting the web pages of the Encyclopedia.

A Wikipedia article is an encyclopedic entry properly entitled that provides information about a particular topic, person, place, date, event, etc. Articles can consist of several sections and can contain images, sounds, videos, and, remarkably, can link to both internal articles and external web pages. Wikipedia editors are encouraged to provide abundant references and solid bibliography in order that readers can contrast the information or widen it in any aspect. Articles are built upon the basis of the *wiki text* or *wiki markup* which consists on a markup language to write and format wiki pages. The wiki markup is a lightweight language with a very simple syntax that allows to produce documents with reduced sizes that make them specially suitable to be massively stored by database servers or other storage solutions. By contrast, wiki-text-based documents usually have to be rendered out by a mediawiki software to generate the corresponding full-featured HTML code to be displayed in web browsers.

The MediaWiki software, responsible of the contents management and in charge of HTML rendering, presents the Wikipedia articles as web pages consisting of two well defined frames. As shown in Figure 1.1, the encyclopedic contents of the article, including image thumbnails, formulae, etc., are placed in the main centered frame, whereas the different options, languages and toolboxes can be found on a bar on the left. Above the content frame there are two tabs on the left side corresponding to the most important namespaces of the article, the main and the discussion ones. There are also other tabs, on the right side, for the most common actions to perform on an article, its edition and

[6] `http://en.wikipedia.org/wiki/History\_of\_Wikipedia` (Retrieved on 13 September 2010)

[7] `http://en.wikipedia.org/wiki/Nupedia` (Retrieved on 13 September 2010)

[8] `http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth` (Retrieved on 22 July 2010)

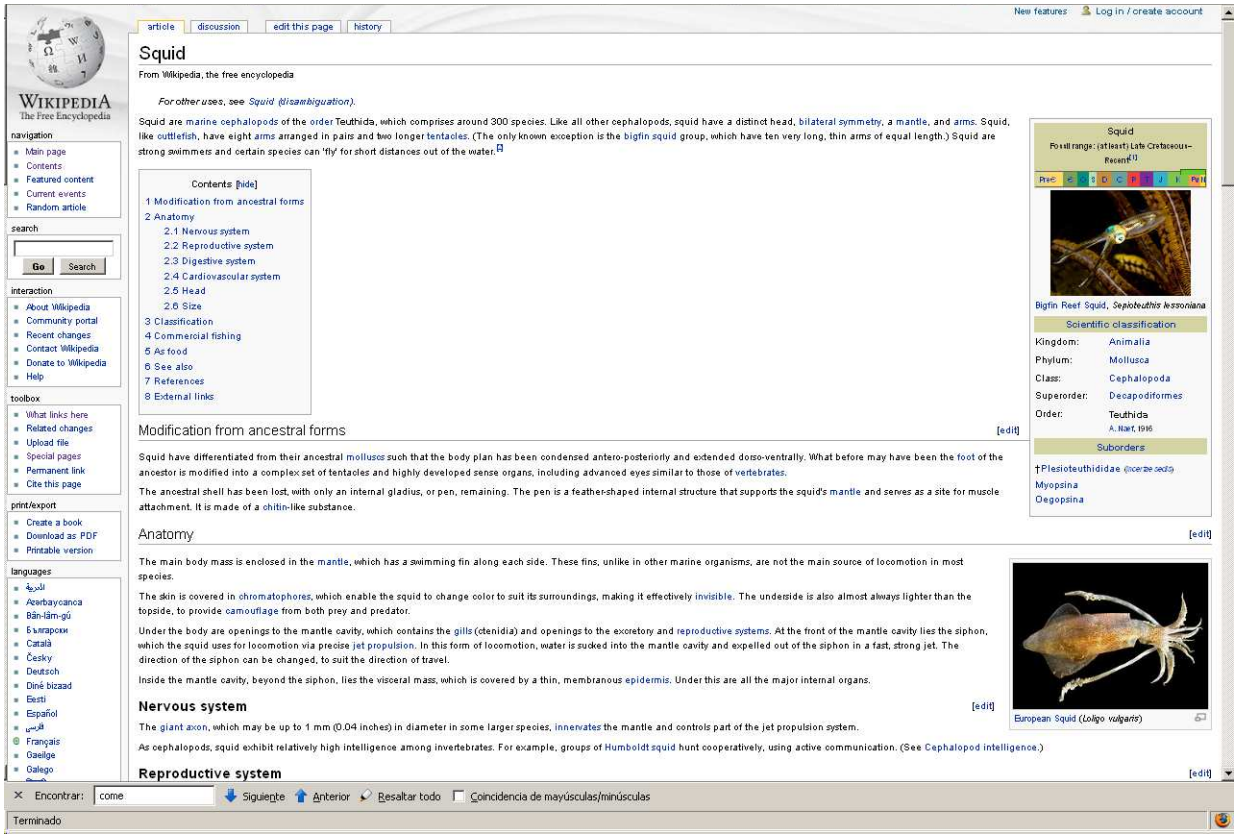


Figure 1.1: *Squid* article in the English edition of Wikipedia.

the viewing of its history log. Next to these tags there is the search input box and its corresponding button. Finally, on the top-right corner of the page there are links for logging-in and creating new accounts.

Most of Wikipedia articles are in the *main namespace* which is the default namespace in which they are created. Visits to these Wikipedia articles are usually for reading (commonly referred as visits or pageviews) and do not specify any special action to be performed. Of course, there are very different ways for users to get to the articles which range from searches in common specialized engines to URLs directly typed in the address bar which every browser include. In any case, all these URLs present the same pattern: the Wikipedia sub-domain according to the referred language edition (such as *http://en.wikipedia.org* for the English version of Wikipedia) followed by the clause *wiki* and the name of the article. As an example, the page shown in the Figure 1.1 (retrieved on 15 September 2010) would correspond to the following URL:

`http://en.wikipedia.org/wiki/Squid`

The *discussion page* or *talk page* of an article contains users' comments and suggestions devoted to improve the quality of that particular article. This page is reached through the corresponding tab

previously mentioned and can be edited in an independent way that its associated article. All the discussion pages corresponding to the Wikipedia articles are grouped under the *Talk namespace* and their URLs add the prefix *Talk* followed by a colon in front of the name of the article. So, the talk page corresponding to the previous Squid article would be pointed by this URL:

`http:/en.wikipedia.org/wiki/Talk:Squid`

In general, namespaces are a method of organizing and categorizing articles according to their nature or to the topics they address. Wikipedia maintains several namespaces for this purpose whose names are added as prefixes in front of the names of the articles in the same way as the Talk namespace explained before. Articles as they are commonly requested are said to be in the main namespace and have no prefix. Most of the information related to the topic developed in common encyclopedic articles is distributed between the main and the talk namespaces. Therefore, other namespaces are used to establish classifications among already available articles, to provide information about static contents such files or images or, even, to provide registered users with a personal page for notifications or messaging. The *Special* namespace [9] deserves a special attention by itself because it corresponds to those pages that do not have any associated wiki text because they are generated as a result of an action requested by the user and involving a given set of arguments. There are several special pages including pages to select an article at random, to obtain the articles referencing a given one, and much more. All of them add the prefix *Special* (followed by a colon) as a part of their corresponding URLs and the name of the requested action. As en example, the following URL would show all the articles referencing the one about squids:

`http://en.wikipedia.org/wiki/Special:WhatLinksHere/Squid`

Given this considerably high number of namespaces, our study will focus on just a few, but the most important, of them:

- The *Main* namespace as it contains most of the contents of the articles.
- The *Talk* namespace because it holds contributions aiming to improve the quality of the article.
- The *User* namespace which corresponds to all the pages allocated for the registered users, and
- The *Special* namespace because search operations correspond to it.

Chapter 3 will present in detail the different issues related to the processing of the URLs belonging to each of the considered namespace.

Talking about actions, users can ask to edit a given article using the corresponding tag. This makes the system to obtain the corresponding wiki text and send it to the user's browse inside a basic editor. The URL submitted to the server, in the case we continue to consider the same article as before, would be:

`http://en.wikipedia.org/w/index.php?title=Squid&action=edit`

Once the proper corrections or contributions have been done, users can preview their changes. In fact, they are encouraged to do so by using the corresponding button. There is also a button for checking the main changes introduced and, of course, another one for saving them to the database. There is a very important issue here, all these three buttons generate URLs similar to the following one:

[9] `http://en.wikipedia.org/wiki/Help:Special_page`

http://en.wikipedia.org/w/index.php?title=Squid&action=submit

In these requests, the user's choice (preview, changes or save) is communicated to the server through the corresponding argument that is sent using the HTTP POST method. This prevent the submitted URL from including any field specifying the particular action. As identifying URLs that cause articles' contents to be saved into the database is crucial according to our aims because these URLs trace users' contributions, the Squid log lines we are receiving include a specific field to indicate when the URL entails a save operation.

Moreover, users may want to access the historical log that reflects all the changes made over an article and presents them chronologically ordered. There is a tag, as previously mentioned, for this purpose and its use generates URLs like the following one:

http://en.wikipedia.org/w/index.php?title=Pope_Benedict_XVI&action=history

Search operations have to be carefully considered because their URLs belong to the *Special* namespace. As a result, they make servers to dynamically compose web pages containing the results provided by the search engine after being queried about a particular topic. The following URL would produce a list with the titles of the articles containing information about the use of Wikipedia:

http://en.wikipedia.org/w/index.php?title=Special\%3ASearch&search=Wipedia+use.

In order to adequately process these URLs, both the namespace and the argument specifying the topic search have to be considered. Different strategies to parse and obtain for these requests of the rest previously described will be largely addressed in Chapter 3.

1.4.3 The Wikimedia Foundation hardware and software server architecture.

Nowadays, all the wiki-based projects supported by the Wikimedia Foundation are running from a set of servers distributed through several facilities based in Amsterdam (The Netherlands) and in Tampa (USA). The structural organization of all these servers has been evolving to meet the requirements in scalability arising from the continuous increase in traffic and content contributions. The last found picture of the overall Wikimedia Foundation architecture corresponds to April 2009 and it is presented in Figure 1.2. Every server in this architecture has a well-defined role and provides a particular service to the rest of the systems.

Technical documentation about configuration internals of the Wikimedia Foundation servers [10] refers to the use of LAMP (Linux, Apache, MySql, Php) environments as the basic software platforms for all the systems. Different services and functionalities are provided by specific software as the ones listed below.

- **Linux**
  Fedora and Ubuntu are the Linux [11]distributions used as operating systems in all the Wikimedia servers with the exception of the image storage systems that run Solaris.

[10] http://www.nedworks.org/\~mark/presentations/san/Wikimedia\%20architecture.pdf (Retrieved on 9 September 2010)

[11] http://linux.org

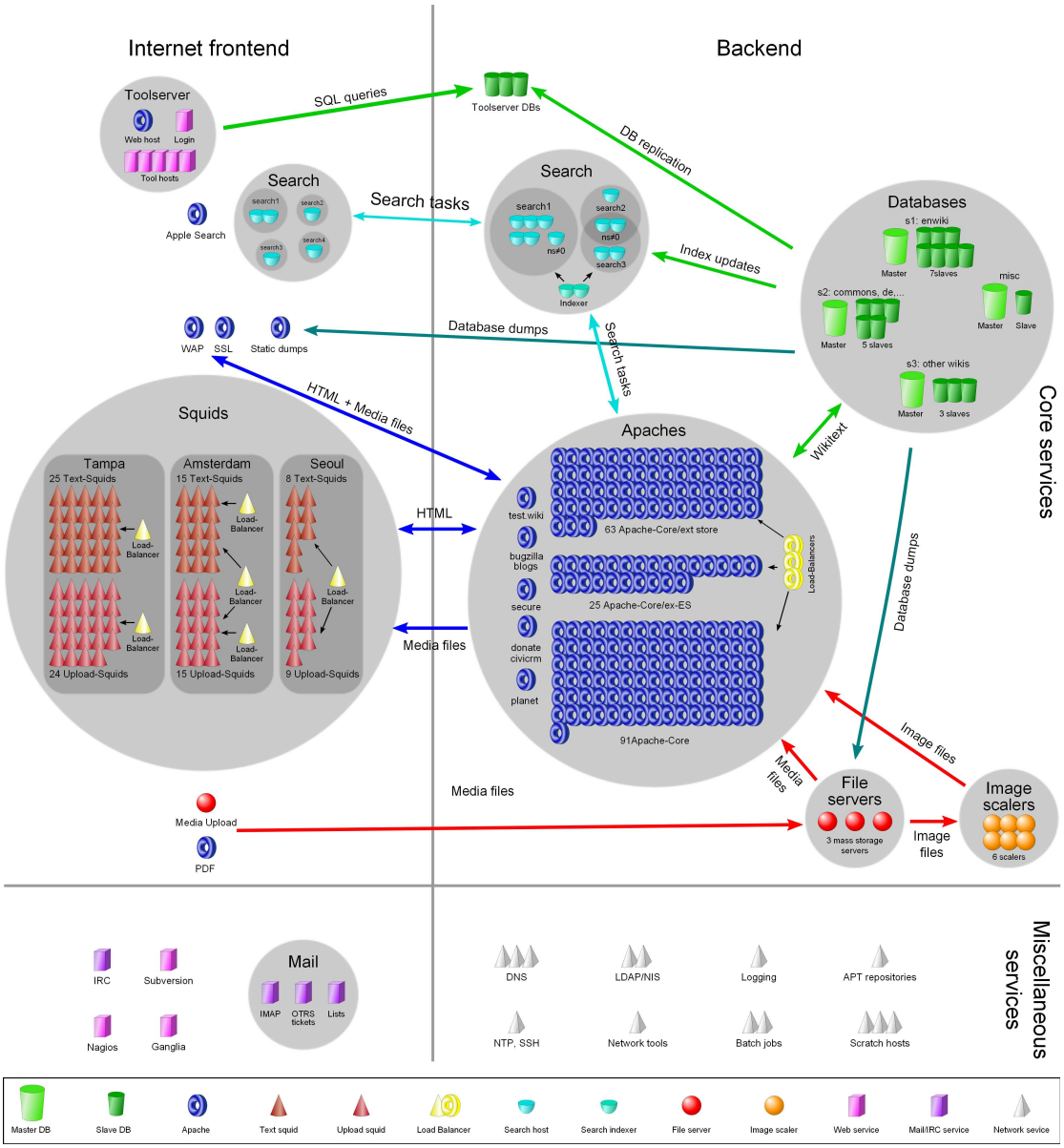


Figure 1.2: Wikimedia Foundation servers architecture (Retrieved from `http://meta.wikimedia.org/wiki/Wikimedia_servers` on 15 July 2010)

- **PowerDNS**
  Provides the DNS resolution [12] to distribute the received requests among the Amsterdam and Tampa facilities according to the geographical location of the users.

- **LVS**
  Linux Virtual Servers [13] are used to balance the workload of both web and cache servers. Load balancing is performed in front of both the Squid servers and the web servers. LVS efficiency is achieved as a result of running at kernel level and establishing a connection count based distribution which also allows a rapid malfunction detection.

- **Squid**
  Squid systems [14] are used to provide reverse proxy caching in order to speed up the content distribution by sending the requested contents directly from the cached elements and, thus, avoiding both database and web server operation.

- **lighttpd**
  Lighttpd web servers [15] are used to serve static files, such as images, as their optimized memory and CPU requirements make them suitable for being used in intensive workload situations and in serving operations which do not involve content dynamically generated.

- **Apache**
  Apache HTTP servers [16] receive the requests submitted by the users, elaborate the appropriate web pages and send them back in response. Web page production usually includes the rendering of the wiki code corresponding to a given article.

- **PHP5**
  Used as the server-side CGI scripting language to produce [17] content of web pages dynamically generated.

- **MediaWiki**
  Core application software [18] implementing all the functionalities of a wiki site. It is written in PHP and allows a high degree of customization through its great number of extensions. PHP execution is accelerated by means of a bytecode cache provided by the APC package [19]. Although PHP offers several and powerful functionalities, some external libraries have been incorporated to manage more types of contents so that wiki articles can result in richer documents. In this way, software support has been added to enhance thumbnailing or to render Tex scientific formulae, as an example. Figure 1.3 shows the relationship between the core MediaWiki application and the rest of external software elements used to improve the quality of the presentation capabilities of the wiki documents.

[12] `http://www.powerdns.com/content/home-powerdns.aspx`
[13] `http://www.linuxvirtualserver.org/`
[14] `http://www.squid-cache.org/`
[15] `http://www.lighttpd.net/`
[16] `http://www.apache.org/`
[17] `http://www.php.net/`
[18] `http://www.mediawiki.org/`
[19] `http://pecl.php.net/package/APC` (Retrieved on 13 September 2010)

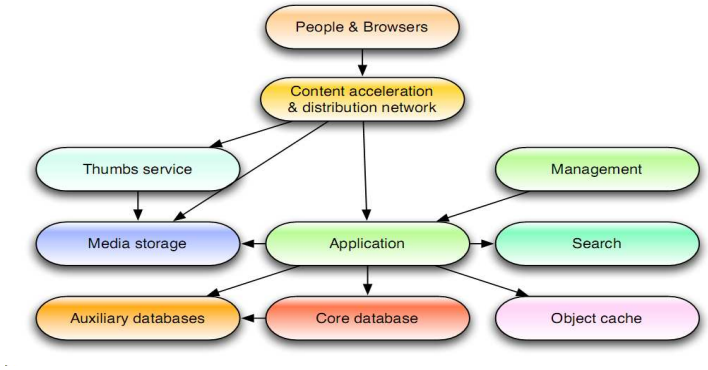


Figure 1.3: MediaWiki core applications and external software components (Retrieved from [Mit07]).

- **Lucene, Mono**
  Tools used [20] [21] for search and indexation. Wikimedia Foundation servers do not run Sun Microsystems Java Virtual Machine because of license issues so .net Lucene server running on top of a Mono .NET compliant framework is used. The search daemon has had to be split for each language edition and the indexes replicated in order the system could scale properly.

- **Memcached**
  Distributed caching system [22] commonly used to improve the performance of web servers by storing in RAM memory objects recently requested and, hence, avoiding delays due to I/O operations. In web sites scaled as much as the Wikipedia one, caching policies become critical. For this reason, the Wikimedia Foundation has arranged several caching systems to improve the performance of its serving systems. In fact, caching is performed at several levels including preprocessed HTML code to accelerate the treatment of contents for users having established the same settings as well as revision text that is not stored in the core databases any longer but in slower distributed storage. The output of some process such as the one requesting the recent changes, the image metadata and the session information are also cached.

- **Media storage**
  Media delivery is commonly performed by the Content Distribution Network but its storage has to be realized in the core systems. Thumb generation is an important and expensive task. In fact, requests for thumbnails are sent to different servers because the whole thumbnail set is scattered through several systems. As previously stated, thumbnail serving, because of the static nature of the images, is performed by lighttpd servers. However, thumbnail generation is done by dedicated servers requested by the application core. These servers have to access the sources images asking for them to the file servers through NFS.

- **Database**
  The MediaWiki Foundation relies on MySQL [23] database servers to be responsible for the main storage facilities. Database servers are split into masters and slaves, the formers can perform writing operations whereas the latters are in charge of only read operations. Furthermore, the

[20] `http://lucene.apache.org/`
[21] `http://www.mono-project.com/Main_Page`
[22] `http://memcached.org/` (Retrieved on 13 September 2010)
[23] `http://www.mysql.com/`

contents belonging to each language edition of Wikipedia are assigned to a particular group of database servers. In this way, each Wikipedia edition is supported by specific database systems that can be shared among other editions. Thus, replication is only maintained at server group level. Queries to the database are balanced across the corresponding group of servers which is determined on-the-fly using the prefix corresponding to the chosen edition. This results in a more efficient and flexible database usage. Queries are sent to the database through an specific API which allows to build more structured queries than using common SQL language. Special functions are used to issue multiple-operation queries that retrieve or insert several data. High level wrappers are used to write index-based offsets. Database servers use RAID configuration and are practically crash-proof due to the their failing management policy and to the robustness of the MySQL innoDB engine. MySQL uses different memory allocation for searching and querying operations (MyISAM for searches and InnoDB for queries). This determines an specialized system set to perform search operations as shown in Figure 1.2. Absolutely all the queries have to use an index and, also, every result has to be index-sorted. Having such a number of database servers allow to split data into several systems. This can be done under different polices or criteria such as data segments, tasks or, even, time. Data compression has been also considered as a way of improving storage efficiency although it can be only applied to text because media formats already include some kind of compression.

It is important to remark that the aforementioned Wikimedia philosophy promotes that not only the access and the contributions to the encyclopaedic contents adhere to the openness policy but also all of the internal documentation so that even purchase orders can be consulted in the Wikimedia web site [24]. Moreover, the overall software architecture used to maintain the wiki-based projects, and including applications as the described above, is based on tools that are released under free licenses. In this way, the core application software, the *Mediawiki* engine, is completely available for the community to use and to improve it [25].

As previously mentioned, the systems supporting Wikipedia have to manage with thousands of million requests sent by its users and, of course, have to keep all its vast compendium of knowledge under some kind of organizational schema. Every offered information has to be made available for the users in an effective and efficient way. Therefore, every issue related to the process of content serving has been always carefully addressed. The fundamental software systems involved in the availability of Wikipedia contents constitute its Content Delivery Network (CDN) which include web caching, HTTP and database servers.

The fact that most of the Wikipedia pages requested by not-logged users can be served avoiding both database and HTTP server operation by means of web caching is considered one of the fundamental improvements for a better performance and scalability. In this way, a Squid front-end system implementing HTTP reverse proxy caching was deployed to directly manage all the traffic generated by users who have not logging into Wikipedia but are browsing it. The basic idea is that the contents requested by this kind of users can be served from cached copies of the web documents previously generated as a result of the operation of both the database and http servers. The Squids also receive and deliver the requests sent by logged in users but this HTML cannot be cached because it includes personal per-user customizations. In any case and given that absolutely every request sent to Wikipedia pass through the Squid layer, its importance for this study is almost critical. Chapter 3 will describe in detail the role of the Squid systems as well as the information they register.

[24] `http://meta.wikimedia.org/wiki/Wikimedia\_servers/hardware\_orders` (Retrieved on 13 September 2010)

[25] `http://www.mediawiki.org/wiki/Download`

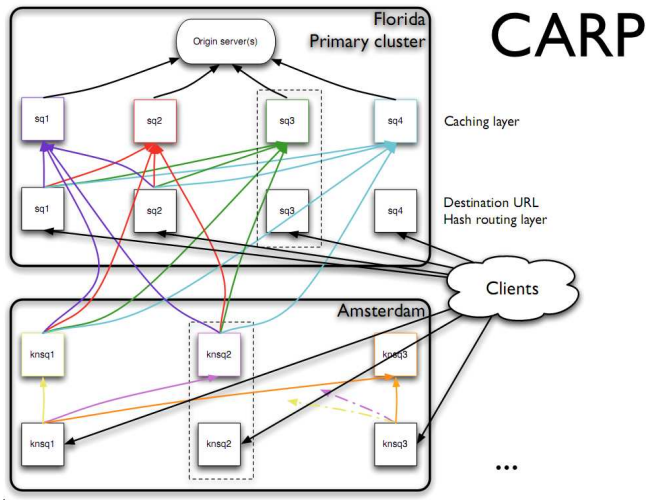


Figure 1.4: Implementation of CARP by two layers of Squid servers (Retrieved from `http://www.nedworks.org/ mark/presentations/san/Wikimedia/%20architecture.pdf` on 13 September 2010)

The Wikimedia Foundation CDN includes two clusters of squid servers, located in Tampa and Amsterdam, that receive the users requests from the DNS server that balance them according to their geographical origin. These Squid servers are running at a hit-rate of approximately 85%-90% multiplying the capacity of the Apache servers behind them so their averaged workload become considerably decreased. This becomes of special interest when traffic is directed towards particular pages via hyperlinks from other sites, as the caching efficiency for these pages will be nearly 100%.

Since text-serving presents a different access/communication pattern than media-serving (such as video or images), each Squid cluster has been split into task-oriented groups [Mit07]. However, the major improvement in the Wikipedia CDN has been the introduction of a multi-tier Squid server schema instead of the previous one using a single tier of servers with neighbor cache coordination through ICP and HTCP protocols. The new deployment sets a first layer of Squid systems to distribute the users' requests according to their corresponding URLs by using CARP (Cache Array Routing Protocol) over a second layer of Squids that properly stores the cached web pages. The CARP protocol allows to perform a hash-based distribution that results in a more reduced set of cached copies of objects and in a more efficient handling of node failures by the redistribution of requests across other active systems. Figure 1.4 illustrates the way in which the two combined Squid layers implements the CARP algorithm to serve the non-logged users' requests from the previously cached web pages.

All related information about HTTP transactions is saved by the Squid servers into a log file where each line corresponds to a served client's request. Each Squid server records the client's IP address (or hostname), the requested URI, the response size, and some other relevant information according to a common logging format. In this way, log lines from Wikimedia Squid servers will constitute our main information source because they contain the requests submitted, among other wiki-based project, to the different editions of Wikipedia.

The normal operation rate of a Wikipedia Squid server is over 1,000 HTTP requests per second (although it is possible to reach peaks of 2,500 HTTP requests/second). Log lines are buffered and

sent to an aggregator host from where a program send them to our facilities. Chapter 3 will explain in detail all the aspects related to the Squid operation, its logging format and the path followed by the log lines until they reach our systems.

1.5 Organization of this thesis

This section is aimed to introduce the rest of the chapters that this thesis consists of. The overall composition obeys to a typical schema consisting on the presentation of the current state-of-the-art related to the main topics addressed by this thesis previous to the development of our research work. The methodology used to conduct this research, the main results obtained and, finally, the most important conclusions and further work will be presented in this order through the corresponding sections.

In this way, chapter 2 consists on a detailed revision of the most important efforts and initiatives previously devoted to study the way in which users from different backgrounds are making use of wikis and, particularly, of Wikipedia. Most of this works have consisted in surveys activities performed on scholar or professional communities but also on the development of ad-hoc tools to perform statistical analysis over data related to this subject.

After this revision, chapter 3 undertakes the description of the methodology followed to develop the work presented in this thesis. Basically, this methodology consists on an empirical study based on the analysis of the log lines registered by the Wikimedia Foundation Squid systems that refer to the URLs submitted by the users. The analysis has involved both the parsing and filtering of the information elements that are part of the aforementioned URLs according to a set of well defined directives. Furthermore, an statistical examination have been performed on the data resulting from this analysis which have been stored, for this purpose, in a database.

Chapter 4 presents the main results obtained from the application of the methodology previously described over the data feed provided by the Wikimedia Foundation. The results are presented in relation to the research questions stated in chapter 2. Finally, most important conclusions and further work are also introduced. This part will present our most efforts, mainly related to the geolocation of the users' requests in order to find out the place from where the request to a given edition of Wikipedia or to a given content are coming.

A website has been set at `http://gsyc.es/~ajreinoso/thesis` to serve as on-line support for this thesis. In this way, it provides the necessary hosting for additional elements such as tables and images that have been separated of this document to avoid excessive length. Moreover, we are offering from here the full code of the application developed to conduct the analysis presented in this thesis and some examples of log files used as information feeding.

Chapter 2

State of the art

> "Tous pour un, un pour tous, c'est notre devise". *Les trois mousquetaires*, Alexandre Dumas, (1844).

> "I still haven't found what I'm looking for". *The Joshua Tree*, Paul David Hewson (aka Bono) U2, (1987).

This chapter introduces the most important research activities related to the use of Wikipedia. Therefore, a detailed and complete description of the state-of-the-art in this topic is also provided. As the work developed here consists in the study of the way in which different communities of users interact and behave when they make use of a collaborative platform, such as Wikipedia, previous initiatives in this area are also examined throughout this chapter. Moreover, given the case that the methodology conducted in this thesis includes the analysis of information from log lines, former searching activities having the same basis are also presented.

As previously mentioned, most of the previous research concerning Wikipedia is devoted to address topics related to the quality of its contents, authors' reputation, reliability and growth tendencies. This chapter also presents a summary of all this research in order to try to set a convenient scenario in which the work performed as a part of this thesis could be considered as complementary and valuable.

The characterization of the Wikipedia traffic can lead to statistical models providing a quantitative and qualitative description of the way in which Wikipedia users are interacting with the Encyclopedia. Previous descriptive initiatives have yet explored this way and offer statistical information about different parameters related to the size and growth of the Wikipedia as well as about its general use. However, although having a great interest for research purposes, this information usually consists in a collection of quantitative data and does not provide any kind of correlation between the different measurements presented. The provided data are not always updated and the specification of temporal ranges or concrete actions for particular analysis is not considered. Moreover, important information elements such as namespaces or topics repeatedly searched are disregarded. In this way, and, as far as we know, such a thorough analysis as presented here has never been undertaken over the traffic formed by the requests submitted to Wikipedia by its users.

2.1 Introduction

This thesis is fundamentally an empirical study about the characterization of the use that users make of Wikipedia and, thus, examines different metrics and measures considered as significant regarding the goal of finding both behavioral and temporal patterns. It is clear that the statistics about both the number of visits and the use of a web site as popular as the Wikipedia one have to become a topic of interest for the community of users of the Internet even from several perspectives. In this way, the Wikipedia is considered as a matter of study, for example, in the area of systems administration and, also, from a sociological point of view because of its dimension as mass phenomenon. Regarding Wikipedia as a tool for gathering and sharing knowledge, several initiatives have been devoted to measure and analyze different aspects considered as descriptive of the way in which users are visiting the different editions of Wikipedia, asking information from them or contributing in any way to their contents. Although some of these initiatives have not been undertaken from an academic approach, as mentioned in chapter 1, they constitute a really valuable source of information since, sometimes, are based on data directly emanated from the own Wikipedia systems. In these cases, the aforementioned sources have been used mainly to assess the validity of our results. In any case, I will present all these previous initiatives here because they deserve to be included as a part of the previous effort to provide a characterization of the use of Wikipedia.

In the aim of providing an adequate context for the work presented in this thesis, we, first, present previous research devoted to the study of the Wikipedia underlaying philosophy that address topics related to the mass collaboration phenomenon, some of its foundational principles and the way of conducting of the communities of users emanated from it. After this, we examine previous efforts also based on wikis and Wikepedia but that focus on different subjects than the ones addressed by thesis. This is intended to provide a wider scope of the topics concerning Wikipedia that are considered of interest by researchers. As this thesis relies on the analysis of users requests, we present some other initiatives that have also considered this feed as their main source of information. Finally, and to get even closer to the subjects developed in this thesis, we include several analysis providing information about the use of Wikipedia from two well-differentiated perspectives:

- Academic studies about the use of wikis and Wikipedia many of them consisting in scholar and academic surveys trying to find out the kind of use that particular groups of students, communities or people in general make of Wikipedia. This also includes works aiming to categorize the Wikipedia topics to which the activity of particular groups of users is directed.

- Initiatives devoted to offer statistical information, generally quantitative, about certain topics related to the use of Wikipedia.

2.2 Communities and generation of knowledge

The main features of the so-called Wiki approach described in chapter 1 situate it in the sphere of the paradigms devoted to provide tools for gathering and producing knowledge as a result of the collaboration of a community of individuals. In this way, mass collaborative authoring tools based on web systems have been previously addressed in studies such as [NKCM90] and [DB92]. The former study addresses the basic features that should be provided by such kind of tools in order to promote the collaborative efforts and to facilitate the interactions between people and the tools as well as among the individuals who are contributing in any way. On the other hand, [DB92] introduces the benefits

of the use of a shared feedback allowing to present the chronological list of changes performed over a given document.

The collaborative philosophy, prior to the building of knowledge, was applied in the environment of the software production resulting in the so-called FLOSS (*Free Libre Open Source Software*) projects. The ideas expressed in manifests emanated from initiatives such as the *Free Software Foundation* or the *Open Source Initiative* attracted a great number of volunteers. As a result, many communities arose around the development of software applications. Apart from the fact that these communities are basically made up of volunteers, they were also special in many other aspects. For example, and unlike traditional working teams, they did not need their members to be next to each other geographically or performing activities during the same periods of time. In this way, they could be considered as a kind of virtual communities and, of course, they had a very important supporting tool for this purpose: the Internet. Another important fact contributing to differentiate these groups was their organizational structure. Opposite to the strongly hierarchical organizations usually adopted by companies and institutions, these communities formed more flexible groups involving the whole community in the decision making process and establishing alternative forms of leadership as the meritocracy or the benevolent dictatorship.

Crowston and Howison present in [CH03] the social structure of open source software development teams. The authors introduce the onion-like model as the characteristic schema defining the development process in FLOSS projects. This model consists in a four-level structure whose core is constituted by the active developers that write the code. The next layer groups together to all the collaborators that provide patches and perform minor changes to the software that have to be reviewed by the core developers. Active users providing wish-list functionalities and informing about errors constitute the third layer. Finally, common users whose role is limited to the merely use of the application would take part of the most external layer. Moreover, Crowston and Howison's study examines the network centrality in the bug-fixing process and determines the non-existence of uniformity in the centralization of decentralization of communication structure of the considered projects. The same authors extend their analysis in [CH03] to the distinct degrees of hierarchical organizations exhibited by the FLOSS projects obtaining the same non-uniformity and stating that larger projects tend to be more decentralized and usually do not present a solid hierarchical organization.

For his own part, Raymond analyzes the organization of these communities in his work entitled *"The Cathedral and the Bazaar"* [Ray01], where traditional, pyramidal, hierarchical and well-structured working groups would correspond to the way of conducting of the cathedral builders whereas flexible, non-centralized, independent and heterogeneous groups would define the activity of a bazaar. Moreover, studies such as [BSKK01] describe the way in which the different members of the community contribute to its overall development and to its main targets and objectives. According to this work, communities leaders usually deal with the organizational issues of the community whereas the rest of the members contribute usually motivated by their own preferences.

Rheingold in [Rhe00] defines a virtual community as a social aggregation emerging from the Internet when enough people carry on public considerably long discussions. In this analysis, he addresses the social implications of relationships established and maintained through the network. If Raymond's *"The Cathedral and the Bazaar"* deserves a relevant place because of its contribution to the disclosure of the benefits of the software development in community, Surowiecki's *"The wisdom of the crowds"* [Sur04] postulates how collaborative efforts can be joined in very different environments to obtain more accurate results than those derived from individual analysis even though they are coming from renowned experts in the matter. Surowiecki introduces the elements acting as the conforming criteria that differentiate the collective movements considered, according to him, as

crowd wisdom. Besides that, the book presents the main advantages of decentralized and unstructured systems for decision making as well as their most important drawbacks.

Focusing on management of knowledge, Benkler addresses in his book "The Wealth of Networks: How Social Production Transforms Markets and Freedom" [Ben06] the revolutionary changes introduced in the production and exchanging of information as a result of the application of the most recent advances in technology, communications and economy. In this way, this book describes how a new concept of information economy based on decentralized and networked paradigms have preempted traditional concepts based on monopolistic information industries. Moreover, the possibility of making every effort available to whole communities through network-based collaborative mechanisms as well as the proved effectiveness of cooperative initiatives, such as the production of high-specialized software, are picturing a new scenario in the access of people to the general culture and information. Benkler considers that the successful communication of knowledge is developed in three phases. First, contents are created, then they become organized and examined, and, finally, they are spread across appropriate channels. This books analyzes the social implications of such kind of changes that come reflected in the new ways of human behaviors and interactions and also in the way in which communities are organizing their operational structure.

Stalder and Hirsh associate the collaborative approach to the term "Open intelligence" in [SH02] that analyzes the applications of the paradigm in three cases of study including Wikipedia. That article is aimed to overcome the boundaries of the application of collaborative efforts in the area of software development by including several socio-technical approaches. Basically, the work focuses on the openness concept and makes a review of its most relevant principles and benefits previous to the presentation of the three cases. Cedergen applies the openness philosophy in [Ced03] to the creation of content for public availability with appropriate permission for re-creation, improvement and re-distribution. The author examines in this work the possible sources and adequate environments for creation of open content as well as the forces governing the communities related to this production.

Quiggin analyzes in [Qui06] the relevance of both blogs and wikis in social innovation and in the process of creative collaboration. On the same line, Kolbitsch and Hermann explore in [KM05] the introduction of new technologies to create non-static knowledge management system that, in addition, are built as a result of a collaborative process by their users. The authors focus on encyclopaedic content as a central point for building communities in conjunction with several elements to establish quality assessment, vote rankings and so on. The same authors analyzse in [KM07] the new mind shift brought by these technologies that encourages individuals to produce their own knowledge and even a sort of collective intelligence. The author even suggests the loss of individuality to favor a king of integrated society maintained by these technological improvements.

Of course, there is a also a place for controversy, Chris Wilson in [Wil08] considers that purportedly collaborative projects present, actually, non-democratic dynamics and, even, non-democratic governance schemas. In his line of argument, he includes the site Digg.com [1], a portal devoted to receive stories from users, who, in addition, are able to rate ("to digg" according to the portal's own terminology) their favorite ones. Wilson suggests that the most ranked stories are determined as a result of the influence of a reduced elite of users and that the same can be applied to Wikipedia, whose contributions would be authored mainly by a little group of users. The analysis presented by Wilson reflected that the 100 most active "'diggers"' contributed in 2007 by the 44% of all the stories. That meant a significant declining from the 56% corresponding to the previous year, so the number of contributions due to the least active users had increased. Thus, even though the raising of the contributions from non-expert users of these portals could be considered as a tendency, their are

[1] `http://digg.com/news`

consisdered as elitist stating that their existence and maintenance are only due to a reduced minority. However, the author agrees to the existence of democracy in the election of the most authoritative users who are, effectively, the most active ones.

Other analyses explain the situation as a change of the predominant group of contributors over time. In this line, Kittur *et al.* study in [KPSM07] the way in which members of particular communities contribute to them by analysing the number of words added and the number of edits performed. One of the communities considered as a subject of study was the one consisting of the users of *del.icio.us.* [2], a portal were users can bookmark web pages using their own tags instead of previously defined categories. The social aspect of this site comes from the fact that users can have access to the result of the tagging process performed by other users. In this way, users can obtain, for example, the web pages most tagged for an specific tag. The study concluded that most of the tag operations were performed by expert users just until a certain date. From then on, novice users started to be the main contributors. Gave that the degree of expertise to classify the users was determined by their number of edits, the number of these operations coming from users with smaller numbers of them was revealed as rising over time. According to the authors, the same situation occurs in Wikipedia as explained later in the following section.

The most currently topical term concerning mass collaboration is, perhaps, the *Web 2.0* approach. The definition and scope of this term has been object of controversy mainly because although it might suggest some upgrade in the protocols and specifications sustaining the *World Wide Web* service on the Internet, the fact is that it does not. Of course, the Web 2.0 carries new software technologies but its main important contributions are related to the new ways in which people can make use of the web resources. Web sites built under this new approach are able to allow users to do a lot more than just obtaining information. According to [O'R05] and [Hin06], the key is to consider the network as a vast computing resource offering different capabilities in the form of web services. Users, now, are encouraged to participate by expressing their opinions, by voting or ranking contents, and, of course, by adding their own information. This new interaction demands a participation-oriented architecture which rely on new interfaces systems resulting in blogs, social portals or networks and, for sure, in wiki sites.

According to the arguments expressed above, the paradigm consisting in the application of collaborative and decentrelized efforts on several types of projects and, particularly, on the creation of knowledge offers considerable benefits. Communities consisting of individuals gathered around a particular project, idea or objective may exhibit very different patterns of behavior and, because of this, their contribution models deserve to be subject of research. Moreover, the principle of collaboration have deeply penetrated on communities of users around the World Wide Web. In this way, the schema of the web documents has been adapted to the new way wave of producing web sites allowing their users to submit their opinions and contributions. As a result, users get more and more involved in the portal's own building process by participating on their contents. This collaboration spirit has made possible very important projects and initiatives as the Wikipedia one. As a curiosity, in 2006 the Time magazine chose the volunteers of collaborative projects and portals as its "*person of the year*".

2.3 The *wikis* and Wikipedia as research topics

The main properties and features of several wiki-based projects have been considered interesting enough to develop academic research to provide a better understanding of their particularities and to

[2] `http://www.delicious.com/`

examine the most relevant characteristics of the communities supporting and feeding them. More in detail, previous studies have addressed several issues related to these projects including their tendencies in evolution, trustability, growth ratios for contents and users and so forth.

As previously mentioned, the collaborative paradigm for the new industry of information demanded effective technologies to implement the supporting platforms. One of the first studies presenting the wikis as a valuable tool for knowledge management and group collaboration was developed in [Wag04] by Wagner. The study introduced the Wiki technology and some implications of its use and applicability in knowledge management and predicted more than linear growth ratios for these systems. Wagner would continue this trajectory later by studying the use of wikis and the applications of other web-based tools in this area ([WB05], [Wag05]).

Later, Ebersbach presented in [EG04] the wikis as a vehicle to fight the one-way information consumption installed on the Internet by offering a tool suitable for receiving users' contributions and suggestions and capable of allowing the characterization of the involved media as emancipatory. His book "Wiki: Web Collaboration" [EGH05] focuses on the same ideas and presents the wikis as the tool driving the production of the majority of the contents contributed to virtual platforms.

The quality of the Wikipedia contents and the finding of methods and measurements to evaluate the authority of contributions to the encyclopaedia constitute one of the most prolific research areas. Korfiatis analyzes these subjects in [KNP$^+$06] and proposes an approach based on the social networking process arose in the building of articles. Credibility is estimated from the metric of centrality of the article's contributors which allows to establish the centrality of the article overall construction process. Chesney also addressed the credibility problem in [Che06] where he determines the authority of several articles by conducting an extensive survey involving research personnel. In [DBWS06], Dondio defines mechanisms to determine trustworthy articles in Wikipedia by computing their trust levels. In this way, articles could be categorized according to these levels. Another way to evaluate the quality of the Wikipedia contents has consisted in comparing its articles with other solid and traditional encyclopaedias such as the *Encyclopaedia Britannica*. In this way, one of the best publicized studies about the credibility of Wikipedia was developed by Giles for the *Nature journal* [Gil05]. According to this study, Wikipedia articles had a similar quality that their equivalents in the *Encyclopaedia Britannica* where they had been put under revision by experts. There were errors found in both encyclopaedias but Wikipedia presented by a 37% more than the Britannica. The same line is followed in [LKSY07] where Luyt *et al.* compare the same two compilations of knowledge in relation with the specific matter of Biochemistry using a well reputed text book as a benchmark reference. Another interesting comparison is performed in [Nie07] where the author compares the references to scientific journals made from Wikipedia articles with the statistics published in the *Journal Citation Reports* in order to find that Wikipedia authors are using a well-structured citation system that points to articles in top-ranked journals. According to the author, this fact can be translated in an increase of the reliability of Wikipedia as an information source. Other studies involving a comparison of Wikipedia with other encyclopaedias or reputed corpus of knowledge are presented in [RH08]. Olleros [Oll08] considers as positive the decentralized Wikipedia quality control and wonders that sustainable success of Wikipedia is in progress as quality dimensions and parameters for encyclopaedias are being redefined, precisely, as a consequence os Wikipedia.

Wilkinson and Huberman establish in [WH07a] and [WH07b] the strong correlation between the number of edits performed on an article and its observed quality. In fact, they found that highest quality articles (referred as "'featured"') had a much larger number of edits and distinct editors than common articles. An interesting analysis is developed by Halavais *at al* in [HL08] where the authors state that the coverage in Wikipedia is not as general as in other encyclopaedias because, according to the authors, the development of the Wikipedia contents does not follow an structured methodology but,

on the contrary, it is driven by personal interests of the contributors. Stvilia and Gasser introduced in [SG08] the concept of an information quality model based on both error detection and error correction as a way of improving the quality of information systems. In [Stv08], Stvilia *et al.* studied the information quality dynamics in Wikipedia.

Author reputation is another aspect considered on previous analyses focusing on Wikipedia. Adler and Alfaro present in [AdA07] a method for estimating the authors' reputation based of the longevity of their edit operations. As a result, changes and contributions coming from well-reputed authors are more likely to remain in the encyclopaedia that the ones coming from authors with low-reputation ratios. The notion of reputation is used to assign trust to the words introduced in the successive revisions of a given article in the work described in [ACdA+08]. In this way, author's reputation is used to qualify the text that he or she introduces in a given article. As a result, it is shown that text assigned with high-trust marks is more unlikely to be edited that text considered low-trust.

The analysis and study of the motivations of people to contribute to the Wikipedia have been addressed in studies such as [Kuz06]. The methodology of this study considers a first phase consisting in a survey conducted in a public university prior to the definition of a set of parameters underlying the motivations. Finally the study analyzes how the wiki technologies affect to these parameters. Nov also studies the motivations of the wikipedians in [Nov07]. Again, the principles that motivates the contributors are studied and even classified in [ON08]. Other studies examining the incentives for participation in the Wikipedia project are developed by Rafaeli *et al.* in [RHA05b], [RHA05a] and [RAH06] where the authors analyze aspects related to the elements motivating the wikipedians and the sense of community perceived by this collective. Hamburger *et al.* present in [HLMH0] and interesting analysis about the personal characters and profiles of a group of people used to contributing to Wikipedia regularly in a similar aim to understand and explain the wikipedians' motivations. This topic is also addressed in other studies such as [CVM09], [SH09] or [HLS+07a].

The evolution in terms of growth ratios and tendencies has been another topic in which the research community has concentrated a notable effort. Capocci *et al.* deal with this topic in [CSC+06] where the growth is statistically modeled by using the topological properties of the graph constituted by the topics and the links among them. A similar approach is presented in [ZBvD06] where Zlatic *et al.* considered articles and hyperlinks among them, respectively, as the nodes and links of a complex network. The study declares to have found several regularities pointing to a unique growth process involving all the Wikipedia editions. Despite the significant growth of Wikipedia that includes an important widen of its scope, its coverage in terms of dealt topics does not seem to be deteriorated. That is the conclusion that Spinellis and Louridas state in [SL08]. Voss developed a quantitative analysis of the German Wikipedia in [Vos05] in which it was found that several parameters such as the number or articles, the active Wikipedians or the total number of links followed an exponential growing rate. On the other hand, Buriol *et al.* presented in [BCD+06] a temporal analysis based on the evolution of the so-called *"WikiGraph"*. This is a graph representing the linking structure of the Wikipedia where articles are represented by nodes and links among them by the corresponding arcs. The main particularity of the graphs lies on the timestamp associated with all the events of each node. This allows a temporal characterization in terms of users, revisions and articles. Moreover the temporal evolution of several topological properties of the "WikiGraph" are also presented. Shyong determines in [TR09] that the distribution of visits to articles in Wikipedia follows a log-normal curve having a so-called long tail distribution and, more important, that article births have reached a peak and may start to decline. Even more, Suh *et al.* suggest in [SCCP09] that Wikipedia growth has slowed and both pages and editors are declining. On the contrary authors advert a raise in coordination, reject of new users' contributions and opposition to new edits.

Particular aspects of the Wikipedia have also deserved previous research efforts. For example,

Viégas *et al.* analyze in [VWM07] the process leading to the promotion of Wikipedia articles to the status of featured and consider that wiki technology, rather than promoting anarchism, tends to produce well structured organizations. David Lindsey developed a very interesting study in [Lin10] in order to assess the quality of featured articles. His methodology mainly consisted in the analysis of a set of Wikipedia featured articles by a group of experts in the subjects developed by the articles. These experts had to assess the general accuracy exhibited by the articles. They also had to determine the conformity of articles with the Wikipedia's own featured criteria and to compare them with other available sources. Finally, they were encouraged to rate the articles in a quantitative scale. The results of this analysis were based on 22 analysis and determined a considerable disparity in the quality of featured articles. Approximately a 54% of the articles complied with the Wikipedia promotion criteria but about 1/3 of them failed in their quality assessment. The author attributes this situation to the lack of experts in several areas among the Wikipedia contributors and, considering the featuring process as unsuccessful, encourages student to be cautious when referring information from Wikipedia. Nevertheless, the author notes that the consulted experts usually have indicated that the Wikipedia contents were usually the best publicly available on the Internet. Another Wikipedia feature considered of interest is its semantic relatedness which is the subject of studies such as [SP06] or [GM07]

Consensus, vandalism and other kind of issues derived from the typical open character of the wiki platforms have been addressed in studies such as [KSPC07], [SCPK07] and [VWKvH07]. The former ones present methods to characterize conflicts in different levels as well as coordination costs. In the latter, the authors present some mechanisms used in Wikipedia to reach consensus when disputes about the content of articles arise. Talking about vandalism, Ciffolilli stands in [Cif03] that the graffiti-type attacks and other non-desirable contributions to Wikipedia are being neutralized with an effective and cost-reduced technology witch may include sporadic authority intervention. The work also enumerates some of the motivations expounded by the Wikipedia contributors and provides indications to sustainable corpus of knowledge virtually managed. Despite of being born to offer free and open contents, Miller discusses in [Mil05] the authority rights over the contents submitted to the Wikipedia and the possibility of apply several mechanisms for controlling them. Lorenzen also deals with this topic in [Lor06] where examines a public system to detect and solve problems emanated from users' behaviors. Priedhorsky, *et al.* developed in 2007 a thorough analysis [PCS+07] over millions of Wikipedia articles to assess vandalism. Surprisingly, they found that a very reduced percentage of pages had been vandalized (approximately the 0,37%). The authors even categorized the types of vandalism into seven categories: misinformation, mass deletion, partial deletion, offensive, spam, nonsense and other. Kostakis analyzes in [Kos10] the problem arose from the peer governance model established in Wikipedia. The author analyzes the conflict between two conflicting policies for content generation: "'inclusionism"' and "'deletionism"'. The former states that Wikipedia has to offer as much information as it can without considering its subject or theme. The latter states that, on the contrary, the presence of information entries not related to traditional academic contents make Wikipedia become less serious and reduces its credibility. This issue is also addressed in [TR09] where topic notability and deletion reasons are studied.

Wikipedia also serves as a test field to develop automatic systems or functionalities. As an example, Wang *et al.* examine in [WWZY07] a collaborative system for annotation and recommendation in Wikipedia. Another example is provided in [LD07] where authors describe an XML retrieval system capable of deal when unpredictable structured documents such as the Wikipedia's articles or the system developed to mine information from pages such as the Wikipedia ones and which is presented in [BFGM]. Other example is given in [RCAC05] where a system to extract entries from Wikipedia and associate them in an ontology or semantical network.

From all the above, it is proved that the wikis and the Wikipedia itself have attracted a lot of attention from the research community that has undertaken several analytic initiatives aimed to provide a better understanding of the new phenomemon constituted by the new form of knowledge generation and management that the wiki technology represents and so does the Wikipedia as its most important representation.

2.4 Workload characterization based on log files analyses and web caching schemes

This section aims to provide an examination of the previous studies involving log files analysis to determine a set of features concerning a certain system or to characterise its use throughout the traffic directed to it. In particular, we will focus on the use of logs generated by Squid web-caching systems as the main data source because, as shown in Chapter 3, they will constitute the main basis for our analysis.

Almeida *et al.* propose in [ABCdO96] models for both temporal and spatial locality of reference in the requests directed to four important web servers corresponding to two relevant supercomputing centers, a research center and a university. The study is based on the web servers log files and the authors present how temporal locality can be characterized from the stack distance metric. On the other hand, spatial locality can be analyzed using the notion of self-similarity.

Other studies devoted to present a detailed workload characterization of the traffic directed to Internet Web server were developed by Arlitt and Williamson in [AW96] and [AW97]. The studies analyzed the workload of six web sites, three from academic environments, two from scientific research institutions, and one from a commercial Internet provider, to study their log files and identified up to ten invariants as constant features in all the considered data sets: success rate, file type, mean transfer size, distinct requests, one time referencing, size distribution, concentration of reference, inter-reference times, remote requests, and wide area usages. According to the authors, these invariants could be assumed as general truths about the Internet and could be used to define possible strategies for the design of a caching system to improve the Web servers performance.

Barford *et al.* analyze in [BBBC98] how certain workload features evolved over time. In this way the study compared two measure sets obtained from the same computing facility at Boston University and separated in time by three years. The obtained results come from the comparison of the statistical distribution of Web client requests and from the study of how the observed differences, mainly in popularity and temporal locality properties, affect the benefits of web caching in the network.

The analysis of log files containing information about the queries submitted to web systems by their users has been developed for a long time. In [SMHM99] the queries submitted to the Altavista Search Engine are analyzed to find some interesting behavioral search patterns exhibited by users when querying the system. Among several others important facts, the authors determined that users rarely modified their queries, did not look beyond the ten first results and used relevant search terms together in phrases.

One of the first studies using the information contained in Squid log files was conducted by Khunkitti *et al.* in [KI01] where the authors examine the life of cached objects in Squid systems. Obtained object life could be used for monitoring web objects in order to eliminate unnecessary validating traffic to the servers.

Bent *et al.* studied in [BRVX04] the properties of a large number of Web sites hosted by a common ISP (Internet Service Provider) and undertook a simulation about the potential benefits in

performance derived from the introduction of content delivery networks (CDNs) for these Web sites. The study found a high degree of uncacheable responses and mandatory cache validations. According to the authors, the main reason is the indiscriminate use of cookies and the disregarding of the HTTP 1.1 cache control features.

Cherkasova and Gupta analyze in [CG04] enterprise media server workloads based on the access logs from two servers at Hewlett-Packard Corporation. Log files were collected during approximately two years and allowed to discover client access patterns, media server access tendencies and the evolution over time of the requests to the media contents. The main goal of the study is the characterization of the dynamics involving the access patterns to the media content and also considers the applications of CDNs for media serving. Other analyses about server workload involving streaming and media access are developed in [GCXZ05], [JHG06] and in [SMZ04]. Almeida *et al.* analyze in [AKEV01] client workloads for educational media servers located in two relevant U.S. universities. In this case, the main goals of the study are to acquire an adequate knowledge of the concerns about designing content distribution networks and to quantify how much server bandwidth could be saved using multicast streaming methods to distribute stored contents.

Baeza-Yates and Poblete also undertake the mining of the queries submitted by users to a certain web server that registers them in an appropriate log file [BYP06]. The analysis considered the queries submitted directly to the server search engine as well as those sent to general search engines and pointing to elements hosted by the server. The main goal of this study was to determine whether the server contents met the users' information requirements and how to collect information helping to improve the overall system quality and, particularly, its usability. The study considers web mining as entailing content, structure, and usage mining. In this way, the authors propose a model aiming to collect information about these three elements in order to define navigation patterns and terms with adequate information scent (IS) values. Furthermore, the documents in the site are classified according to the way in which they are reached and queries are classified as successful or unsuccessful depending on whether they lead or not to subsequent visits. According to the authors, web use mining has proven to be a useful approach to analyze several aspects such as isolated pages and needs of re-organization. The study concludes that the introduction of adequate IS elements in links or description fields lead to an increase of successful external queries and to a decrease of internal queries as well to a great number of accessible documents. The authors address the same topics in [PY06] and [PY08].

Query classification is also studied in [BJL+07] where authors introduce a system for automatic query classification based on the content of log files. In this case, the aim is to improve the search service in order to make it achieve a better performance and accuracy and to reduce its operational costs. This thorough analysis present several classification techniques that are evaluated according to the precision/recall measurement. The classification system proposed by the authors combines manually classification with techniques ranging from machine learning to computational linguistics. Another study involving query analysis is developed in [BJC+04]. Here, the authors explore the changes and evolutions of the queries to a general commercial site across the hours of the day. The article concludes that the total traffic of queries experiments variations in magnitude and correlation among the queries received in a particular hour and those of the next one. However, it also states that the distribution of frequency of the queries in an hour remains constant throughout the overall day.

Wolfram *et al.* use cluster analysis in [WWZ09] to determine whether different groups of sessions can be obtained from the log information collected from three differentiated web systems and, more important, if the same types of groups are present in all the web sites. The findings of this work present several common types of sessions observed in all the environments as well as common session transformations over time.

Web caching approach is considered one of the most effective technologies to improve Web traffic

delivery and to reduce bandwidth consumption. Aggarwal *et al.* presented in [AWY99] the main characteristics of the web caching and the main differences with traditional caching. Liu *et al.* present in [LWZ04] a wide revision of some techniques used to implement web caching such as heterogeneous caching network structures, and dynamic content caching. Database backed web systems have been largely addressed in several studies and analysis. Luo *et al.* analyze in [LNX08] two caching schemes consisting, respectively, in passive and active request caching. Passive queries are keyword-based queries, whereas active ones embed some kind of functionality. The study shows how passive caching results in a great gain of performance but active caching, on the contrary, cannot be worthwhile. Labrinidis *et al.* review in [LLXX10] several caching techniques to improve performance, scalability, and manageability in web systems relying fundamentally on database support. Tailored solution for particular web systems having to deal with a great amount of traffic have also been proposed. For example, Candan *et al.* presented in [CLL+01a] and [CLL+01b] an architectural framework for enabling dynamic content caching for database backed e-commerce sites.

2.5 Characterising the use of *wikis* and Wikipedia

As previously mentioned, most of the previous research involving the Wikipedia has focused on aspects concerning the quality of its contents, its evolution, reputation or any other more particular features or properties. By contrast, this section is aimed to provide a review of the previous efforts focusing on the use of Wikipedia. According to the stated in the introduction of this chapter, these works will be examined from two very distinct perspectives.

Firstly, I will present academic works and research focusing on topics involving the use of Wikipedia and having a basis consisting fundamentally in surveys carried out in specific communities of users or on inquiries performed on non-related independent users. Then, I will introduce the initiatives and studies devoted to provide some kind of information, both qualitative or quantitative, about the use of Wikipedia and its traffic. This kind of information is generally offered from web sites that dynamically generate tables and graphs that are generally updated. Thus, I will examine previous developed works from these two different perspectives.

2.5.1 Academic research on the use of *wikis* and Wikipedia

When wikis appeared on scene, several publications presented their main features and the benefits derived from their use to the scientific community but also to the particular collectives considered as specially adequate to take advantage of the new tool. As an example, McKiernan examines in [Mck05]the use of wikis for librarians and professionals related to information management. The important role of wikis to support a critical attitude towards the information offered by the media is discussed by Barton in [Bar05]. Gillmor [Gil04] analyzes the possible effects of collaborative working groups over the classical perception of centralized journalism. The use of Wikipedia as a method for cooperative journalism can also be found in [Lih04]. Müller *et al.* explore in [MMB08] the main aspects of the wikis as an appropriate tool for knowledge management. This article analyzes existing wiki-based networks under the approach of Social Network Analysis (SNA) and Dynamic Network Analysis (DNA). SNA considers that networks can be translated into a graph G(N,L) with a finite number of vertices (N) and edges (L) where two vertices are adjacent if there is an edge between them. In this scenario, the authors consider the degree centrality and the betweenness centrality as the fundamental metrics. The first concept is used in SNA to investigate the activity of individuals, considering that a vertix is central if it has many relations to adjacent vertices. On the other hand,

betweenness centrality establishes important vertices if they lie on a shortest path between other two ones. Dynamic Network Analysis, in his two variants of cumulative analysis and sliding-window based analysis, allow to study the process of a network transformation over time. All these metrics were analyzed in a wiki system created to serve as the knowledge management tool for a company. Authors found interesting facts such as the progressive increase of centrality as the networks growths (meaning that more authors join progressively the wiki project and begin to contribute). Density of the network (ratio between the number of existing edges and all the possible ones) seems to be negatively correlated with the average path length (average numbers of node between any two nodes) whereas a positive correlation was found between the article count (total number of articles) and the average degree (A node's degree is defined as its number of direct edges to neighbor nodes). The analysis concludes with the degree centrality obtained for the network that determines the existence of so-called *hubs*, people that contribute in an active way and have an almost complete understanding of the way in which wikis operate. This kind of users, due to their early adoption of the technology, encourage the rest to contribute as well as represent a fundamental role, specially in the early stages of the wiki evolution. Other uses or applications of wiki technologies include translation, as presented by Désilets *et al.* in [DGPS06].

The inner structure of the Wikipedia community was explored in [EH05], where Emigh and Herring found that there existed several correlations between the level of post-production editorial control and the degree of compliance with the standards of the collaborative documents stemming as a result. Moreover, Pfeil *et al.* state in [PZA06] that the cultural differences among the Wikipedia's authors have their reflect in their contributions and in the use they made of the on-line encyclopaedia. A very interesting survey paper is presented in [MMLW09] where different uses of the Wikipedia are discussed. In this way, Medeylan *et al.* consider four main categories which correspond to natural language processing, ontology building, and both isolated and combined use with other information sources. Other general study is conducted in [Fal08] in order to analyze the epistemic results of the use of the Wikipedia. In this study, Fallis concludes that in terms of knowledge obtaining, that is epistemology, the access to Wikipedia offers very valuable properties and possibilities. McGrady presents in [McG09] several concepts related to the credibility, rules and spirit of Wikipedia. The author stands that authority is expertise-based and comes from verifiable information and accurate references. In this way, the author highlight that experts do not create authority in Wikipedia but helps editors throughout adequate information sources. McGrady also remarks that to favor a neutral point of view research is disallowed in Wikipedia. Rhetoric is another possibility of inaccuracy but the author considers that is controlled by the generic revision processes, capable of detecting false or tendentious facts. Finally, the author describes several bad uses that take advantage of the Wikipedia's own rules and spirit to thwart its main aims. Reagle studied in [Rea05] and [Rea07] different examples of social interactions manifested themselves inside the Wikipedia and, basically, related to adequate and proper behavior as well as to leadership roles.

The use of wikis in specific environments involving collaborative dynamics and developments, such as scholar or academic ambients have been largely addressed. In 2004, Buffa *et al.* showed the benefits of derived from the introduction of wiki technologies in a collaborative process of software development [BSG04]. The participant students where geographically distributed and become rapidly adapted to the use of the tool. Forte and Bruckman analyze in [FB06] the possibilities of introducing Wikipedia in the development of activities involving collaborative writing. The study analyzes the applications of publishing tools, such as Wikipedia, in programs and curricula in order to improve several aspects like authenticity, disciplinary and assessment. Konieczny makes a brilliant and complete review of the use of wikis and Wikipedia in the university scenario in [Kon]. In this article, Konieczny presents the major advantages and benefits of the use of wikis as a teaching

tool as well as discusses how the Wikipedia itself can facilitate and foment the students' activities and assignments by offering them all its services of course imbued with its open and collaborative character. Another example can be found in [Sch08], in this case a new survey is conducted in order to determine the accessibility to the Wikipedia's contents related to Psychology from common Internet search engines. Moreover, this study inquires how students are using Wikipedia for both scholarship and personal interests. Addressing the use of Wikipedia in academic environments, some arguments for controversy were presented by Waters in [Wat07] where the author, using a real case of misinformation, recommends not to consider any encyclopaedia as the receptacle of the absolute truth although it may include reliable and trustable references. On the contrary, Waters suggests always to assess and contrast the information obtained from a certain source using any available possibility. The Wikimedia Foundation maintains the same attitude towards its project and encourages not to use Wikipedia in direct citations.

Willinsky analyzes the important question of the external references provided from Wikipedia articles. In this way, he examines in [Wil07] the number of citations to research or scholarships works found in the Wikipedia articles and the possibility of having an open access to them. The author used a sample of 100 articles and concluded that a very poor ratio (2%) of them included references to open access research works despite they was considerably easy-to-find previous open related productions reachable through Google scholar or other search engines.

Kittur *et al.* stated in [KPSM07] that the contents of Wikipedia were being produced mainly by a little elite of administrators only until 2004 (*The power of the Few*). From then on, most of the contributions were sent by individuals not belonging to the elite group (*Wisdom of the Crowd*). Another interesting point is introduced in this article, according to its authors people with highest numbers of editions are the ones who contributes in a more prolific way to the contents of Wikipedia because they add, in average, twice as many words as they delete. On the other hand, users whit lowest numbers of edits are deleting more words that they add. This means that most of Wikipedia content is contributed by a few users whereas the great majority of them just perform precise corrections or get involved in minor changes. The same conclusions are presented in [Chi07] where it is shown that the number of users with lowest ratios of edits becomes a larger part of the total contributions over time. This analysis also shows how the participation in Wikipedia fits a long tail structure as a result of a power-law distribution governing the ranking of of edits per user over several months.

One of the first works considering the analysis of the visited namespaces as an indicator of the activity of Wikipeadia users is [ELB08]. Here the authors analyze the relationship between the content of the Talk page of a set of articles and their edit activity. The importance of Talk pages had been also stated by Viégas *et al.* in [VWKvH07]. This study concluded that in most cases the discussion entries in the Talk pages were accompanied by editing activity. Vandalism was the factor that invalidated the correlation because it did mean non-contributing to the article. Ehmann's research also refers to the so-called *advantage of the first mover* postulated by Viégas *et al.* in [VWD04] that states that the original content of an article would remain over time. This was found for articles with high scientific content. However, the article also enunciated an inverse relationship between the age of an article and the permanence of the original text, fact that may incur into controversy with the previous statement. Finally, the article determines a strong difference in quality aspects from articles belonging to different disciplines. In this way, articles related to high-level scientific contents would have been considered in a top quality level although they were usually written in a such kind of style that limited their access by the overall community.

Head and Eisenberg developed a very interesting examination [HE10] based on a survey about the use of Wikipedia in several colleges and universities from the United States. The survey considered the responses of about 2,000 students and focused on the frequency of the visits to Wikipedia, the

students' motivations, the stages of research in which the Wikipedia was used the relationships between the use of Wikipedia and other resources. Among other results, the study found that students in architectures, engineering and other scientific disciplines were likely were more likely to use Wikipedia that students from other degree. A 22% of the student declared to use the Wikipedia frequently, whereas a 23% said to use it ocassionally and a 13% rarely. The most important motivation found for the use of Wikipedia was the obtaining of general background information about a particular subject in the initial stages of research.

Spoerry analyzes the most popular topics in Wikipedia during a five-month period in [Spo07b]. The methodology of this study consists in determining the 100 most visited articles in Wikipedia for each considered month. Then, the titles of these articles are submitted to general Internet search engines and the ranking position of the corresponding Wikipedia articles in the result lists is registered. A previous study by the same author had reported that a very few percentage of the most visited articles in Wikipedia corresponded to typical academic contents. On the contrary, these articles were related to entertainment shows, fictitious characters, TV series, sexuality or celebrities. According to the author, approximately a 70% of the traffic directed to Wikipedia come from result lists generated by portals acting as search engines. Furthermore, Spoerry's article examines the ranking position achieved in the lists of results by the most visited Wikipedia articles to determine the impact of the search engines on the Wikipedia articles most requested by its users. The author uses the *WikiChart* tool to obtain the most visited articles in Wikipedia. This tool is not currently in use and its most important successors, will be described in the next section. The developed examination includes a categorization of the most visited articles according to a set of established categories that can be considered as tags assigned to the articles. The study merge together the most visited articles corresponding to the five months to produce a list of the unique articles visited through all the months. The first result is the distribution of the number of unique articles from the aforementioned list found in each of the months. Then, the total number of articles corresponding to each category is presented as well as the distribution of the articles in each categories over the different months. The article also refers the high degree of overlap between the most visited articles in Wikipedia and the most repeatedly submitted queries to the search engines. This despite of the fact that the lists with the most searched topics, regularly provided by the corresponding engines, are previously sanitized to avoid the inclusion of subjects related, for example, with explicit sexuality or drugs, the authors find a high degree of overlap between the most visited articles in Wikipedia and the most repeatedly searched queries. More in detail, when the author determines the ranking position of the Wikipedia's most visited articles in the result lists from the web engines, he found that moren than the 90% of the most visited Wikipedia articles appeared among the top ten positions when the corresponding engine was queried about a topic similar to the title of the Wikipedia articles. The studied search engines included Yahoo, MSN and, of course, Google. These findings could be used to confirm that the visits to Wikipedia articles were being fueled by the common Internet search engines.

Urdaneta *et al.* performed in [UPvS07b] an analysis over the traffic directed to all the Wikipedia editions and, particularly, over the requests directed to the English one. The analysis was performed on a sample consisting in the 10% of the traffic corresponding to 108 days. Requests involving read and write operations were considered and the analysis examined the load variations and the URLs requesting non-existent pages. The main aim was to offer alternative supporting architectures and data management techniques allowing an adequate scalability of the server system supporting Wikipedia. The authors consider three main approaches: Replication, caching and distribution. This analysis also identified several types of request and presented their relative frequency. Moreover, it grouped the URLs according to the targeted Wikipedias finding that more than the 90% of the traffic was directed to ten most popular Wikipedias. A deeper analysis was undertaken with the English Wikipedia. In

this case, the authors studied several variables at the page level such as the distribution of popularity in terms of number of requests and number of save operations, the format in which pages are read, and the ratio between save and read operations. One interesting aspect of this study is the possibility of comparing the traffic sample with a snapshot of the database obtained the day after the last one of the considered period. In fact, this was done in order to asses the validity of the sample. In this way, for each page, the number of save operations found in the sample was compared with the number of the same operation reported by the database snapshot. As a result, pages with highest numbers of save operations presented an small difference between the two measures. Regarding the popularity of articles, this work concludes that the number of visits to the Wikipedia articles does not follow exactly a typical Zipf distribution. Instead, the group made up of the more visited articles presents number of visits not decreasing as fast as in the Zipf distribution, the bulk group of articles complies with the Zipf law and, again in the bound, the articles with less visits have smaller number of visits than the predicted by the Zipf distribution. In the case of save operations, its distribution decreasingly ordered per page is closer to a typical Zipf law. When studying the number of read and save operations over each particular page, the analysis concludes that both numbers are correlated, so the most popular articles are also the most updated ones. The article also presents the correlation between the number of requests for reading articles in any format and the number of them asking for the default HTML version of the targeted articles. Finally, the impact of indirect save operations caused in cache misses because of write operations over included pages and URLs requesting non-existing articles is also discussed. The distribution approach to improve the Wikipedia Foundation supporting architecture is addressed in [UPvS07a] where a decentralized system is proposed.

Viégas *et al.* developed a new method to study the evolution of the contributions submitted to a given Wikipedia article over time. This new method was presented in [VWD04] and is based on a software tool, the *History Flow* application, capable of translating to a colored map the different additions, deletions and modifications performed on the contents of a given Wikipedia article through its revision history. Using this tool, Viégas *et al.* studied in [WVH07] different patterns describing the activity and interactions of the Wikipedia users when performing their different contributions. To do so, they used a new data visualization called *chromograms* which consisted in diagrams picturing users contributions over time and where different interactions are plotted in different colors.

Adler *et al.* analyze in [AdAPV08] the use of different measures to determine how users are contributing to Wikipedia. The authors introduce an approach based on the consideration of both quality and quantity measures as the parameters characterising authors' contributions. In this way, two measures related to quality, text longevity measure and edit longevity measure, are incorporated to the analysis. These measures were found able to reward properly quality contributions as well as to cause that short-term ones get low ranking marks. This particularity makes them suitable of being used to model user behavior and, because of this, can be used to detect and quantify deliberate introduced vandalism or to consider contributions devoted to repair vandalized articles. Moreover the idea is that such kind of measurement system could be easily integrated in a content-driven reputation system such as the aformentioned one described in [AdA07].

Ortega *et al.* presented in [OGBR07] a classification of the Wikipedia articles according to their length in bytes. Authors estimated that two great subpopulations of articles co-existed inside the Encyclopaedia: tiny articles (less than 200 bytes in length) and standard (greater or equal than 200 bytes in length) ones. In this way, authors found a direct relationship between the contribution level of authors in a given edition (measured in terms of their number of edits) and the resulting length of the articles corresponding to that language edition. Ortega and Barahona analyzed in [OGB07] the production process followed to build the Wikipedia articles. Moreover, they identified the nucleus of authors responsible of the majority of the changes introduced on the Wikipedia articles and determined

the way in which their behavior evolved over time. In this way, they validated previous results obtained by Kittur and Chi and obtained new activity patterns when classifying authors by their contributions in particular periods of time instead of considering their whole activity since the Wikipedia inception. According to the authors, although the number of contributions stemming from the users with the least contribution rates are increasing, more than the 90% percent of contributions corresponding to each month were being sent by a corpus of very active users. Continuing this research line, the same authors concluded in [OGB07] that there was an important inequality on the contributions sent to Wikipedia and, specifically, by the 15% of the authors would be responsible of approximately the 80-85% of all the contributions submitted to Wikipedia. In his doctoral thesis, Ortega completed the quantitative analysis of the top-ten Wikipedias according to their number of articles. Among his most important findings, Ortega concluded that several parameters such as the number of active logged authors, the number of articles and the number of revisions have reached and steady state from approximately summer 2006. Talking about coordination about authors in the top-ten Wikipedias, Ortega found very different ratios of talk pages that indicate very different attitudes to the discussion of the contents exhibited by the corresponding communities of users. The survival analysis developed as a part of this thesis revealed an important difference between the authors that stop contributing and leave the project and the new ones enrolled in the content production. According to the author, this difference could even be used to explain the steadiness in the evolution of the aforementioned parameters. Regarding the featured articles, Ortega states that these articles are older than the common articles and present higher numbers of participant authors and revisions. In summary, Ortega presents an scenario in which the inequality level of the contributions is biased towards the core of active authors and the lack of new core members constitutes a considerable risk for the scalability of the Wikipedias.

Reinoso *et al.* started to analyze the users' requests making part of the traffic directed to Wikipedia in [RGBOR08]. This study consisted in an initial examination of the possibilities, in terms of traffic characterization, brought by the analysis of different information elements contained in the Squid log lines offered by the Wikimedia Foundation to research institutions. Although these log lines were completely anonymized and did not include all the data registered by the Squid servers, some of their fields provided specific information such as the date, the HTTP method or whether the request caused a write operation. A part from these information elements directly obtained from the log lines, the study identified several others that could be parsed from the URL field included in each log line. These information elements basically consisted in the language edition pointed by the URL, the corresponding Wikimedia Foundation project, the targeted namespace, the requested action and the title of the article involved in the request. In order to obtain all the aforementioned elements from the log lines, a tailored application, which constituted the origin of the *WikiSquilter* project was developed. The retrieved data were used to perform a quantitative analysis that presented the daily and weekly distribution of the users' requests as well as the ratios of the requests directed to each namespaces and the percentages corresponding to each requested action. Moreover, the study concluded that there was a strong correlation between the total number of requests and the ones directed to articles in the main namespace. As the abovementioned work covered a time period of only a week, the authors extended the analysis in [RGBRO09]. This new study included the analysis of the log lines corresponding to six weeks, each belonging to a different month from November 2007 till April 2008, and submitted to the twenty most visited editions of Wikipedia in the same period. In this case, the main goal was to verify that the temporal distributions of the users' requests corresponding to the weekly intervals taken from different months were similar to the ones found for the week considered in [RGBOR08]. In effect, the averaged distribution of the requests throughout the hours of the day presented the same shape, as expected, in all the considered weeks. The same occurred regarding the evolution of the number of requests through the days of the week in all the considered weekly periods. Moreover, the article

concluded that the distribution of the users' requests according to the targeted namespace as well as the percentages of the different types of actions presented a very similar tendency in all the considered weeks. In this way, this thesis is intended to broaden even more the time period of the analysis so that it covers a whole year. In addition, to obtain new metrics describing behavioral patterns followed by users, other information elements, such as the searched topics or the articles' titles, have been added to the analysis. The titles of the articles were extracted, although disregarded, in the previous works but included in the analysis performed as part of this thesis to relate all the requests involving the same article. As previously mentioned, this research work pays an special attention to the featured articles dynamics. This topic was previously addressed in [ROGBH10] where the authors analyze the influence of the promotion of articles to the featured status in their subsequent number of visits and editions. The analysis considered featured articles in different editions of Wikipedia and found that only in the English edition the consideration of an article as featured had a relevant impact over its number of visits and editions.

2.5.2 Initiatives to provide statistic information about the use of Wikipedia

Several initiatives have been developed to provide accurate and descriptive enough information about Wikipedia because of its dimension of phenomenon and its popularity among the users of the Internet. These initiatives present statistical information about several aspects of the web site: traffic volume, growth evolution, number of articles, most frequently visited pages, different ranking positions of the site, etc. All this information is really valuable even though some of the initiatives are not maintained any more, cover very specific sets of articles or time periods or concerns very few information elements as representative of the interaction between Wikipedia and its users. However, I will consider in a very special all the information emanating from the own Wikipedia supporting system because it can be used for assessing the validity of the results of our analysis.

The Wikimedia Foundation system staff has set special pages [3] devoted to collect statistical information not only about the Wikipedia itself but also about the rest of the supported wiki-based projects. Information accessible from this page covers visits counts, number of articles, traffic rates, size comparisons, populars pages and several other topics. For example, Figure 2.1 presents a page providing information about the number of articles, administrators, registered users etc. in the English Wikipedia. As another example, there is a page [4] that offers information about the most active Wikipedians of the English Wikipedia according to their number of editions.

Information about raw traffic is offered in several pages automatically updated and it can be obtained in several time scales. This information can be useful in order to assess the traffic observations obtained from our study and to check particular non-regular situations such as specific traffic peaks. However, these graphics are usually offered exactly as this and there is no possibility of any kind of customization in order to study traffic variations in more specific periods. Figure 2.2 shows the workload of the Wikimedia Foundation servers in terms of the number of received requests in different time scales. As we are receiving log lines corresponding to requests sent to all the projects maintained by the Wikimedia Foundation, the results from our traffic characterization could be use as an estimation factor about the overall traffic amount and composition.

Domas Mituzas, hardware officer at Wikimedia Foundation and a member of its advisory board, set up a system to gather information about the most visited pages in Wikipedia. This information is

[3] `http://en.wikipedia.org/wiki/Wikipedia:Statistics` (corresponding to the English Wikipedia)
[4] `http://en.wikipedia.org/wiki/Wikipedia:List_of_Wikipedians_by_number_of_edits`

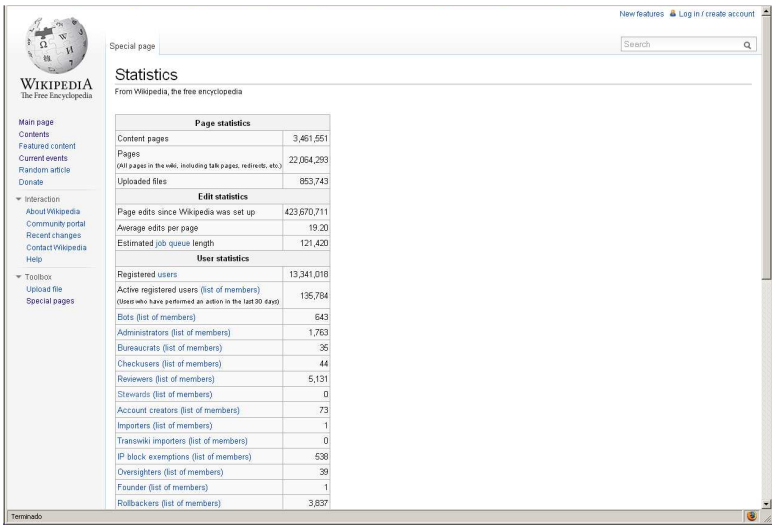
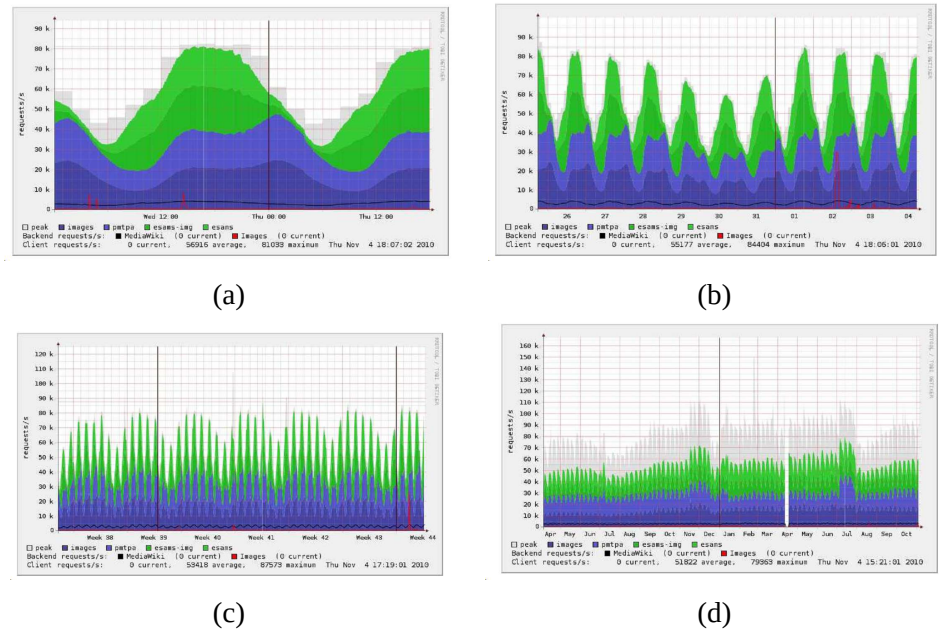


Figure 2.1: Available information about the number of articles, registered users and so forth in the English Wikipedia

offered from Mituzas's portal [5] and consists in per-page view counts taken hourly. In this way, and according to the information provided in its availability announcement (December 2007), registered information reflects the number of pageviews, or visits, corresponding to articles in all the Wikipedia editions that have been requested during each hour and is obtained by applying a regular expression to the URLs logged by the Wikimedia Foundation Squid systems. We have confirmed that these logs are not being sampled or filtered in any way, so the figures based on them that offer several portals and web pages can be considered as absolutes and, consequently, the sampling factor used for our feed can be applied to them for comparison purposes. As far as we are concerned, these data result of great usefulness regarding, for example, the evolution over the time of the number of visits or the differences among the amount of requests directed to each considered Wikipedia edition. However, these data do not offer any information about the requests asking for any type of action or about the topics involved in the search queries submitted by users.

On the other hand, the *Wikimedia Toolserver* [6] is a collaborative platform devoted to support initiatives and software tools involving the wiki-based projects maintained by the Wikimedia Foundation. The *WikiTrends* [7] portal is one of these initiatives and presents the articles with important differences (both positive an negative) in their number of visits. Most trendy articles can be obtained for about 25 Wikipedia editions and in three different periods: current day, week and month. Results are based on the Mituza's pageviews compilation. In a similar way, another tool [8] allows to get the temporal evolution of the number of visits to any article of any Wikipedia edition. As an example, Figure 2.3 shows the evolution of the visits to the *Squid* article in the English Wikipedia during December 2009.

[5] `http://dammit.lt/wikistats/`
[6] `https://wiki.toolserver.org/view/Main_Page`
[7] `http://toolserver.org/~johang/wikitrends/english-uptrends-this-week.html`
[8] `http://toolserver.org/~emw/wikistats/`



(a) (b)

(c) (d)

Figure 2.2: Number of requests per second directed to all the projects supported by the Wikimedia Foundation in different time scales: (a)Daily, (b)Weekly, (c)Monthly and (d)Yearly

As all the Wikipedia contents are licensed under Creative Commons Attribution-ShareAlike 3.0 License (CC-BY-SA) and under the GNU Free Documentation License (GFDL) everyone is permitted to distribute them complying with the terms specified by the licenses. This is not necessary true for images that may be published under privative or copyrighted licenses or their use may be forbidden out of the Wikipedia scope. Because of this, the Wikimedia Foundation regularly offers database dumps containing all the wiki-text basing its articles but do no provide any automatic system for downloading images. Content dumps have been analyzed using different software tools that offer separate visualizations of the data. Some of these visualizations involving Wikipedia topics have been gathered together by Erik Zachte and they are presented in his portal [9].

Precisely Erik Zachte, currently data analyst at the Wikimedia Foundation, maintains one of the most interesting sites devoted to offer statistical information about all the projects supported by the Wikimedia Foundation [10]. This site is monthly updated and, for all the information collected and presented, deserves to be considered, perhaps, as the most exhaustive effort to quantitatively describe the Wikimedia Foundation projects and, particularly, Wikipedia. Among several others, information about the following topics is provided:

- Number of pageviews (i.e. visits), their evolution and their distribution over the different editions of Wikipedia.
- Number of new and total articles and growing evolution.
- Number of new and active registered users (wikipedians) and contributors.
- Database size evolution.

[9] `http://infodisiac.com/Wikimedia/Visualizations/`
[10] `http://stats.wikimedia.org/`

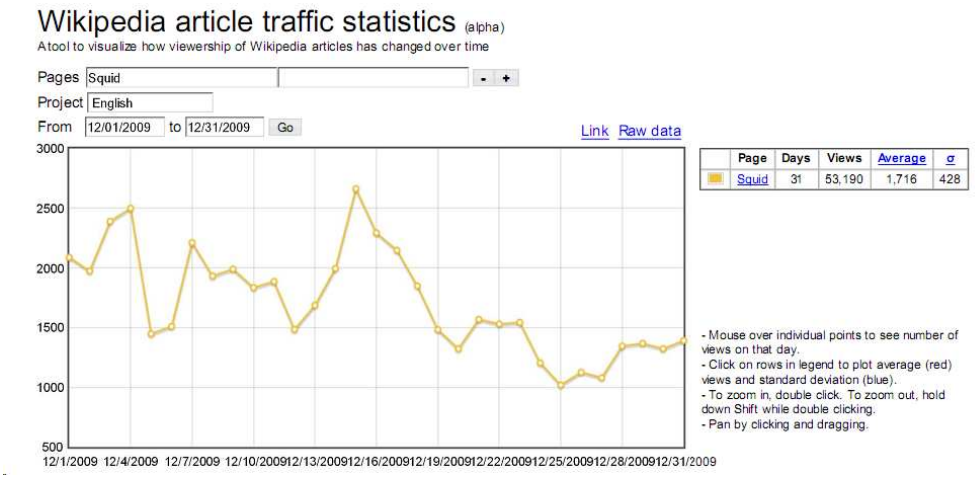


Figure 2.3: Number of visits to the *Squid* article from the English edition of Wikipedia through December 2009

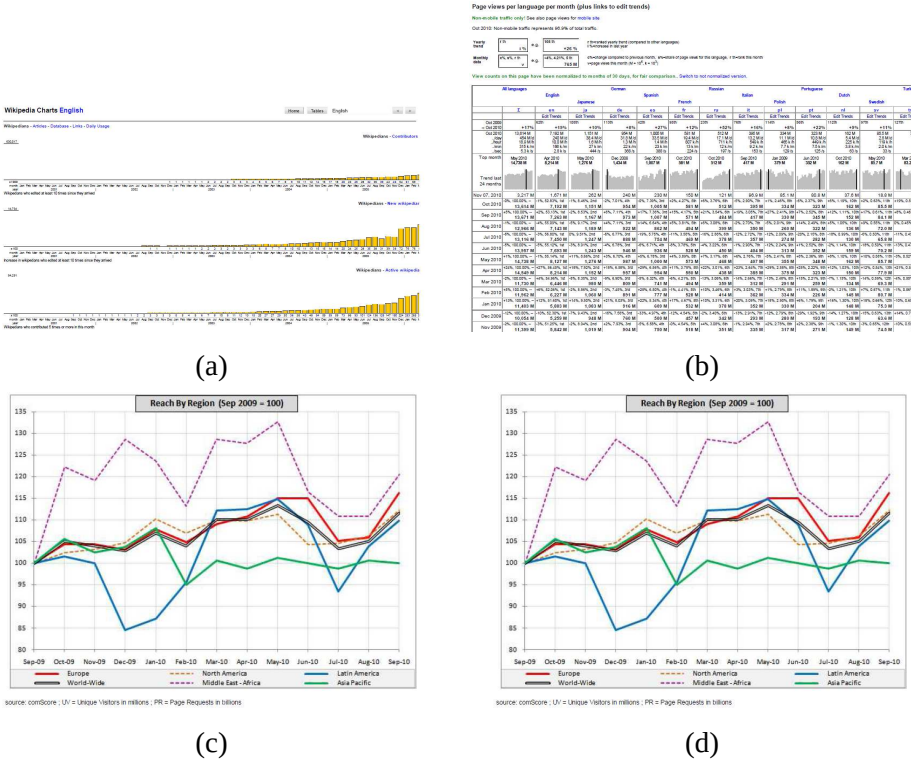
- Evolution of edit operations and their distribution over users.
- Size of articles, number of links, words, etc...

Figure 2.4 presents several examples of both charts and tables available at this site (one of them referring to the statistics provided by Comscore [11])

There is also available a compact version covering exactly one year [12] (from September 2009 till September 2010 when retrieved in November 2010) displaying charts with similar information as the described above. Since January 2010, the site also offers geographical characterization of the origin of the visitors who are visiting or editing Wikipedia articles [13]. In this way it is possible to determine the percentage of all the visits and edits to the Wikipedia issued from each country as well as the editions most targeted from every individual world country or region. This has been done using a 1/1000 Squid log sample covering a period from July 2009 till October 2010. Other informations such as the HTTP requests types, the most popular users' browsers or the number of files daily requested is also provided from these pages. However, this information covers much more smaller periods such as a month or just a fortnight. Moreover, Zachte develops many other activities all related with the analysis of different aspects concerning the Wikimedia Foundation projects and all linked from his portal [14]. Here, one can found graphical animations presenting the growth evolution of the Wikipedia editions, a blog devoted to publish relevant announcements related to the development of his activities and analysis and, of course, the scientific works based on his data.

Featured articles were introduced in chapter 1 and correspond to those articles considered as having an exceptional quality and, thus, deserving the promotion to this status. Information about quantitative aspects related to promotion and demotion of Wikipedia articles is provided from pages such as the one presented in Figure 2.5. Dynamics characterizing featured articles deserve a particular interest because they are relevant indicators of the participation an degree of involvement exhibited by the community of users of particular editions of Wikipedia.

[11] `http://www.comscore.com`
[12] `http://stats.wikimedia.org/reportcard`
[13] `http://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryOverview.htm`
[14] `http://infodisiac.com/`



(a) (b)

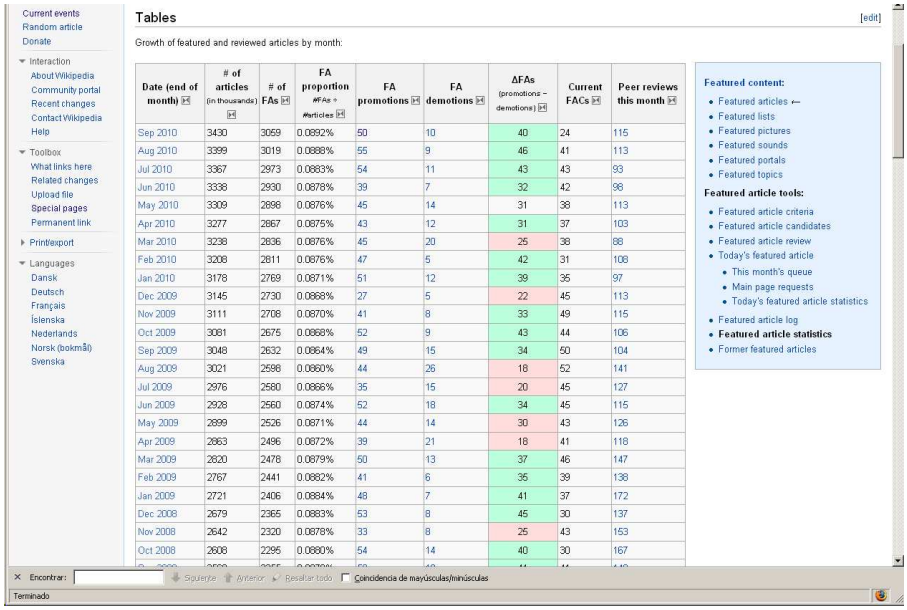
(c) (d)

Figure 2.4: Different visualizations of the data available at the site maintained by Erik Zachte: (a)Chart picturing the growing of contributors, wikipedians and active users (b)Table presenting the number of visits to the different editions of Wikipedia corresponding to each month(c)Number of visits per region(d)Reach values per region (according to Comscore)

Wikipedia also provides a page to present the most visited pages [15] presented in Figure 2.6. The offered list of pages was last updated in 2009 so it is not providing recent information. A smilar page devoted to provide information about the most frequently edited pages is also available [16]. The problem again comes from the fact that the page is not up-to-date. Curiously it is possible to find the most popular (according to their number of visits) per Wikiproject. Wikiprojects is an available space specially thought for collaboration among wikipedians. In fact, a Wikiproject consists in a group of users, usually contributors or editors, that manage a set of articles belonging to an specific topic such as medieval history, industrial design, and so on. The aim is to combine and join efforts to produce quality articles or to improve the existent ones by providing a coordination and organization area for users. Apart from this, the Wikipedia version 0.7, a test release made up of approximately 30,000 articles belonging to all the knowledge areas provides a monthly list with the most visited pages. However, and as previously mentioned, this is not provided for the common Wikipedia. Another not updated attempt is [17] where a list with the most popular articles in the last hour is provided. As its predecessors, the page is not updated any more.

There are some sites providing visualizations of the data collected by the Wikimedia Foundation

[15] `http://en.wikipedia.org/wiki/Wikipedia:Popular_pages`
[16] `http://en.wikipedia.org/wiki/Wikipedia:Most_frequently_edited_pages`
[17] `http://en.wikipedia.org/wiki/User:Emijrp/Popular_articles`



Current events
Random article
Donate
Interaction
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Help
Toolbox
What links here
Related changes
Upload file
Special pages
Permanent link
Print/export
Languages
Dansk
Deutsch
Français
Íslenska
Nederlands
Norsk (bokmål)
Svenska

Tables [edit]

Growth of featured and reviewed articles by month:

Date (end of month)	# of articles (in thousands)	# of FAs	FA proportion #FAs ÷ #articles	FA promotions	FA demotions	ΔFAs (promotions − demotions)	Current FACs	Peer reviews this month
Sep 2010	3430	3059	0.0892%	50	10	40	24	115
Aug 2010	3399	3019	0.0888%	55	9	46	41	113
Jul 2010	3367	2973	0.0883%	54	11	43	43	93
Jun 2010	3338	2930	0.0878%	39	7	32	42	98
May 2010	3309	2898	0.0876%	45	14	31	38	113
Apr 2010	3277	2867	0.0875%	43	12	31	37	103
Mar 2010	3238	2836	0.0876%	45	20	25	38	88
Feb 2010	3208	2811	0.0876%	47	5	42	31	108
Jan 2010	3178	2769	0.0871%	51	12	39	35	97
Dec 2009	3145	2730	0.0868%	27	5	22	45	113
Nov 2009	3111	2708	0.0870%	41	8	33	49	115
Oct 2009	3081	2675	0.0868%	52	9	43	44	106
Sep 2009	3048	2632	0.0864%	49	15	34	50	104
Aug 2009	3021	2598	0.0860%	44	26	18	52	141
Jul 2009	2976	2580	0.0866%	35	15	20	45	127
Jun 2009	2928	2560	0.0874%	52	18	34	45	115
May 2009	2899	2526	0.0871%	44	14	30	43	126
Apr 2009	2863	2496	0.0872%	39	21	18	41	118
Mar 2009	2820	2478	0.0879%	50	13	37	46	147
Feb 2009	2767	2441	0.0882%	41	6	35	39	138
Jan 2009	2721	2406	0.0884%	48	7	41	37	172
Dec 2008	2679	2365	0.0883%	53	8	45	30	137
Nov 2008	2642	2320	0.0878%	33	8	25	43	153
Oct 2008	2608	2295	0.0880%	54	14	40	30	167

Featured content:
- Featured articles
- Featured lists
- Featured pictures
- Featured sounds
- Featured portals
- Featured topics

Featured article tools:
- Featured article criteria
- Featured article candidates
- Featured article review
- Today's featured article
  - This month's queue
  - Main page requests
  - Today's featured article statistics
- Featured article log
- **Featured article statistics**
- Former featured articles

Encontrar: Siguiente Anterior Resaltar todo Coincidencia de mayúsculas/minúsculas
Terminado

Figure 2.5: Information about the number of promotions, demotions and other quantitative data related to the featured articles in the English Wikipedia

tools, specially of the data about pageviews collected by Mituzas. One of the sites [18] currently in use (November 2010) provides information about the number of visits to articles in every Wikipedia edition. From this page we can obtain the number of visits to a certain article in an particular month. As an example Figure 2.7 shows the number of visits to the Squid article (main namespace) in October 2010 for the English Wikipedia. Curiously, you are prompted to get the most visited articles for a given month and a given edition of Wikipedia, but at the moment of trying to get these articles for October 2010, the page refers to the results corresponding to December 2009. Even if we ask for the top articles in January 2009 the page again presents the ones corresponding to December 2009. That means that, as stated in the page itself, this functionality is not working at this moment. Fortunately, as the covered period finishes in December 2009, the results offered by these pages can be compared with the derived from our analysis for the same month.

There are also several sites meant to present the most visited Wikipedia articles or the most popular topics in the enclyclopaedia again after the squid log files collected by Mituzas's. One of these sites is *THEWikistics* [19] that presents the most visited Wikipedia articles till July 2009 but is not updated any more. Figure 2.8 presents the most visited articles in August 2009 as reported by this portal. Another site [20] also presents emerging Wikipedia topics and its initial seed is also the Mituza's log files compilation. Other similar initiatives use their own sources, such as the *Wikirage* site [21] which uses the Special Wikipedia page *RecentChanges* to get the most recently edited articles and, then,

[18] `http://stats.grok.se/`
[19] `http://wikistics.falsikon.de/`
[20] `http://www.trendingtopics.org`
[21] `http://www.wikirage.com/`

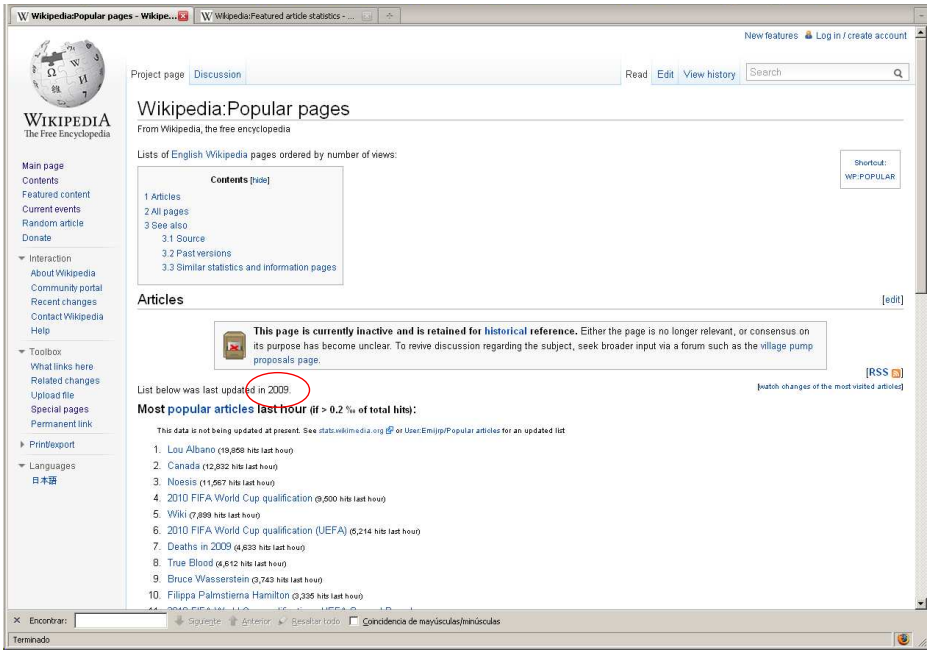


Figure 2.6: Wikipedia popular pages

uses their history page to obtain the type of revision performed. As our analysis also includes a categorization of both the most requested articles and the most repeatedly topics submitted for search to Wikipedia, the information offered by this kind of portal constitutes a very important comparison and reference element.

There are also statistics based on external (non-Wikimedia) data providing valuable information about the requests submitted to Wikipedia. An interesting information source about the traffic received by the Wikimedia Foundation wiki-based projects, and about the Wikipedia in particular, is offered from the *Alexa* web site. This portal provides statistical information about several features of the traffic directed to web pages. To gather all this information, Alexa is constantly crawling the public web sites to periodically build snapshot of the Web status. Moreover, Alexa gets information related to web usage from toolbars or sidebars voluntary installed by users on their browsers and that send to the Alexa servers the URLs they visit. With this information, Alexa offers a traffic rank over the traffic aggregated in a temporal sliding window consisting in the last three months. The rank of a site is determined by combining the measure of *reach*, which is defined as the number of different Alexa's users who visit a page in one day, and *pageviews*, which consists in the number of URLs from Alexa's users requesting the same site considering that the repeated similar URLs sent by the same user in the same day are counted as one *pageview*. That means that two URLs such as `http://mysite/a` and `http://mysite/b` sent by the same user count as two *pageviews* for the site `http://mysite.com`. However if the same user send again in the same day any of the two URLs the number of *pageviews* will remain unchanged. Alexa's accuracy has been object of controversy [22] and this way of determine traffic rank is questioned as susceptible of provide wrong

[22] `http://www.seobook.com/alexa-relevant-2010`

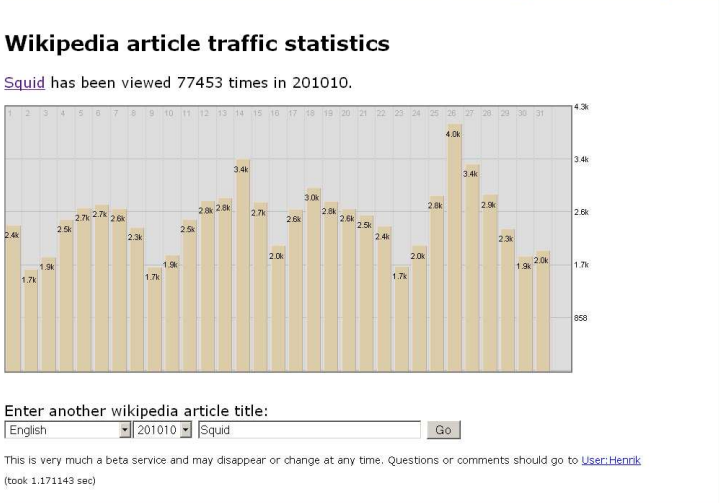


Figure 2.7: Evolution of the number of visit of the “Squid” article in the English Wikipedia during October 2010

values [23].

Figure 2.9 shows both *reach* and *pageviews* values from October 2010 and were retrieved from the Alexa web site on 5 November 2010. So far, as our analysis is concerned, Alexa does not provide absolute values about *reach* or *pageviews* but percents relative to all the data it collects. This fact prevents that we can compare our traffic measures with the ones it publishes. However, Alexa offers valuable information about the most targeted sub-domains of a site. This is interesting because it is related to the amount of traffic received by each edition of Wikipedia and can be used as an element for comparison. Another interesting fact is that Alexa offers the specific queries sent to general search engines that more traffic attract to Wikipedia. This can be useful because we can evaluated if the same terms are also searched using the Wikipedia internal search engine.

In the same line, *comScore* is another company devoted to collect information from joined individuals when they browse the Internet. ComScore users also have any kind of tracking software installed on their systems that regularly reports to the central servers information about different parameters concerning the visited sites. The company estimates approximately in two millions the number of users providing information to the aggregation systems. In order to ensure a representative sample of the different communities of users, comScores uses different recruitment policies as well as demographic validation techniques. As a result, the portal offers important information about the

[23] `http://techcrunch.com/2007/08/13/alexa-says-youtube-is-now-bigger-than-google-theyre-wrong/`

Wikipedia article traffic statistics

Most viewed articles in 200912

Rank	Article	Page views
1	Special:Search	167901984
2	Main Page	141908600
3	404 error/	64616022
4	Special:Random	43880763
5	Search	10962587
6	Brittany Murphy	5255255
7	Avatar (2009 film)	4563366
8	index.html	4380799
9	Wiki	3207697
10	Special:Watchlist	2609126
11	Script kiddie	2452569
12	Lady Gaga	2231952
13	Deaths in 2009	1923763
14	Christmas	1806479
15	HTTP 404	1780605
16	Tiger Woods	1686639
17	YouTube	1654255
18	Avatar	1565635
19	The Beatles	1473478
20	Special:Randompage	1356146
21	Special:Export/Apple Inc	1325587
22	Elin Nordegren	1302223
23	Glee (TV series)	1280753
24	Hanukkah	1213995
25	United States	1144760
26	Michael Oher	1055440
27	Michael Bublé	1045408
28	Facebook	1015937
29	Sex	1005832
30	Pear	994041

Figure 2.8: Most visited articles in December 2009 according to the *THEwikiStics* portal

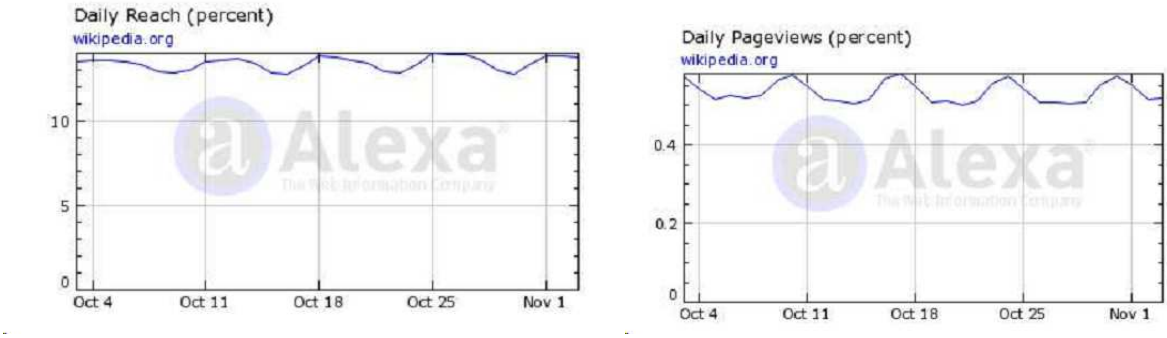
traffic directed to Wikipedia. Part of this information is publicy available [24].

Another site offering information about Wikimedia Foundation projects is the *WikiStatistics* site [25]. In this case users can get the number of both total and new articles, edits, users, files and administrators for all the Wikipedia projects. The temporal period of interest can be adjusted using an intuitive graphical interface or via a parametrized URL. Data used to build the graphs are also provided [26], but without any information about their origin or way of obtaining. Figure 2.10 shows the temporal evolution of the number of edit operations across two different months of 2009.

The *Wikicheker* site [27] focuses on the number of edits performed on the articles of the different editions of Wikipedia and shows articles that may be involved in a war of edits. Moreover, the site offers graphs, as the one shown in Figure 2.11, comparing the number of edits to the different Wikipedia editions. This portal also offers interesting graphs picturing the percentages of edits due to the top 10% frequent edit users and to the rest of users, as well as the differences ratios of edits submitted by logged and non-logged users. Finally the portal shows graphs detailing the evolution over the time of the edit operations submitted by most active users. However, not much information is offered about the way in which edits operations are observed.

An analysis showing the decrease in the number of edit operations was presented by the Wikipedia

[24] `http://meta.wikimedia.org/wiki/User:Stu/comScore_data_on_Wikimedia`
[25] `http://www.wikistatistics.net/`
[26] `http://www.wikistatistics.net/data/`
[27] `http://en.wikichecker.com/`



(a) (b)

Figure 2.9: Reach (a) and pageview (b) values for the site Wikipedia.org according to the Alexas's statistic services

Figure 2.10: Evolution of the number of edits throughout April and May 2009 according to the *Wikistatistics* portal

user *DragonFly* who presented it in a subpage of his Wikipedia user page [28]. According to this work, edit operations had been growing at an exponential rate until April 2007 where they had started to decrease. This analysis is based on a 118,000 article edit sample compiled from the September 2007 database dump. The author suggests that this fact could be related to the so-called *"'Essjay controversy"'* [29] that made it to the headlines in February 2007 when a prominent Wikipedia administrator recognized to have falsified data about his curriculum and also to have used his influence to bias the content of some Wikipedia articles. A similar analysis was conducted by another Wikipedia user [30] who presented an extrapolation of the number of edits performed on the sandbox to approximate the total number of edits to the Wikipedia. This study supported the previous findings by *DragonFly* and perceived a new growing edit tendency, although linear.

[28] `http://en.wikipedia.org/wiki/User:Dragons_flight/Log_analysis`
[29] `http://en.wikipedia.org/wiki/Essjay_controversy`
[30] `http://en.wikipedia.org/wiki/User:Ais523/Stats`

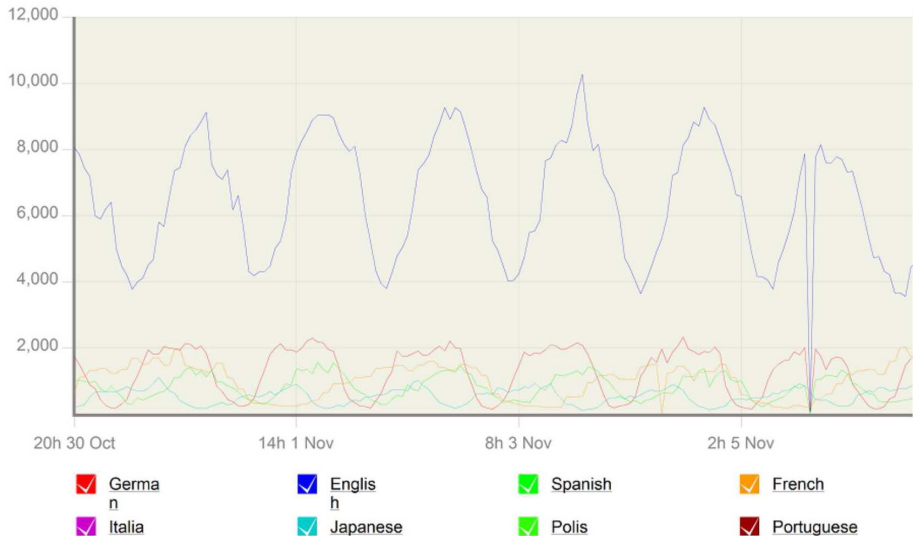


Figure 2.11: Number of edit operations for the most active Wikipedias from 20 October 2010 to 5 November 2010 according to the *Wikichecker* portal

Chapter 3

Methodology

> "Method is essential, and enables a larger amount of work to be got through with satisfaction". Samuel Smiles, (1888).

3.1 Introduction

This chapter describes the methodology conducted to carry out this empirical research and the subsequent study. As presented in Chapter 2, most of the previous research activities involving Wikipedia topics focuses on criteria such as reputation, trust or growth tendencies. Moreover, the very few analysis found dealing with issues related to the use of Wikipedia in different scenarios were considerably restrictive in terms of the considered editions (languages), the size of the taken samples or the cases of use at which researchers paid attention to. For this reason, this work is intended as a wider exam of the ways in which users are interacting with Wikipedia and it is aimed to provide a complete and detailed framework which gathers and discusses the different patterns both temporal and behavioral exhibited by the Wikipedia users when requesting information from it.

The methodology used in this work is mainly based on the analysis of log files containing a large sample of the requests sent to the most relevant editions of Wikipedia during a whole year. The relevance of the considered editions has been regarded according to both size and traffic volume. Such a great number of requests constitutes a meaningful part of the overall traffic directed to the these editions of Wikipedia as a result of the interaction with their users. The analysis of the traffic consists in a characterization based on a parsing process to extract the relevant information elements prior to a filtering one according to the study aims and directives.

Throughout this chapter, an special attention is paid to the different elements of measure identified and used to provide a suitable answer to the main research questions that have motivated this thesis. This set of metrics will lead to our results and will allow us to establish valid models and to obtain right conclusions.

Other different approaches and metrics considered valuable for our research interests but, in the end, impossible to be included as a part of this work are also presented. Most of the drawbacks in this sense are due to technical issues, as in the case of the geo-location tracking of the requests, or to legal questions concerning the individuals' privacy. However, some tools devoted to work with this kind of information have been already developed and wait for the corresponding agreement to be signed. Their contribution to the analysis of the traffic and their research interest are considered valuable and, because of this, they are also discussed here.

So, the rest of this chapter is aimed to provide an exhaustive report about the methodology used to perform the work developed for this thesis. First of all, the general workflow and its most important stages is presented. Then, a rigorous description of the data feed we have used follows. Later, the chapter includes a detailed introduction to the `WikiSquilter` project. This project constitutes the software tool designed and developed to perform the analysis over the log lines and its main features and capabilities as well as the results it can provide are described in detail. Finally, the statistic models suitable of being applied to the obtained data are also introduced.

3.2 Methodology general workflow

The empirical approach used in this thesis is based on the analysis of a sample of log lines containing information about the requests submitted to Wikipedia by its users. This kind of information is offered by the institution supporting Wikipedia, the Wikimedia Foundation, to universities and education centers interested in it for research purposes.

The size of the sample and the way in which it is obtained make it suitable for being considered as meaningful and, hence, the results derived from its analysis can be assumed as descriptive enough to be thought as patterns modeling the way in which users, in general, are using Wikipedia. In order to assure the robustness of our analysis, the results presented here have been validated by comparing them with corresponding metrics provided by Wikipedia own supporting sytem staff as well as by other particular initiatives.

Once the log lines are received in our facilities they are stored conveniently and become ready to be analyzed by the tool developed for this aim: The WikiSquilter project. The analysis consists in a parsing process devoted to obtain the relevant information fields from the log lines. Then, these information elements are filtered according to a set of directives specifying the ones in which the analysis has to focus on. As a result of both processes, necessary data to conduct a characterization are obtained and stored in a relational database for further analysis.

The most important aspects of the methodology conducted to perform the analysis carried out as a part of this thesis are developed throughout the sections of this chapter. These aspects include:

- An exhaustive description of the data sources involved in the analysis.
- The most relevant issues related to the implementation of the parsing and filtering tool as they may result of interest for further developments in the area of wiki-engines log processing or in the case that other researchers may want extend the functionalities of the application by themselves.
- The statistical models used to characterize the users' visits to Wikipedia.

3.3 Data feeding

This section is aimed to describe in-depth the data feeding considered for the analysis presented in this thesis. This feeding consists, basically, in the log lines from the Wikimedia Squid servers because they constitute a valuable way of studying the interaction between Wikipedia and its users. Because of this, they are considered as fundamental for the research presented here.

Therefore, the following sections present the principal issues related to how these log lines are registered, their path to our storage systems and the most important information elements that they contain.

3.3.1 The Wikimedia Foundation Squid system

Chapter 1 presented a general overview of the architecture of the server systems arranged by the Wikimedia Foundation to support Wikipedia and the rest of its other wiki-based projects. In this architecture, the layers containing the Squid servers play a decisive role because they have to manage with all the traffic directed to all the Wikimedia projects.

Squid servers are usually used as proxy servers performing web caching. In this way, they can cache the contents browsed by a group of users to make them available for further requests. This results in an important decrease of the bandwidth consumption and in a more efficient use of the network resources. Furthermore, Squid servers may be used to speed up web servers by caching the contents repeatedly requested to them. Under this approach, Squid servers are said to work as reverse proxy servers because they try to reply to the received requests using the cached contents. This lead to a considerable reduction of the workload of both web and database servers placed behind the Squid systems.

In this way, the Squid operation is based on web caching and it is aimed to avoid the participation of the rest of the database and web server systems in operations for content serving. Thus, when there is a hit and the requested page can be found on a Squid server and it is up to date, the page is directly served from the Squid and neither the database server nor the web server have to be involved in the delivery process. Otherwise, the request is sent to the web servers which elaborate the corresponding HTML code and submit it to the Squid for its delivery to the user. If the page is cacheable, the Squid stores a copy of it for further requests.

The Wikimedia Foundation server architecture places, from the users' perspective, two layers of Squid servers in front of its Apache and database servers. In this way, most of the requested content can be directly served from the Squid subsystem without involving the Apache servers nor the databases in the operation. In particular, Squid servers are able to manage all the requests from non-logged users as the corresponding web pages can be cached because they do not include, in their HTML code, any customization such as the user name or particular skins to be applied when displaying the page. As the Wikimedia Foundation maintains several wiki-based projects, such as Wikipedia, Wikiversity or Wikiquote, the Squid layers have to deal with all the traffic directed to these projects.

Currently, there are two large Squid server clusters: a primary cluster, located in Tampa (Florida, USA), is placed in front of the Apache web servers, databases and media storage systems which are supporting all the wiki projects. Another secondary cluster, located in Amsterdam, performs only web caching. These Squids servers usually run at a hit-rate of approximately 85% for text and 98% for media using CARP (Cache Array Routing Protocol). Users' requests are firstly routed to one of the Squid clusters using a DNS balancing policy.

As a part of their job, Squid systems do log information about every request they serve, disregarding if the sent content comes from the cache or is provided by the web servers. In the end, Squid servers register a log line with different information for each served request and these lines are written to a file or sent to another process through a pipe, as in the case of the Wikimedia Foundation.

3.3.2 The Squid log lines management

Every Squid system deployed as a part of the Wikimedia server architecture puts its log lines into 1450-byte packets and sends them to a central aggregator host, `locke.wikimedia.org` [1]. A program called `udp2log` is running there and is able to log the received lines to several destinations

[1] `http://wikitech.wikimedia.org/view/Squid\_logging`

which may include log files as well as pipes to other processes acting as log processors. Its configuration file, (`/etc/udp2log`), specifies in each one of its lines the destination, file or pipe, and the sampling factor of the corresponding registered log processor.

When this program starts, it reads the configuration file and instantiates a set of log processors initializing each of them with the arguments provided in the corresponding line of the file. Therefore, every log processor will be an instance of the class defining the file processors or of the one defining the pipe processors. After this, the program enters in a loop waiting for packets consisting in buffered log lines and sending each line to the instantiated log processors.

Every log processor checks whether each received line has to be logged according to its sampling factor and, if so, it writes the line to the corresponding file, in case of a file processor, or to the specified pipe in case of a pipe processor.

The log lines used in this analysis are sent from the `udp2log` program to another one called `log2udp` using a pipe processor with a 1/100 sampling factor. The `log2udp` program, in turns, sends a UDP-packet stream made up of the lines to a set of destination hosts belonging to different universities or research institutions as ours. This program includes a reference number in each line that may be used to track possible packet losses.

In the end, a `syslog-ng` client running in our facilities receives the UDP stream containing the log lines and writes them to a log file which is daily rotated. Every rotated file is stored including the rotation date as a part of its name. Thus, there is a log file containing the log lines received since the last rotation, i. e., every line received since the time at which the rotation corresponding to the previous day was performed. Such log files storing the traffic received during a whole day have an averaged size of 900 MB. and contain approximately 40 million log lines.

For our research purposes, it is very important to remark that we are receiving a sample from the central aggregator host of the Wikimedia Foundation. This means that the sample is taken from the log lines sent by the whole set of Squid servers. This assures that we are avoiding the influence of local effects such as, for example, the derived from receiving solely the requests submitted to certain editions of Wikipedia. In this way, we were not able to determine the percentage of the considered types of requests directed to each edition of Wikipedia with respect to the total traffic neither to establish comparisons among the different metrics obtained as they would not be referred to a common portion of the overall traffic.

3.3.3 The Wikimedia Foundation Squid logging format

Every time a Squid server replies to a user request sending the corresponding content, it writes down to a log file the URL submitted by the user or sends it to another process depending on its configuration specifications. Squid servers do not register only the URLs but also some other important data concerning the users' requests. In this way, each Squid log line contains several information fields related to a particular request and can be used as an effective way to trace and to characterize it.

A general purpose Squid server, working as a reverse proxy, provides several log formats to set the information logged as a result of its activity. The Wikimedia Foundation Squid servers use a customized format for generating their log lines which is summarized in Table 3.1. However, we do not receive all this information but just those fields marked as received in the aforementioned Table 3.1.

These fields are conveniently described hereafter:

- **Squid hostname**
  Name of the Squid server sending or writing the log line and responsible of serving the

Field	Description	Received
Squid Hostname	Squid server generating each log line	
Sequence number	Unique sequence number per log line	
GMT time	Current GMT time	Yes
Request service time (ms.)	Total time spent to serve the logged request	Yes
Client IP address	Client source IP address	
Squid request status	HTTP Status code ICP specific	
Reply size including HTTP	Number of bytes transferred to the client (includes overheads) because of TCP/IP headers	
Request method	Request method (GET, POST, etc.)	Yes
URL	URL containing the request.	Yes
Squid hierarchy status	Information about the ICP management	
MIME content type	MIME header corresponding to the URL	
Referer header	URI from where the URL was obtained	
User-Agent header	Information about the agent sending the request	

Table 3.1: The Wikimedia Foundation Squid log format.

corresponding content.

- **Sequence number**
  Unique number generated for each of its log lines by the a particular Squid server.
- **GMT time**
  GMT time according to the Squid own clock. The time is obtained when writing the log line and, therefore, just when the requested content has been sent to the user.
- **Request service time**
  Number of milliseconds that the transaction lasted and, therefore, involved the use of the cache. In the case of an HTTP transaction, this time refers to the interval between the time in which the request was received and the time at which the Squid server finished sending the last byte of the response.
- **Client IP**
  IP address of the client sending the request.
- **Squid request status/HTTP status code**
  This field consists of two code numbers separated by a slash. The first one corresponds to the transaction result whereas the second one. The second token is the HTTP response status code (e.g, 200, 304, 404, etc.). These status codes normally come from the origin server. In some cases, however, Squid may be responsible for selecting the corresponding status code. These codes are defined by the HTTP RFC.
- **Reply size**
  Size in bytes of the response sent to the client. It includes the bytes corresponding to the HTTP headers

- **Request method**
  Specifies the HTTP request method (GET, POST, HEAD, ...) used by the client to request a certain resource.

- **URL**
  The URL submitted by the client specifying a particular content or the requested action. /

- **Squid hierarchy status**
  The hierarchy information consists of three items:

  – A prefix indicating a timeout for the the ICP replies.
  – The way in which the request was handled.
  – The IP address or hostname of the peer node to which the request was forwarded in case of a miss when searching for a given object in the local cache.

- **MIME type**
  The type of the requested content as included in the HTTP reply header.

- **Referer header**
  As specified in the HTTP definitions, the `Referer` field indicates the URI of the resource (site, document, ...) from where the submitted URL was requested.

- **X-Forwarded-For**
  IP address of a client requesting contents through an HTTP proxy or load balancer. It can be used to avoid the anonymisation derived of the use of a proxy server and in order to prevent abuse or malicious behavior.

- **User Agent**
  As specified in the HTTP definitions, this field contains information about the user agent originating the request. This information can be used to produce tailored responses that fit particular users requirements.

In this way, the log lines used as a base for the analysis developed in this thesis are made up of the fields marked as received in Table 3.1. Moreover, the `log2udp` program used to send the aggregated log lines adds to these fields its own sequence number which is independent of the sequence number registered by each Squid server. As a result, every two consecutive lines packed and sent by this program will also have consecutive sequence numbers. This numbers are, thus, received as a part of the log lines and can be used to look for packet looses in the UDP stream containing them.

On the other hand, we are receiving an special field which it is not included in the default Squid logging format and which indicates whether the request caused a write operation to the database. It is a really valuable field because it may be used to identify the URLs requesting editing operations over Wikipedia articles.

Finally, the syslog-ng client that receives the UDP packet stream adds to every incoming line the date and time in which the line is received according to its own clock. This field appears in the first position of the final format of the log lines used in our analysis. As the Squid servers always write their dates and times in GMT, the datetime field added by our system, which operates in the CET time zone, just differs from it in one or two hours depending on the consideration of the daylight saving time. This field is not considered in any way and is disregarded automatically for the analysis in favor of the time indicated by the Squid server. Apart from the current time, the syslog-ng client also registers

the IP of the host from where each line is received. As expected, this IP belongs to the Wikimedia aggregator host `locke.wikimedia.org` and does not change so it is neither considered in any way.

In summary, the log lines received in our facilities are similar to the one presented next. All its fields have been identified and briefly commented to provide a complete description of the final format of the analyzed log lines.

```
(1)May  6 13:46:04 (2)208.80.152.138 (3)22260437 (4)2010-05-06T13:42:43.827
(5)http://en.wikipedia.org/wiki/Arbil (6)- (7)2 (8)GET
```

- (1) syslog-ng datetime
- (2) Wikimedia Foundation aggregator IP
- (3) Sequence number included by the `log2udp` program
- (4) Squid datetime
- (5) Requested URL
- (6) Field indicating a save operation (save) or a read one (-)
- (7) Response time
- (8) HTTP request method

It is also important to note that the log lines we are receiving do not contain any private information susceptible of compromising the users' privacy, such as their IP addresses or any other data suitable of being tracked and resulting in any form of identification. Such kind of information has never been included in the log files used in our analysis. Thus, log files used in this work has been completely anonimised in such a way that they preserve individuals privacy and confidentiality.

3.3.4 Namespaces and actions

Every article in Wikipedia is said to be in a given namespace according to the prefix in front of its title. A Wikipedia namespace defines a set of articles whose title begin with a particular prefix (like *User*, *Wikipedia* or *Talk*) and related because of their nature or purpose. For example, the namespace *Wikipedia* includes all the articles describing important concepts, rules as well as the organization of Wikipedia itself, whereas the *User* namespace gathers all the articles corresponding to the registered users' pages.

Although new namespaces can be added, the number of namespaces in most wiki engines is typically low. In fact, Wikipedia uses ten built-in namespaces[2]: the main namespace, in which every new article is created by default and which has no prefix, and still other nine, each with its own prefix. Moreover, every article in any of these namespaces has its own *Talk* page, which keeps all the discussion issues related to the changes introduced in the contents of the article. All the "Talk" pages corresponding to the articles in a given namespace add to their namespace's prefix the clause *_Talk*, so each namespace is considered to have its corresponding *Talk* namespace. Finally, there are two virtual namespaces, *Special* and *Media* not properly related to articles. In fact they correspond,

Namespace ID	Namespace
-1	**Special**
0	**Main**
1	**Talk**
2	**User**
3	**User_talk**
4	Wikipedia
5	Wikipedia_talk
6	Image
7	Image_talk
8	MediaWiki
9	MediaWiki_talk
10	Template
11	Template_talk
12	Help
13	Help_talk
14	Category
15	Category_talk

Table 3.2: List of namespaces in the English edition of Wikipedia.

respectively, to the pages dynamically generated in response to certain users' requests and to pages providing information about the uploaded files. Table 3.2 summarizes all the general namespaces.

In the following we provide a brief presentation of the namespaces not yet described:

- **Portal**
  This namespace gathers links related to a particular subject and is intended as an organization space devoted to assist the users when browsing and reading the encyclopaedia.

- **File or Image**
  It is the namespace of the pages providing information about the files (images, audio, video, ...) referred from the articles.

- **Mediawiki**
  It is a restricted namespace which associates the pages containing the textual elements to be displayed as a part of the web interface. Users cannot modify articles in this namespace to preserve the web appearance and integrity.

- **Template**
  It is the namespace corresponding to the general code snippets ready to be inserted in articles to make a set of information appear in a common format. For example, articles about rugby teams or noble gases usually include templates to summarize important information.

[2] `http://en.wikipedia.org/wiki/Wikipedia:Namespace`

- **Category**
  It corresponds to a particular category of articles according to a particular classification criterion such as musical or film genres.

- **Book**
  It is the case of collections of Wikipedia articles which can be easily saved or exported to a printable version.

- **Help**
  It includes the articles describing the use of the main features and functionalities of Wikipedia itself and its supporting software. It also serves as reference manual describing the proper ways to perform the most common actions as well as the advanced operations. Moreover, it presents the appropriate behavioral guidelines.

In any language edition, the titles of the Wikipedia articles consist of two parts, an optional namespace and the title properly said, separated by a colon {:}. As previously said, articles in the main namespace do not include any prefix and, because of this, if the title of a page contains a colon, but its initial part is not one of the pre-defined namespaces, that page is considered to be in the main namespace.

Namespaces are usually translated to the language corresponding to each edition of Wikipedia. Therefore, the *Talk* namespace is referred as the *Talk* namespace in the English Wikipedia but as the *Diskussion* namespace in the edition corresponding to the German language. This is extremely important, specially in the filtering process because the namespace has to be checked accordingly to the language edition to which the URL is directed.

Apart from visits requesting the contents of articles in any given namespace, users usually ask Wikipedia to perform different types of actions. The most common ones are listed below:

- **Edit requests**
  An edit request is submitted every time a user clicks on the *edit* tab of any Wikipedia article. In response, the user gets the wiki text of the article inside a basic editor that allows to change its content or to add any contribution in an easy way.

- **Edits**
  Actions resulting in write operations to the database. They correspond to the creation of new content and, also, to the revisions and contributions made by the users to the contents of existing articles in Wikipedia.

- **Submits**
  They correspond to the requests for previewing the result of the changes performed in an article, for highlighting the changes introduced in a particular revision or between two given ones. Article previewing involves rendering its given wikitext in the corresponding HTML code so it can be displayed in a web browser. The overall process does not include database operation but just needs the web server support. Usually, these actions are submitted to obtain a preview of the introduced changes and prior to ask for the save operation.

- **History**
  These actions are requested to obtain a page summarizing the consecutive versions of an article caused by the introduction of the users' contributions. The dates of the revisions are also presented allowing to picture the time-line of the evolution of the article.

3.3.5 Featured articles

Featured articles are considered the best articles all over Wikipedia. In order to be promoted to this status, articles, first, have to be nominated and included in an special page as candidates to featured articles. Usually and prior to the their nomination, future candidate articles pass through a peer revision process in which reviewers make suggestions to improve their quality.

Featured article have to meet a set of criteria apart from the requirements demanded to every Wikipedia article. These criteria cover from a clear and comprehensive writing of the article to a proper structure and organisation. Other aspects such as stability, neutrality as well as length and citation robustness are also considered.

When an article is nominated for the status of featured article, editors and reviewers must built a consensus on whether the article satisfies or not the established criteria. The Wikipedia featured articles director (or one of his delegates) determines if the consensus has been reached and, consequently, the nomination has to be promoted or archived.

After having been promoted, featured articles which no longer meet the described criteria will face a two-step reviewing process aimed to cover their lacks or to extinguish their consideration as featured. In the first step, reviewers make suggestions about how the article could be improved in aspects such as formatting, comprehensiveness or accuracy but without pronounce on its permanence in the featured list. If there is not consensus after this first stage, the article has to face the second step in which participants have to declare their position in favor or against the removal of the article from the featured article list. Every pronunciation has to be presented accompanied by the corresponding arguments are will likely be subject of discussion. Finally, when participants reach a consensus, the article will be removed from the set of the considered as featured or let in this group.

As far as our research is concerned, the consideration of an article as featured can have a notable influence over its number of visits during the period near its promotion and may also affect to the number of contributions received during the same period. In this way, a promotion to the featured status may result in a meaningful alteration of the pattern of accesses to the page of the article. Moreover, the analysis carried out to obtain some kind of patterns which can serve to model the traffic to an article after being considered as featured is also presented here. Finally, this work evaluates the main differences among the several access patterns to the featured articles found in the considered editions of Wikipedia and also the propagation of this kind of articles across them.

3.3.6 The data feeding in detail

The analysis presented here is based on a sample of the Wikimedia Foundation Squid log lines corresponding to the whole year 2009. As the used sampling factor has been 1/100, it means that this study has involved the analysis and characterization of the 1% of the overall traffic directed to all the projects maintained by the Wikimedia Foundation during that year. In general terms, aproximately more than 15,000 million log lines have been processed in order to be characterized accordingly to the directives of the analysis. In order to avoid storage problems derived from this huge amount of information, the log lines have been processed, filtered and stored month by month. This makes the analysis easier and results in more manageable database tables which, in any case, hold, each of them, about 90 million rows.

The analysis developed as a part of this thesis has focused on the traffic directed to the Wikipedia project. In order to ensure that the analysis involved mature and highly active language editions of Wikipedia, the requests corresponding to the ten largest editions in January 2009, according to their number of articles, have been considered. Moreover, these editions were also the top-ten ones

Code	Language	Articles	Monthly pageviews (in Millions)
EN	English	2,700,000	5,615 M
DE	German	888,000	1,271 M
FR	French	757,000	489 M
PL	Polish	571,000	379 M
JA	Japanese	563,000	1,020 M
IT	Italian	540,000	324 M
NL	Dutch	516,000	154 M
PT	Portuguese	453,000	174 M
ES	Spanish	436,000	526 M
RU	Russian	354,000	244 M

Table 3.3: Top-ten editions of Wikipedia according to their volumes of articles (January, 2009).

regarding their volumes of traffic (also in January, 2009) which represented by the 91% of the overall traffic directed to all the editions of Wikipedia. These editions are summarized in Table 3.3 ordered decreasingly by their number of articles.

As in other previous analyses such as [RGBOR08] or [RGBRO09], this thesis focuses on the *Main*, *Talk*, *User* and *User_Talk* namespaces. Additionally, and given the case that this study focuses on the search operations submitted to Wikipedia, the *Special* namespace has been also included in the analysis because it is the one corresponding to the pages generated in response to the users' requests asking for this type of action.

In respect to the actions, this analysis focuses on the ones consisting in *edits*, *edit*, *history* and *submit* requests because they represent the most common types of interaction between Wikipedia and its users. URLs specifying search operations for particular topics are not considered properly actions because we are assuming that actions have to be requested over concrete articles. In this way, search actions are always filtered associated to the *Special* namespace, whereas the rest of actions are filtered considering the article and namespace to which they are being applied.

Although only the normalized information corresponding to the namespaces and actions abovementioned is stored into the database, the application performs a complete characterization of the overall traffic providing quantitative results in result files. This information allows to determine the percentage of the overall traffic directed to each project maintained by the Wikimedia Foundation and, more important, the number of request pointing to the different editions of Wikipedia. Moreover, we also estimate the amount of traffic received by the different Wikipedia editions in each day of 2009 and, even, the distribution of this traffic according to the day of the week in which it is generated.

3.4 The WikiSquilter application

This section presents the main features of the software tool designed and developed to process the data feeding used in this analysis in the aim of characterizing the requests submitted to Wikipedia. It is a Java written application which, basically, parses the log lines from the Wikimedia Squid systems to obtain several information elements contained in their fields. When these information elements comply with the analysis directives, the corresponding lines are considered of interest and, thus, they

are filtered by the application. Information elements from the filtered log lines are finally normalized and stored into a MySQL database for further analysis. The Java Language has been chosen as the implementation language because of its maturity and popularity as well as because it offers a complete and powerful API to develop multithreaded applications. On the other hand, the proved efficiency of the Java drivers when communicating with databases allows excellent performance ratios for operations consisting in massive insertions and data recovery. This two capabilities are decisive for this work because of the huge amount of data to be processed.

The application has been developed under the name of *The WikiSquilter project* after a capitalization of *WIKImedia SQUId Log filTER* and considering the fact that a "'skilter"' is some kind of filtering system for water commonly used in aquariums and fish tanks. The *WikiSquilter* project has been released under the GPL v.3 license and it is available at:

`http://sourceforge.net/squilter`.

This tool has been developed with a strong adherence to four important principles of modern Software Engineering: robustness, extensibility, flexibility and efficiency. The application is really robust and, indeed, it has been able to classify and characterize every single log line contained in the log files used in this thesis. That means that it has rightly parsed and filtered more than 14,600 million log lines. That it is very important because each log line contains the corresponding URL submitted by a user and its analysis results in a really intricate task because of questions such as the language, the translation of the namespaces, the use of different sets of characters (including oriental and Arabic alphabets) and the complexity of the ones requesting different kind of actions such as searches or editions.

Extensibility has been another leading argument. The module devoted to the data definition and management, and the ones devoted to the processes of parsing and filtering have been completely differentiated and their coupling reduced to the minimum. This results in an easy-to-extend application with on a modular design based on the fundamental principles of the Object Oriented Programming such as inheritance and polymorphism. In this way, if new fields are added to the logging format in the future, their processing by the application will require a data definition for the database (in case they are supposed to be stored), an entry for the parser so they can be itemized, and, finally, a filtering directive which specifies the elements considered interesting for the research.

Flexibility is achieved by making the analysis parameters fully configurable. In fact, when the application starts, it builds a logical structure according to the specifications given in a XML file. This file contains the elements which have to be filtered because of their consideration as meaningful for the analysis. The logical structure will serve as the basis for the filtering operations but also as a counting mechanism capable of manage several measurements which will permit the application to provide a useful set of quantitative results just when the analysis finishes.

Efficiency is gained, fundamentally, in two ways. First of all, the application runs under a multithreaded approach in which an activity thread is launched for each log file to be processed. In this way, an independent thread undertakes the analysis of each particular file. This improves notably the overall performance of the application mainly because it allows to take advantage of multi-processing platforms. Moreover, each thread maintains a dedicated connection with the database in order to avoid possible bottlenecks or contentions when multiple access to store the filtered data are needed. The other decisive issue is the performance when filtering the parsed information. In this case, efficiency is achieved with the use of the logical structure supporting the filtering process that is able to determine if an element has to be filtered with O(1) complexity due to its hash-based internal mechanism.

The class diagram corresponding to the *WikiSquilter* application detailing the different implemented classes and the relationships among them is presented in Figure 3.4. In addition, a description of the most important functionalities developed by each class is also included next.

- **WikiMediaProjectSAXParser**
  Defines the SAX parser to be used to process the configuration file specifying the information items considered of interest for the analysis and, thus, to be filtered. Its most significant method parses the configuration file to extract the information elements to be filtered and stores them in the *Filter* object.

- **SquidLogFileProcessor**
  Class corresponding to the thread objects devoted to process the log files containing the log lines from the Squid server systems to be analyzed.

  Its constructor instantiates a new thread to process and analyze the log lines from the Squid server systems contained in a given log file. The analysis of the log lines consist on both a parsing and filtering process to extract a set of information elements from the log lines and to validate them according to the analysis directives.

- **Main**
  Defines the main function of the application which specifies the actions and steps of the algorithm it implements.

- **FiltrableItem**
  Defines the types of information elements forming the URLs submitted by the users in which the analysis will focus on. The application will parse and filter these types of information items according to the directives of the analysis. This class also establishes the maximun number of information items of each type that can be considered of interest and, thus, susceptible of being filtered.

- **FilteredWMProject**
  Defines a Wikimedia Foundation project whose URLs are considered of interest for the analysis and specifies the information to be filtered for this project. This information is referred to the general namespaces, the languages, the actions and the request methods considered of interest for the project.

- **FilteredLanguage**
  Defines the information attributes for a particular language whose URLs are considered of interest and, thus, are going to be processed to filter their information elements according to the analysis directives. As the number of URLs corresponding to each filtered language has to be counted the class inherits from the *FilteredCountedItem* class. The *FilteredLanguage* class includes the set of namespaces objects corresponding to the translation into the defined language of the general namespaces specified for the project for which the language is considered of interest. Each namespace will be represented by a *FilteredCountedItem* object associated to a namespace name inside a *Map* structure. This name corresponds to the translation into the defined language of the corresponding general namespace name and will serve as the identifying string of the namespace for the language. Each *FilteredCountedItem* object representing a namespace name will hold the database code corresponding to the general namespace and the string identifying the translation itself. All the namespaces names consisting on translations of

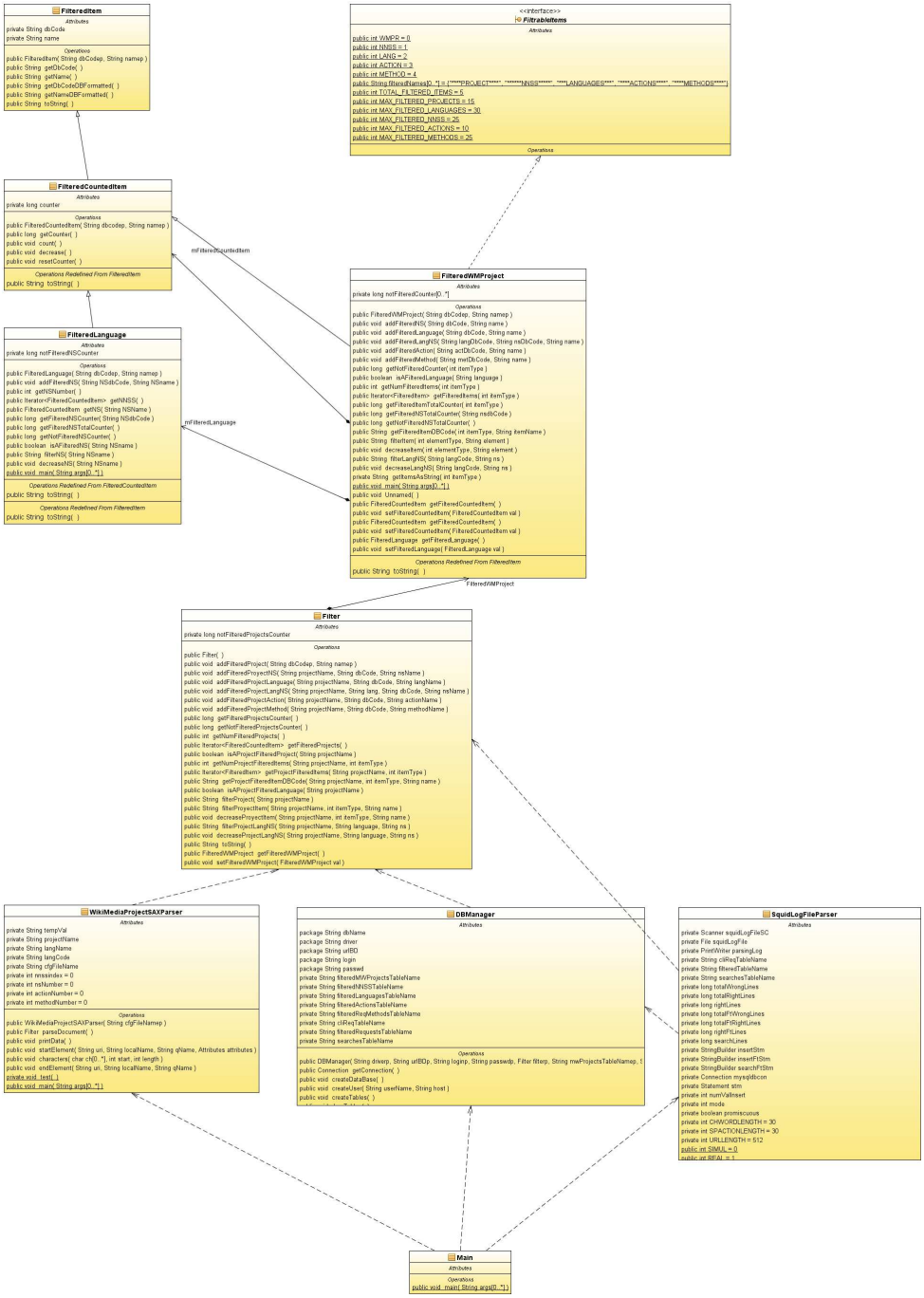


Figure 3.1: WikiSquilter application class diagram.

the same general namespace will be normalized into the same database code. Remember, that general namespaces to be filtered are specified for each particular project.

- **FilteredItem**
  Defines an information element to be filtered.

- **FilteredCountedItem**
  Defines an information element to be filtered and whose number of occurrences is going to be counted for statistical purposes

- **Filter**
  Class holding the different information elements considered of interest for the log lines analysis and, thus, to be filtered. The *Filter* class organizes this information using a set of *FilteredWMProject* objects, defining each the information to be filtered for the corresponding Wikimedia Foundation project.

  It contains a *Map* structure storing the set of objects corresponding to the different projects considered of interest for the analysis. Each Wikimedia project will be represented by a *FilteredWMProject* object and will be stored in the Map structure associated to the string of characters corresponding to its name.

- **DBManager**
  This class is responsible of the database management opereation required to perform an analysis of the log lines from the Wikimedia Squid servers. It includes the creation and destruction of the tables involved in the processing of the log information.

The next sections discuss in detail more questions related to the algorithm used to implement the parsing and filtering operations and also provide a suitable description of the data model applied in the design of the database.

3.4.1 The application workflow

The application receives a set of arguments specifying, among several others settings, the files containing the log lines to be processed. The program, then, launches an independent thread for each indicated file to parse, filter and store the information elements contained in its log lines. The parsing process basically consists in extracting the information elements directly from the log lines fields and, apart from this, it also entails the parsing of the URL contained in the each line. Then the elements are filtered according to the analysis directives and, as a result, only those of interest are stored in the database.

The lines received from the Wikimedia Foundation offer a really valuable information source but they do not include specific information elements to describe certain features of the corresponding requests. However, these elements can be obtained from the URL embedded in each line which, therefore, has to be parsed looking for specific data serving as characterization elements.

More in the detail, the application parser is devoted to obtain the following information elements:

1. The Wikimedia Foundation project, such as Wikipedia, Wiktionary or Wikiquote, to which the URL is directed.

2. The corresponding language edition of the project.

3. When the url requests an article, its namespace.
4. The action (edit, submit, history review...) requested by the user (if any).
5. If the URL corresponds to a search request, the searched topic
6. The title of every requested article or user page name.

From the elements above, both the Wikimedia project and the language can be used to find out the requests directed to each Wikipedia edition whereas the requested namespaces and the performed actions may be put in relation with the aim of the corresponding visits. Determining the title of the articles is specially relevant because it can be used as the linking element able to relate all the URLs requesting the same article in different namespaces or involving it in different actions.

The parsing process often relies on the use of regular expressions for verifying whether an URL, or a part of it, matches a given pattern. If so, its components can be obtained using common functions for string manipulation. For example, when determining the Wikimedia project to which the URL points to, this is the regular expression used to check if it corresponds to the *Wiktionary* project:

```
http://[a-z[A-Z]]{2,3}/.wiktionary.org/.*
```

This suggests that it is absolutely necessary to get an appropriate knowledge about the manner in which URLs are formed and, furthermore, about some of their specific components. On the other hand, URLs requesting articles in a given namespace, such as the *Talk* one, present the following format:

```
http://en.wikipedia.org/wiki/Talk:Squid
```

Apart from the coupling between the article's title and its namespace, the easiest identifiable elements from the URL are the language and the project. However, URLs requesting specific actions or contents can vary significantly and, as a consequence, the task of recognizing all of them become really complex and intricate. As an example, URLs requesting search operations can present different syntactical structures. This supposes a considerable difficulty when obtaining the searched string. These are two different types of URLs asking for a search operation:

```
http://en.wikipedia.org/wiki/Special:Search?search=Linux\&go=Go
```

```
http://en.wikipedia.org/wiki/Special:
Search?search=Linux\&go=Go\&search=Android\&fulltext=Search
```

The parser functions have been developed to be aware of special characters with may cause processing errors because they are special characters (i.e.: meta-characters) in the Java language or in the syntax of the MySQL querying language. Moreover, a major problem is due to the fact that browsers may issue URLs using characters of a given alphabet or their corresponding Unicode representation. The following URLs use, respectively, the colon character (':') and its Unicode codification ('%3A') to separate namespace and article's name can serve as an excellent illustration of this situation:

```
http://fr.wikipedia.org/wiki/Utilisateur:Ajreinoso
http://fr.wikipedia.org/wiki/Utilisateur\%3AAjreinoso
```

URLs belonging to language editions of Wikipedia such as Russian or Japanese are logged using the Unicode representation of their characters. In this way, we have had to obtain the corresponding Unicode representation of the namespaces considered of interest for the analysis as they have to be compared with the ones extracted from the URLs to determine if they have to be filtered or not. As previously mentioned, these namespace names as well as the rest of information elements having interest for the analysis are specified in the XML configuration file (`cfgWPFilter.xml`).

Users request actions by submitting URLs that look like the following one:

```
http://de.wikipedia.org/w/index.php?title=Diskussion:Berlin\&action=edit
```

In a first parsing stage, they are assigned to the fictitious *Index* namespace. This is a namespace used by the application to assign a first characterization to the URLs requesting any action. Once the requested action has been tracked (in the case above, a request for editing the content of the talk page of an article), the application filters it if the action is considered of interest (like in the presented case as it is an edit request). At this point, the title is re-parsed and the proper namespace (the Talk namespace for the German edition of Wikipedia) is obtained. If the action has not interest for the analysis, the URL will remain characterized as in the INDEX namespace and it will not be included in any further statistical calculation nor stored in the database.

The filter process consists in assessing whether an URL has to be considered of interest for the analysis according to directives given for it. This is accomplished by checking whether the information elements it contains, once parsed, has been indicated to be filtered in the configuration file.

The application uses an special hash structure as a part of its filtering entity which is widely described in the next section. This structure gathers all the elements to be filtered as well as their corresponding normalized database codes. The application queries the filter about each information element parsed from the URL to determine if it has to be filtered according to the specifications of the analysis. In case the filter finds the element in its supporting structure, it is considered of having interest and, thus, to be filtered. Then, the filter returns the database code corresponding to the particular information element that will be used in the subsequent insert request to the database. The queries to the filter are issued in such an order that allows to determine the validity of the URL as soon as possible.

The pseudo-code describing the algorithm for the overall parsing and filtering process is presented below.

```
get_reference_number_from_log_line
get_date_from_log_line
get_response_time_from_log_line
get_request_method_from_log_line
get_URL_from_log_line
parse_Wikimedia_project_from_URL
if ( it_is_a_filtered_Wikimedia_project ){
parse_Language_from_URL
filter_Language
if ( it_is_a_filtered_Language ){
parse_NameSpace_from_URL
filter_NameSpace
get_save_field_from_log_line
if (it_is_a_save_action){
```

```
action= 'SAVE'
}else if (NameSpace == 'INDEX' || NameSpace == 'SPECIAL'){
parse_requested_action_or_search_from_URL
filter_Action
if (it_is_a_filtered_Action){
parse_title_from_URL
     re-parse_NameSpace_from_URL
     if (it_is_a_parsed_NameSpace){
     insert_into_Database
     }
     }
    }else if (it_is_a_filtered_NameSpace){
     insert_into_Database
    }else {
     discard_URL
   }
}else{
discard_URL
}
}else{
discard_URL
}
```

Regarding the efficiency and the performance, the application has been developed to optimize as much as possible both the parser and the filtering processes. To do so, the parser operations rely on the effectiveness of the Java regular expressions. These expressions are compiled once into pattern objects which are used, from then on, in every subsequent string verification. The pattern objects consist in a programmatically optimized representation of the regular expression and, because of its immutable nature, are thread-safe so there are not special concerns about synchronization when they are accessed. The optimization of the filtering process, on the other hand, is attained with the use of a hash-based structure as the main part of the filter object. This hashing support allows a O(1) complexity when querying the filter. Moreover, as the structure holds the information elements as well as their corresponding database codes, the validation of an element as filtered results, when successful, in obtaining the normalized value to be used for ins insertion into the database.

According to the analysis directives, the information elements of the URLs considered of interest can be stored in two database tables. One of them stores most of the information elements found in the log lines whereas the other one just registers information related to search operations. There is another table which is used to record general information about absolutely all the processed lines and is populated when the application runs in promiscuous mode. In this mode, the application registers information about all the requests submitted to the Wikimedia Foundation projects apart from the data corresponding to the log lines complying with the analysis directives. Once all the threads have finished, the resulting tables are indexed by the fields more often used when querying the database.

The figures corresponding to the application running times can serve as the best indicators of the efforts made on it to take advance of the benefits of the multiprogramming and to improve the overall performance. In this way, processing the traffic corresponding to a whole month takes approximately 1 day and 6 hours in a quad-core CPU system with 8 GB. of RAM memory. Such kind of traffic

involves more than 1,300 million log lines stored in about 31 or 32 files. Log files are rotated daily so there is a file related to each day of the month. However, to be more accurate, the file corresponding to the next day to the monthly period is also included. In this way, the requests submitted in a given day but stored in the file rotated the next day, because of time differences, are also considered. It is important to remark that the previous running time includes the creation of the indexes for the database tables. Due to the considerable number of rows stored on the different tables and the several indexes to be created in order to speed up future queries, as described in the next section, the indexation process takes approximately 1 day which represents by the 80% of the overall processing time. In summary, the parsing, filtering and storing of the traffic corresponding to a whole month is accomplished, on average, in 6 hours which means a processing speed of more than 60,000 log lines per second.

3.4.2 The filter structure

The most important element taking part in the filtering process is the logical structure containing the elements to be filtered. It is a special type of map structure called
em LinkedHashMap which offers the Java Collections API. The map structures, or associative arrays, hold pairs consisting in a key and the corresponding value. So, given a key, the map can be asked for the associated value. A map can be supported by several types of underlying structures, ranging from arrays to ordered trees, which allow different performance ratios. The most efficient one is the hash table that stores each value of a given set of pairs in a table using the hash code of the key as index. This provides constant-time operations of insertion and recovery over the map. On the other hand, the Java Collection interface allows to get an iterator object which can be used to navigate through the different elements of a particular collection. In the case of maps, the order of the elements returned by two different iterators obtained from the same map can vary if there is not an additional structure to specify a particular order. This is the main feature of the *LinkedHashMap* that maintains a *LinkedList* whose elements point to the objects of the map. In this way, the order of the elements in the list correspond to the order in which the elements of the map were inserted into it. Any iterator requested over the map will navigate through the list, so the iteration order will be always the same. In the Wikisquilter application, the order in which elements are recovered has to be constant because sometimes it is related to the normalization values used for the database operations.

Once the WikiSquilter project Main class is started, it parses the XML configuration file to built up the abovementioned *LinkedHashMap* filter structure. The XML file allows to specify the different elements to be filtered making the application flexible and fully configurable to meet the aims of each specific analysis. The parsing of the XML file is done with the Java implementation for the SAX (*Simple API for XML*) parser interface.

The XML configuration file contains a *WikimediaProject* tag for every project supported by the Wikimedia Foundation whose URLs are relevant for the analysis. For each opening tag corresponding to a Wikimedia project, a database code and a name are assigned as attributes. Following this tag, the set of namespaces considered of interest for the given project are specified. As previously mentioned, namespaces are translated to every particular language but, here, they are specified as a generic list using the names given in the English version of Wikipedia. In order to make the application and the future queries to the database more efficient, namespaces will be stored using a code which does not depend on the language but only on the namespace itself. More in detail, the code for each namespace will consists in its position in the aforementioned list. That means that two URLs requesting articles in the *Talk* namespace for the English Wikipedia and in the *Diskussion* namespace for the German one are both stored with the database code corresponding to the position of the generic *Talk* namespace in the namespaces list. Of course, the languages corresponding to the different editions of Wikipedia

included in the analysis have to be specified as well as their translations for every targeted namespace. Finally, the requested actions to be filtered are established along with the HTTP requesting methods. The additional *INDEX* namespace, which does not belong to the set of namespaces of Wikipedia, is maintained to be assigned to the URLs requesting actions in which the analysis is not interested or in the case that the action has to be filtered but the namespace to which it is referred is not in the filtered list. In any case, URLs assigned to that namespace will not be stored on the database unless the application runs in promiscuous mode. In this case, every URL, filtered or not, is stored in an special table for further analysis.

The content of the XML configuration file used for the analysis performed as a part of this thesis is presented next:

```
<Filter_cfg>
<WikiMediaProject dbCode="0" name="WIKIPEDIA">
  <NNSS_INDEXES>
    <NSINDEX>ARTICLE</NSINDEX>
    <NSINDEX>INDEX</NSINDEX>
    <NSINDEX>ARTICLE_TALK</NSINDEX>
    <NSINDEX>USER</NSINDEX>
    <NSINDEX>USER_TALK</NSINDEX>
    <NSINDEX>SPECIAL</NSINDEX>
  </NNSS_INDEXES>
  <Language dbCode="EN" name="ENGLISH">
        <NameSpaces>
        <NS>Talk</NS>  <NS>User</NS>
        <NS>User_Talk</NS> <NS>Special</NS>
        </NameSpaces>
  </Language>
  <Language dbCode="DE" name="GERMAN">
    <NameSpaces>
      <NS>Diskussion</NS>  <NS>Benutzer</NS>
      <NS>Benutzer_Diskussion</NS> <NS>Spezial</NS>
    </NameSpaces>
  </Language>
  <Language dbCode="ES" name="SPANISH">
    <NameSpaces>
      <NS>Discusi%C3%B3n</NS>  <NS>Usuario</NS>
      <NS>Usuario_Discusi%C3%B3n</NS> <NS>Especial</NS>
    </NameSpaces>
  </Language>
  <Language dbCode="JA" name="JAPANESE">
    <NameSpaces>
      <NS>%E3%83%8E%E3%83%BC%E3%83%88</NS>  <NS>%E5%88%A9%E7%94%A8%E8%80%85</NS>
      <NS>%E5%88%A9%E7%94%A8%E8%80%85%E2%80%90%E4%BC%9A%E8%A9%B1</NS>
<NS>%E7%89%B9%E5%88%A5</NS>
    </NameSpaces>
  </Language>
  <Language dbCode="PL" name="POLISH">
    <NameSpaces>
```

```
    <NS>Dyskusja</NS> <NS>Wikipedysta</NS>
    <NS>Dyskusja_wikipedysty</NS> <NS>Specjalna</NS>
  </NameSpaces>
</Language>
<Language dbCode="FR" name="FRENCH">
  <NameSpaces>
    <NS>Discuter</NS> <NS>Utilisateur</NS>
    <NS>Discussion_Utilisateur</NS> <NS>Special</NS>
  </NameSpaces>
</Language>
<Language dbCode="IT" name="ITALIAN">
  <NameSpaces>
    <NS>Discussione</NS>  <NS>Utente</NS>
    <NS>Discussioni_utente</NS> <NS>Speciale</NS>
  </NameSpaces>
</Language>
<Language dbCode="PT" name="PORTUGUESE">
  <NameSpaces>
    <NS>Discuss%C3%A3o</NS>  <NS>Usu%C3%A1rio</NS>
    <NS>Usu%C3%A1rio_Discuss%C3%A3o</NS> <NS>Especial</NS>
  </NameSpaces>
</Language>
<Language dbCode="NL" name="DUTCH">
  <NameSpaces>
    <NS>Overleg</NS>  <NS>Gebruiker</NS>
    <NS>Overleg_gebruiker</NS> <NS>Speciaal</NS>
  </NameSpaces>
</Language>
<Language dbCode="RU" name="RUSSIAN">
  <NameSpaces>
      <NS>%D0%9E%D0%B1%D1%81%D1%83%D0%B6%D0%B4%D0%B5%D0%BD%D0%B8%D0%B5</NS>
      <NS>%D0%A3%D1%87%D0%B0%D1%81%D1%82%D0%BD%D0%B8%D0%BA</NS>
      <NS>%D0%9E%D0%B1%D1%81%D1%83%D0%B6%D0%B4%D0%B5%D0%BD%D0%B8%D0%B5
      _%D1%83%D1%87%D0%B0%D1%81%D1%82%D0%BD%D0%B8%D0%BA%D0%B0</NS>
      <NS>%D0%A1%D0%BB%D1%83%D0%B6%D0%B5%D0%B1%D0%BD%D0%B0%D1%8F</NS>
  </NameSpaces>
</Language>
<Actions>
  <Action>edit</Action>
  <Action>history</Action>
  <Action>save</Action>
  <Action>submit</Action>
  <Action>search</Action>
</Actions>
<Methods>
  <Method>GET</Method>
  <Method>HEAD</Method>
  <Method>POST</Method>
```

```
    <Method>LOCK</Method>
    <Method>NONE</Method>
    <Method>OPTIONS</Method>
    <Method>CONNECT</Method>
    <Method>PROPFIND</Method>
    <Method>PURGE</Method>
    <Method>PUT</Method>
  </Methods>

</WikiMediaProject>
</Filter_cfg>
```

In this thesis we focus only on the URLs directed to ten editions of Wikipedia and in a particular set of both namespaces and actions. However, the analysis can be easily extended to other projects, languages, namespaces or actions simply by including them in the XML configuration file. This feature makes the WikiSquilter project a versatile tool in order to analyze the overall traffic directed to all the Wikimedia Foundation projects.

As previously stated, after having extracted each information element during the parsing process, the filter is queried in order to determine if the given element is considered of interest and, in consequence, has to be stored in the database. The filter is called under a common function invocation although it is important to note that the filtering of each information element is based on the previous filtered ones. As an example, the targeted Wikimedia Foundation project will be the first information element to obtain and filter. However, the filtering process of the language edition of the project has to considerer the project itself because it is possible, for example, to filter the URLs addressed to the japanese edition of Wikipedia but not to filter them if the project is Wikiversity. Moreover, the WikiSquilter application will allow to parse and to filter specific namespaces, actions and methods for each particular project.

Apart from being used in the filtering operation performed by the WikiSquilter application, the filter structure also serves for accounting purposes and, in fact, keeps the number of filtered items corresponding to each information element: language, namespace, action, etc. This is done by maintaining a counter for each element to be filtered which is increased each time that the given element is found in a submitted URL. As a result, the application is able to offer statistical information about the log files processing immediately after it finishes. This information became available as a summary of the overall processing and does not involves any query to the database. Of course, during the development stage the information obtained in this way has been contrasted with the one held by the database. An example of the data offered directly by the WikiSquilter application is presented below.

```
Total Elapsed Time: 2 days 19 h. 44 min. 14 sec.
TOTAL FILTERED PROJECTS: 1
******NNSS*****
--dbCode: 0 name: ARTICLE Total: 86522371
--dbCode: 1 name: INDEX Total: 116980804
--dbCode: 2 name: ARTICLE_TALK Total: 369946
--dbCode: 3 name: USER Total: 208097
--dbCode: 4 name: USER_TALK Total: 210489
--dbCode: 5 name: SPECIAL Total: 13165404
***LANGUAGES***[
```

```
dbCode: EN Name: ENGLISH NSNumber: 6 Total NS Filtered: 55463209 NNSS: [
--dbCode: 0 name: ARTICLE counter: 48014271,
--dbCode: 1 name: INDEX counter: 57621876,
--dbCode: 2 name: TALK counter: 202555,
--dbCode: 3 name: USER counter: 101850,
--dbCode: 4 name: USER_TALK counter: 96680,
--dbCode: 5 name: SPECIAL counter: 7047853] No Filtered NSS: 4266744,
dbCode: DE Name: GERMAN NSNumber: 6 Total NS Filtered: 12905529 NNSS: [
--dbCode: 0 name: ARTICLE counter: 9589448,
--dbCode: 1 name: INDEX counter: 11127385,
--dbCode: 2 name: DISKUSSION counter: 30568,
--dbCode: 3 name: BENUTZER counter: 31858,
--dbCode: 4 name: BENUTZER_DISKUSSION counter: 18482,
--dbCode: 5 name: SPEZIAL counter: 3235173] No Filtered NSS: 875851,
dbCode: ES Name: SPANISH NSNumber: 6 Total NS Filtered: 5062410 NNSS: [
--dbCode: 0 name: ARTICLE counter: 4510621,
--dbCode: 1 name: INDEX counter: 13044626,
--dbCode: 2 name: DISCUSI%C3%B3N counter: 22164,
--dbCode: 3 name: USUARIO counter: 13036,
--dbCode: 4 name: USUARIO_DISCUSI%C3%B3N counter: 14520,
--dbCode: 5 name: ESPECIAL counter: 502069] No Filtered NSS: 609726,
dbCode: JA Name: JAPANESE NSNumber: 6 Total NS Filtered: 10252415 NNSS: [
--dbCode: 0 name: ARTICLE counter: 9225701,
--dbCode: 1 name: INDEX counter: 10750417,
--dbCode: 2 name: %E3%83%8E%E3%83%BC%E3%83%88 counter: 30389,
--dbCode: 3 name: %E5%88%A9%E7%94%A8%E8%80%85 counter: 11354,
--dbCode: 4 name: %E5%88%A9%E7%94%A8%E8%80%85%E2%80%90%E4%BC%9A%E8%A9%B1 counter: 11430,
--dbCode: 5 name: %E7%89%B9%E5%88%A5 counter: 973541] No Filtered NSS: 515453,
dbCode: PL Name: POLISH NSNumber: 6 Total NS Filtered: 3747542 NNSS: [
--dbCode: 0 name: ARTICLE counter: 3408563,
--dbCode: 1 name: INDEX counter: 6403491,
--dbCode: 2 name: DYSKUSJA counter: 6677,
--dbCode: 3 name: WIKIPEDYSTA counter: 8990,
--dbCode: 4 name: DYSKUSJA_WIKIPEDYSTY counter: 4167,
--dbCode: 5 name: SPECJALNA counter: 319145] No Filtered NSS: 289966,
dbCode: FR Name: FRENCH NSNumber: 6 Total NS Filtered: 4215999 NNSS: [
--dbCode: 0 name: ARTICLE counter: 4076718,
--dbCode: 1 name: INDEX counter: 5876883,
--dbCode: 2 name: DISCUTER counter: 33111,
--dbCode: 3 name: UTILISATEUR counter: 15982,
--dbCode: 4 name: DISCUSSION_UTILISATEUR counter: 21915,
--dbCode: 5 name: SPECIAL counter: 68273] No Filtered NSS: 1106882,
dbCode: IT Name: ITALIAN NSNumber: 6 Total NS Filtered: 3214070 NNSS: [
--dbCode: 0 name: ARTICLE counter: 2855021,
--dbCode: 1 name: INDEX counter: 3278845,
--dbCode: 2 name: DISCUSSIONE counter: 13792,
--dbCode: 3 name: UTENTE counter: 7835,
--dbCode: 4 name: DISCUSSIONI_UTENTE counter: 14360,
```

```
--dbCode: 5 name: SPECIALE counter: 323062] No Filtered NSS: 310118,
dbCode: PT Name: PORTUGUESE NSNumber: 6 Total NS Filtered: 1713328 NNSS: [
--dbCode: 0 name: ARTICLE counter: 1522917,
--dbCode: 1 name: INDEX counter: 3866358,
--dbCode: 2 name: DISCUSS%C3%A3O counter: 7751,
--dbCode: 3 name: USU%C3%A1RIO counter: 4805,
--dbCode: 4 name: USU%C3%A1RIO_DISCUSS%C3%A3O counter: 13287,
--dbCode: 5 name: ESPECIAL counter: 164568] No Filtered NSS: 183072,
dbCode: NL Name: DUTCH NSNumber: 6 Total NS Filtered: 1533098 NNSS: [
--dbCode: 0 name: ARTICLE counter: 1325065,
--dbCode: 1 name: INDEX counter: 1714624,
--dbCode: 2 name: OVERLEG counter: 7101,
--dbCode: 3 name: GEBRUIKER counter: 4838,
--dbCode: 4 name: OVERLEG_GEBRUIKER counter: 9933,
--dbCode: 5 name: SPECIAAL counter: 186161] No Filtered NSS: 159050,
dbCode: RU Name: RUSSIAN NSNumber: 6 Total NS Filtered: 2368707 NNSS: [
--dbCode: 0 name: ARTICLE counter: 1994046,
--dbCode: 1 name: INDEX counter: 3296299,
--dbCode: 2 name: %D0%9E%D0%B1%D1%81%D1%83%D0%B6%D0%B4%D0%B5%D0%BD%D0%B8%D0%B5 counter: 15838,
--dbCode: 3 name: %D0%A3%D1%87%D0%B0%D1%81%D1%82%D0%BD%D0%B8%D0%BA counter: 7549,
--dbCode: 4 name: %D0%9E%D0%B1%D1%81%D1%83%D0%B6%D0%B4%D0%B5%D0%BD%D0%B8%D0%B5_
%D1%83%D1%87%D0%B0%D1%81%D1%82%D0%BD%D0%B8%D0%BA%D0%B0 counter: 5715,
--dbCode: 5 name: %D0%A1%D0%BB%D1%83%D0%B6%D0%B5%D0%B1%D0%BD%D0%B0%D1%8F
counter: 345559]
No Filtered NSS: 374073]
****ACTIONS****[
--dbCode: 0 name: EDIT counter: 1513211,
--dbCode: 1 name: HISTORY counter: 310864,
--dbCode: 2 name: SAVE counter: 109230,
--dbCode: 3 name: SUBMIT counter: 103053,
--dbCode: 4 name: SEARCH counter: 9363612]
****METHODS****[
--dbCode: 0 name: GET counter: 99704420,
--dbCode: 1 name: HEAD counter: 398210,
--dbCode: 2 name: POST counter: 367783,
--dbCode: 3 name: LOCK counter: 200,
--dbCode: 4 name: NONE counter: 0,
--dbCode: 5 name: OPTIONS counter: 5121,
--dbCode: 6 name: CONNECT counter: 0,
--dbCode: 7 name: PROPFIND counter: 563,
--dbCode: 8 name: PURGE counter: 0,
--dbCode: 9 name: PUT counter: 6]
```

As shown, the total elapsed time as well as the number of URLs corresponding to each particular namespace are presented. Following, for each considered language, the number of URLs found in the different namespace are provided. Finally the totals for the different analyzed actions and HTTP requesting methods (without grouping by language) are presented.

3.4.3 The database schema

Because of the enormous amount of data to be processed as a part of this study, the role played by the underlying databases became specially relevant. In fact, the database is intented as the main storage support for all the information elements filtered by the *WikiSquilter* application and the base for the subsequent queries devoted to extract the data involved in the analysis developed in this thesis.

In this way, to conduct properly our analysis I decided to set up two databases. The *squidlogs* database is the largest and most important one and it is meant to store all the information elements from the URLs considered as important according to the directives of our analysis. This database is filled by the *WikiSquilter* application after parsing and filtering the Squid log lines contents. On the other hand, the *analysis* database, is much more smaller and was conceived as the result of a grouping and aggregation process over the data stored in the previous one and involved in the set of statistical calculations developed as a part of this thesis. The main goal pursued with this second database was, of course, the acceleration of all the queries to be issued as a part of the statistical examinations.

Among the different database management systems, the MySQL server was chosen because of its release as free software under the GNU General Public License and because of the availability of a highly optimized driver allowing Java applications to access and manipulate databases through the *Java Database Connectivity (JDBC)* API. Moreover, MySQL offers the special *Innodb* storage engine, specially designed to achieve adequate performance ratios in situations that require to process a large amount of data.

The Entity-Relationship (E-R) diagram of the database storing the information elements from the Squid log lines is presented in the figure 3.4.3. Again, it is important to remark that all the information fields extracted from the Squid log lines are adequately normalized prior to their storage on the database. This result in a great saving of space and improves the performance of the subsequent queries involving those fields.

All the database tables are conveniently described next:

- **FilteredMediaWikiProjects**
  This table holds the Wikimedia Foundation projects considered of interest for an specific analysis and their corresponding database codes as specified in the XML configuration file. The value of the code will be used as the primary key.

- **FilteredLanguages**
  This table keeps the language editions to be filtered for each Wikimedia project defined as an object of the analysis. Languages are stored using ISO 639 2-letter codes and both the project code and the language one form the primary key.

- **FilteredNNSS**
  This table stores the namespaces corresponding to each project in which the analysis focuses on. The namespace codes are assigned basing on the order in which they are specified in the XML file. As in the case of the languages, the primary key consist of both the project code and the namespace one.

- **FilteredActions**
  Table holding the actions submitted by the users that the application will filter. The action codes are also assigned from their specification order in the XML file and constitute a part of the primary key together with the project code.

- **FilteredRequestMethods**
  This table contains all the methods for submitting requests supported in the HTTP protocol and

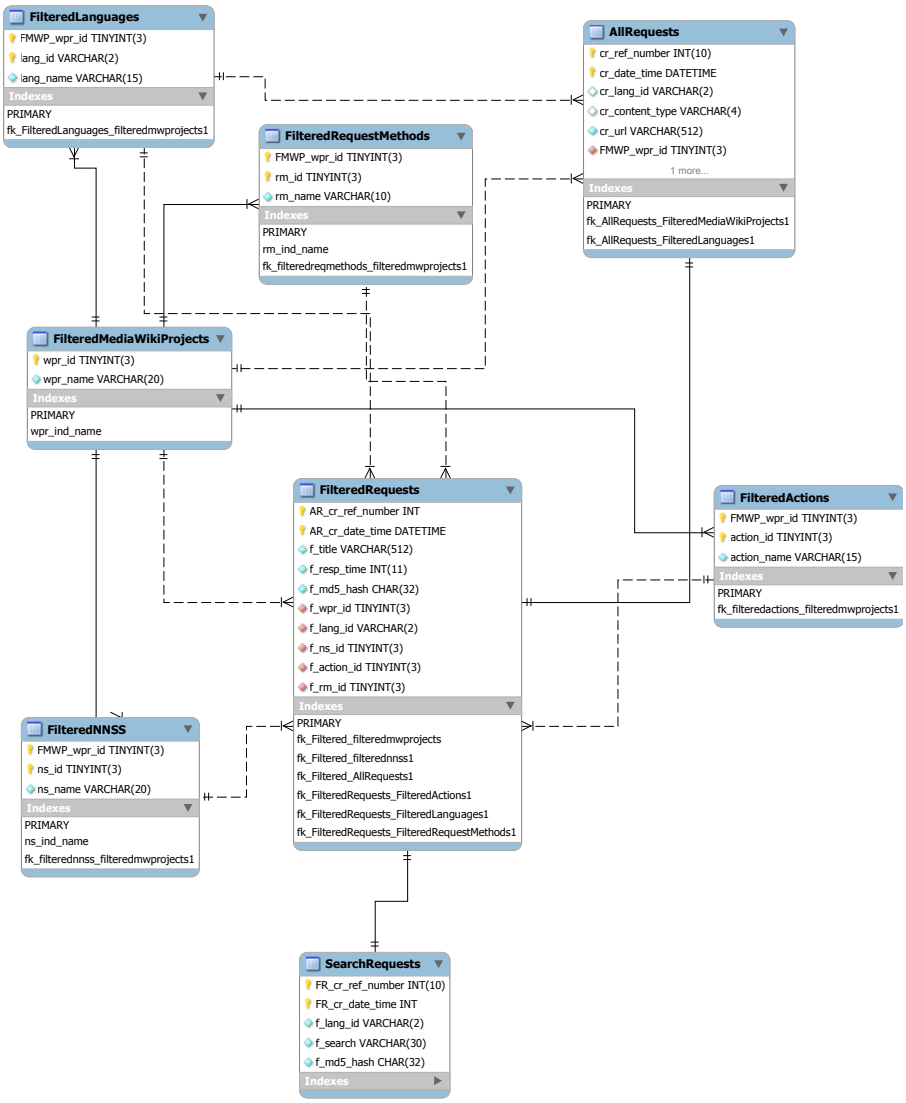


Figure 3.2: Entity-Relationship Diagram for the database used to store the information elements considered as relevant for our analysis.

their corresponding database codes after the specification given in the XML configuration file. As the request methods are specific for each project, their database code and the project one will make the table's primary key.

- **FilteredRequests**
  One of the most important tables because it is used to store information from the URLs filtered by the application as it estimates that their fields meet the criteria established for a particular analysis. Different information elements extracted from the URL are always normalized during the filtering process prior to their storage and, therefore, most of fields in this table act as foreign keys to the previously described tables. It is important to note that as the title of the requested page is stored to correlate different types of requests involving the same article, its computed md5 hash is also stored to speed up the queries having to group the table rows by the article title.

- **SearchRequests**
  This table keeps the strings submitted by the users in the search operations. These strings are held separately because their storage in the *FilteredRequests* table will produce a vast amount of NULL values in the table rows, just one for each URL not requesting a search operations.

  As the *FilteredRequests* table will participate in most of the queries, several indexes are created over its fields. Specifically, an index will be created over each foreign key to another table. Indexes are created after all the rows are inserted in order to avoid excessively slow insert operations.

  In order to improve the insertion process, there are several connections to the database which are maintained separately. In this way, each thread in charge of processing a log file will maintain its own connection to the database to prevent bottlenecks and row blocking issues as a result of the concurrent operations performed by the other threads. Moreover, the insert operations are not sent individually but in 500 row packages in order to achieve a better I/O performance due to the use of larger written operations instead of several individual ones.

- **AllRequests**
  This table maintains basic information about all the requests directed to every Wikimedia Foundation project which is registered when the WikiSquilter application runs in promiscuous mode.

Figure 3.4.3 shows the Entity-Relationship diagram corresponding to the *analysis database*. As previously mentioned, the tables of this database are filled with different results from grouping queries involving the data stored in the *squidlog* database described above. This process has been completely automated by using bash and MySql scripting.

In the following an adequate description of the tables contained by this database is provided:

- **Visited2009**
  This table stores the number of Wikipedia articles corresponding to the different namespaces and language editions considered visited in every day of 2009.

- **Saved2009**
  This table holds the number of Wikipedia articles that have been object of an edit operation resulting in a write operation to the database in each day of 2009. The articles may also correspond to any of the different namespaces and language editions considered in the analysis.

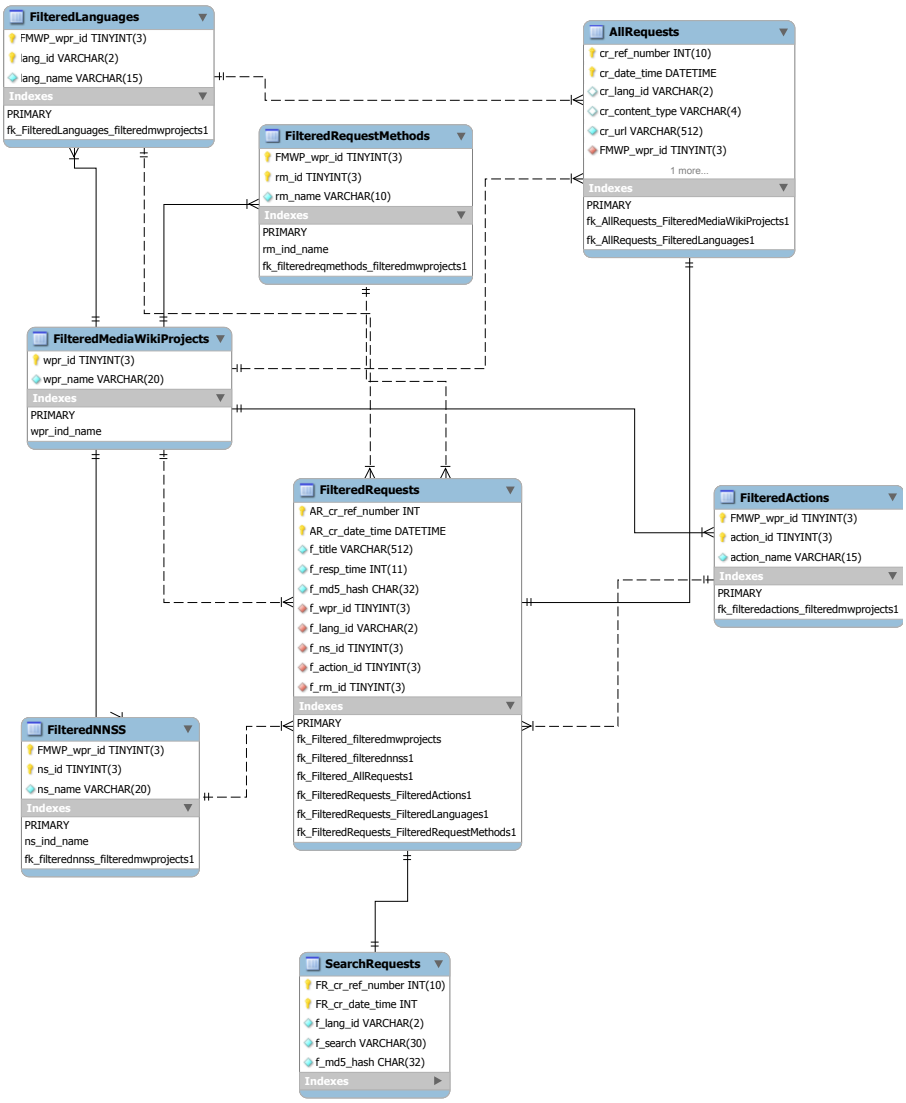


Figure 3.3: Entity-Relationship Diagram corresponding to the database arranged to improve the statistical analysis.

- **Actions2009**
  This table stores the number of the different actions considered in the analysis (with the exception of the edit operation registered in the previous table) that have been performed on the Wikipedia articles again for each day of 2009. As in the previous tables, the targeted articles may correspond to any of the different namespaces and language editions considered in this work. Actions reported by this table may consist in edit requests, history reviews, search requests or submit requests.

- **Articles2009**
  This table stores the number of times that a certain article has been visited or involved in any of the considered actions during each month of 2009. The huge amount of information does not allow to register this information in a daily basis so I chose to provide it by month. To improve subsequent queries performance, articles are just referred by the md5 digest of this title. In this way, results from queries involving this table can be easily crossed with the *filtered* table from the *squidlog* database to obtain the sources titles of the corresponding articles. In addition to the month, the namespace and language edition of each article is also stored in this table.

3.5 Validation and statistical examination

Following sections introduce the methodological developments conducted to validate the results presented in this thesis as well as to offer a suitable answer to the research question stated in chapter 1. Therefore, the different procedures together with the statistical examinations and tests used to perform our anaylsis are coveniently described in the following.

3.5.1 Validation

To ensure the validity of the sample we are receiving and, more important, of the processing of the log lines performed by the WikiSquilter application, we have compared some of our results with the ones offered by the Wikimedia Foundation itself, because the data emanating from it can be considered as the most reliable information source. It is important to consider that the sampling factor used to take the sample we are receiving is 1% and that Erik Zatche's site is based on the logs collected by Mitouzas which, as mentioned above, have not been sampled or filtered in any way. Thus, all of our results have to maintain the same ratio with respect to the corresponding to the overall traffic. And that is just what I have confirmed by comparing the number of visits and edit operations filtered by the WikiSquilter application with the information provided in the Erik Zatche's site about Wikipedia [3].

Moreover, I have compared the results obtained by the *WikiSquilter* application with the findings of previous works such as Ortega's doctoral thesis [Ort09]. To get the data involved in his thesis, Ortega developed a software tool called *WikiXRay* that allows to automatize the analysis of the dump files containing the Wikipedia articles and their different editions over the time. So, although this work quantitatively analyzes the Wikipedia contents and some other important topics such as quality, reputation or authoring dynamics related to them, its analytical software tool allows to find out some measures such as the number of edits (edit operations) that can be compared with the ones obtained from the analysis presented in this thesis. More in detail, I have compared the number of edits performed on articles of the considered Wikipedias corresponding to each month of 2009 and belonging to the main namespace. As Ortega's data are obtained from Wikipedia dump files, they

[3] `http://stats.wikimedia.org`

refer to the total number of these operations performed by the Wikipedia users. Of course, each one of these operations is requested by the corresponding URL. Because of this, the relation of our data respect to the Ortegas's is expected to be equivalent to the sampling factor used for the data feed we receive, i. e., the 1%

In addition, I have compared our traffic estimations with some of the traffic statistics provided by third-party sites such as Alexa. In this case, our interest is focused on the traffic attracted by each particular edition of Wikipedia. Thus, Alexas's figures about sub-domain traffic for a three month period from October till December 2010 have been compared with the traffic characterization performed by the WikiSquilter application.

Finally, I have also compared the results offered by some of the initiatives described in 2 with the ones obtained in this analysis. Most of these results are based on the Mitouzas's logs which are also the source of the abovementioned Zatche's portal. In order to avoid redundancy, a reduced set of this information has been considered for comparisons purposes.

In any case, if a high degree of similarity is obtained when comparing the different measures, we may guarantee the validity of the data involved in the analysis as well as the procedural developments performed as a part of it. In this case, the used sample would be proved as significant enough for the aims of the analysis and the method for obtaining it could be considered as reliable. Regarding the *WikiSquilter* application, a positive match between its results and the ones provided by other initiatives and analysis would validate its operations of parsing and filtering and would permit us to affirm that very few, if any, of the URLs that are objective of the analysis have been disregarded. Summarizing, a positive assessment of the part of our results offered by other sources would allow us to be more confident about the validity and accuracy of the rest of them.

3.5.2 Traffic characterization

To analyze the traffic directed to the considered editions of Wikipedia in the aim of determining the different types of requests comprising it and their respective frequencies, we have processed the log files containing the requests registered by the Wikimedia Foundation Squid servers using a software tool included as a part of the WikiSquilter project. This tool uses regular expressions to characterize and compute the different URLs contained in the file. Characterization here in not undertaken in such as thorough way as for the filtering process. In this way, we determine the Wikimedia Foundation project pointed by the URL's as well as the specific edition of Wikipedia targeted. In this way, requests for images and other resources are computed as the same level as the Wikimedia Foundation projects. This is due because these resources have to be uploaded first to the platform and, from them on, they can be referred from articles belonging to Wikipedia but also from articles corresponding to other projects like Wikiquote, Wikiversity and so on. Apart from the Wikimedia Foundation project and the edition of Wikipedia, we have also obtained, for each Wikipedia, the amount of traffic consisting in visits to articles in any namespace or in editions over them. The purpose, in this case, is to compare the number of filtered requests and the amount of traffic to verify that most of traffic is directed to very particular namespaces, the ones considered in this thesis, in both cases of visits and editions. The amount of traffic specifying any action is, in the same way, calculated in purpose of assessing the proportion of the actions considered for this thesis. Traffic involving search operations is also accounted, again to determine the percentage of the overall traffic corresponding to these operations. Ideally, considering that search questions are not applied to any article in particular because they are issued to recovery the list of the ones covering a certain topic, it is expected that our application filters all of them and stores the corresponding information into the database.

3.5.3 Temporal patterns

The finding of temporal patterns presenting how users' requests are distributed over time is one of the main aims of this thesis. Temporal patterns have been considered as repetitive sequences of a certain distribution of requests throughout different time units. In this way, we have used the information stored in the analysis database as the main data sources as its tables and fields were defined considering a subsequent temporal characterization of requests. As an example, the day-of-the-week field was added to the tables in order to allow faster queries at this temporal unit level. So, we obtained the distributions of the different types of requests throughout several time periods such as months, weeks and, of course, the whole year. This analysis was carried in terms of general traffic as well as separately for each considered edition of Wikipedia in the aim of determining similarities and differences in the temporal habits of accessing Wikipedia. Squid time is registered always using GMT time, so requests from different time-zones are grouped as having the same time although their issues, regarding their local time-zones, were performed at very different times.

3.5.4 Behavioral patterns

In addition to the finding of temporal patterns, this thesis is aimed to study the users' behavior when interacting with Wikipedia. In this way, we have obtained different correlations devoted to analyze whether some kinds of behaviors are related in any way. For example, we have studied if visits and edits present some kind of correlation because. If so, this may be intended as the result of a collaborative attitude in which visitors act also as contributors. We have also studied the behavior of users when submitting contributions to the different Wikipedia editions. In this way, we have analyzed the differences among the percentages of edit requests that are not finished with the corresponding write operation because these measurements can serve as an indicator of users' reluctance when contributing contents. The attention to the different kind of contents has been measured in terms of the targeted namespaces. In the same way, the ratios corresponding to the different types of requested actions have been also analyzed and compared, again in the aim of determining different types of conducts. The analysis of different pairs of measurements to determine the degree of relationship between them has involved the application of an statistical test usually consisting in the calculation of the Pearson's product moment correlation coefficient for the two compared set of values. This coefficient takes values in the range $[-1, 1]$ and closeness to 1 means highly related measurements while 0 indicates no association. The Pearson's product moment correlation coefficient (r) can be computed using the following expression:

$$r = cor(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

In addition, we always provide the p-value of the statistical test that consists in the probability of getting a result as extreme as the obtained if the null hypothesis (no association) were true. Usually p-values under a certain threshold (usually 0.05 or 0.01) allow to reject the null hypothesis.

3.5.5 Featured contents

Featured articles are considered the best articles all over the Wikipedia. In order to be promoted to this status, these articles , first, need to be nominated as candidates. Featured articles have to meet a set of demanding criteria apart from the requirements which are expected in all the Wikipedia article.

These criteria range from clearness and comprehensiveness in the article's writing to an adequate organizational structure. Other aspects such as stability, neutrality, length and references are also taken into account.

In what our research is concerned, we have analyzed the impact of featured articles in two very different ways. First, we have considered the influence of the promotion of articles to the featured status in their subsequent number of visits. Then, we have also studied the impact of the presentation of featured articles as examples of high quality content in the main page of some editions of Wikipedia. From this point of view, our main goal has been to determine some kind of patterns which can serve to model the influence of the promotion to the featured status in the traffic directed to the articles awarded with this consideration.

In order to evaluate the impact of the consideration of articles as featured, we obtained the articles promoted to the featured status during April and October 2009 by browsing special pages of each Wikipedia edition devoted to its featured contents. Furthermore, we extracted the featured articles selected to appear in the main page of different Wikipedia editions during the same months. Then, we queried the database resulting from the processing of the Squid log lines to look for the number of visits corresponding to those articles during the aforementioned months as well as during the previous and the following ones. In this way, two groups, each made up of three months were established, one around April and the other centered on October.

To determine the statistical test to be applied for comparing the number of visits received in the different months, we used the Shapiro-Wilcox statistic to assess the normality of the distributions of visits corresponding to each month. Normal distributions can be compared using the mean measure. On the other hand, to compare non-Normal distributions the use of the median is more appropriate as this statistic is more robust to skewed set of values with may also present extreme values (outliers). Given that certain distributions are found to be non-Normal, the Wilcoxon rank-sum test (also known as Mant-Withney-Wilcoxon test) becomes an appropriate tool to determine whether they are different because this test is not sensitive to the normality of the data.

3.5.6 Popular topics

Apart from the quantitative analysis of the information elements involved in the common interaction with Wikipedia, this thesis is also devoted to provide a categorization of the most popular subjects and topics in Wikipedia according to the users' requests. Thus, in order to determine and classify the most visited and edited articles, we, first, inserted into the database the md5 hash representation of every article's title whose request were considered of interest according to our analysis directives. This was also done with the character strings submitted as a part of the search operations. The purpose of the use of the hash code is to speed the subsequent queries devoted to determine the most accessed articles by grouping the database rows with information about the requests by the md5 hash field, always 32 characters long, instead of by the original title which is arbitrarily long. MD5 algorithm guarantees that two similar character strings will always obtain the same hash code. So this solution leads to a fast computation of the articles involved in the visits and in the edits requests. In order to characterize visits, edits and search requests, we have used a classification based on the one proposed by Spoerry in [Spo07b]. The author established a set of main categories to assign to the requests, in the same way as tag systems do. I did not used all the Spoerry's categories because I considered that some of them could be joined to form more representative groups. In other cases, I decided to extend the scope of certain categories in order to cover related topics or subjects. Although some articles or search topics may easily correspond to more than one category, we have assigned each article to just an unique category. In the following, we detail the different categories constituting our characterization scheme:

1. Main (MAIN): Just refers to the main page of every edition of Wikipedia considered for this thesis.

2. Entertainment (ENT): It includes books, comics, films, games, music, performers, TV series and video games.

3. Politics + War: This category covers those articles exploring topics about political figures and conflicts.

4. Geography: Articles dealing with countries, cities, villages, natural surroundings, etc.. correspond to this category.

5. Sexuality: Includes sex-related terms and pornography

6. Science: Include the articles presenting topics related to any scientific discipline such as Mathematics, Astronomy, Physics, Chemistry, Biology as well as the ones covering subjects in the environment of Technology and Industrial development. Weapons and military technology are also assigned to this group of articles.

7. ICT (Information and Communication Technologies): Articles about Computer Science, Internet, programming languages, operating systems, databases, as well as communication standards, protocols, mobile devices and technologies, smartphones, means of transmission and signal theory and processing among others belong to this group.

8. Arts: It includes articles belonging to disciplines such as art, painting, sculpture, religion, literature, history, Religion, humanities and so forth.

9. Current Events: This category is devoted to gather the articles related to events of certain relevance during a given period of time. In this way, articles related to any kind of competitions or championships during their development, to particular people or celebrities after their death or to topics involved in mass media because of its dimension (such as the episodes of NH1 gripe) are assigned to this category.

In order to analyze to which subjects correspond the articles receiving highest numbers of visits, we have classified the top-65 most visited articles corresponding to different months (January, February, June, July, August and November) and for certain editions of Wikipedia (German, English, Spanish and French). The same classification has been performed for the top-65 most edited articles, again in the same months and corresponding to the same editions of Wikipedia that were involved in the visits categorization.

As *WikiSquilter* also computes the md5 hash of every string submitted as a search topic, we have been able to group and obtain the strings more repeatedly involved in search operations. In this way, we have got the top-65 most searched topics also in the German, English, Spanish and French Wikipedia and, again, for January, February, June, July, August and November 2009. Then we have performed the same categorization applied to the articles's titles in order to determine the different subjects most frequently searched by users.

When determining the impact of search operations in the number of visits and contributions to the articles, we found the serious drawback that the two md5 hashes corresponding to the articles' titles and to the searched topic do not match if just one character differs in the two strings. This happens unless title and search string consist in a sole word with no differences in capitalization. Articles' titles with two or more words separate them using underscores (_) where in the search strings different

names usually appear as separated by the plus symbol (+). Due tho this fact, all the categorizations have had to be manually performed.

Previous categorization entails the articles belonging to a set of Wikipedias which received more visits end edits during certain moths and also includes the topics submitted as search string. It can be complemented with the analysis of the distribution of the requests among the different categories of articles and search topics. In this way, we have grouped the requests to the the top-65 most visited and edited articles and to the top-65 most searched topic according to the established categories in order to determine how many requests correspond exactly to each category in order to precise which ones of them are being requested more frequently by users. The influence of search operations on visits has been assessed by correlating the two observations corresponding to each category of topics.

Chapter 4

Statistical Analysis and Results

"One shaft of light that shows the way" *A Kind of Magic*, Roger Taylor, (1986).

4.1 Introduction

Next sections present the most important results obtained from the empirical study conducted as part of this thesis. As the main aim of this work is to explore both the temporal and behavioral patterns in the use of Wikipedia as well as to provide a characterization of the traffic directed to the encyclopaedia, the following is devoted to introduce our most relevant findings in this area.

Results will be presented as the proposed answers for the research questions stated in chapter 1 and will include appropriate evidences in the form of graphs or tables. In this way, every supporting element will be coupled with the corresponding explanation and discussion. When a deeper analysis or study is recommended, the line of its development and further work will be introduced.

In general, results will be usually presented related to the measures or parameters being studied according to its consideration as representative and descriptive enough to deserve the corresponding analysis. In this way, days of week, months, language editions, namespaces, actions and general articles will be the common articulatory elements of the presented graphs and tables.

4.2 Validation of our study

According to the stated in chapter 2, requests issued by users have been previously used to analyze the queries submitted to web systems in order to determine the effectiveness of the current descriptive terms they use. In addition, such kind of analysis is useful to provide the web systems with the necessary contents to satisfy their users' information needs.

In this way, and given that our approach is based on the analysis of the users' requests contained in the log lines received from the Wikimedia Foundation Squid systems, we considered that the best way to assess the validity of such kind of study was to obtain specific measures also offered by trusted sources and establishing the corresponding comparison with them in the aim of finding a high matching degree. If so, we will be in position of guaranteeing not only the trustworthy of analyses based on this kind of feed but also the results emanating from this particular one.

We introduced in chapter 2 several initiatives devoted to offer statistical information about the number or articles, users, etc., belonging to the different editions of Wikipedia, the traffic directed

Lang.	Jan.	Feb.	Mar.	Apr.	May.	Jun.
DE (Ours)	10,821,625	6,833,171	8,034,636	6,945,878	7,612,949	7,249,244
DE (Mituzas)	1,271 M	982 M	978 M	817 M	875 M	909 M
Quotient	0.009	0.007	0.008	0.009	0.009	0.008
EN (Ours)	47,369,841	43,136,627	51,845,199	48,242,580	48,085,156	43,950,168
EN (Mituzas)	5,615 M	5,944 M	6,092 M	5,989 M	6,066 M	5,819 M
Quotient	0.0084	0.0073	0.0085	0.0081	0.0079	0.0076

Table 4.1: Comparison of the number of pageviews from Mituzas's log files (Rows indicated with 'Mituzas') related to the German and English Wikipedias and corresponding to the first semester of 2009 with our results (Rows heading by 'Ours'). The quotient (Rows with 'Quotient') between the two measures is also presented. M stands for Million.

to them and the evolution over time of their numbers of pageviews and edit operations. All this information results of enormous interest in order to have a reference element to compare with and became a really useful help to assess the validity of the conducted analysis. In particular, we consider specially valuable the statistical information stemming from the Wikimedia Foundation and other relevant companies such as Alexa or comScore.

Here it is important to recall that our data feeding is made up of the 1% of all the traffic directed to the Wikimedia Foundation projects. Considering that it is not a very large sample, although it includes thousands of millions of log lines, we have to be very effective and accurate when obtaining the information elements from it. As explained in chapter 3, our analysis focuses on several editions of Wikipedia as well as on certain namespaces and actions. In the following it will be shown that, in effect, we are not disregarding relevant information and that the data basing our analysis is consistent with the total figures about Wikipedia requests.

We will include here, for clarity purposes, only a sample of the exhaustive comparison performed on all the considered editions. In this way, we will include the results of the assessment related to some particular Wikipedia editions. Readers can find tables with the whole set of results corresponding to all the examined editions in Appendix A.

Let start by comparing the number of visits obtained from our analysis with the figures presented in `http://stats.wikimedia.org/EN/Sitemap.htm`. This information, compiled and presented by Erik Zachte, is obtained from the pageviews collected by Domas Mituzas and, in consequence, is one of the most reliable sources. Thus, table 4.1 and table 4.2 present the comparison between these pageviews and the ones observed from own our results for the German and English Wikipedias.

As introduced in chapter 3, pageviews or visits correspond to those URLs requesting articles in any namespace and not involving any type of action on them. Thus, pageviews or visits are considered as requests issued to retrieve information from Wikipedia. Tables 4.1 and table 4.2 also include the quotient between our own figures ant Zachte's ones. As Zachte's information stems from Mitouzas's log files which are not filtered in any way, the ratio between the two measures should correspond to our 1% sampling factor if both sampling and processing have been correctly driven. As both tables present, quotients are really close to that factor. The small difference respect to it correspond to the articles in the namespaces not considered in this thesis and, thus, not filtered by the *WikiSquilter* application.

Lang	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
DE (Ours)	6,626,701	6,942,208	7,404,872	7,223,746	7,615,539	7,102,197
DE (Mituzas)	819 M	813 M	889 M	885 M	904 M	760 M
Quotient	0.008	0.009	0.008	0.008	0.008	0.009
EN (Ours)	44,451,649	48,426,122	49,713,090	49,392,482	49,738,157	47,687,869
EN (Mituzas)	5,614 M	5,604 M	5,938 M	6,041 M	5,842 M	5,259 M
Quotient	0.0079	0.0086	0.0084	0.0082	0.0085	0.0091

Table 4.2: Comparison between the number of pageviews from Mituzas's log files corresponding to articles in the German and English Wikipedias for July till December (Rows indicated with 'Mituzas') and the number of edits obtained from our results (Rows heading by 'Ours'). M stands for Million.

Lang.	Jan.	Feb.	Mar.	Apr.	May.	Jun.
DE (Ours)	11,041	9,457	10,341	8,361	8,052	7,754
DE (Zachte)	876 K	752 K	802 K	655 K	684 K	701 K
DE (Quotient)	0.0126	0.0126	0.0129	0.0128	0.0118	0.0111
EN (Ours)	53,121	46,778	54,564	47,921	47,692	42,282
EN (Zachte)	4,300 K	4,200 K	4,400 K	4,000 K	4,300 K	4,000 K
EN (Quotient)	0.0124	0.0111	0.0124	0.0120	0.0111	0.0106

Table 4.3: Comparison of the edit operations reported by Zachte's site for the German and English Wikipedias during the first semester of 2009 with the results of our analysis. (Rows indicated with 'Zachte') and the number of edits obtained from our results (Rows heading by 'Ours'). K stands for thousands. M stands for Million.

Once the results related to the visits or pageviews have been checked, we proceed to assess the validity of the measures about edit operations as their rates and frequencies are also offered from Zachte's site. These values are also trustworthy because they are computed from the dump files offered by the Wikimedia Foundation. Table 4.3 and table 4.4 present, therefore, the comparison between the number of edit operations reported from Zachte's site, which correspond to the German and the English Wikipedias for every month of 2009, and the ones observed as a result of our own filtering process. Again, the quotient between the two measures is included for validation purposes. In the case of edit operations, the ratio is even closer the sampling factor in practically all the cases and even slightly surpasses it. This is surely due to the fact that Zachte's data are considerably rounded. In fact, all his values are exact multiples of the Kilo or thousand (K) and Mega or million (M) units. This means that edit operations rarely involve articles in namespaces other than the considered in this thesis.

After this, we are going to compare the number of edit operations after our analysis and after the *WikiXRay* tool used by Ortega in [Ort09]. This kind of comparison is an unparalleled opportunity because it allows to put in relation data resulting from the analysis of the Wikipedia dump files with the information obtained from the logs reporting users' requests to the encyclopedia. In this way, Tables 4.5 and 4.6 summarize the number of edits performed on articles from the German and English

Lang	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
DE (Ours)	7,688	8,393	8,111	7,968	7,942	7,581
DE (Zachte)	688 K	729 K	680 K	714 K	716 K	714 K
DE (Quotient)	0.0112	0.0115	0.0119	0.0112	0.0111	0.0106
EN (Ours)	41,087	45,492	43,969	38,631	37,641	36,568
EN (Zachte)	3,800 K	3,900 K	4,000 K	4,000 K	3,900 K	4,400 K
EN (Quotient)	0.0108	0.0117	0.0110	0.0097	0.0097	0.0083

Table 4.4: Comparison between the number of edits from Zachte's site corresponding to articles in the German and English Wikipedias for July till December (Rows indicated with 'Zachte') and the number of edits obtained from our results (Rows heading by 'Ours'). K stands for thousands. M stands for Million.

Lang.	Jan.	Feb.	Mar.	Apr.	May.	Jun.
DE (Ours)	11,041	9,457	10,341	8,361	8,052	7,754
DE (Ortega)	1,227,017	1,069,725	1,148,209	962,561	987,244	1,013,734
DE (Quotient)	0.0090	0.0088	0.0090	0.0087	0.0082	0.0076
EN (Ours)	53,121	46,778	54,564	47,921	47,692	42,282
EN (Ortega)	6,195,518	5,926,109	6,614,845	5,876,645	6,166,014	5,702,894
EN (Quotient)	0.0086	0.0079	0.0082	0.0082	0.0077	0.0074

Table 4.5: Comparison between the number of edits on articles of the German and English Wikipedias obtained from our results (Rows heading by 'Ours') for January till June 2009 and the same number of operations reported by Ortegas's tool *WikiXRay* (Rows indicated with 'Ortega') for the same period. Both data correspond to articles in the main namespace. Rows headed by 'Quotient' correspond to the quotient between the two measures.

Wikipedias as determined by the Ortega's tool and the number of save requests to the database servers according to the results obtained from our own analysis.

Now, we are going to validate our results involving the number of visits, or pageviews, to particular articles by comparing these numbers with the ones provided by initiatives based on reliable sources such as `http://stats.grok.se/`, which is again built on Mituzas's logs. In this case we have compared the number of visits to the *Squid* article through the days corresponding to two different months of 2009 (April and May). Thus, Figure 4.1 shows the evolution of the number of visits received by the article in both months as reported by the site `http://stats.grok.se/` as well as the same information obtained from the results of our analysis. As it is shown, both time-line evolutions are practically similar and present the same dips on 4, 5, 12, 13, 18, 19, 25, 26 April 2009 and on 2, 8, 9, 16, 17, 23, 24, 30, 31 May 2009. In the same way, relevant peaks appears in the same days (8, 9, 15, 16, 21, 22, 27, 28 April 2009 and 13, 14, 18, 19, 26, 27 May 2009) in both charts. As a result, these graphics show how the number of visits to the article follows the same evolution in both months. This constitutes another endorsement to the reliability of the results obtained by our application because the comparison is established at a finer grain than the previous involving the

Lang	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
DE (Ours)	7,688	8,393	8,111	7,968	7,942	7,581
DE (Ortega)	993,866	1,048,137	975,990	1,056,171	1,091,001	1,073,048
DE (Quotient)	0.0077	0.0080	0.0083	0.0075	0.0073	0.0071
EN (Ours)	41,087	45,492	43,969	38,631	37,641	36,568
EN (Ortega)	5,492,827	5,557,041	5,762,412	5,747,647	5,497,166	6,060,027
EN (Quotient)	0.0075	0.0082	0.0076	0.0067	0.0068	0.0060

Table 4.6: Comparison between the number of edits on articles corresponding to the German and English Wikipedias obtained from our results (Rows heading by 'Ours') for July till December 2009 and the same number of operations reported by Ortegas's tool *WikiXRay* (Rows indicated with 'Ortega') for the same period. Both data correspond to articles in the main namespace. Rows headed by 'Quotient' correspond to the quotient between the two measures.

overall pageviews.

Again making comparisons at the level of particular articles, if we consider another site such as `http://toolserver.org/~emw/wikistats/`, also based on Mituzas's logs, we will observe a similar correspondence between the number of pageviews reported by the site and the one obtained from our own results. This is shown in Figure [1] which presents the evolution of both pageviews for the *Spain* article in May 2009. Finally, even if we considerer an external site, such as `http://www.wikistatistics.net/`, and we compare the number of edits, because most of the other information is referred to quantitative data about aspects of Wikipedia such as articles or users, we will find a new match in the presented evolutions[2].

Alexa site offers a distribution of the requests to Wikipedia by sub-domains that describes the percentage of visits that every edition attracts. As explained in chapter 1, the different editions of Wikipedia are referred through corresponding URLs that point to specific sub-domains of the `wikipedia.org` general one. In this way, we compare in Table 4.7 the composition of our traffic sample to check whether it has a similar distribution to the presented in the Alexa web site. One important issue in this sense is the fact that Alexa does not allow to get this kind of information from a period prior to the last 3 months. Thus, although our analysis is based on the Squid log lines corresponding to 2009, we have analyzed the traffic corresponding to the period from October till December 2010 to perform the appropriate comparisons.

Although figures seem not to completely match, if we put them together in a chart (Figure 4.2) we will appreciate how both distributions of visits over the editions of Wikipedia present very similar shapes. However, it is important to remark that Alexa's main data source consists in the information sent by the toolbars installed by its users and may no reflect the overall traffic to Wikipedia. Similarity between this two lines can be interpreted as a significant use of Wikipedia on the side of Alexa's users.

After having assessed the correction of both the sample and the data processing leading to the information stored in our database, we consider of great importance to address the question of representativeness. Here, representativeness is dealt in terms of verifying whether the information elements considered of interest by our analysis correspond to a relevant part of the overall traffic

[1] `http://gsyc.es/~ajreinoso/thesis/figures/spain.pdf`
[2] `http://gsyc.es/~ajreinoso/thesis/figures/wikistats.pdf`

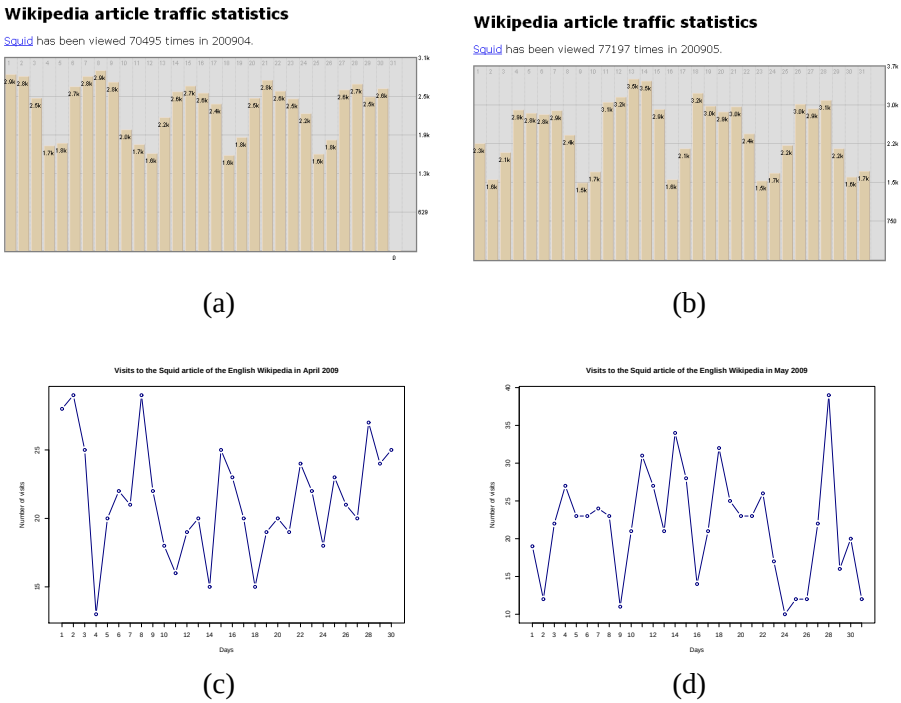


Figure 4.1: Comparison of the information reported by the site *stats.grok.se* about the number of visits to the *Squid* article in the English Wikipedia with the data obtained after our own analysis.(a) Number of visits to the *Squid* article (EN) in April 2009 according to the *stats.grok.se* site. (b) Number of visits to the *Squid* article (EN) in May 2009 according to the *stats.grok.se* site. (c) Number of visits to the *Squid* article (EN) in April 2009 according to our results. (d) Number of visits to the *Squid* article (EN) in April 2009 according to our results.

directed to Wikipedia and, therefore, do constitute a representative approximation to it. Therefore, and according to our traffic estimations, conveniently developed in the next section, the editions considered for this thesis attract more than the 90% of the overall traffic directed to Wikipedia. If we compare the number of requests filtered by our application with the requests making up the general traffic, we will find that, for each considered Wikipedia, the requests to the namespaces used in this analysis correspond, in average, to the 85% of the total requests asking for articles in any namespace for these Wikipedias. So, apart from disregarded requests, we can conclude that very few of the requests issued to visit an article in the studied Wikipedias are not directed to the namespaces we have considered. Edits are even easier to trace, and the edit operations we have filtered, for each Wikipedia, do constitute more than the 94% of the total requests found in the traffic to each Wikipedia soliciting a save operation. That means that almost all the edit operation are done on articles in the namespaces considered by this analysis. In the case of search operations, our filtered requests correspond to the 99% of the observed traffic involving such kind of operations. These actions have to be filtered in any case because they are not applied on any specific article. On the contrary, they are submitted to retrieve articles containing a particular topic. Its high percentage is a good indicator of the accuracy of our filtering process. These percentages are given to illustrate that we are focusing on the most relevant editions of Wikipedia as well as on the most significant namespaces. Moreover, they also serve as validation facts to support the reliability of our work.

Edition	Alexa traffic	Sampled traffic
DE	8.1%	7.95%
EN	54.0%	45.71%
ES	5.7%	8.23%
FR	3.5%	4.57%
IT	2.9%	2.65%
JA	10.3%	7.86%
NL	0.7%	1.49%
PL	1.5%	2.99%
PT	1.5%	2.58%
RU	3.5%	5.83%
OTHER	10.13%	5.56%

Table 4.7: Comparison between the traffic volumes per Wikipedia project reported by Alexa for October-December 2010 and the ones extracted using the WikiSquilter application.

Next section will provide a quantitative analysis of the traffic composition in the aim of providing an appropriate characterization of all the requests directed to Wikipedia. This kind of analysis may lead to a better comprehension of the kind of use given to the Wikipedia by its users. In addition, the obtained results may be used as an estimation of the overload imposed to the server architecture deployed by the Wikimedia Foundation to support all its wiki-based projects.

4.3 Traffic characterization

As described in chapters 1 and 3, this study undertakes the analysis of the traffic directed to the 10 most active editions of Wikipedia in terms of their volumes of requests and number of articles. In this way, this section is aimed to provide a quantitative analysis of the composition of the traffic directed to the Wikipedia project, as a whole, as well as to the considered editions of Wikipedia in particular [3].

Therefore, we will present the distribution of the different types of requests comprising the traffic to these editions. Apart from this, we also present information about the general traffic directed to all the Wikimedia Foundation projects. Traffic information is always computed in terms of number of requests disregarding, by the moment, considerations about amount of information or transference rates. In addition, we are usually presenting the daily average of the requests for each month because absolute values will introduce a biased perception due to the different number of days corresponding to each month. Moreover, technical problems have prevented us from obtaining the traffic corresponding to absolutely all the days of the year. Fortunately, we have only failed to get the traffic of just 4 days, what is an absolute success in terms of the reliability of our receiving infrastructure.

In the aim of determining how the overall traffic to the Wikimedia Foundation was distributed among its projects during 2009, Table 4.8 provides the percentages of the total traffic corresponding

[3]There is a summary of data related to the quantitative analysis of traffic available at `http://gsyc.es/thesis/tables/tabTraffic.pdf`

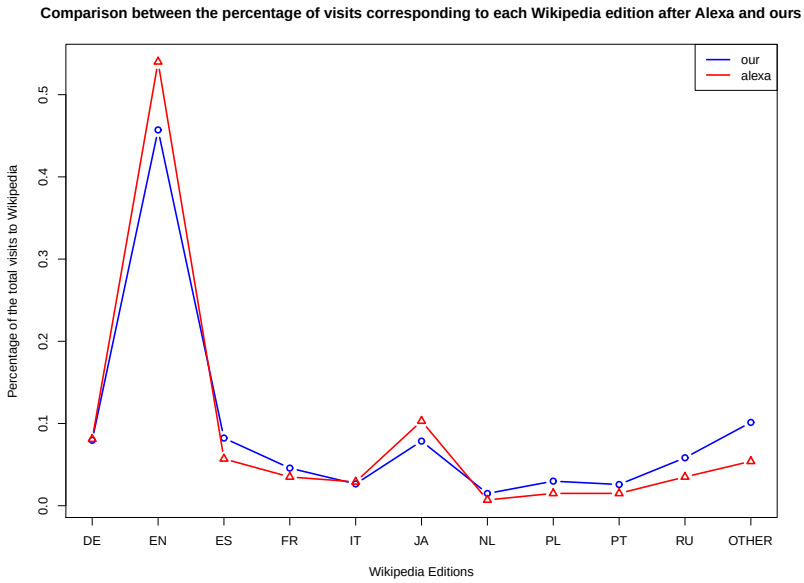


Figure 4.2: Percentage of the overall traffic attracted by each considered edition of Wikipedia after Alexa statistics and after our own analysis.

to each particular project. As it is clearly seen, the requests for Wikipedia pages and, interestingly, for images and other resources uploaded to the platform in order to be referenced later from the articles do constitute by the 96% of all the traffic received by the Wikimedia Foundation servers. Here, we have to remark that the later kind of requests are issued when browsing articles not only from Wikipedia but also from other Wikimedia Foundation project. In this way, images and other resources act as a kind of central repository and articles in any of the Wikimedia Foundation projects can refer to them. Figure 4.3 shows the relevance of these two types of requests in the traffic and also includes the amount of it corresponding to each Wikipedia edition.

Figure 4.4 shows the evolution of the traffic directed to all the Wikimedia Foundation projects for every month of 2009. The vertical edge shows the daily average of requests corresponding to each particular project and to the resources, mainly images, requested by users. In order to adequately examine these figures, it is important to remark that they correspond to the daily average for our sample, which is the 1% of the total traffic, so real ones would be, for instance, 30 * 100 times higher in the case of months having 30 days. From Figure 4.4 we can examine the yearly evolution of the total amount of traffic to all the Wikimedia Foundation projects as well as to the Wikipedia one in particular.

We proceed to characterize the traffic corresponding only to the Wikipedia project. In this way, our first aim is to determine the amount of traffic attracted by each of its editions and, particularly, by the ones considered in this thesis. Thus, Figure 4.5 shows the distribution of the Wikipedia traffic over its different editions during each month of 2009. The English Wikipedia appears as the most popular one with a traffic volume much higher than the rest of editions. Besides this, we have considered appropriate to summarize the daily average of the traffic corresponding to the editions of Wikipedia throughout the whole year and to present their respective percentages of the overall traffic. Table 4.9 presents this information. As we can see, considered Wikipedias attract more than the 91% of the total traffic directed to the Wikipedia project. This is important in terms of the relevance of the

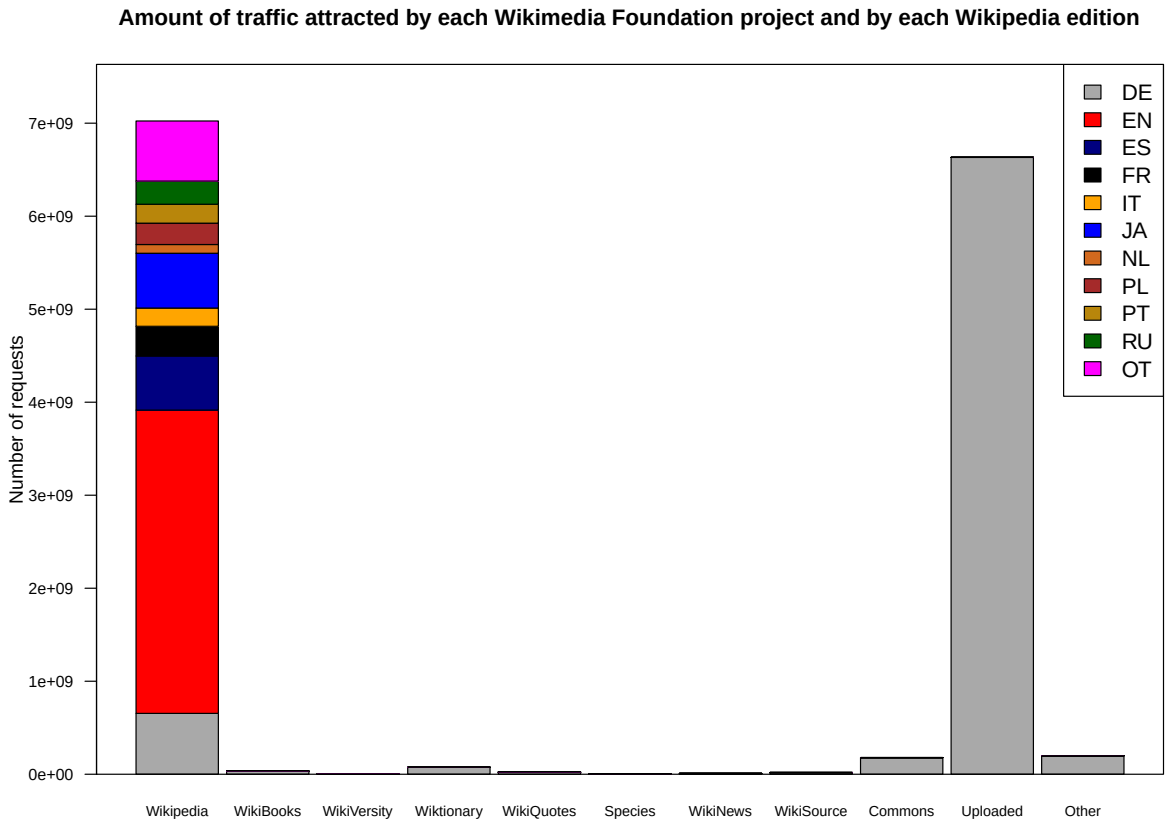


Figure 4.3: Amount of traffic corresponding to each Wikimedia Foundation project and to each edition of Wikipedia.

WMF project	Percentage of traffic attracted
Wikipedia	49.47%
Wikiversity	0.03%
Wikibook	0.23%
Wiktionary	0.52%
Wikiquote	0.16%
Species	0.01%
Wikinews	0.06%
Wikisource	0.13%
Commons (images)	1.26%
Uploaded resources	46.72%
Other	1.41%

Table 4.8: Percentage of the overall traffic to the Wikimedia Foundation servers during 2009 attracted by each of its projects.

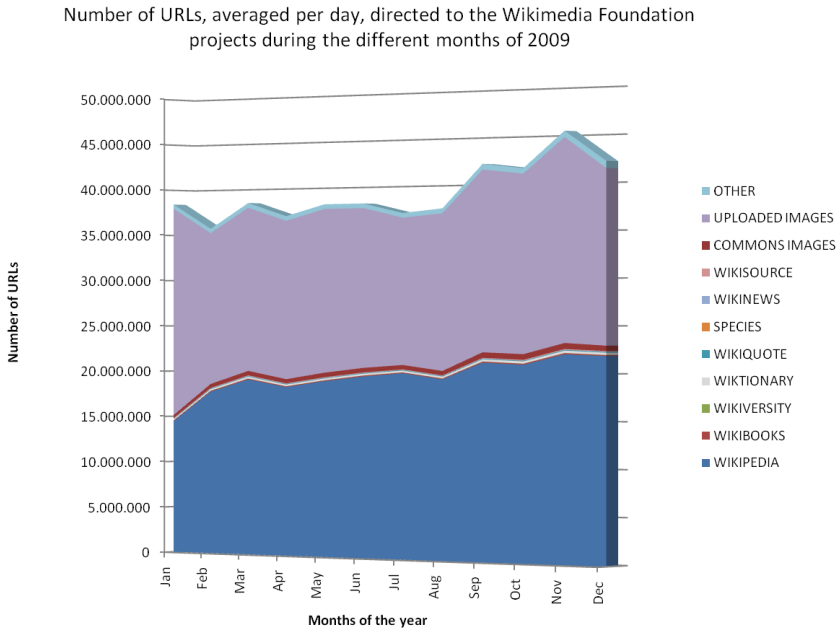


Figure 4.4: Evolution of the overall traffic to the Wikimedia Foundation projects during 2009.

considered sets of Wikipedias. The particular evolution of the daily average of traffic for each edition of Wikipedia during each month of 2009 is presented in Figure 4.6. As it is shown, not all the editions of Wikipedia follow the same distribution of their traffic over time, which means different temporal patterns of use.

On the other hand, we can compare the evolution of the traffic to the different editions of Wikipedia with the evolution of their respective sizes. Ideally, larger Wikipedias should attract a higher amount of traffic but this is not always true according to the Figures 4.7 and 4.8 which present, respectively, the amount of traffic attracted by each Wikipedia during each month of 2009 and their sizes expressed in number of articles during the same months. The vertical axis in both figures is in logarithmic scale because the English Wikipedia is several orders of magnitude larger than the other editions and this makes their data not to be properly displayed. As it is shown, the size of the different editions of Wikipedia is quite stable throughout the overall year. The largest Wikipedia edition corresponds to the English language whereas the smallest corresponds to the Russian one. The English and the German Wikipedias are the largest according to their number of articles and also are the ones that receive the greatest amount of traffic. However, the size of the Spanish Wikipedia, for instance, situates it among the three editions with less volume of articles but, regarding its traffic, it ranges from the fourth to, even, the second most requested edition. The same occurs with the Russian Wikipedia. Having the smallest number of articles, its traffic its larger than the attracted by many other editions. This is interesting, because the relative growth of all the editions remains quite similar throughout the year, so differences in traffic are not resulting in differences in number of articles. Next section will analyze in detail the different temporal patterns found in the use of each considered edition of Wikipedia and, in particular, it will deal with the evolution of both visits and edits in all the considered Wikipedias.

Probably, it is more interesting to obtain a characterization of the traffic directed to each edition of Wikipedia in order to compare their respective percentages of the different types of requests. This

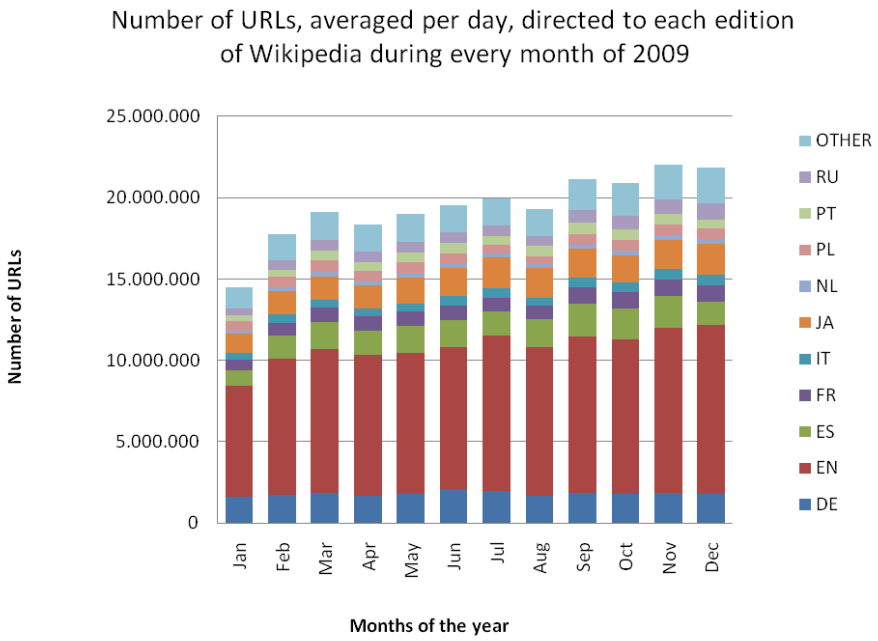


Figure 4.5: Comparison of the traffic directed to each edition of Wikipedia during each month of 2009.

Wikipedia edition	Daily average of the traffic attracted	Percentage
DE	21,767,176.73	9.40%
EN	108,407,534.61	46.45%
ES	19,336,747.61	8.25%
FR	10,622,527.01	4.54%
IT	6,516,987.21	2.79%
JA	19,591,570.27	8.38%
NL	3,128,496.65	1.34%
PL	7,628,743.39	3.30%
PT	6,755,424.08	2.87%
RU	8,269,484.01	3.51%
REST	21,467,547.49	9.17%

Table 4.9: Summarized daily average of the traffic attracted by each considered edition of Wikipedia corresponding to the whole year 2009. The traffic corresponding to the rest of disregarded editions is presented together in the entry REST.

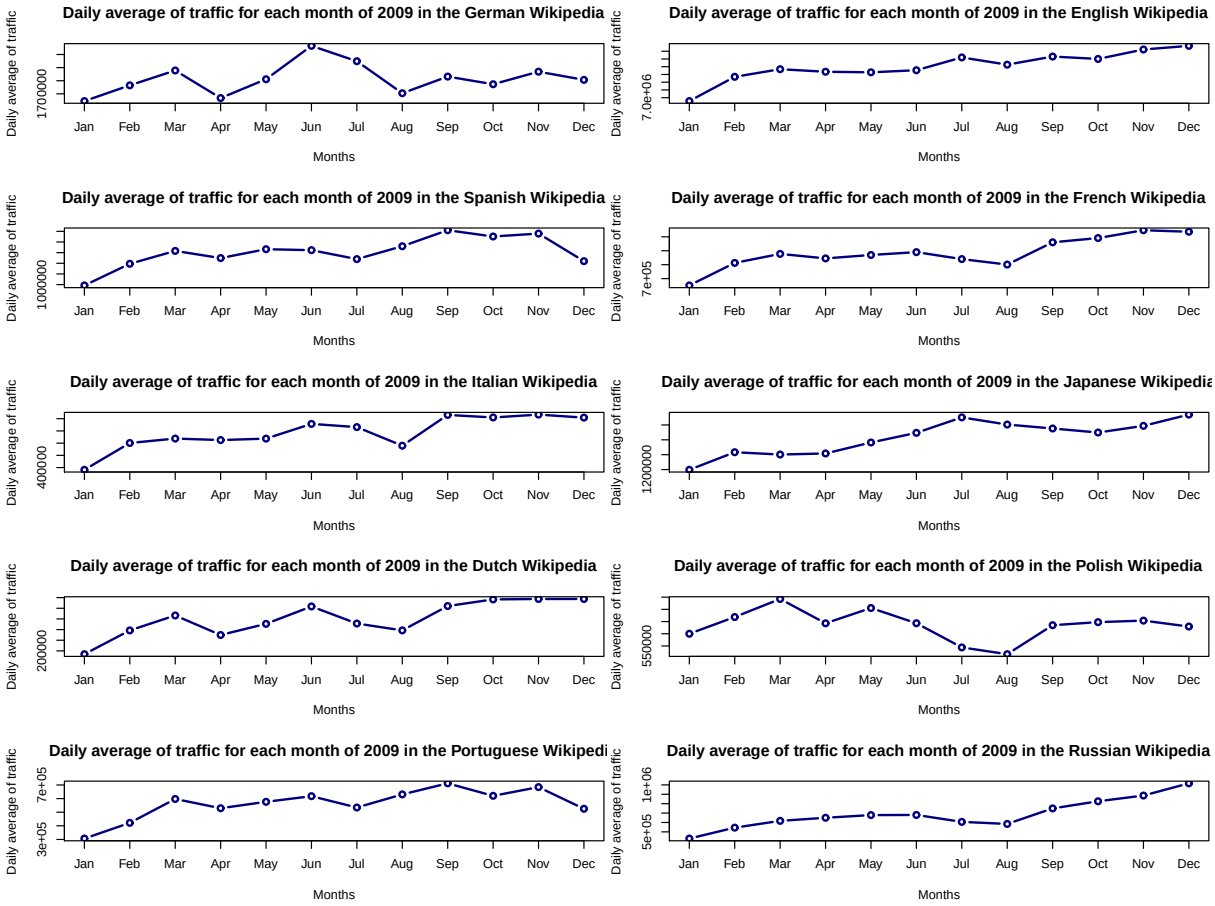


Figure 4.6: Evolution of the daily averaged traffic directed to each edition of Wikipedia during each month of 2009.

kind of information will provide an approximation to the utilization that the users of the different editions make of them. Table 4.10 shows the percentage of the general traffic directed to each edition of Wikipedia that corresponds to visits to articles, to requests for edit operations, for general actions performed on articles, for search operations and for css files used to present tailored pages as well as for the Wikipedia own icon.

In section 4.2 we saw that we were discarding very few requests, if any, consisting in article views or edits on them. Percentages presented in Table 4.10 are referred to the general raw traffic without applying any kind of filtering and are obtained as a result of a line counting process using regular expressions. So visits to articles, for example, refer to requested articles in any namespace, included, of course, the ones not considered by our analysis. The same can be applied to actions, whose column in Table 4.10 entails any type of requested action (except searches).

Once the analysis of the traffic directed to the editions of Wikipedia considered for this thesis has been performed, Section 4.4 will present the temporal patterns found in the general traffic as well as in the filtered requests.

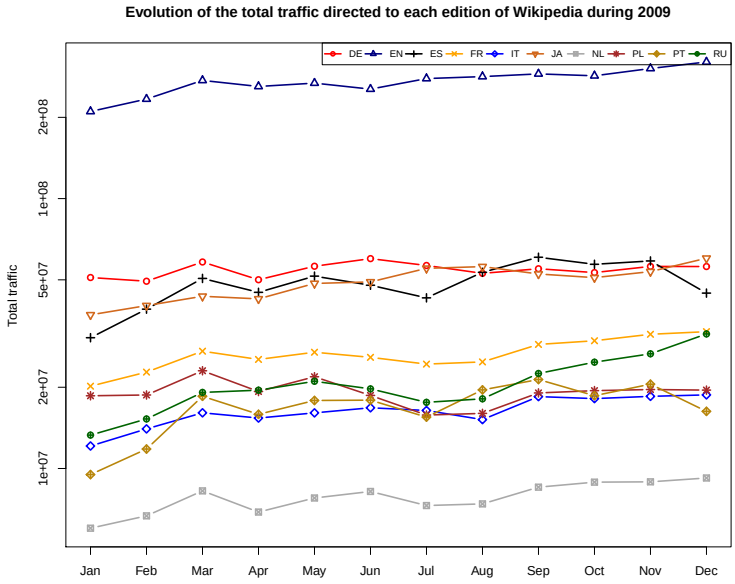


Figure 4.7: Evolution of the total traffic directed to each edition of Wikipedia throughout 2009.

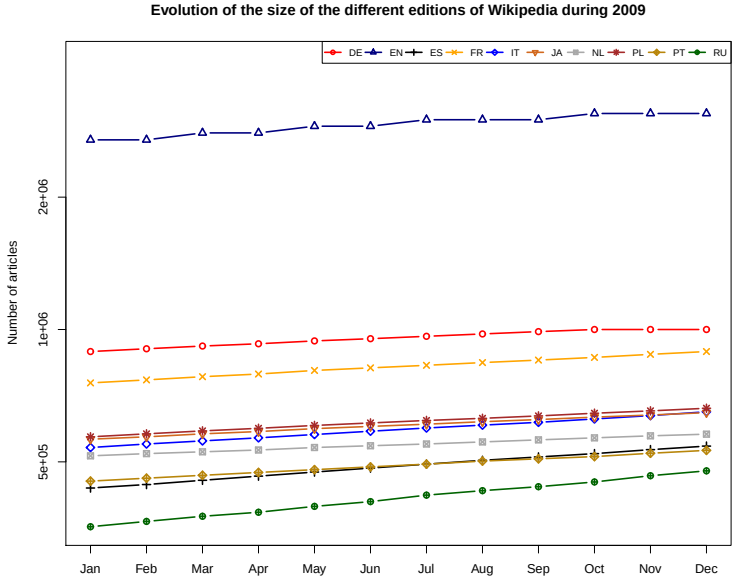


Figure 4.8: Evolution of the size of the different editions of Wikipedia throughout 2009.

Ed.	Visits to articles	Actions (except search)	Edit op.	Search op.	Api calls	Skins /css	icons	mw ext.	Total
EN	21.51%	22.52%	0.27%	4.75%	6.53%	34.62%	4.38%	3.47%	93.05%
DE	16.54%	20.87%	0.23%	4.09%	7.69%	30.74%	3.46%	14.72%	94.02%
ES	13.58%	33.90%	0.31%	4.12%	6.02%	32.13%	3.68%	3.89%	93.20%
FR	18.24%	23.15%	0.33%	4.00%	6.05%	36.87%	4.42%	4.23%	92.96%
IT	19.80%	21.81%	0.43%	4.44%	5.77%	37.57%	4.49%	3.07%	90.31%
JA	20.69%	25.15%	0.37%	4.22%	3.95%	36.01%	4.19%	2.81%	90.78%

Table 4.10: Characterization of the traffic directed to some particular editions of Wikipedia in terms of the percentages corresponding to visits to articles in any namespace, to edit operations, to actions requested by users, to search operations, to api functions calls, to skins and css files for tailored visualizations of articles, to the icon itself of the Wikipedia, and, finally, to mediaWiki extensions.

4.4 Temporal patterns describing the use of Wikipedia

As stated in chapter 1, our analysis has considered the requests submitted by users throughout the year 2009 to the Wikipedia editions having the highest volumes of both articles and traffic. Therefore, in order to find temporal patterns related to the use of Wikipedia, we have studied how the number of different types of requests submitted evolve throughout several time periods different in length such as days, weeks, months and, even, the whole year.

In the following, we are going to present a temporal characterization of the traffic directed to the set of Wikipedia editions analyzed in this thesis as well as to the overall traffic directed to all the Wikimedia Foundation projects. As we saw in section 4.3, the editions studied in this thesis constituted by the 91% of the overall traffic directed to Wikipedia. Considering that we are not filtering all the traffic to these Wikipedias but only the requests asking for certain namespaces and actions, we have considered appropriate to assess if the filtered traffic temporarily evolves in the same way that the general traffic to the Wikipedia project does. In this way, Figure 4.9 presents the yearly evolution of the traffic directed to the aggregated set of the editions of Wikipedia in order to compare it with the overall traffic directed to all the projects maintained by the Wikimedia Foundation. Moreover, Figure 4.9 also plots the number of requests filtered [4] after our analysis. As we can see, all three lines, each in its corresponding scale, present a relative similar behavior over time. The decrease appreciated since November till the end of the year is documented in [5] and is due to a problem in the reception of the UDP packets containing the Squid log lines at the Wikimedia Foundation aggregator host. The slumps in the number of visits that appear in February, June, July and October correspond to the days in which we were not able to receive and store the log lines from the Wikimedia Foundation Squid systems due to technical problems related to our system's storage capacity.

In order to examine more accurately the relationship between the traffic to Wikipedia and to all the Wikimedia Foundation projects, Figure 4.10 shows the correlation between the daily measures of both traffics corresponding to the entire year. As it is shown, there is a positive correlation between

[4] There is a summary of data related to the quantitative analysis of filtered requests available at `http://gsyc.es/thesis/tables/tabFilterReq.pdf`

[5] `http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm`

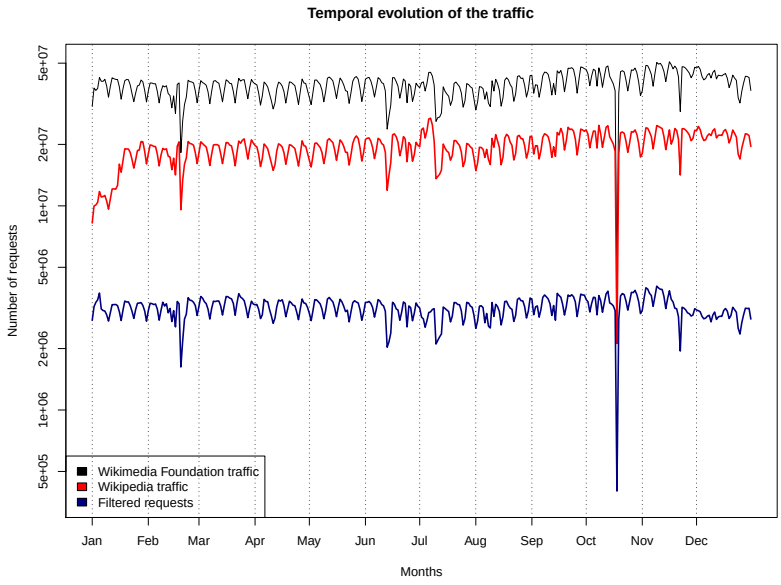


Figure 4.9: Evolution of the traffic throughout 2009.

the two variables so, effectively, Wikipedia traffic can a serve as model of the overall received by the Wikimedia Foundation. This means that temporal variations involving Wikipedia requests will have a proportionally repercussion in the traffic to all the Wikimedia Foundation projects.

If we consider the information about traffic reported by the Erik Zachte's portal and based on the Mitouzas logs, we can compare the evolution of the monthly number of visits and requests. Using this source of data we cannot obtain information related to a more precise period of time so the number of requests has to be studied month by month. Figure 4.11 presents the evolution of the traffic to several editions of Wikipedia for every month of 2009 as reported by Zachte's portal and by our own analysis. Zachte's data corresponds to the lines in the top of chart, those plotted using circles, as they represent the total number of visits without performing any sampling process. In the same way, the lines in the bottom of the chart, drawn with triangles, correspond to the results obtained from the sample we are receiving. The data, both Zachte's and ours, corresponding to the same Wikipedia edition have been plotted using the same color for comparison purposes. The chart confirms that our data follow a similar temporal evolution than the general ones and also serves to validate the filtering process as, in the logarithm scale, our data correspond to the 1/100 of the overall requests.

In respect to the edits, Figure 4.12 present the monthly evolution for these operations as reported by Zachte's portal as well as by our own analysis. Again, both evolutions are parallel, for all practical purposes.

Once we have checked that our filtered requests evolve in a similar way than the general traffic, we undertake the analysis of their distribution over time. In this way, we will examine separately the behavior over time of different kinds of requests. Hence, Figures 4.14 and 4.13 show the evolution of the different types of request during the entire year 2009 and corresponding to all the considered Wikipedias. It is important to recall that we are considering a visit to an article as its page request for reading and without involving any other action. In turn, edit operations are intended as modifications over the content of articles that are finally saved to the database. The difference between edit requests and edit operations is that the first are issued when users just click on the ”edit“ tab placed on top

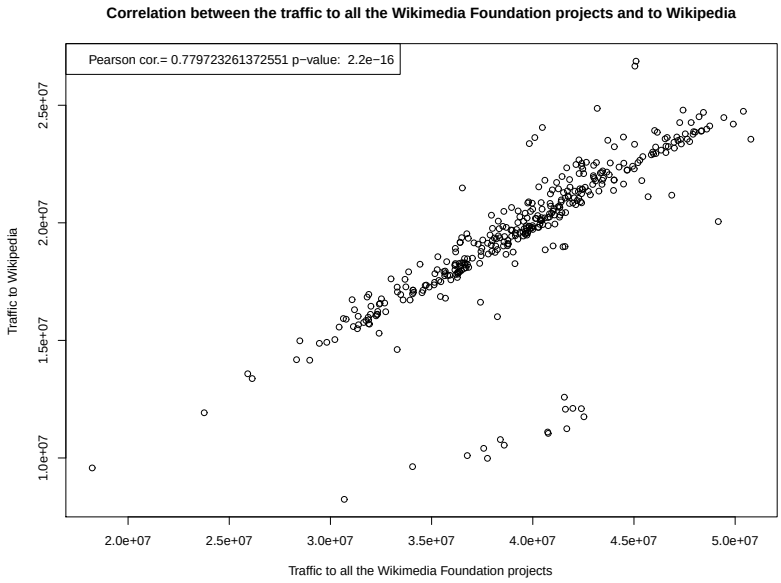


Figure 4.10: Correlation between the traffic to Wikipedia and to the whole set of Wikimedia Foundation projects throughout 2009.

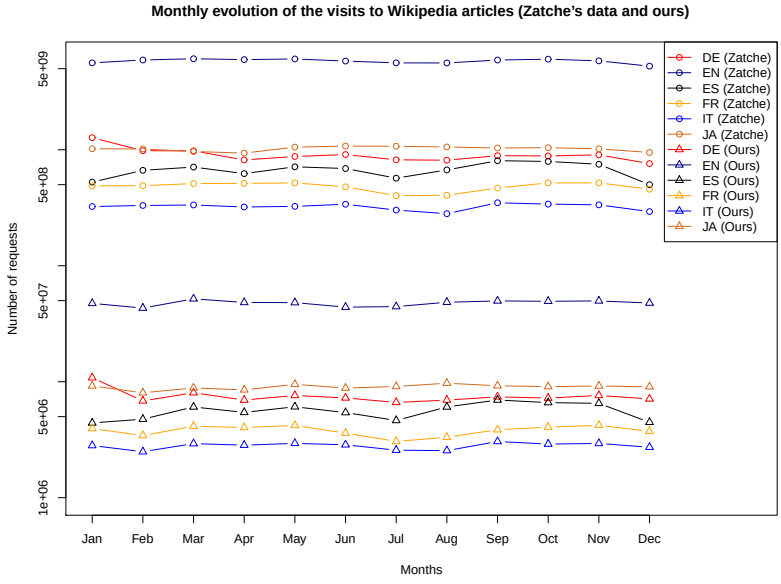


Figure 4.11: Comparison of our results about the evolution of visits to Wikipedia articles throughout 2009 with Zachte's data.

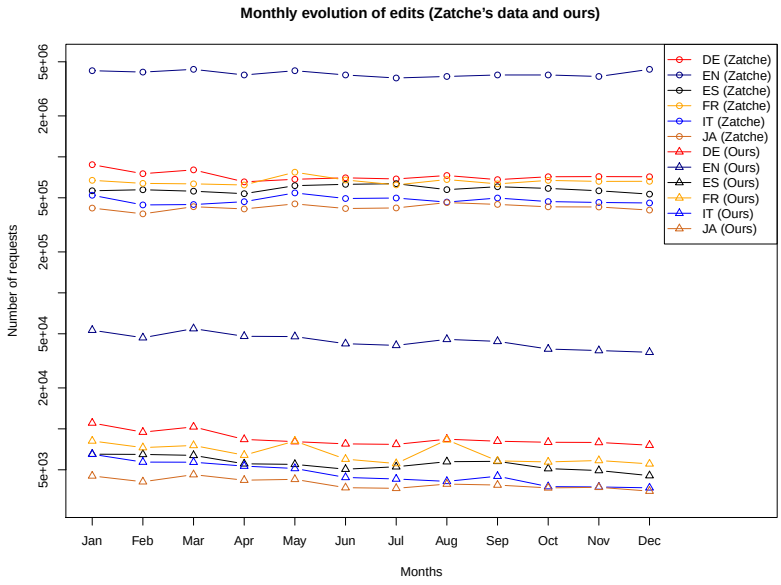


Figure 4.12: Comparison of our results about the evolution of edits on Wikipedia articles throughout 2009 with Zachte's data

of the articles' pages whereas the latter are generated when users indicate a write operation to the database to save their changes or their contributed contents. Submit operations are those directed to preview the result of the modifications performed on the current content of an article or to highlight the differences introduced by a given edit operation in curse. History requests present the different revisions (edit operations) performed on an article's content and leading to its actual version and state. In accordance with this chart, only those URLs involving visits, searches and edit requests would exhibit temporal repetitive patterns. Other types of requests such as edits (save operations), history reviews or submits for previewing changes would present an irregular distribution over time. In the aim of performing a thorough examination, in the following we analyze and compare the temporal evolution, both monthly and weekly, of the aforementioned types of requests.

Let us compare, first, the monthly evolution of visits and edits and, after, the different types of filtered actions. Edits and visits are always considered as belonging to a certain Wikipedia edition because of our interests in patterns corresponding to particular communities of users. In this way, Figure 4.15 shows the monthly evolution of the visits and edits submitted to the English and German Wikipedias [6]. Moreover, visits presented in Figure 4.15 correspond to articles in the main namespace which is the one involved in common read operations. The idea, here, is to compare, not the figures, but the tendency during the different months analyzed and, as it can be observed, visits and edits follow considerably similar temporal evolutions. In order to disaggregate monthly data and to obtain a closer perspective, we have analyzed these requests at the level of days of the week. This analysis is described below and reveals that visits and edits present, in practice, almost identical tendencies in some of the considered Wikipedia. This fact can be interpreted as a strong correlation between the evolution of visits and edit operations which indicates that general visitors, in a moment, tend to become contributors. Of course, a deeper quantitative examination will be performed later to explore

[6]For the rest of considered Wikipedias: `http://gsyc.es/~ajreinoso/thesis/figures/monthVisEd.eps`

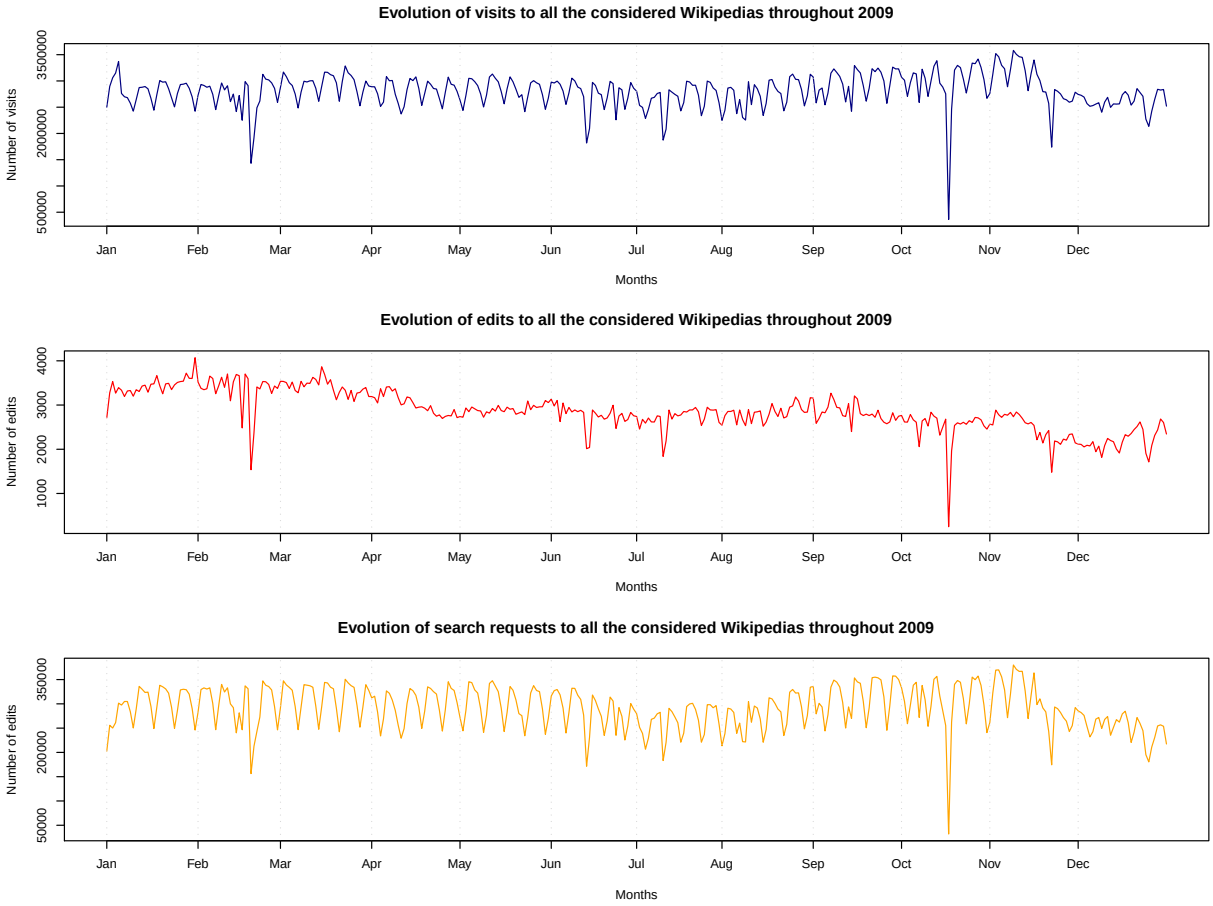


Figure 4.13: Evolution of visits, edits and search requests aggregated for all the considered Wikipedias throughout 2009 .

the rates of both visits and edit operations to analyze the existence of a elite of contributors or, on the contrary, the increase of contributions from users with a small number of edit operations submissions (the power of the fews).

Now, let consider the monthly distribution of the different types of actions addressed in this thesis. Therefore, Figures 4.16 presents the monthly evolution of edit requests, edit operations as well as history, submit and search requests for the German, English, Spanish and French Wikipedias [7]. All these figures corresponding to the different types of actions are very similar in scale. However, we have preferred to present them using a logarithm scale in order to obtain more differentiated lines and, by means of this, a higher level of detail. As it can be observed from the chart, search operations are the most numerous actions followed by the edit requests. As we can see, edit requests are considerably higher in number than edit operations. This means that an important number of edit requests are not finished by the corresponding write request to the database. Moreover, edit (write) operations are always very near the submit ones, which means that most of users regularly preview their changes before indicating their permanent storing to the database.

We undertake now the same analysis although focusing on weeks. The aim is to determine whether

[7] For the rest of considered Wikipedias: `http://gsyc.es/~ajreinoso/thesis/figures/monthAct.eps`

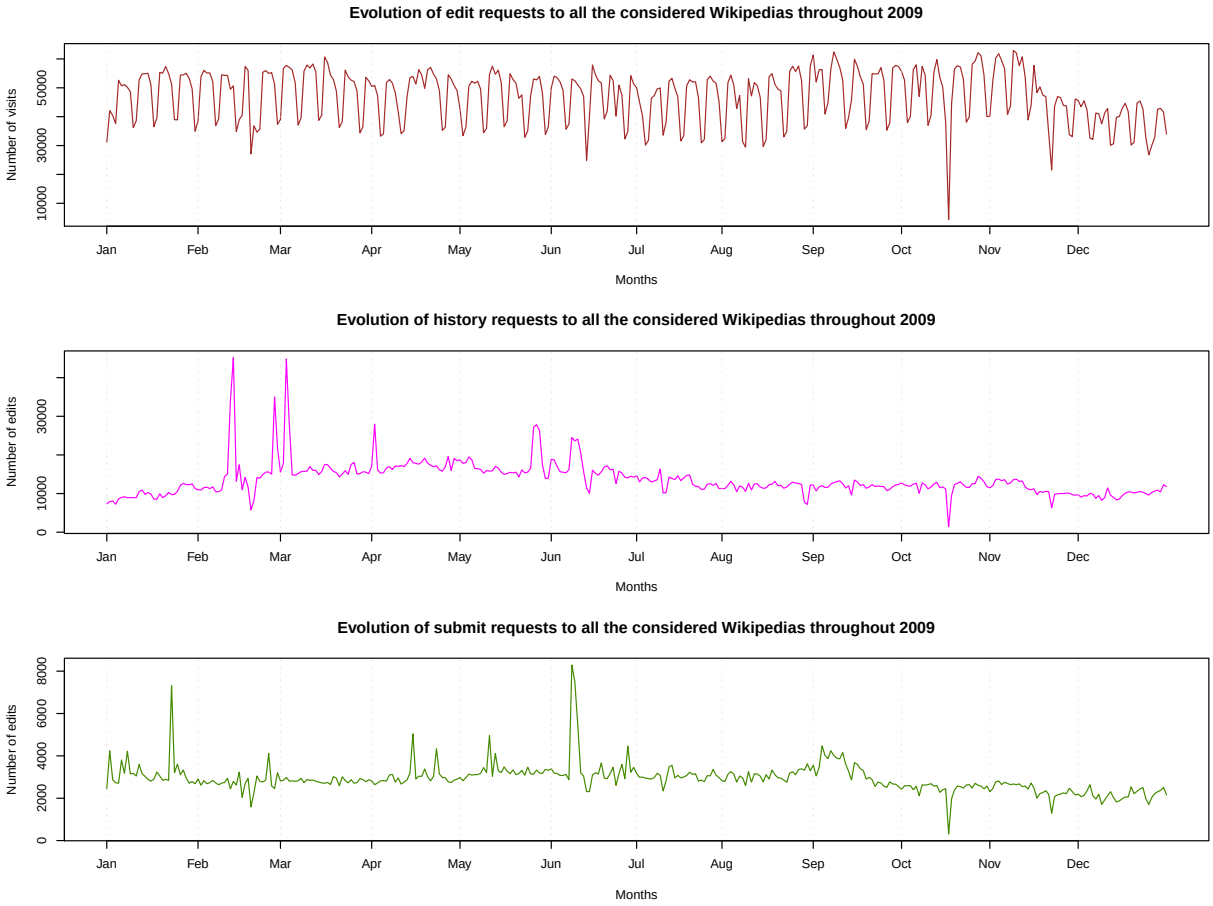


Figure 4.14: Evolution of submits, edit requests and history reviews aggregated for all the considered Wikipedias throughout 2009.

there are patterns involving any type of requests that are repeated throughout the days of the week disregarding changes of month. First we present, in Figure 4.17, the aggregated number of requests corresponding to every day of the week considering all the requests received during the year. In the Figure 4.17 it is patent a continuous decrease of the number of requests as the week advances. In this way, on Saturdays users would submit the lowest number of requests to Wikipedia as on Sundays there is a little increase. In the following, we will see if this same tendency is maintained during different months as well as for the requests corresponding to each particular considered Wikipedia.

First of all, we are going to present the evolution of visits, edits and other actions regarding all the whole weeks (from Monday to Sunday) corresponding to 2009. This is done, for example, in Figure 4.18 that provides a closer perspective and confirms the similar weekly evolution of visits, searches and edit requests in contrast to the spurious and irregular nature of the requests consisting in edit operations, history and submits. Charts reflect very different temporal distributions depending on each edition of Wikipedia. Because of this, we are presenting the charts corresponding to the Spanish and Japanese Wikipedias as the former presents relatively well-defined and identifiable

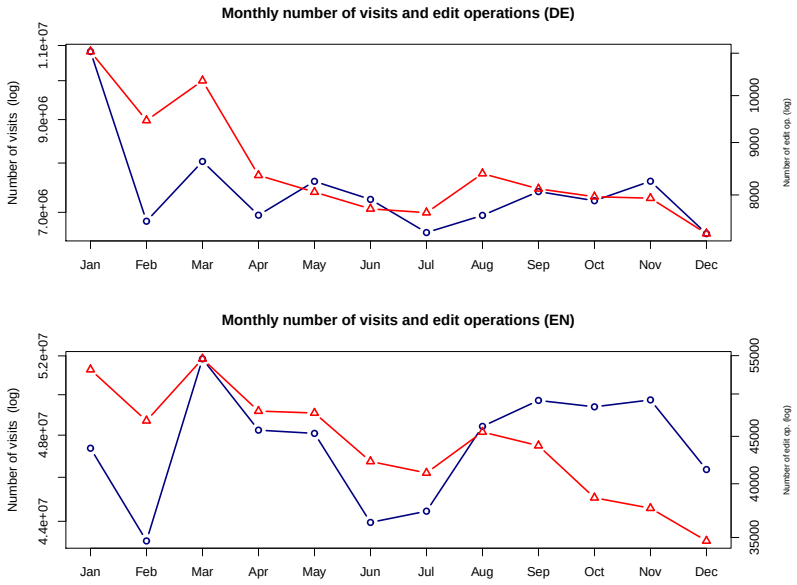


Figure 4.15: Number of monthly visits to articles and monthly edit operations in the considered Wikipedias throughout 2009. The blue line reflects the visits while the red line is related to the save operations. Left y axis corresponds to the scale for visits where as the right one corresponds to the scale for edit operations. In this way, values for the visits line have to be transported to the left y axis and the ones for the edits line are in the right y axis. The graph is presented in this way because visits and edits operations are very different in scale so presenting them together will cause a considerable loose of detail in the tendency examination.

patterns whereas the latter shows more irregular distributions[8]. More in detail, all the editions present a weekly repetitive pattern for visits except the Japanese, French and Polish Wikipedias which do not show such well defined patterns. Considering the requests for searchs, again these two Wikipedias present the most irregular distributions. For all the Wikipedias, submits and edit operations present very similar evolutions. However, it is interesting to check how, only for the German Wikipedia, the number of submit operations is always higher than the save ones. This indicates that in this Wikipedia almost all the changes are previously assessed.

If we aggregate the different types of requests and analyze their distributions over the days of the week, as presented in Figures 4.19 for the German Wikipedia [9], we can study the differences and similarities among the distributions of the diverse types of requests in the considered Wikipedias.

[8]For the rest of considered Wikipedias: `http://gsyc.es/~ajreinoso/thesis/figures/week1.eps` and `http://gsyc.es/~ajreinoso/thesis/figures/week2.eps`

[9]For the rest of analyzed Wikipedias:
`http://gsyc.es/~ajreinoso/thesis/figures/weekEN.eps`
`http://gsyc.es/~ajreinoso/thesis/figures/weekES.eps`
`http://gsyc.es/~ajreinoso/thesis/figures/weekFR.eps`
`http://gsyc.es/~ajreinoso/thesis/figures/weekIT.eps`
`http://gsyc.es/~ajreinoso/thesis/figures/weekJA.eps`
`http://gsyc.es/~ajreinoso/thesis/figures/weekNL.eps`

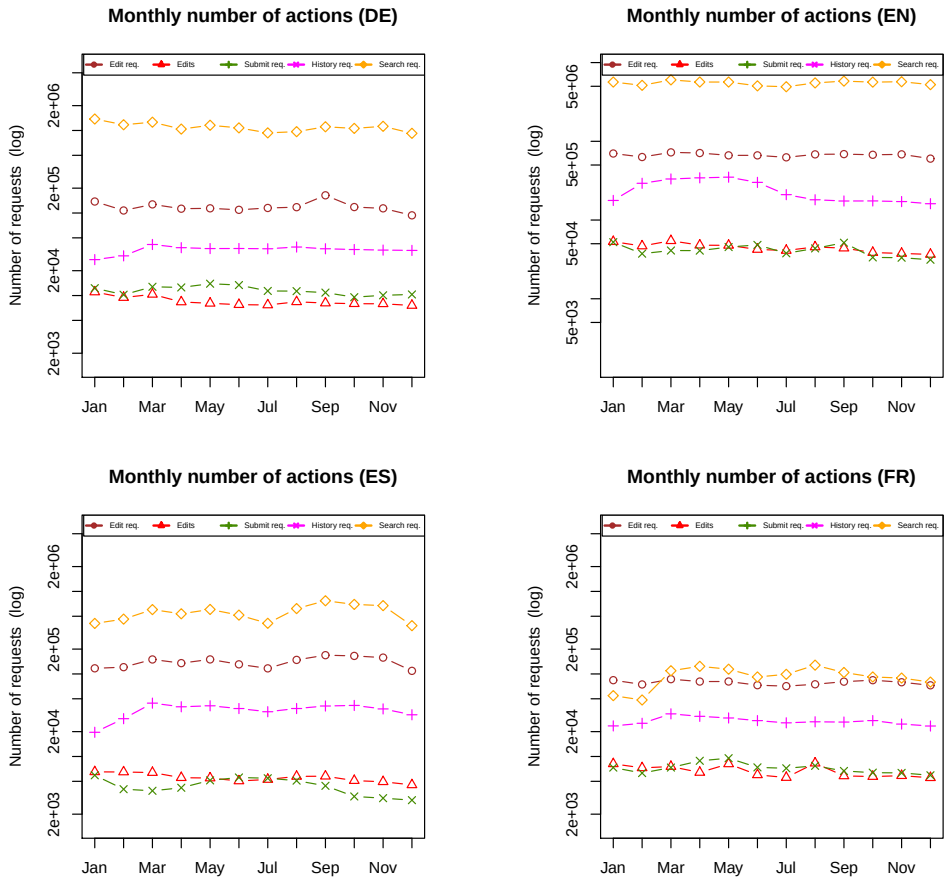


Figure 4.16: Monthly aggregation of the different types of actions in some of the considered Wikipedias.

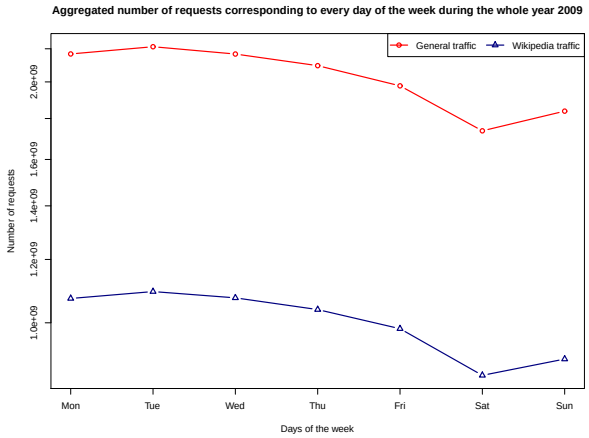


Figure 4.17: Aggregated number of requests corresponding to each day of the week.

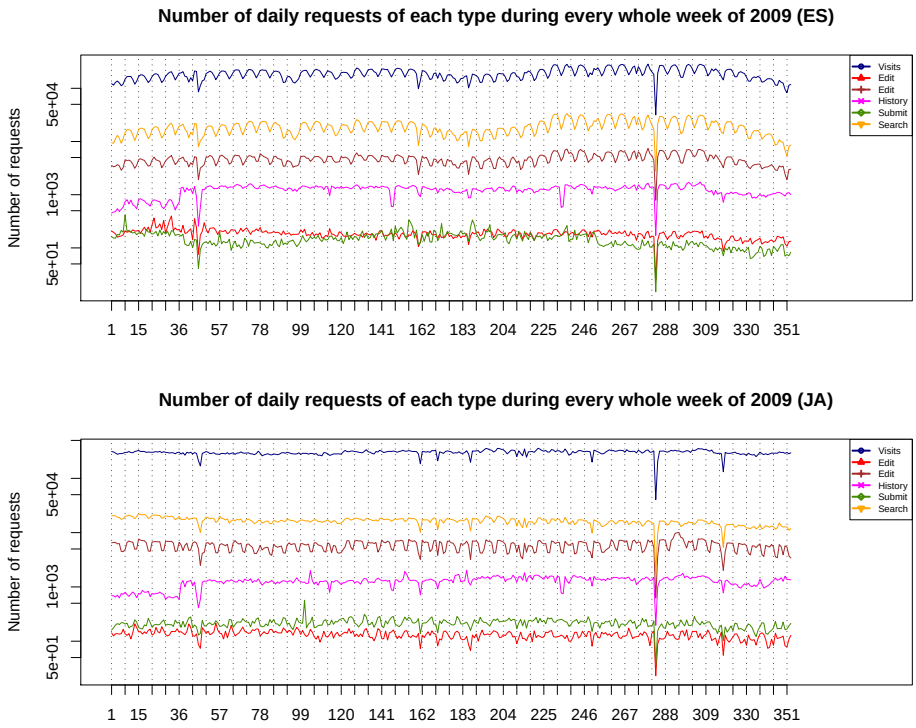


Figure 4.18: Number of daily requests of each different type issued for every complete week of 2009. This chart presents the evolution of each kind of request during every whole week of 2009 in different editions of Wikipedia. X-axis begins with the first Monday of the year and finishes with the last Sunday and each vertical pair of divisions delimit an entire week.

After studying this type of charts, we can state that all the types of requests are similarly distributed throughout the days of the week in the considered Wikipedias. Submit and edit requests are the ones with outstanding differences among all the editions and, consequently, they adopt more different patterns. However, in the case of the German, English and Spanish Wikipedias they conserve a relatively similar shape that also match the evolution of visits.

We decided to undertake the study of the evolution of visits and edits at the level of the days of the week in the aim of finding a meaningful closeness between their two temporal variations. As a result of such kind of analysis, Figure 4.20 presents the evolution of both types of requests throughout the days of the week for all the considered Wikipedias. Visits and edits, in each Wikipedia edition, correspond to the entire year and have been grouped by their day of issue. So, Figure 4.20 presents their compared progressions and shows a considerably closeness in the evolution of both types of requests in several Wikipedias. Nevertheless, the number of edits tends to raise in weekends for a group of them (French, Japanese, Dutch and Polish). That could mean that, in those editions, editors are not part of the great mass of people visiting the articles but just a minor group devoted to contribute

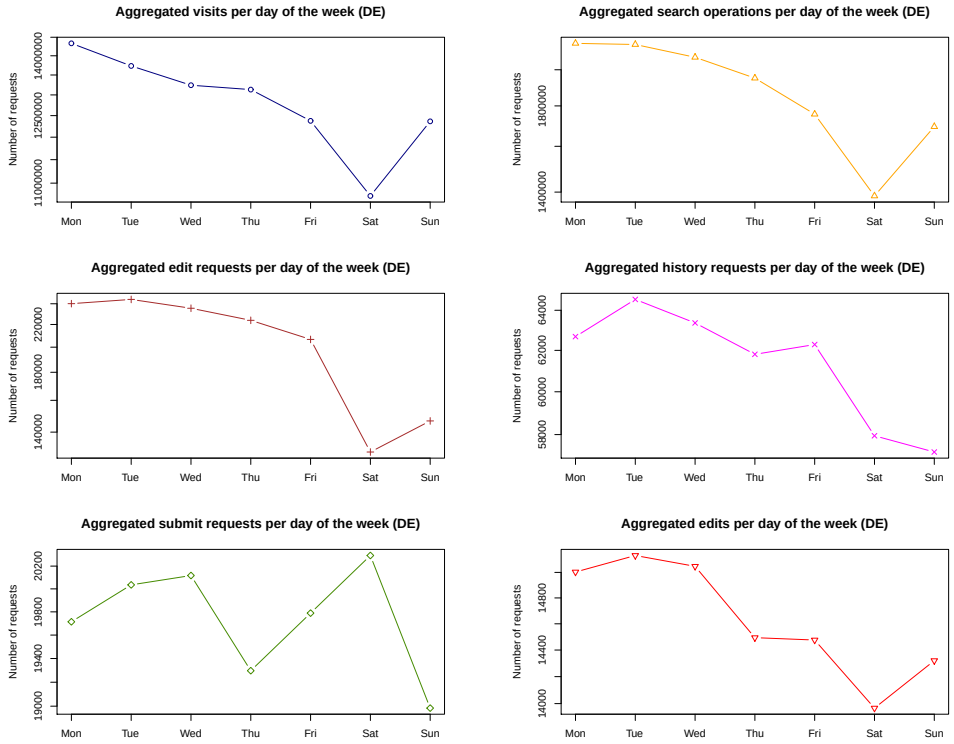


Figure 4.19: Evolution of the different types of requests throughout the days of the week (DE).

or to maintain them[10].

Moreover, Figure 4.21 presents the weekly distributions of edits (saves) and edit requests. Again, we have to pay attention to the different edges and scales for each type of action. In other case, it would seem that there are more edit operations on some days (specially on Saturdays) than edit requests (impossible situation because every edit operation has to be preceded by the corresponding edit request). The graph shows how edit requests and completed edits are closer on Saturdays than in any other week day for some of the considered Wikipedias. This is due to the fact that on Saturdays in the French, Japanese, Polish and Dutch Wikipedias, edit requests decrease whereas completed edit operations raise. In other words, almost every edit request submitted on Saturday in these Wikipedias ends with the corresponding edit that implies a write operation to the database. In the same way, Figure 4.22 analyzes the evolutions of edits and submit operations. Continuing with the presentation in two axes, first of all we can see that there is less difference between the scales corresponding to each axis than in the case of the comparison of edits and edit requests. This is due to the fact that edits and submit operations are more similar in number that edits and edit requests. Moreover, whereas edit requests are always higher in number than edits operation, this not true in the case submit operations and edit operations. In this way, in the German and Japanese Wikipedias, submit requests are always greater in number than edit operations whereas in the Russian or the Italian Wikipedias present the opposite situation.

[10]The same comparison between visits and edit requests, history requests and search requests, respectively, is presented in `http://gsyc.es/~ajreinoso/thesis/figures/viEdReq.eps`, `http://gsyc.es/~ajreinoso/thesis/figures/viHiReq.eps` and `http://gsyc.es/~ajreinoso/thesis/figures/viSeReq.eps`. As it can be observed, all the requests corresponding to actions are temporarily distributed in the same way than visits do.

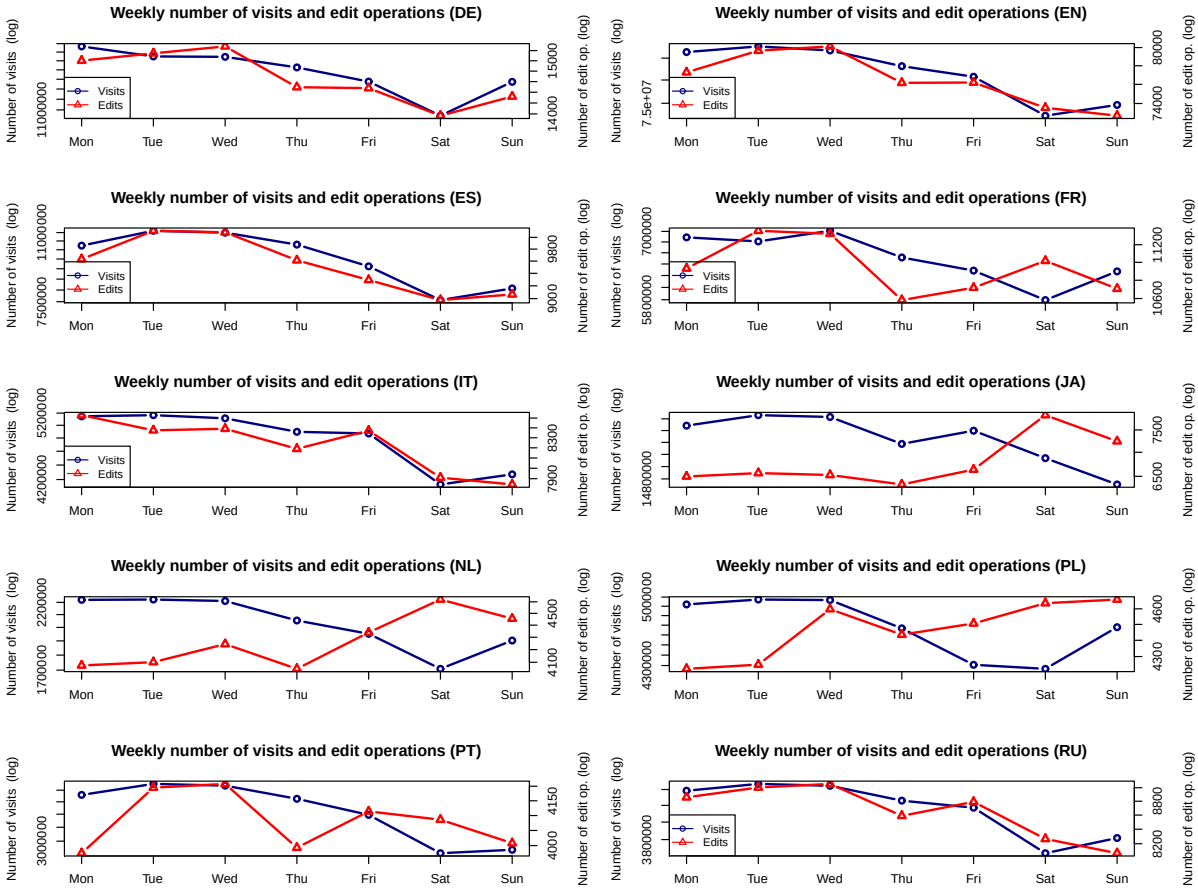


Figure 4.20: Evolution of visits and edits throughout the days of the week in the different editions of Wikipedia.

In a deeper level of detail, Figures 4.23 and 4.24 present the evolution of visits and edits in different editions of Wikipedia throughout the days of the week corresponding only to two months. As it can be shown, both types of requests coincide in their temporal evolution. Again this type of chart does not focus on the differences in magnitude but in evolution tendency during the days of the week.

4.5 Behavioral patterns

As we introduced in chapter 1, one of the aims of this thesis is to describe behavioral patterns related to the use of Wikipedia. Of course, behavior is a wide concept and may involve a great variety of information elements. Here, we will focus on some of them, specially from the perspective of the comparison between the number of visits that the considered Wikipedia editions receive and the number of edits performed on them. In this way, our objective is twofold. On the one hand, we want to determine whether the contributions to the different Wikipedia editions come from the bulk of users or just from a minority group of them. On the other hand, we also aim to obtain different quantitative parameters about the type of participation of each community of users when browsing the

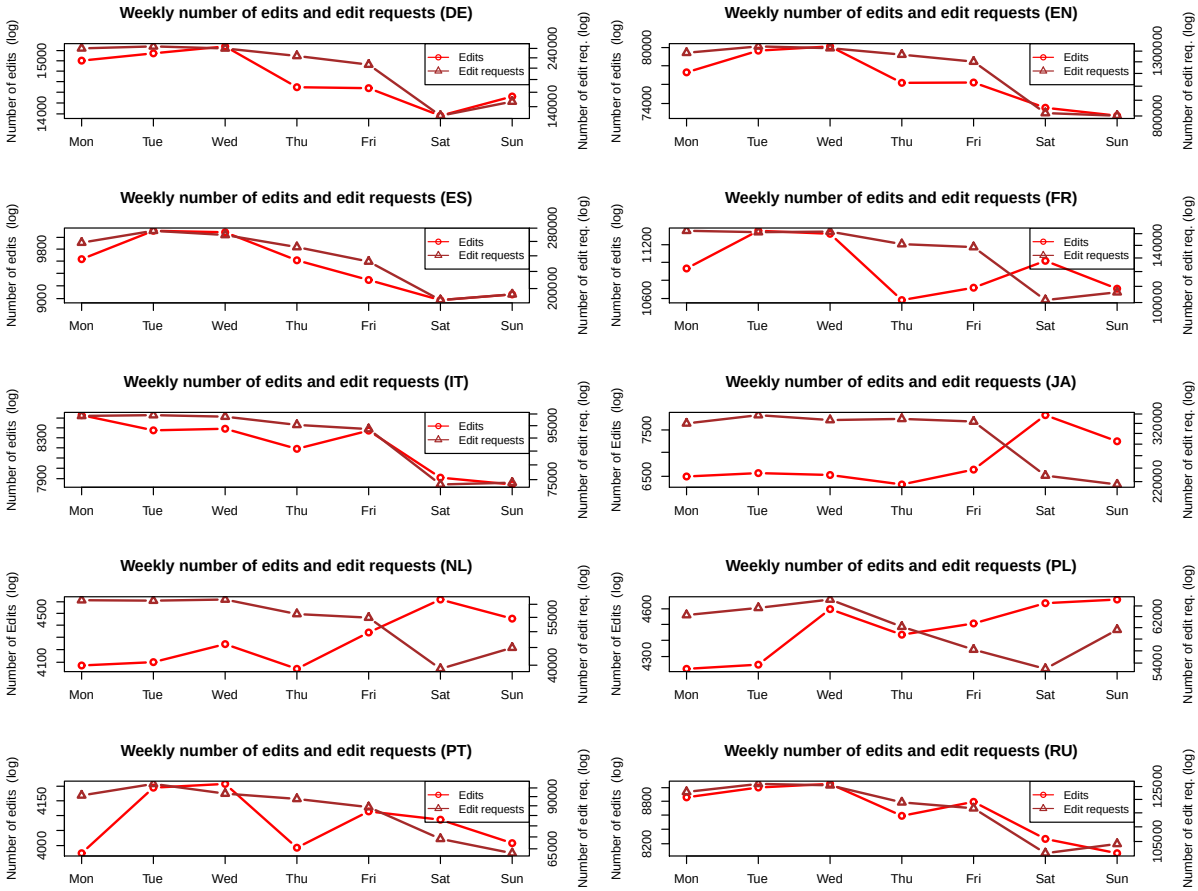


Figure 4.21: Evolution of edits and submit requests throughout the days of the week in the different editions of Wikipedia.

Encyclopedia.

First, and as a continuation of the analysis of temporal patterns, Figures 4.25 and 4.26 show the correlation between the number of visits and edits corresponding to the days of the week. As images show, the German, English, Spanish, Italian and Russian Wikipedias do present positive correlations between visit and edits throughout the days of the week. The rest of Wikipedias present low correlation values between the two types of requests or even negative ones that indicates that the two kind of requests are inversely correlated. This is the case of the Japanese and Dutch Wikipedias where visits and edits follow completely opposed tendencies.

If we compare other types of requests to assess if they evolve in a similar way than visits do, we find that search requests and visits are highly correlated [11] in absolutely all the considered editions. The issue of edit, history and submit requests [12] is also positively correlated to visits in all the considered editions.

If we focus now on the relationship between edits and edit requests (Figure 4.27) we can appreciate

[11] `http://gsyc.es/~ajreinoso/thesis/figures/corViSe1.eps`
`http://gsyc.es/~ajreinoso/thesis/figures/corViSe2.eps`

[12] `http://gsyc.es/~ajreinoso/thesis/figures/corViEdReq1.eps`
`http://gsyc.es/~ajreinoso/thesis/figures/corViEdReq2.eps`

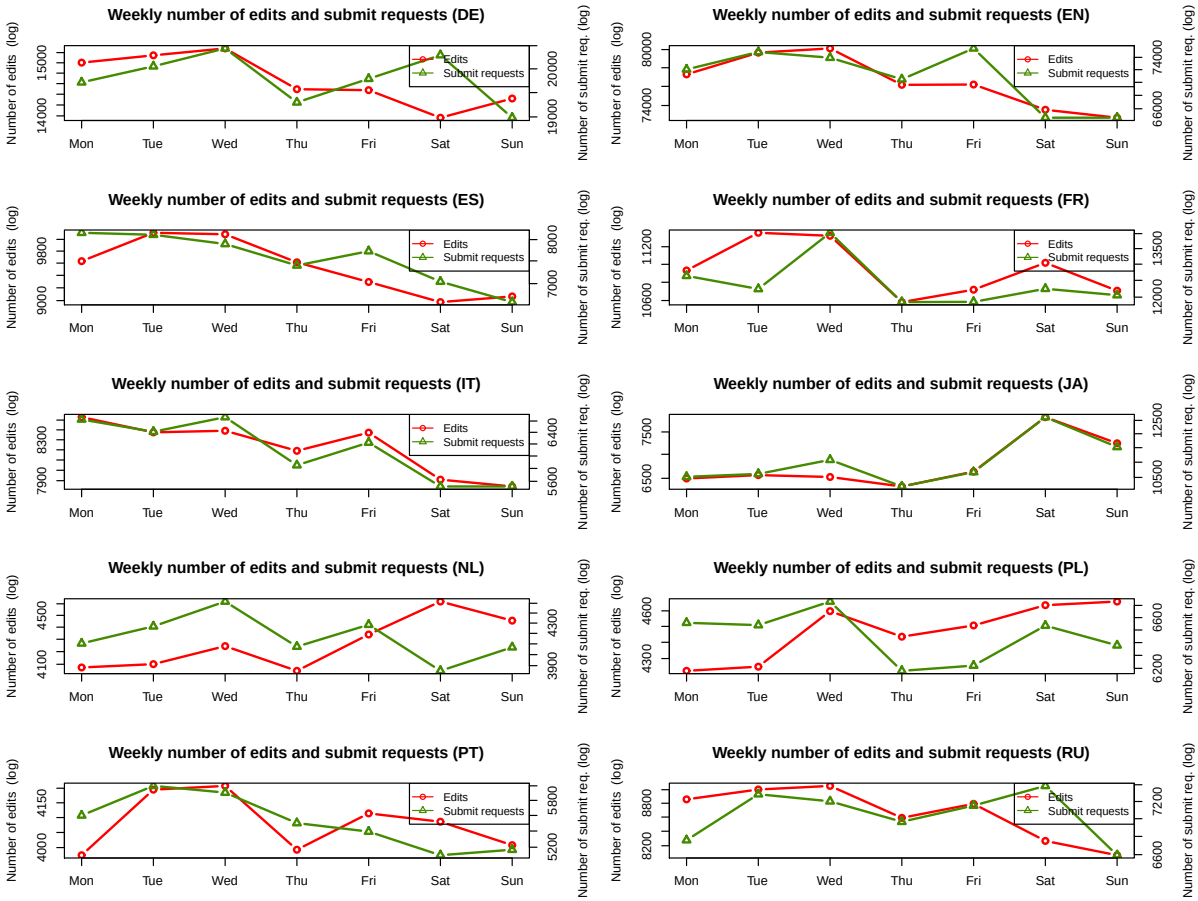


Figure 4.22: Evolution of edits and edit requests throughout the days of the week in the German and English Wikipedias.

that both variables are positively correlated in the German, English, Spanish, Italian and Russian Wikipedias.

Regarding the evolutions of edits and submit requests, we find that only the English, Italian and Russian Wikipedias present correlations between the two measures. That would mean that only the users of these Wikipedias would issue similar values of edits and submit requests in the same days.

Considering that a correlation between the visits and edits for a certain Wikipedia edition can be intended as the participation of a broad group of users in the contributions to its contents and, by so, the result of a more proactive and collaborative community where users acting as visitors, at a given moment, decide to become editors, we have analyzed the ratio between edits and visits for all the considered Wikipedias. Our purpose, in this case, is to assess whether this ratio remains unchanged throughout the year in the different editions and to establish the editions presenting the highest ratios, which could be also considered as the ones having the most participative communities of users. Figure 4.29 presents the evolution of the ratio edits over visits throughout the entire year. In this figure we can see three groups of editions. The first one is made up of the the Dutch, Polish, Italian, French and Russian Wikipedias that present higher rations, a second group would consist of the Spanish, Portuguese, English and German Wikipedia with a lower ratio. Finally, the Japanese is the only one in the third group with the least ratio. Interestingly, only the Russian and Italian editions, which

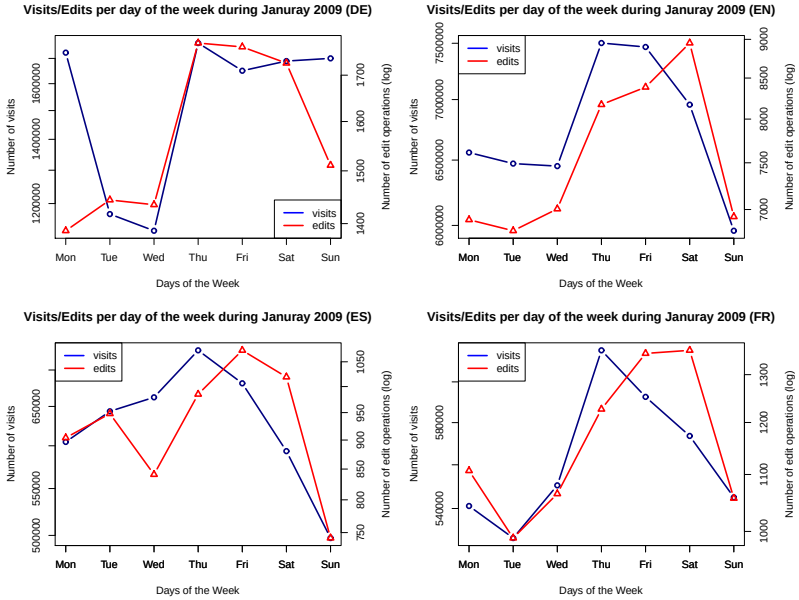


Figure 4.23: Evolution of visits and edits throughout the days of the week in the different editions of Wikipedia during January 2009.

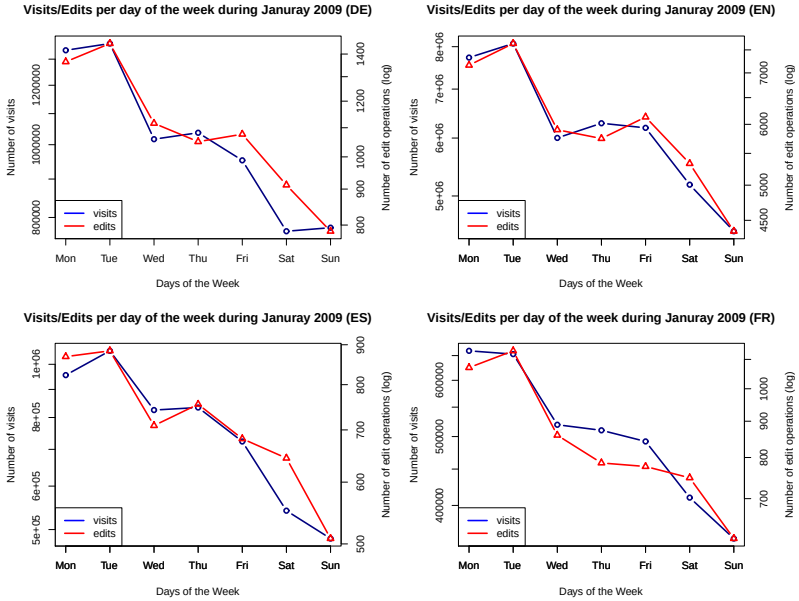


Figure 4.24: Evolution of visits and edits requests throughout the days of the week in the different editions of Wikipedia during June 2009.

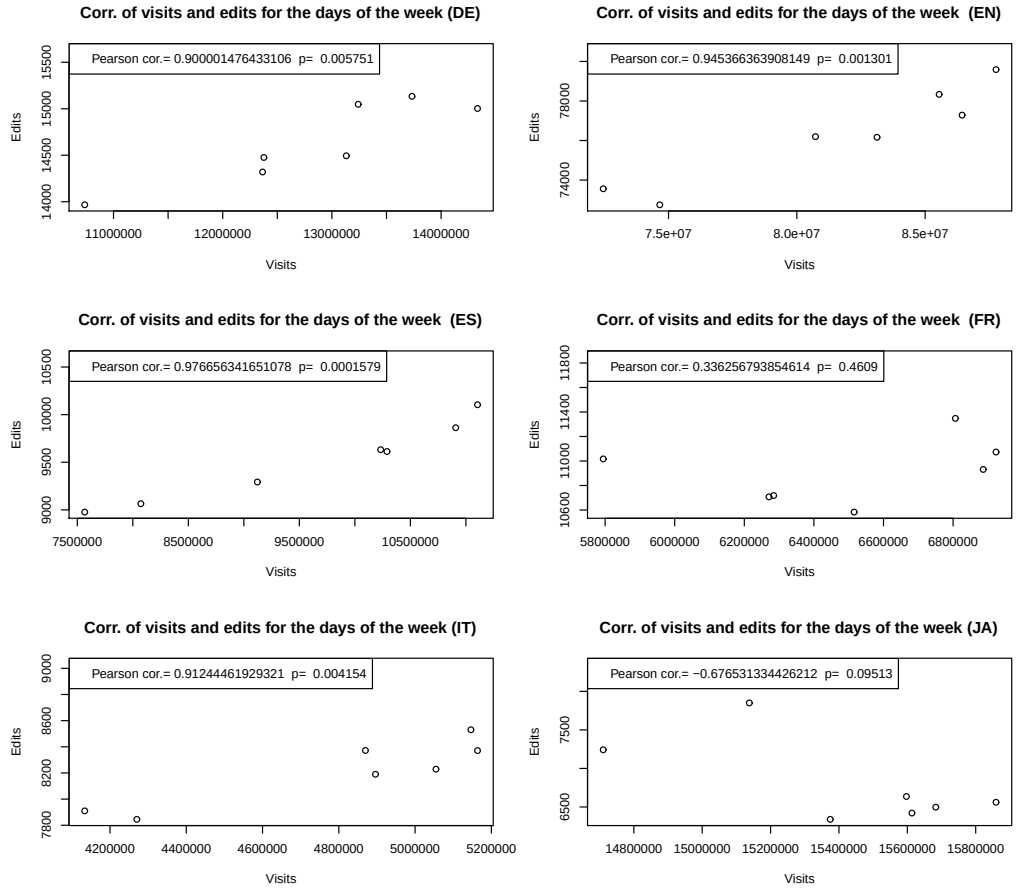


Figure 4.25: Correlation between visits and edits through the days of the week for the German, English, Spanish, French, Italian and Japanese Wikipedias.

presented positive correlations between edits and visits, are included among the editions with higher edits/visits ratios. Regarding the evolution of the ratio edits over visits for the different Wikipedia editions, although there are differences in the plots of each one of them, we found a relative similarity in their shapes. Effectively, most of them decrease, although with different inclines, from January till May-June and they start raising after these two months. Again, there is a general drop after September with an slightly increase in December for most of the editions except the Russian, English and Japanese ones.

Another interesting parameter can be the ratio of edits performed over edit requested as we have noticed that there is a great number of edit requests that are not finished with the corresponding save operation to the database. In this way, Table 4.11 presents the ratios corresponding to the different editions of Wikipedia decreasingly ordered. In this case, we do not considered of interest to analyze the evolution of the ratios throughout the time, so we present them aggregated for the entire year. If we compare 4.29 and Table 4.11 we would observe that Wikipedia with higher ratios of edits over visits are the ones with lower percentages of abandoned edit operations, which is an absolutely interesting finding.

Now, we are going to focus on the number of requests involving the different namespaces and actions in the different Wikipedias. The purpose, again, is to compare behavioral habits exhibited

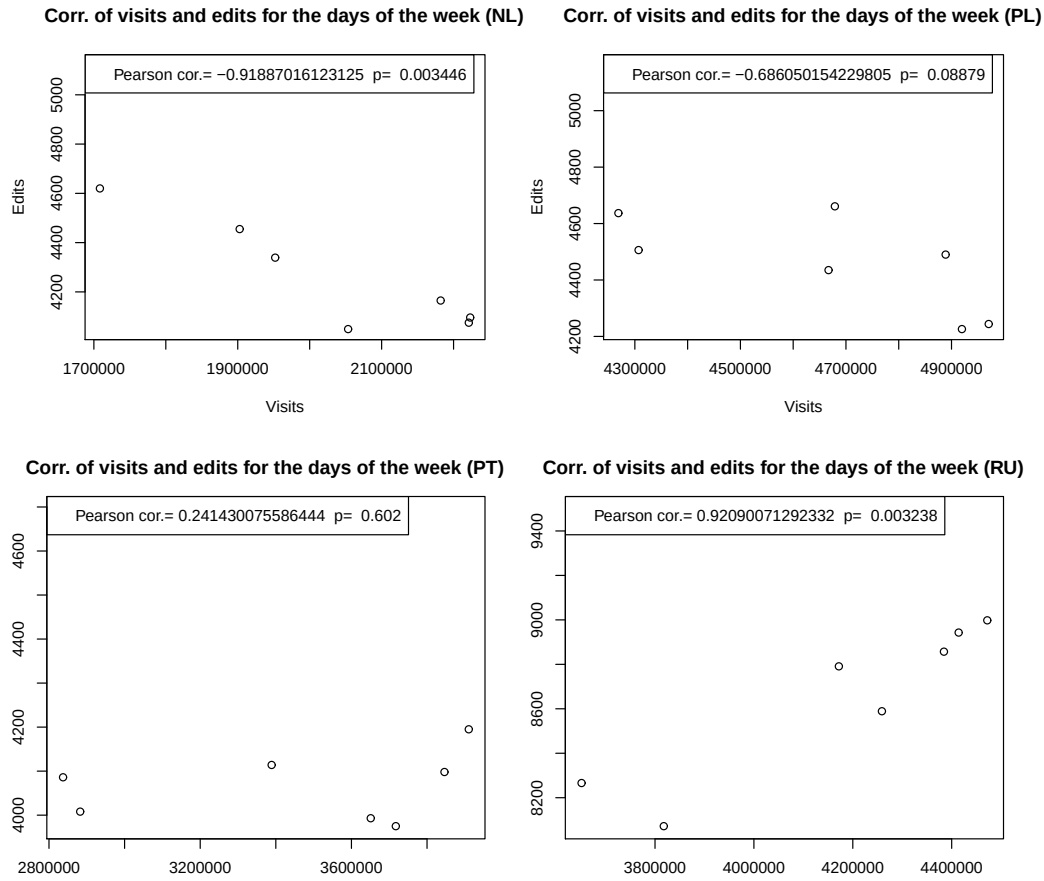


Figure 4.26: Correlation between visits and edits through the days of the week for the Dutch, Polish, Portuguese and Russian Wikipedias.

by the different communities of users. In this way, Figure 4.30 shows the yearly aggregated number of requests asking for each specific namespace from the total of visits to each considered Wikipedia. As expected, articles in the *Main* namespace are the most requested ones followed by special pages created in response to particular users' demands. Because of their order of magnitude, these two namespaces practically cover all the visits to the considered Wikipedias, so the rest of them may appear as negligible. In order to illustrate the different ratios of visits corresponding to namespaces other than the *Main* one, we present, in Figure 4.31, the amount of requests involving each one of them. It is, perhaps, remarkable the few requests involving the *Special* namespace in the French Wikipedia. On the contrary, these requests are hegemonic in the Japanese Wikipedia. The *User_Talk* namespace is mainly used as a communication tool to facilitate coordination and collaboration among users, so higher rates may indicate more collaborative attitudes.

Considering edit operations, Figure 4.32 show the different namespaces to which correspond the edit operation performed in each considered Wikipedia. Most of edits correspond to articles in the *Main* namespace, whereas, higher ratios in edits performed on *User* and *User_talk* namespaces would mean more communicative editions. Interestingly the French Wikipedia presents a high volume of visits to the *User* and *User_Talk* namespaces, however the number of edits to the same namespaces do not preserve the same ratio.

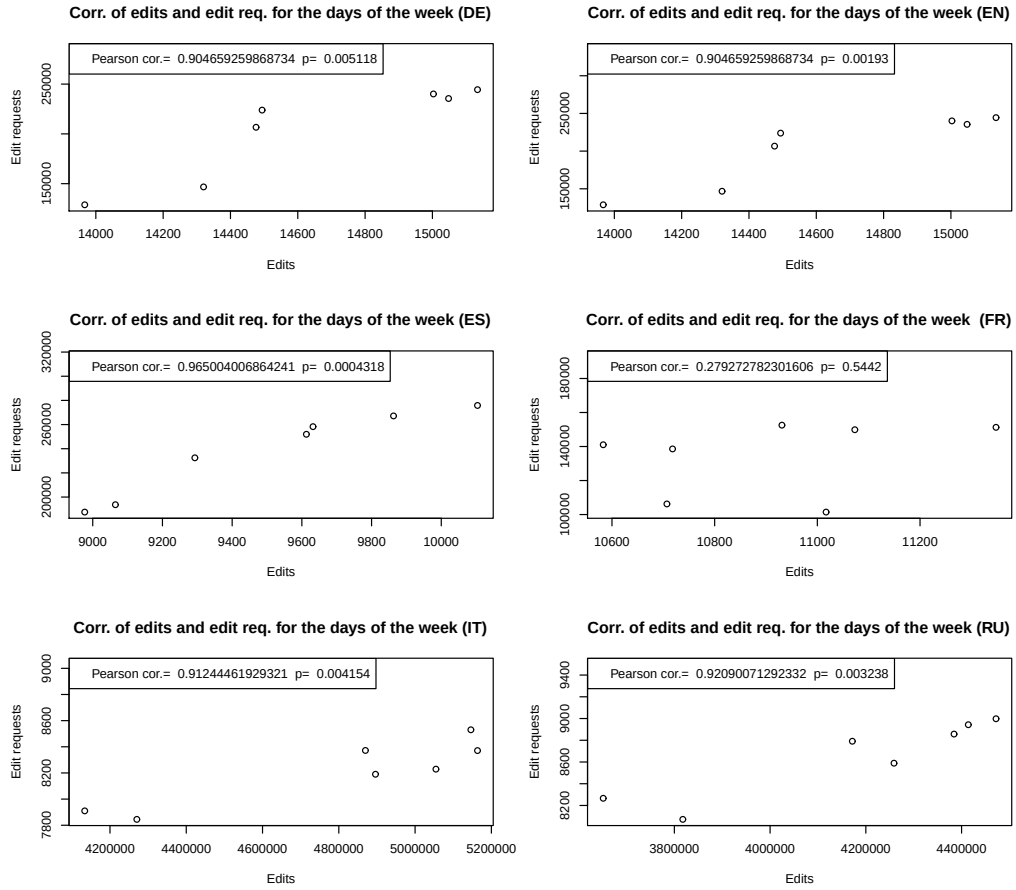


Figure 4.27: Correlation between edits and edit requests through the days of the week for the German, English, Spanish, French, Italian and Russian Wikipedias.

Regarding the different kind of actions that users requests, we have considered of interest to compare the amount of them solicited to every Wikipedia edition. Figure 4.33 shows how many requests involving each different type of action are submitted in each considered Wikipedia. Again, it is specially remarkable the case of the French Wikipedia. Considering that it has the lower ratio of requests to the *Special* namespace according to Figure 4.30 and searches operations are issued as special demands to this namespace, it would be expected that searches had a lower ratio for the French Wikipedia. This fact is confirmed by Figure 4.33.

4.6 Featured contents

In this section we present a statistical analysis of the impact that the promotion of high-quality articles to the featured status has on the attention they receive. Moreover, we also analyze the effect of the appearance of featured articles as examples of quality content in the main pages of several Wikipedia editions in the number of visits they attract. We will use different tests to study these questions. Although they are standard statistical tests, for the sake of completeness, we are citing an introductory text on the topic [Cro05] that can be used to find the full details about them.

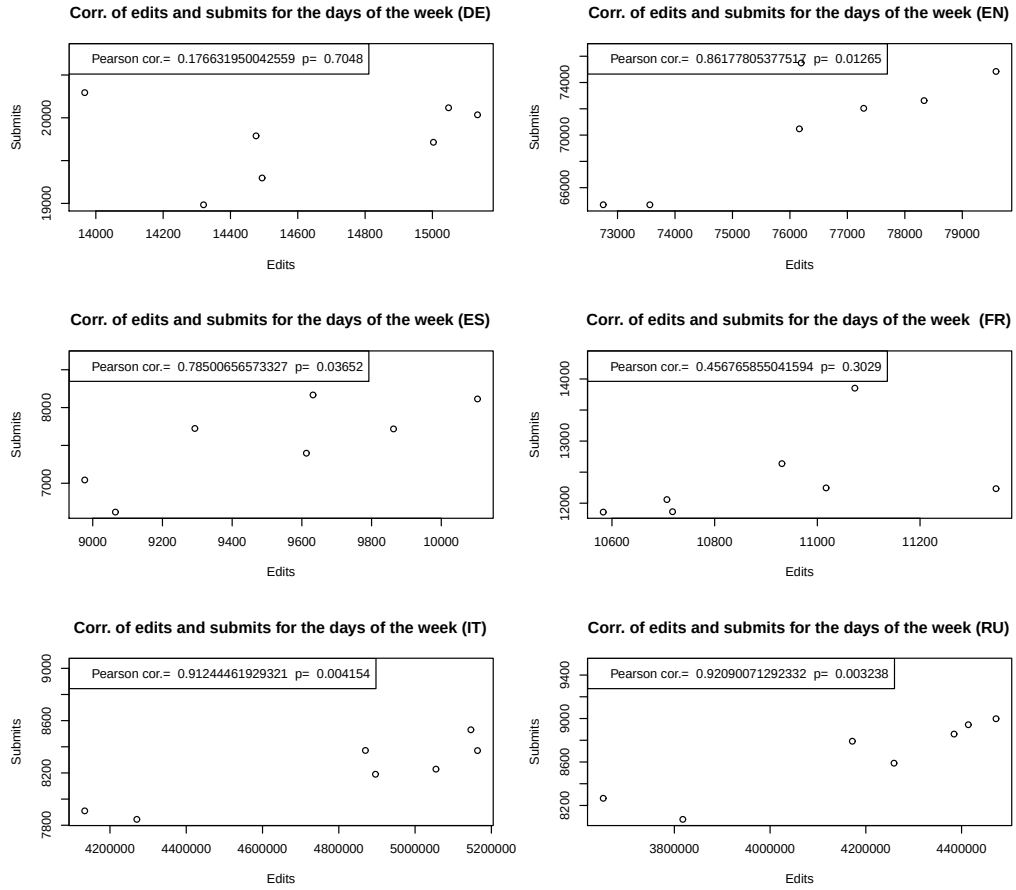


Figure 4.28: Correlation between edits and submit requests through the days of the week for the German, English, Spanish, French, Italian and Russian Wikipedias.

To begin with, we will analyze the attention attracted by the articles presented in the main pages of several Wikipedia editions. In this way, Figures 4.34 and 4.35 show the average number of visits (or mean) for the featured articles presented in the main page of the English Wikipedia during April and November. At a first glance, it seems clear that the so-called "today's featured articles" attract a great amount of attention in the month they are included in the main page, compared with the previous and the following ones.

If we analyze now the same metric applied to the articles just promoted to the featured status in April and November, we obtain that those articles do not receive always the highest number of visits in the month they are promoted as today's featured articles did. This is probably due to the effect of the internal mechanism for promotion that entails a reviewing, a nomination and a consensus process. In this way, the different dynamics exhibited by each community of users in the promotion process are reflected in the visits that the articles attract. As an example, Figure 4.36 presents the evolution of the number of visits for the April's featured articles in different Wikipedias during this month as well as during March and May.

Figure 4.37 shows a boxplot of all the visits to the featured articles presented in the main page of the Wikipedias under study during both considered periods. In the boxplots, the main box shows the bulk of data (those values between the 25 and 75 percentile), and the median is highlighted with

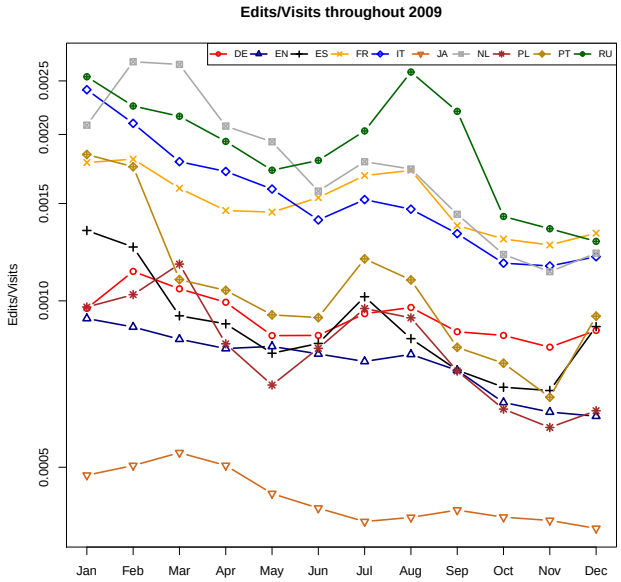


Figure 4.29: Evolution of the ratio edits over visits throughout 2009 for all the considered Wikipedias

Edition	Edits	Edit requests	Percentage of finished edits
IT	57447	632295	9.09%
FR	76377	941017	8.12%
NL	29799	379450	7.85%
PL	31199	419411	7.44%
RU	60516	814103	7.43%
DE	102442	1426027	7.18%
EN	533879	8026886	6.65%
PT	28469	584498	4.87%
ES	66547	1666890	3.99%
JA	47546	2079305	2.29%

Table 4.11: Edit requests finishing with a write operation to the database.

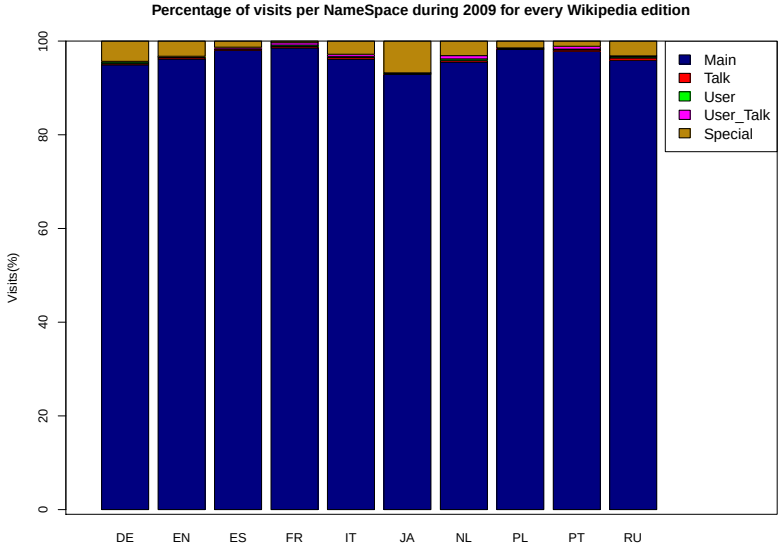


Figure 4.30: Yearly aggregated visits to each namespace in the different Wikipedias.

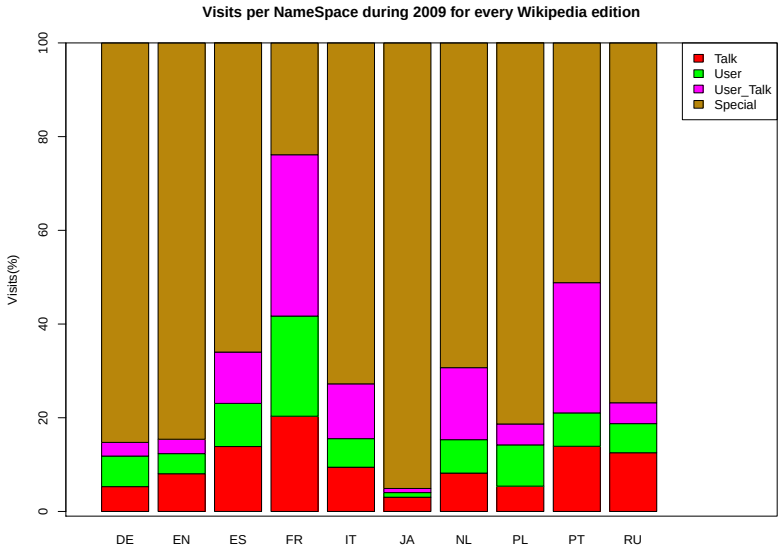


Figure 4.31: Yearly aggregated visits to each namespace (except the *Main* one) in the different Wikipedias.

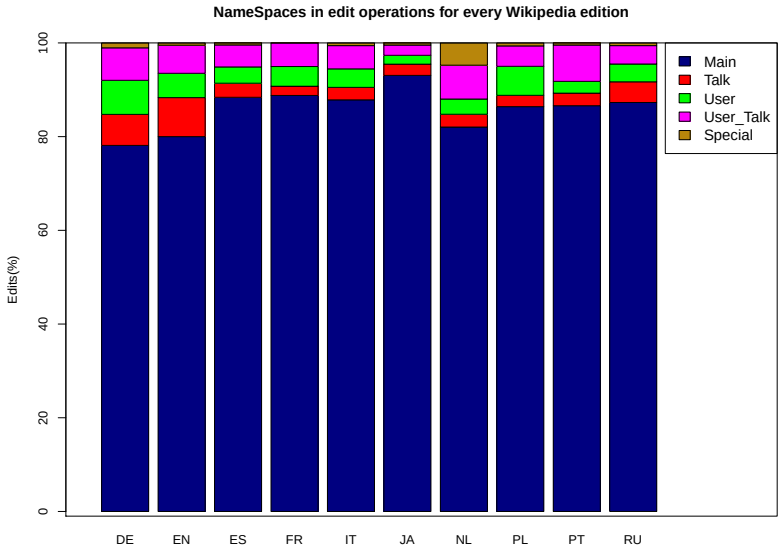


Figure 4.32: Yearly aggregated ratios of namespaces involved in edit requests.

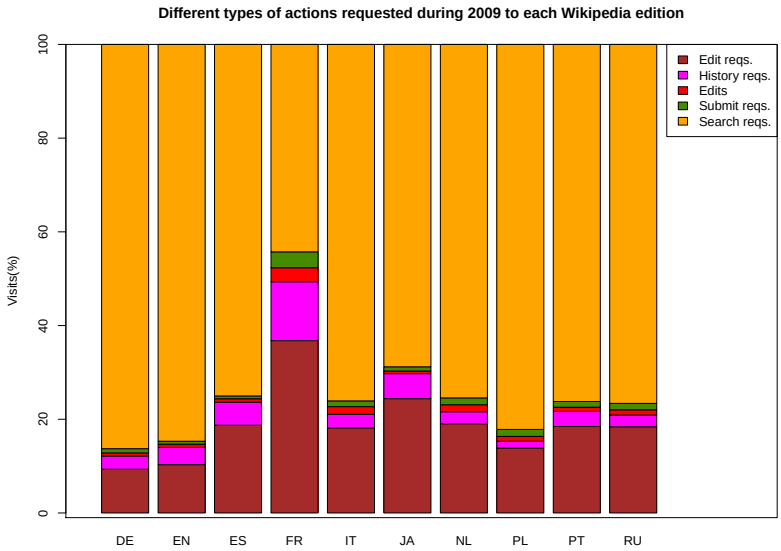


Figure 4.33: Yearly aggregated ratios of requested actions for every Wikipedia edition.

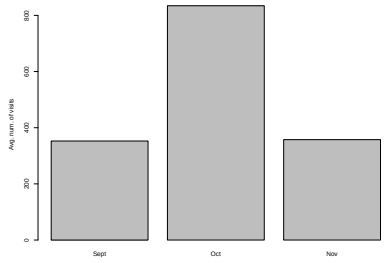


Figure 4.34: Average number of visits for today's featured articles in the English Wikipedia during November 2009.

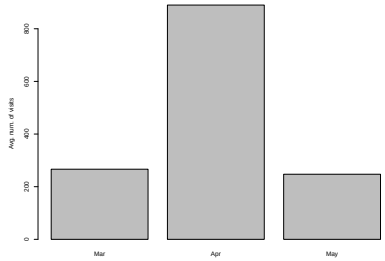


Figure 4.35: Average number of visits for today's featured articles in the English Wikipedia during April 2009.

a line inside the box. Outliers (values with very extreme values) are marked with circles outside the box. For instance, if we focus on the case of the English Wikipedia, at a first glance, it seems that level of visits during April and October was higher than it was during the corresponding previous and following months, when the level of visits remained quite similar. It seems that, in both periods, the bulk of visits correspond to the months when articles are displayed in the main pages in all the Wikipedias except the Spanish one that presents a similar behavior in all the months.

In the same way, if we plot the visits, Figure 4.38, to the promoted articles during the two sets of months we could appreciate the different dynamics exhibited during the featured promotion processes.

To find out whether the differences in the median values for all the samples are negligible or not, we will use a statistical test. Because the median values seem to be highly skewed in the box, the first step is testing whether the samples are extracted from a Normally distributed population. Depending on the result, we will choose a different statistical test to compare visits in different months.

Tables 4.12 and 4.13 show the results of the Normality test for the visits to the featured articles displayed in the main page of the English (EN), Spanish (ES), German (DE) and French (FR) Wikipedias during the two considered sets of months. The value of the W column is the Shapiro-Wilcox statistic, which indicates whether the sample is normal if and only if the p value is lower than a certain threshold (0.05 most often). In the case of these samples only the distributions corresponding to the month of April for the English and the German Wikipedias, the month of September for the French Wikipedia and the month of October for the English Wikipedia present Normal distributions.

The same tests applied to the promoted articles revealed that only the ones corresponding to the English Wikipedia followed a Normal distribution in the months of March, April, September and October. The rest of distributions were all non-Normal.

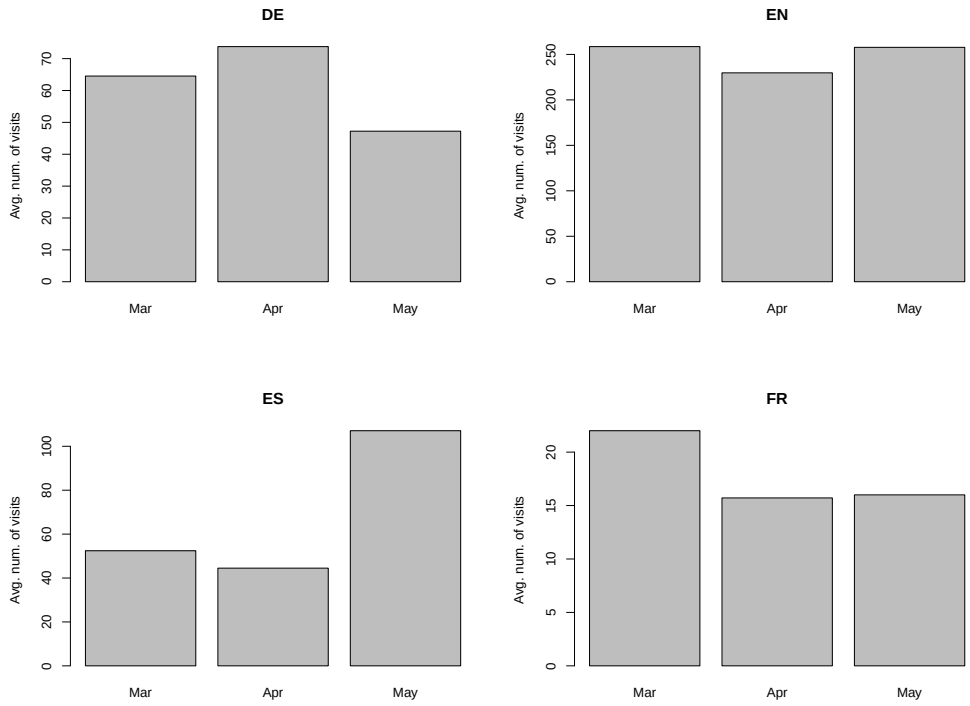


Figure 4.36: Different patterns of visits for the featured articles corresponding to April 2009 in different Wikipedias.

This non-normality of the samples implies that we have to test the median rather than the mean values, because the mean is highly biased for this kind of samples. If the histograms of the samples are highly skewed, the mean value can be affected by extreme values. For the samples under study, during the two central months, it is likely that we find articles with very high values of visits (outliers), which will increase the mean value even though the rest of featured articles remain with a similar level of visits. In such cases, the median value is more robust to outliers.

Because of this issue, we decided to use a Wilcoxon rank-sum test (also known as Mann-Whitney-Wilcoxon test) to find out whether or not the appearance of a featured article in the main page implies a greater number of visits to those articles. This test is not sensitive to the normality of the data.

Tables 4.14 and 4.15 show the results of the test. The column labeled U shows the value of the statistic, and the column p shows the level of significance. A high value of U with $p < 0.05$ indicates that the level of visits of the two samples under comparison is different; otherwise, it is similar. For instance, in the case of the English Wikipedia, the months of September and October have different levels of visits, as October and November have. But when comparing September with November, the level of visits is similar. Interestingly, these results indicate that featured articles displayed in the main pages attracted more visits during October only in the case of the English Wikipedia. However, in April both the English and the German Wikipedias attracted more visits over the featured articles presented in their main pages. When examining the promoted articles, none of the central months attracted a number of visits significantly higher than the next ones. Again the explanation may reside in the different way of conducting when developing the promotion process.

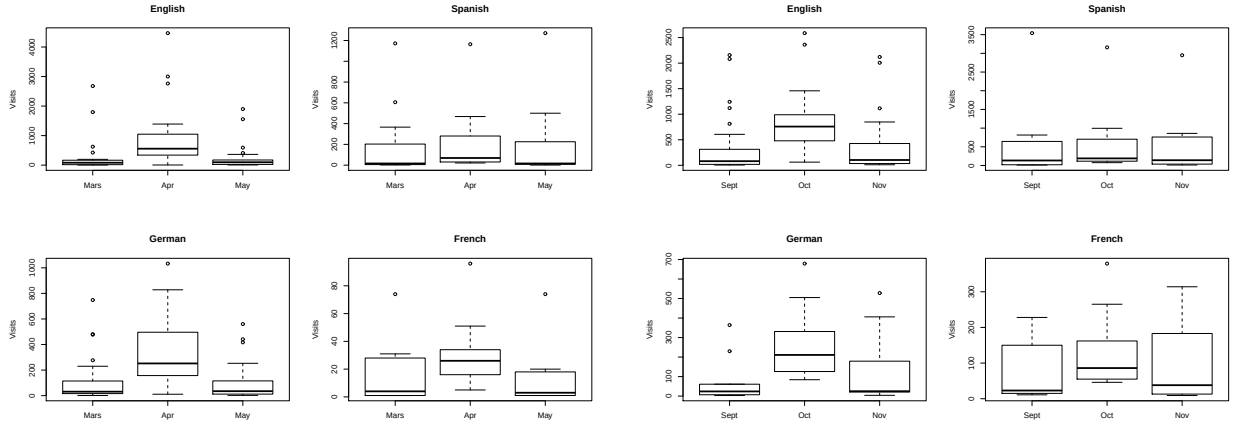


Figure 4.37: Boxplot of the visits to featured articles included in the main pages of the considered Wikipedias.

Lang	Mar		Apr		May	
	W	p	W	p	W	p
DE	0.97	0.83	0.90	**0.02**	0.97	0.77
EN	0.97	0.65	0.83	**0.00**	0.95	0.35
ES	0.89	0.12	0.87	0.08	0.95	0.75
FR	0.86	0.10	0.98	0.98	0.86	0.10

Table 4.12: Normality tests for featured articles displayed in the main pages. The month of April for the English and German Wikipedias seems to be Normal ($p < 0.05$). The rest of distributions are non-Normal.

4.7 Most visited, contributed and searched topics

As we exposed in chapter 1, nowadays there is no possibility of having access to a reliable and updated information about both the most visited and edited articles in the different editions of Wikipedia. Chapter 2 described several initiatives in this line, but all of them are, presently, out of service or unmaintained.

As far as this study is concerned, the most visited and edited articles are of a great importance because they can serve as a good indicators of the uses given to the different editions of Wikipedia by their corresponding communities of users. Apart from identifying and categorizing the most popular topics, such kind of study can allow to evaluate if articles' popularity and certain other habits are transmitted among the different editions of Wikipedia.

First of all, we are going to compare one of the few list with the most visited articles that we have managed to obtain with the results after our own analysis. In this way, Figure 4.39 presents the 50 most visited articles during August 2009 in the German and English Wikipedias according to the portal `http://wikistics.falsikon.de`. Again pursuing a validation of our results and conclusions, we compared our results with the ones obtained from this portal. Considering that its information is based on Domas Mituzas pageviews, it can be regarded as a reliable element to

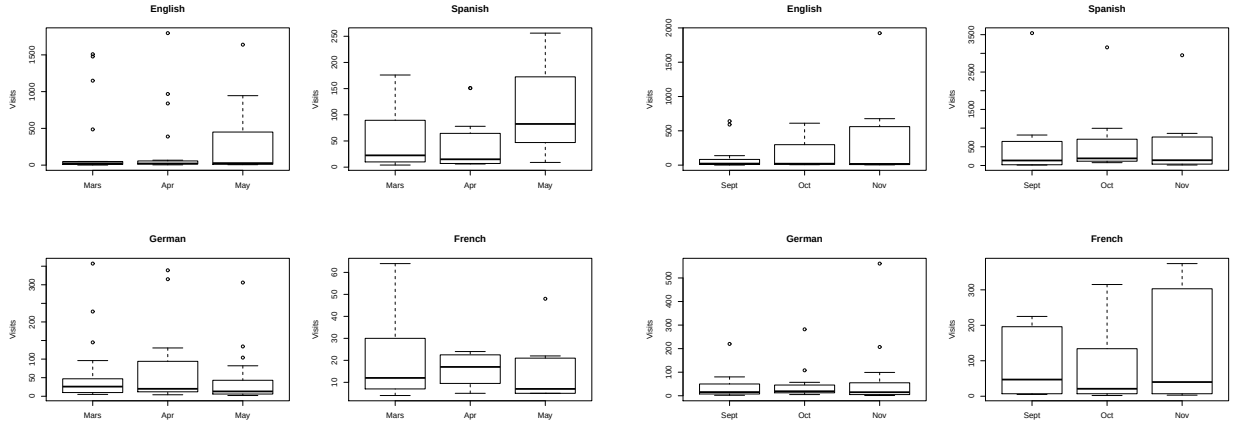


Figure 4.38: Boxplot of the visits to articles promoted to the featured status in the considered Wikipedias.

Lang.	Sept		Oct		Nov	
	W	p	W	p	W	p
DE	0.94	0.64	0.96	0.85	0.91	0.33
EN	0.95	0.19	0.92	**0.03**	0.95	0.16
ES	0.94	0.63	0.91	0.30	0.97	0.85
FR	0.82	**0.03**	0.89	0.22	0.87	0.13

Table 4.13: Normality tests for featured articles displayed in the main pages. Only the month of September for the French Wikipedia and the month of October for the English one seem to be Normal ($p < 0.05$). The rest of samples are non-Normal.

compare with. In this way, Tables 4.16 and 4.17 present, respectively, two lists made up of the 50 most visited articles in the German and English editions of Wikipedia. As we are interesting only in static articles, we are not considering *Special pages* dynamically generated on demand in response to specific users' requests such as random article, articles linking to a given one and so forth. Focusing on articles, if we compare our lists with the ones in Figure 4.39 we can see that the rank position of almost all articles match.

After having validated the obtaining of the articles which receive the highest numbers of visits, we undertake now their classification according to a categorization based on the one proposed by Spoerry in [Spo07b] and which has been adequately described in chapter 3. In short, we have classified the top-65 most visited and edited articles corresponding to the German, English, Spanish and French Wikipedia during six months of 2009. The different categories used for the classification are enumerated bellow:

1. Entertainment (ENT)
2. Politics + War (POL)
3. Geography (GEO)

Lang	M / A		A / Y		M / Y	
	U	p	U	p	U	p
DE	119	**0.00**	351.5	**0.00**	351.5	0.83
EN	100	**0.00**	617	**0.00**	336	0.62
ES	39	0.06	100	0.11	68	0.83
FR	21.5	0.10	64	0.04	46.5	0.62

Table 4.14: Results of the Wilcoxon rank-sum test for all the samples. In the English Wikipedia, the month of October gets more visits ($p < 0.05$). According to the figures, the English and German Wikipedias receive more visits in April as well as the French one. In the rest, the level of visits is similar in all the months. M: March, A: April, Y: May

Lang	S / O		O / N		S / N	
	U	p	U	p	U	p
DE	13	0.01	63	0.05	32	0.47
EN	140	**0.00**	645	**0.00**	337	0.37
ES	33	0.53	46.5	0.62	36	0.72
FR	25	0.19	52	0.34	38	0.86

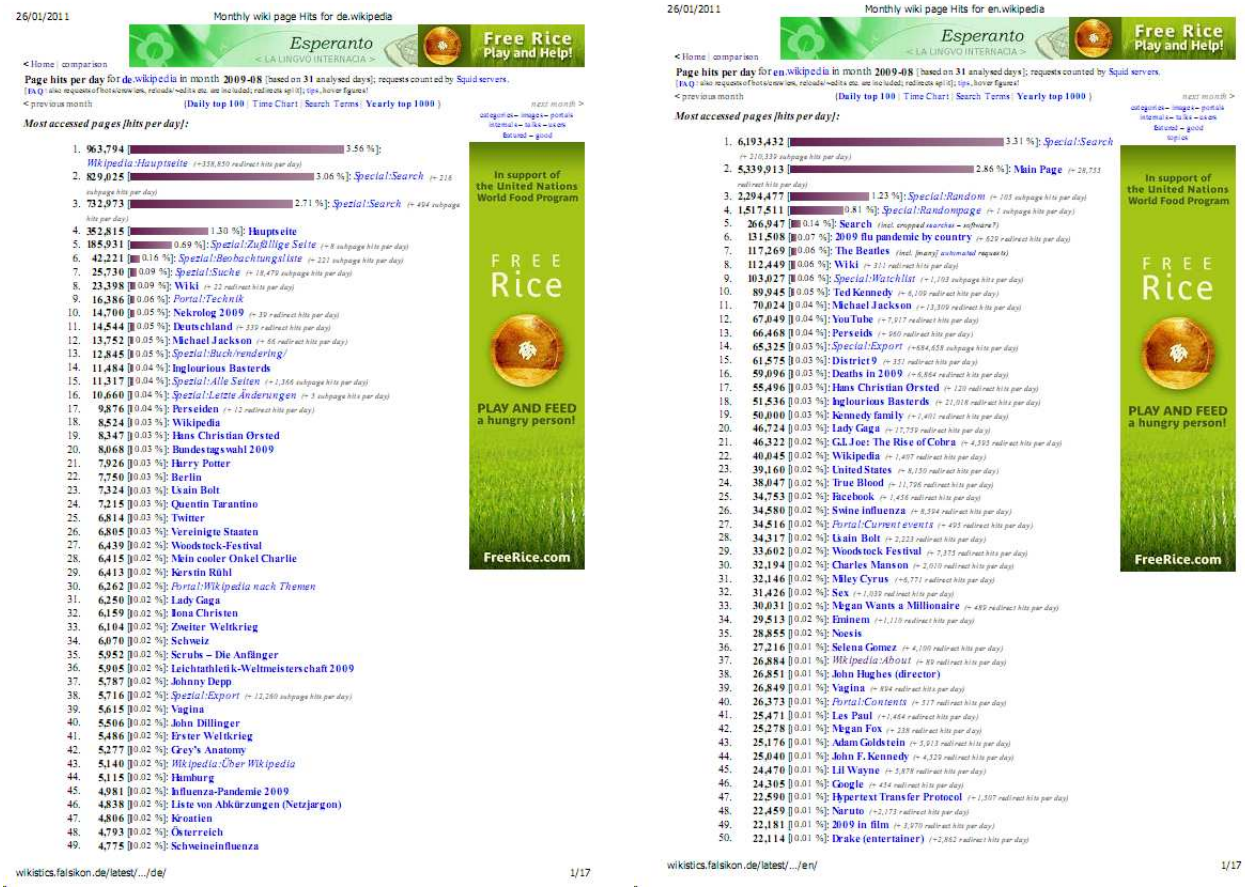
Table 4.15: Results of the Wilcoxon rank-sum test for all the samples. In the English Wikipedia, the month of October gets more visits ($p < 0.05$). In the English and German Wikipedias, the month of October receives more visits. In the rest, the level of visits is similar. S: September, O: October, N: November

4. Sexuality (SEX)
5. Science (SCI)
6. Information and Communication Technologies (ICT)
7. Arts (ART)
8. Current Events (CUR)

Table 4.18 presents the result of the categorization of the most visited and edited articles. The different categories considered in our analysis are presented in the table's rows whereas the visits and edits corresponding to each edition correspond to the columns. Here, it is important to note that *Main Page* articles are the unique in their category and, because of this, they have such low percentage. Looking at the table, it is clear that there exist important differences in the subjects that attract more attention in each considered Wikipedia. For example, topics related to the entertainment category do constitute the 44.92% in the English Wikipedia, whereas in the Spanish edition the same kind of articles attract only a 16.00%. Interestingly, again in the Spanish Wikipedia, these type of articles are the ones that receive most contributions from users. It is also noticeable than articles concerning sex topics gain more attention than the ones dealing with scientific or humanistic contents in the English Wikipedia. Scientific articles are the most requested in the Spanish Wikipedia (24.00%) followed by

Title	Pageviews	Title	Pageviews
Hauptseite	107261	Johnny_Depp	2165
Wiki	56801	Vagina	2101
Deutschland	26799	Grey%E2%80%99s_Anatomy	2075
Nekrolog_2009	22034	Erster_Weltkrieg	2004
Michael_Jackson	13105	John_Dillinger	1991
Inglourious_Basterds	12202	Hamburg	1963
Perseiden	10602	InfluenzaPandemie_2009	1937
Hans_Christian_%C3%98rsted	8901	%C3%96sterreich	1906
Wikipedia	7097	Liste_von_-Abk%C3%BCrzungen_-(Netzjargon)	1868
Bundestagswahl_2009	6720	Adolf_Hitler	1845
Harry_Potter	4483	Kroatien	1821
Berlin	4470	Europ%C3%A4ische_Union	1819
Usain_Bolt	4402	Schweineinfluenza	1789
Quentin_Tarantino	4175	Sex	1787
Vereinigte_Staaten	3897	%25s	1695
Twitter	3872	Bud_Spencer	1674
Mein_cooler_Onkel_Charlie	3566	Liste_der_Pornodarstellerinnen	1670
Kerstin_R%C3%BChl	3245	Schweinegrippe	1540
WoodstockFestival	3046	Papierformat	1537
Ilona_Christen	2573	Borderline-Pers%C3%B6nlichkeitsst%C3%B6rung	1480
Zweiter_Weltkrieg	2550	Penis	1466
Lady_Gaga	2467	Gossip_Girl	1461
Schweiz	2467	Frankreich	1444
Leichtathletik-Weltmeisterschaft_2009	2327	Irland	1433
Scrubs_%E2%80%93_Die_-Anf%C3%A4nger	2275	Twilight_%E2%80%93_Bis(s)_-zum_Morgengrauen	1426

Table 4.16: Most visited articles in the German Wikipedia (August, 2009).



(a) (b)

Figure 4.39: Lists with the 50 most visited articles and Special pages in the German (a) and English (b) Wikipedias during August 2009 according to the *http://wikistics.falsikon.de* site.

the ones dealing with humanistic topics (20.92%) such as literature or arts. Articles related to current events present significant visiting ratios in the English and French Wikipedias which could mean that their users would use Wikipedia as a kind of reference tool after a certain new or event becomes a subject of interest. Interestingly, articles devoted to current events or facts receive important numbers of contributions also in the German Wikipedia. The Spanish edition, on the contrary, present low rates corresponding to this kind of articles.

Regarding the most searched topics, we have classified the strings submitted by users when issuing search operations using the same categorization that was applied to the articles's titles. Table 4.19 presents the percentages corresponding to the different categories of searched topics in several editions of Wikipedia. According to this table, a high number of search operations involves entertainment-related topics in all the considered editions. This number is particularly high in the English Wikipedia. Spanish Wikipedia most searched topic corresponds to the Geography category and holds the highest numbers of searched topics related with scientific and humanistic disciplines. It is noticeable that in the French Wikipedia it is a high number of undetermined topics because they do not correspond to existing articles and they seem to be individuals' names and surnames.

Let us consider now how the requests to the top 65 most visited and edited articles are distributed

among the different categories. Table 4.20 presents this information and, according to it, most of the visits correspond to the main pages for all the editions except for the Spanish one. In the German and English Wikipedias the entertainment category has more visits than the rest of them. In the French edition is Geography the category which attracts more visits. Finally, in the Spanish edition, scientific and humanistic related articles are the most requested by users. The considerable low percentage to the main page in the Spanish Wikipedia may be due do the fact that its users mainly access the Encyclopedia through external search engines and by directly pointing to its URL from their web browser.

In the same way, we also obtained the distribution of the search requests throughout the considered categories. The results are presented in Table 4.21. This table shows how entertainment related topics are the most searched in all the considered editions of Wikipedia except in the Spanish edition where Geographical topics are the most frequently submitted. In turns, this group of topics is the second most frequent in the rest of editions whereas in the Spanish Wikipedia entertainment related topics are in the second position.

As we aimed to assess whether search requests involving particular topics could influence the number of visits to articles related to the same topics, we correlated search requests and visits to each category of subjects. As a result, we found that, from the four considered Wikipedias (German, English, Spanish and French) only the German and English one exhibited positive correlation between the two measures. This is shown in Figure 4.40 and means that, at least in these Wikipedias, there is a well-defined impact of search requests in the subsequent visits to articles. Reason explaining the opposite situation may include a not-generalized use of the Wikipedia built-in search engine in favor of external engines or that users did not get the appropriate results when querying the Wikipedia engine.

4.8 Summary of results

We are summarizing here our most important and significant results. These results are stemming from analysis of a sample of the log lines registered by the Wikimedia Foundation Squid servers. All the lines corresponding to 2009 have been parsed and classified in order to obtain a characterization of the overall traffic. In addition, those containing the most relevant namespaces and actions have been filtered to perform a thorough analysis.

In the following, we are presenting our most relevant achievements:

- First of all, we have validated the results obtained from the study realized a as a part of this thesis and which is based on the analysis of the requests submitted to Wikipedia by its users. This kind of analysis is innovative not only because of the nature of the sample of data used but also because of the results that it allows to obtain. It has been possible to validate our results due to the availability of trusted data sources to compare with. The different comparisons we have performed have shown the reliability of our analysis in general terms, related to whole editions or categories of contents, as well as at a higher level of detail involving particular articles or actions.

- The results from the validation process permit us to conclude that most of the visits to Wikipedia articles correspond to the namespaces considered in this thesis: *Main*, *Talk*, *User*, *User_talk* and *Special*. In the case of edit operations, our results allow to infer, even more certainly, that, in practice, they only involve the aforementioned namespaces.

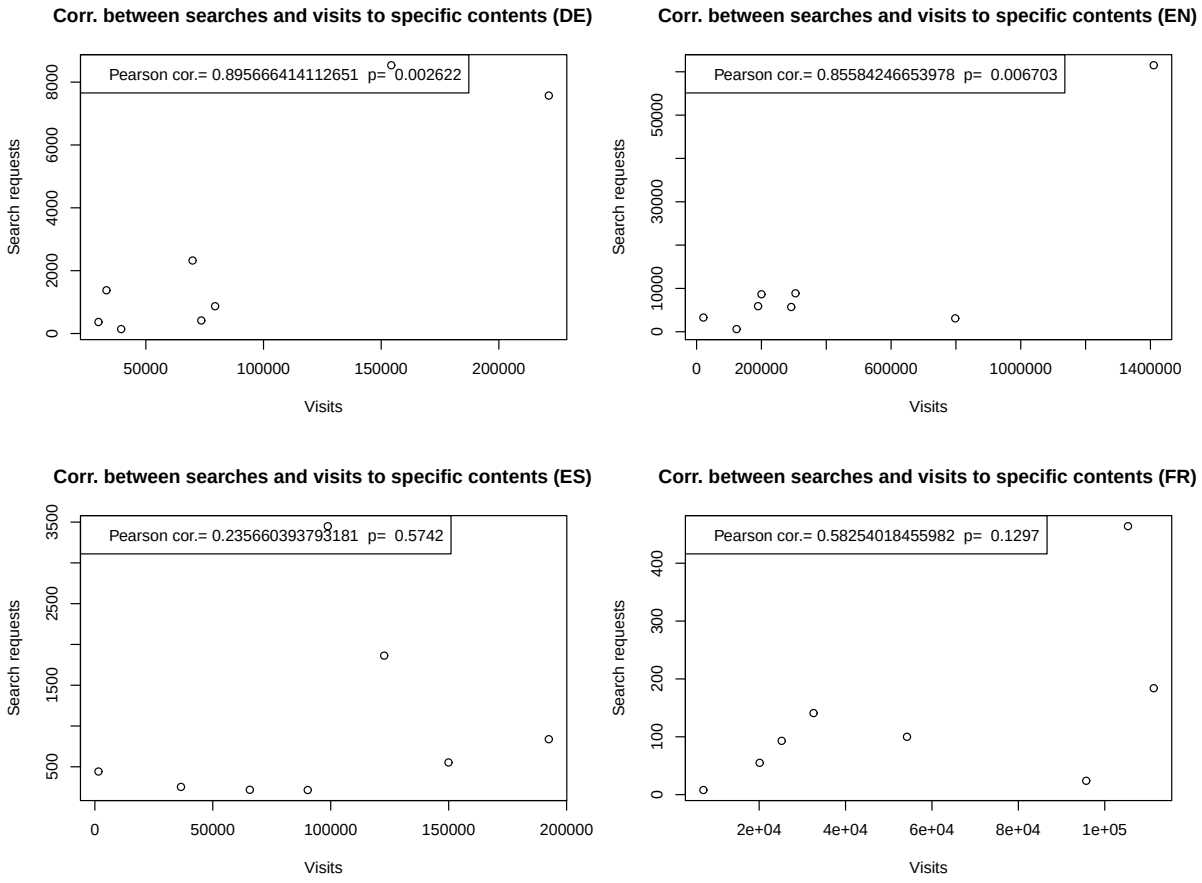


Figure 4.40: Correlation of visits and search operations involving specific topics in the German, English, Spanish and French Wikipedias.

- After comparing the size of the different editions of Wikipedia with the amount of traffic they attract, we can conclude that higher volumes or articles do not correspond to higher amounts of traffic. This means that resources related to the storage and to the contents delivery scale with ratios completely different.
- The examination of the traffic to all the Wikimedia Foundation projects shows that approximately the 96% of the requests correspond to Wikipedia pages and to uploaded resources such as images or videos.
- The characterization of the traffic to Wikipedia reveals that about a fourth of it correspond to visits (or pageviews) to articles. Requests specifying some kind of operation constitute approximately another quarter. Talking about actions, it is remarkable the high percentage of search operations and the considerably low number of edits. Another interesting fact, is the number of requests related to tailored presentation and visualization customizations that reach approximately the 35% of all the requests.
- The ten editions of Wikipedia considered in this thesis attract more than the 91% of all the requests directed to whole set of Wikipedias.

- The study of the temporal patterns has shown, firstly, that the traffic consisting in the request filtered for our analysis can serve to model the general traffic to Wikipedia. Filtered requests are those involving the previously mentioned *namespaces* and consisting in visits, edits, searches, history reviews or edit or submit requests. In addition, we have assessed that only visits and search operations follow patterns repeated over time whereas the rest of them present a irregular tendencies.

- Regarding users' behavior, we have, firstly, determined that a significant number of edit requests are not finished by the the corresponding write operation to the databases. This means that users decide to abandon, at a given time, the process of editing started with the corresponding requests. In this line, we have obtained the different ratios of incomplete edit requests corresponding to the different Wikipedia editions. On the contrary, we have verified that submit and edit requests are very close in number in most of the editions. This can be seen as the generalized use of the changes preview before committing permanently the changes performed or the contributions submitted.

- Correlation between edits and visits has shown that both requests are only related in some Wikipedia that, interestingly, also present a positive correlation between edit requests and edits (entailing a write operation to the database). These editions are the German, Englisn, Spanish, Italian and Russian ones. Correlation between visits and the other types of requests is positive in all the cases and for all the considered editions.

- Evaluating the impact of featured articles has allowed to sate that articles displayed during specific periods of time in the main pages of the different Wikipedia editions, as examples of high-quality contents, surely attract more attention from users in the English Wikipedia. On the other hand, visits to featured articles during their promotion process has permitted to highlight the differences in the dynamics exhibited by the corresponding communities of users when looking for consensus.

- We have categorized the most visited and edited articles in the different Wikipedia editions. As a result, articles related to entertainment are the most visited in the English Wikipedia whereas articles related to scientific or humanistic topics are the most requested in the Spanish edition. We have also classified the topics most repeatedly searched using the same previous categorization. In this case, all of the editions present high ratios of searches corresponding to the entertainment category. In the English Wikipedia they are the most frequently looked at whereas geographical are the ones most repeatedly queried about in the Spanish Wikipedia. After correlating both visits and search requests, we have found that there is only a positive correlation for the German and English Wikipedias.

Title	Pageviews	Title	Pageviews
Main_Page	1835745	Eminem	9058
2009_flu_pandemic_by_country	40524	Noesis	8960
The_Beatles	35804	Selena_Gomez	8359
Wiki	34558	John_Hughes_(director)	8228
Ted_Kennedy	27434	Vagina	8056
Michael_Jackson	21368	Les_Paul	8013
YouTube	20593	Adam_Goldstein	7782
Perseids	20567	Megan_Fox	7659
District_9	18855	Lil_Wayne	7515
Deaths_in_2009	18132	Google	7270
Hans_Christian_%C3%98rsted	17174	Hypertext_Transfer_Protocol	6997
Inglourious_Basterds	15759	Naruto	6955
Kennedy_family	15386	2009_in_film	6834
Lady_Gaga	14543	Penis	6834
Wikipedia	12170	Drake_(entertainer)	6777
United_States	12056	Barack_Obama	6714
True_Blood	11753	Human_penis_size	6620
Usain_Bolt	10662	Quentin_Tarantino	6602
Facebook	10559	List_of_sex_positions	6551
Swine_influenza	10471	Avatar_(2009_film)	6545
Woodstock_Festival	10302	Julia_Child	6484
Charles_Manson	10041	Harry_Potter	6484
Miley_Cyrus	9746	United_Kingdom	6478
Sex	9615	Chappaquiddick_incident	6477
Megan_Wants_a_Millionaire	9385	Vanessa_Hudgens	6462

Table 4.17: Most visited articles in the German Wikipedia (August, 2009).

Category	DE (Visited)	DE (Edited)	EN (Visited)	EN (Edited)	ES (Visited)	ES (Edited))	FR (Visited)	FR (Edited)
Category	DE (Visited)	DE (Edited)	EN (Visited)	EN (Edited)	ES (Visited)	ES (Edited)	FR (Visited)	FR (Edited)
MAIN	1.54%	0.00%	1.54%	0.00%	1.54%	0.00%	1.54%	0.00%
CUR	9.23%	19.69%	17.85%	25.23%	5.23%	5.23%	11.08%	9.23%
GEO	24.62%	15.38%	7.69%	9.85%	13.23%	17.54%	21.85%	23.69%
ICT	7.08%	7.69%	5.23%	2.15%	12.31%	1.85%	6.15%	0.92%
ENT	31.08%	14.77%	44.92%	36.31%	16.00%	46.46%	27.69%	25.23%
POL	9.85%	12.62%	8.92%	9.54%	5.23%	6.46%	6.77%	7.38%
SCI	5.54%	7.38%	3.38%	1.54%	24.00%	7.08%	4.31%	4.62%
ART	4.31%	17.23%	0.92%	14.46%	20.92%	13.85%	15.38%	27.69%
SEX	6.77%	0.31%	8.92%	0.00%	0.31%	0.00%	2.77%	0.31%
UN	0.00%	4.92%	0.62%	0.92%	1.23%	1.54%	2.46%	0.92%
TOTAL	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 4.18: Result of the categorization of the most visited and edited articles in the German, English, Spanish and French Wikipedias during January, February, June, July, August and November 2009

Category	DE	EN	ES	FR
CUR	9.49%	4.36%	4.36%	2.31%
GEO	29.74%	5.64%	26.41%	7.18%
ICT	5.64%	5.13%	3.59%	3.33%
ENT	31.54%	68.46%	25.13%	24.10%
POL	8.21%	2.56%	2.05%	2.56%
SCI	2.82%	0.77%	17.18%	5.38%
ART	2.82%	4.36%	14.36%	10.00%
SEX	4.62%	4.62%	5.13%	1.28%
UNDETERMINED	47.93%	43.61%	49.45%	54.55%

Table 4.19: Categorization of the 65 most searched topics in the German, English, Spanish and French Wikipedias during January, February, June, July, August and November 2009

Category	DE (Visited)	DE (Edited)	EN (Visited)	EN (Edited)	ES (Visited)	ES (Edited))	FR (Visited)	FR (Edited)
Category	DE (Visited)	DE (Edited)	EN (Visited)	EN (Edited)	ES (Visited)	ES (Edited)	FR (Visited)	FR (Edited)
MAIN	47.28%	0.00%	74.05%	0.00%	7.41%	0.00%	57.77%	0.00%
CUR	5.53%	20.27%	6.18%	28.30%	7.76%	5.94%	8.18%	11.58%
GEO	11.60%	14.40%	1.55%	11.16%	11.66%	18.47%	9.51%	24.73%
ICT	5.97%	7.64%	2.26%	2.27%	10.66%	1.17%	2.79%	0.58%
ENT	16.64%	16.17%	10.92%	31.63%	14.48%	50.53%	9.00%	23.74%
POL	5.25%	13.18%	2.36%	10.37%	4.31%	4.88%	2.15%	6.29%
SCI	2.97%	6.42%	0.95%	1.36%	22.72%	6.16%	1.72%	3.72%
ART	2.25%	17.17%	0.16%	12.33%	17.70%	12.10%	4.63%	28.21%
SEX	2.50%	0.22%	1.47%	0.00%	0.18%	0.00%	0.61%	0.25%
UNDETERMINED	0.00%	4.54%	0.09%	2.57%	3.13%	0.74%	3.63%	0.91%

Table 4.20: Distribution of the requests to the most visited and edited articles in the German, English, Spanish and French Wikipedias during January, February, June, July, August and November 2009

Category	**DE**		**EN ES**	**FR**
CUR	1.00%	1.79%	1.41%	1.02%
GEO	20.58%	5.00%	22.27%	7.82%
ICT	2.09%	3.30%	1.39%	5.99%
ENT	18.25%	35.53%	12.02%	19.73%
POL	5.60%	5.13%	1.63%	3.95%
SCI	0.34%	0.35%	5.41%	2.34%
ART	0.88%	1.89%	3.57%	4.25%
SEX	3.32%	3.41%	2.85%	0.34%
UNDETERMINED	47.93%	43.61%	49.45%	54.55%

Table 4.21: Distribution of the requests to the most searched topics in the German, English, Spanish and French Wikipedias during January, February, June, July, August and November 2009

Chapter 5

Conclusions and Further Research

> "The Things to do are: the things that need doing, that you see need to be done, and that no one else seems to see need to be done.". *Letter to Micheal*, Buckminster Fuller, (1892).

Wikipedia, the largest wiki-based platform available on the Internet, is a source of information for millions of people around the world. Due to this relevance, Wikipedia has become a profuse subject of research during the last years. However, all this research has been usually concerned with the quality and reliability of the Wikipedia's contents or with its growth and evolution tendencies. Other aspects such as authors' reputation or survival of contributions have been also regularly addressed. However, the examination of the use that the different communities or users are giving to Wikipedia has received little attention by the research community. The characterization of the traffic made up of users's requests can lead to patterns describing the interaction among them and the platform. Moreover, in conjunction with the study of the temporal distribution of requests, such kind of characterization may help to improve the response of the Encyclopedia to the imposed workload, both in general terms as well as on particular situations of system stress. In the aim of performing such traffic analysis, we have parsed and filtered the Squid logs lines containing information about the requests submitted to the most active editions of Wikipedia by their communities of users.

In the following, we will review the most important results after our analysis putting them in relation to the research questions presented in chap 1.

5.1 Summary of results

After having introduced the main goals of this thesis in chapter 1 in the form of several research questions, we presented in chapter 3 the most relevant aspects of the information elements composing our data feed and the methodological development conducted to process and examine all of them. Furthermore, we introduced in chapter 4 the analysis leading to obtain both behavioral and temporal patterns characterizing part of the use of Wikipedia.

As a result, we summarize here the most important achievements and the main conclusions we have reached after the work described in this thesis. In order to provide a well-structured and easy to follow compendium, we will use the search questions introduced in chapter 1 as organizational and articulatory items.

1. **Can we trust the results obtained from the analysis of requests sampled from the Wikimedia Foundation Squid servers?**
   We have based all our research in the analysis of the log files registered by the Wikimedia Foundation Squid servers. These servers save relevant information about each served request and, in particular, they store the URLs, which are the form in which requests are expressed, submitted by users. The analysis of log files aiming to determine particular aspects of web sites is an interesting alternative to the examination of dump files storing all their contents. Analyses based on log information can be fastest and less-resource demanding because log files do not contain as much data as dump files. This feature allows that log processing could be done even on-the-fly avoiding the downloading of heavy files prior to their analysis. In addition, dump files are offered by the companies or institutions supporting the web sites with a determined, or some times undetermined, periodicity. So, there is a dependence on their availability to perform a particular examination. Of course, log files also come from the institutions' facilities hosting the web sites but, at least in the case of the Wikimedia Foundation, once the corresponding agreement was established, we are receiving a continuous stream of data. In this way, we can perform any study regarding all the information received at any time with absolute independence.

   Although log information about any project maintained by the Wikimedia Foundation can contain as many information elements as configured in the corresponding server, it is obvious that Wikipedia dump files include important data not offered in the log format such as the contents contributed in a certain revision, its author, etc... However, log files contain the requests sent by users to browse the Wikipedia articles and to contribute to them. This is a really valuable data source not comprised in dump files that offers interesting research possibilities. Thus, in order to exploit as maximum these data, one of our most important concerns has been how to extract the largest amount of information from Squid log lines. In this way, URL parsing has allow us to identify relevant information elements such as the Wikipedia edition, the namespace and action requested or the title of the article involved in the requested. Then, in the filtering step, the application has determined if the request was considered of interest for our analysis. Perhaps the other most important concern for us has been the assessment of the validity and reliability of such an analysis based on log files. This assessment has been done by comparing some of our results with other well reputed sources, such as available information provided by the Wikimedia Foundation itself as well as with other previously developed tools such as Ortega's *WikiXRay*. After this comparison, we have obtained that our results nearly match the sampling factor used to construct our data feed. Moreover, we have also verified our results at a considerable finer grain by making the same comparisons at the level of particular articles or very specific periods of time. Again the matching has been positive and has maintain its closeness to the sampling factor. So, log analysis, being properly conducted, is reliable enough to trust its results. Hence, it may be considered as a complement or, even, an alternative to the analysis of dump files.

2. **Can we obtain a characterization and quantification of the types of requests composing the traffic to the different editions of Wikipedia?**

   Perhaps, one of the most relevant particularities of this thesis is the twofold analysis performed on the general, raw, traffic to Wikipedia as well as on the particular set of requests which have been filtered as a result of being considered of interest after the directives of our study. This

thesis has entailed the characterization of the general traffic directed to Wikipedia and made up of all the requests sent by users when browsing the Encyclopeda and asking for different kinds of contents and services. Moreover, we also have analyzed at the level of requested project all the traffic towards the Wikimedia Foundation servers. In relation to this subject, we have found that approximately the 96% of the overall traffic directed to the Wikimedia Foundation servers is composed by requests to Wikipedia as well as by requests for previously uploaded media content, mainly images. Interestingly the proportion of both types of requests is quite similar. Focusing on Wikipedia, the editions considered for this thesis attract more than the 91% of all the requests and the English Wikipedia maintains an important hegemony with more than the 46% of the traffic. The considered editions of Wikipedia present different distribution of their respective traffic throughout the months of 2009 which is a first indicator of different habits when the Encyclopedia is visited.

The examination of the composition of the general traffic directed to Wikipedia has allowed us to determine the types of requests directed to the each considered edition. Concerning this topic, about a 20-25% percent of all the requests correspond to visits to articles and almost the same ratio correspond to URLs requesting any type of action. However, edit requests are by two magnitude orders less than visit ones, approximately a 0.03% of all the requests. Search requests, in turns, constitute by a 9.5% of all the requests. Surprisingly, requests for CSS, skins and other visualization or customization elements suppose, in average, by the 35% of all the requests for the considered Wikipedias.

The obtained information resulting from the traffic characterization may be of a great interest because, as we have previously mentioned, almost a half of all the traffic directed to the Wikimedia Foundation servers correspond to the Wikipedia project. In this way, the characterization of such significant volume of traffic can lead to improvements in the systems in charge of their management and processing. The largest part of the rest of the traffic correspond to images and other media resources whose treatment may be much more homogeneous than the traffic made up of users' requests submitted as they can ask for a great variety of resources and actions.

3. **Is there a proportional relationship between the size of the Wikipedia editions and the amount of traffic they attract?**
   We have related the size of the different editions of Wikipedia with the amount of traffic they attract. As a result, we have observed that there is no a relationship of proportionality between them. In fact, smaller editions of Wikipedia such as the Spanish of the Russian are able to obtain greater ratios of the overall traffic than others editions having much more articles. In addition, while growth tendencies are stable in all the Wikipedia editions during 2009, traffic evolution presents important fluctuations in different periods of time.

4. **Can we identify patterns temporarily repeated which involve specific types of requests to Wikipedia?**
   Our analysis of the temporal distribution of the requests submitted to Wikipedia begun with the comparison between the evolution over time of the overall traffic to all the Wikimedia Foundation projects and the traffic attracted only by the Wikipedia project. As almost the half of this general traffic consists of requests to Wikipedia, the two traffics presents, as expected, a very similar temporal behavior. Furthermore, we added to the comparison the traffic composed solely by the requests filtered by our application. This was done in order to assess whether our filtered traffic and the real one showed a similar temporal evolution. As we obtained a positive result,

conclusions inferred from our analysis can be extrapolated to the traffic directed to Wikipedia. We have also checked our temporal distributions of visits and edits with the resulting from trusted information, such as the data from the Wikimedia Foundation itself presented by Zachte in his portal, obtaining evolutions that positively match.

Having assessed that the temporal distributions we have obtained can model the real ones, we have analyzed how the different kind of requests are distributed over time. In this way, we have found that only the URLs consisting in visits, searches and edit requests show temporal repetitive patterns. Other types of requests such as edits (save operations), history reviews or submits for previewing changes present considerably irregular temporal distributions.

Considering the evolution over time of visits and edits, we have found that both types of requests present considerably similar tendencies not only in the scale corresponding to the overall year but also but also when considering more in-depth ratios such as the corresponding to months and weeks. In the case, of edit operations, we consider specially relevant the fact that a substantial number of edit requests do not finish with the corresponding edit (save) operation. This result has been obtained from the great difference in number between the two types of requests. On the contrary, requests for edit and submit operations present very similar rates indicating that most of users issue a preview of their changes before committing them.

If we put in decreasing order the number of requested actions, we have found that search operations are the most requested ones, followed by edit requests, history requests and, alternatively, edits and submits. This ordination is maintained in all the considered Wikipedias. Interestingly, although in a half of the considered Wikipedias history requests appear considerably higher in number than edits and submits, in the other half they are much nearer the latter two actions, that would mean that in these Wikipedias, users are visiting the history review of the articles while making a contribution.

Studying the different kind of requests submitted during all the complete weeks of the year, we have confirmed that only visits, edit requests and searches present repetitive patterns. On the contrary, requests consisting in submit, history and edit operation follow a more spurious and unpredictable tendency. However, the found repetitive patterns are easier to observe in certain editions whereas in the rest present more irregularities. Interestingly, edit and submit requests present the nearest plots and their respective lines are coincident or, alternatively, one is slightly higher than the other. Curiously, only for the German Wikipedia submit operations are always over the edit ones. In general terms, there is continuous decrease in the number of requests as the week advances with the exception of Sundays when received requests experiment a little increase. This tendency is maintained by most of the different types of requests. Edit requests are the ones that adopt more different patterns. However, in the case of the German, English and Spanish Wikipedias they conserve a relatively similar shape.

As previously mentioned, visits and edit requests present very similar temporal progressions. We have shown that this also occurs between visits and the rest of actions. However, it is interesting to analyze how visits and edits mainly differ in weekends, when visits tend to decrease whereas edits increase their ratio. This can be attributed to the existence of a small elite of contributors which spend part of their spare time to produce contents for the Encyclopedia. As we do not have any kind of information to track authors and distribute the edit requests over them, we cannot examine to which authors the edits performed in weekends correspond to. In any case, it is patent, that contributions in weekends do not follow the same tendency than visits in several Wikipedias indicating that or edits are not coming from the bulk of visits or

that a significant number of visits were issued prior to an edit operation. Weekends also present another interesting fact, on Saturdays and Sundays edit requests present a descent whereas edits raise in number. This suggests that more edit requests finish with a subsequent save operation, so this fact may serve to reinforce the idea of contributors providing contents or looking after them in weekends.

5. **Are visits to the Wikipedia contents related with edits and the other type of actions in any way?**
   When examining behavioral aspects exhibited by users when interacting with Wikipedia, we found that regular users tend to turn intro contributors only in some of the considered editions: the German, English, Spanish, Italian and Russian ones. Interestingly, the same editions present a positive correlation between edits and edit requests, which could mean that the latter type of requests are finished by the expected write operation to the database. In general terms, a positive correlation between visits and edits may suggests collaborative and participative attitudes from the corresponding communities of users. The analysis of the ratio between edits and visits revealed that only the Italian and Russian editions were in the group having the highest values. On the other hand, all the considered actions where positively correlated to the visits. This results is consistent with previous obtained patterns resulting from the examination of the temporal evolution of the different types of requests.

6. **Does the promotion of articles to the featured status affect to the number of visits that they receive?**
   Articles considered as excellent because of their high quality and accomplishment of the most demanding criteria in terms of writing, neutrality, expression, completeness and references are recognized with the promotion to the featured status. Moreover, the inclusion of featured articles in the main pages of the different Wikipedia editions during a period of time pursues the attraction of attention to those articles, again as a sort of prize for its continued effort to achieve a level of quality.

   As we have analyzed the impact of the promotion of articles to the featured status in their subsequent number of visits as well as the attention attracted by featured articles presented in the main pages of several editions of Wikipedia, we have been able to appreciate that featured articles exposed as examples of quality contents in the front pages of different editions attract much more traffic in the month of their appearance than in the previous and following ones. The same do not necessarily occurs when considering promoted articles, perhaps due to the different internal mechanisms developed in the different editions when looking for a consensus in the promotion process. We have evaluated these two perspectives about featured articles during two temporal periods, each consisting in 3 months, and focusing on the articles promoted to the featured status during the central month and on the ”todays’s featured“ articles corresponding to this month.

   In the case or featured articles presented in the main pages of different editions of Wikipedia, the boxplots obtained with the number of visits received during each month indicated that most of visits had been paid during the month corresponding to their presentation for all the considered editions except the Spanish one. Promoted articles, on the contrary, exhibited more different distributions of visits as a result of different promotion processes. As boxplots reflected skewed medians, we estimated if visit evolutions fitted Normal distributions. This was aimed to apply adequate statistical tests to determine whether visits corresponding to different months were, actually, different in number. The results of such tests revealed that only the distributions

corresponding to a few months followed Normal distributions. Thus, we selected the Wilcoxon rank-sum test, which is not sensitive to the normality of data, to determine whether or not the appearance of a featured article in the main page implied greater number of visits to those articles. The results from these tests permitted us to state that, analyzing the German, English, Spanish and French Wikipedias, only the featured articles presented in the main page of the English Wikipedia attracted a greater number of visits during the month of their appearance in the two considered periods of time. In the German Wikipedia, articles featured in the main page received more visits only in the central month of one of the two periods.

7. **What are the topics to which correspond the articles that receive the highest numbers of visits and edits?**
   Presently, there are not updated services about the most visited and edited articles. We have tried to overcome this lack preparing our application for that purpose. In this way, among other information, we have stored the title of filtered articles as well as the topic involved in search requests. As a result, we have been able to determine the articles receiving the greatest numbers of visits and edits and also the topics most frequently submitted as part of search operation. We have classified both of them to determine the categories of articles attracting more attention from users of the different Wikipedia editions. In the same way, we have also obtained the kind of topics most often searched by the community of users corresponding to each edition. Apart from the categories or articles and search topics themselves, we have analyzed the distribution of visits over them.

   Among other results, we have seen how topics related to the entertainment category do constitute the 44.92% in the English Wikipedia, whereas in the Spanish edition the same kind of articles attract only a 16.00% or that scientific articles are the most requested in the Spanish Wikipedia (24.00%) followed by the ones dealing with humanistic topics (20.92%) such as literature or arts. Regarding the most searched topics, a high number of search operations involves entertainment-related topics in all the considered editions. This number is particularly high in the English Wikipedia. Spanish Wikipedia most searched topic corresponds to the Geography category and holds the highest numbers of searched topics related with scientific and humanistic disciplines. Considering the distributions of visits over the different categories, most of the visits correspond to the main pages for all the editions except for the Spanish one. In the German and English Wikipedias the entertainment category has more visits than the rest of them. In the French edition is Geography the category which attracts more visits. Finally, in the Spanish edition, scientific and humanistic related articles are the most requested by users. The distribution of the search requests throughout the considered categories shows, for example, how topics related to entertainment are the most searched in all the considered editions of Wikipedia except in the Spanish edition where Geographical topics are the most frequently submitted.

8. **Do search requests involving particular subjects have an impact on visits to articles related to same topics ?**

   In order to determine the impact of search operations involving particular contents on the visits to articles related to them, we have correlated search operations and visits corresponding to articles belonging to different categories of subjects. Our results show that only in two of the four analyzed Wikipedias search operations involving different categories of topics had a verifiable influence on the subsequent visits to the corresponding articles.

5.2 Further work

It is clear that this thesis is not the end of a way but just the beginning of several ones. In fact, there are several aspects that deserve deeper research and analysis. In particular, I am outlining here the ones that , in my opinion, constitute the natural steps after this work:

1. The study of distributions which fit visits and edits to articles deserves important efforts. Although this matter has been addressed by other researchers, our approach of analyzing both visits and edits from the perspective of requests sent by users constitutes a promising challenge. In fact, I have started this kind of analysis as a part of this thesis but presently I have not been able to find a distribution that fit visits or edits to Wikipedia. I have checked the power law and log normal distributions, two of the fittings more commonly related to Wikipedia accesses by previous literature. However, up to the present date the results have not allow us to model visits nor edits. Perhaps, requests to Wikipedia follow an special combination of the two distributions and, because of this, this subject has to be explored in the future. Furthermore, I want also examine the possible relationships and correlations between different information elements such as namespaces and actions.

2. Geolocation is surely one of the most promising ways of continuing our research. Any form of request geolocation would allow us to determine the geographical origin of the requests sent to the different editions of Wikipedia. This information could be used to determine where users of the different editions of Wikipedia are from. Moreover, we could assess if it is normal to browse the same article in different Wikipedias and, if so, to determine the first choice for particular communities of users. I have started to work in this area and an initial version of the software needed to register users location has been already sent to the Wikipedia technical staff. Because confidentiality and privacy of users have to preserved, this software has to be run on the Wikimedia Foundation systems and, of course, it results has to come in a completely anonymized format.

3. Featured articles deserve, of course, a further research. To begin with, featured articles of more editions than the considered in this thesis could be included in the analysis. The analysis of the evolution of the process leading to consensus in the consideration of a certain article as featured in the different Wikipedia editions is an absolute undertaking for us. In this way, we would be in the position of study how different communities of users behave when considering the promotion or demotion of articles and we could analyze the existence of trending tendencies propagating among different editions of Wikipedia.

4. Technical improvements will be done to offer all, or at least, an important part of all the information obtained as a result of this thesis. In particular, the database used for most of the statistical analysis will be publicly opened soon through a web interface. Moreover, this interface will include the possibility of generating customizable graphs and charts for researchers of other less technical areas.

5. Squid systems can register several features describing the type of client requesting Wikipedia. Existing plugins or browsers having specials features to facilitate and made more comfortable the navigation through Wikipedia may notably influence in the users choices to visit Wikipedia. In this way, accesses from mobiles devices are of an special relevance for us because the application of mobile technologies to browse Wikipedia may have an effect in the design of contents specially planned for this kind of devices.

6. The process of categorization of articles' titles and searched topics has to be automatized and improved to allow an efficient classification of the topics attracting the attention of users'. In this line, the correlation between both topics has to be established as a result of a computational process. Ontologies and automatic tagging systems may be an excellent choice in this field.

VALE

Bibliography

[AA05] David Aumüller and Sören Auer. Towards a semantic wiki experience - desktop integration and interactivity in wiksar, 2005.

[AB00] A.B. Atkinson and F. Bourguignon, editors. *Handbook of Income Distribution*, volume 1 of *Handbook of Income Distribution*. Elsevier, 2000.

[ABCdO96] Virgílio Almeida, Azer Bestavros, Mark Crovella, and Adriana de Oliveira. Characterizing reference locality in the WWW. In *DIS '96: Proceedings of the fourth international conference on on Parallel and distributed information systems*, pages 92–103, Washington, DC, USA, December 1996. IEEE Computer Society.

[ACdA+08] Thomas B. Adler, K. Chatterjee, Luca de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to wikipedia content. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, 2008. ACM Press.

[ACH05] Kevin C. And, Kevin Crowston, and James Howison. Hierarchy and centralization in free and open source software team communications. *Knowledge Technology & Policy*, 18:65–85, 2005.

[AdA07] Thomas B. Adler and Luca de Alfaro. A content-driven reputation system for the wikipedia. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 261–270, New York, NY, USA, 2007. ACM Press.

[AdAPV08] Thomas B. Adler, Luca de Alfaro, I. Pye, and Raman V. Measuring author contributions to the wikipedia. In *Proceedings of the 4th International Symposium on Wikis*, New York, NY, USA, 2008. ACM Press.

[AdR05] Sisay F. Adafre and Maarten de Rijke. Discovering missing links in wikipedia. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 90–97, New York, NY, USA, 2005. ACM Press.

[AKEV01] Jussara M. Almeida, Jeffrey Krueger, Derek L. Eager, and Mary K. Vernon. Analysis of educational media server workloads. In *Proceedings of the 11th international workshop on Network and operating systems support for digital audio and video*, NOSSDAV '01, pages 21–30, New York, NY, USA, 2001. ACM.

[AMC07] Rodrigo Almeida, Barzan Mozafari, and Junghoo Cho. On the evolution of wikipedia. In *International Conference on Weblogs and Social Media*, Boulder, Colorado, USA, March 2007.

[AMP06] Inmaculada B. Aban, Mark M. Meerschaert, and Anna K. Panorska. Parameter estimation for the truncated pareto distribution. *Journal of the American Statistical Association*, 101(473):270–277, March 2006.

[Aro02] L. Aronsson. Operation of a large scale general purpose wiki website. In J. Carvalho, A. Hübler, and A. Baptista, editors, *Proceedings of the 6th International ICCC/IFIP Conference on Electronic Publishing*, Berlin, 2002. Verlag für Wissenschaft und Forschung.

[ASW] Denise Anthony, Sean Smith, and Tim Williamson. Explaining quality in internet collective goods: Zealots and good samaritans in the case of wikipedia. Electronically.

[Aue05] S. Auer. Powl: a web based platform for collaborative semantic web development. In *Proc. of 1st WS Scripting for the Semantic Web*, May 2005.

[Aum05a] D. Aumüller. Semantic authoring and retrieval in a wiki(wiksar), May 2005.

[Aum05b] D. Aumüller. Shawn: Structure helps a wiki navigate. In W. Mueller and R. Schenkel, editors, *Proceedings of the BTW-Workshop WebDB Meets IR*, March 2005.

[AW96] Martin F. Arlitt and Carey L. Williamson. Web server workload characterization: the search for invariants. *SIGMETRICS Perform. Eval. Rev.*, 24(1):126–137, May 1996.

[AW97] Martin F. Arlitt and Carey L. Williamson. Internet Web servers: workload characterization and performance implications. *IEEE/ACM Trans. Netw.*, 5(5):631–645, October 1997.

[AWY99] Charu Aggarwal, Joel L. Wolf, and Philip S. Yu. Caching on the world wide web. *IEEE Trans. on Knowl. and Data Eng.*, 11:94–107, January 1999.

[Bar05] Matthew D. Barton. The future of rational-critical debate in online public spheres. *Computers and Composition*, 22(2):177–190, 2005.

[BB04] F. Bellomi and R. Bonato. Lexical authorities in an encyclopedic corpus: a case study with wikipedia. Electronically, 2004.

[BB05] F. Bellomi and R. Bonato. Network analysis for wikipedia. In *Wikimania 2005*. Wikimedia Foundation, 2005.

[BBBC98] Paul Barford, Azer Bestavros, Adam Bradley, and Mark Crovella. Changes in web client access patterns: Characteristics and caching implications. Technical report, Boston, MA, USA, 1998.

[BCD+06] Luciana S. Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. Temporal analysis of the wikigraph. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 45–51, Washington, DC, USA, 2006. IEEE Computer Society.

[Bee06] Angela Beesley. How and why wikipedia works. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, pages 1–2, New York, NY, USA, 2006. ACM Press.

[Ben01] Yochai Benkler. Coase's penguin, or linux and the nature of the firm, October 2001.

[Ben06] Yochai Benkler. *The Wealth of Networks : How Social Production Transforms Markets and Freedom*. Yale University Press, May 2006.

[BFB05] Susan L. Bryant, Andrea Forte, and Amy Bruckman. Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *GROUP '05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 1–10, New York, NY, USA, 2005. ACM Press.

[BFGM] Abhijit Bhole, Blaž Fortuna, Marko Grobelnik, and Dunja Mladenić. Extracting named entities and relating them over time based on wikipedia. *Informatica*.

[Bis07] Andreas Bischoff. The pediaphon - speech interface to the free wikipedia encyclopedia for mobile phones, pdas and mp3-players. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA 2007)*, pages 575–579, 2007.

[BJC+04] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, and Ophir Frieder. Hourly analysis of a very large topically categorized web query log. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 321–328, New York, NY, USA, 2004. ACM.

[BJL+07] Steven M. Beitzel, Eric C. Jensen, David D. Lewis, Abdur Chowdhury, and Ophir Frieder. Automatic classification of Web queries using very large unlabeled query logs. *ACM Trans. Inf. Syst.*, 25(2):9+, April 2007.

[BKdR06] Sigurbjörnsson Börkur, Jaap Kamps, and Maarten de Rijke. Focused access to wikipedia. In *Proceedings DIR-2006*, 2006.

[BR07] Davide Buscaldi and Paolo Rosso. A bag-of-words based ranking method for the wikipedia question answering task. In *FALTA*, pages 550–553. FALTA, 2007.

[BRG07] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering short texts using wikipedia. In *SIRIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788, New York, NY, USA, 2007. ACM Press.

[BRVX04] L. Bent, M. Rabinovich, G. M. Voelker, and Z. Xiao. Characterization of a large web site population with implications for content delivery. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 522–533, New York, NY, USA, 2004. ACM Press.

[BSG04] Michel Buffa, Peter Sander, and Jean-Claude Grattarola. Distant cooperative software development for research and education: three years of experience. In *Proceedings of the International Conference on Computer Aided Learning in Engineering Education (CALIE 04)*, 2004.

[BSKK01] B. Butler, L. Sproull, S. Kiesler, and R. Kraut. Community effort in online groups: Who does the work and why, 2001.

[BYP06] Ricardo Baeza-Yates and Barbara Poblete. A Website Mining Model Centered on User Queries. pages 1–17. 2006.

[CCT04] S. E. Campanini, P. Castagna, and R. Tazzoli. Platypus wiki: a semantic wiki wiki web. In *Semantic Web Applications and Perspectives, Proceedings of 1st Italian Semantic Web Workshop*, December 2004.

[Ced03] Magnus Cedergren. Open content and value creation. *First Monday*, 8(8), August 2003.

[CFTR06] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1037–1046, New York, NY, USA, 2006. ACM Press.

[CFTR07] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. Suggestbot: using intelligent task routing to help people find work in wikipedia. In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*, pages 32–41, New York, NY, USA, 2007. ACM Press.

[CG04] Ludmila Cherkasova and Minaxi Gupta. Analysis of enterprise media server workloads: access patterns, locality, content evolution, and rates of change. *IEEE/ACM Trans. Netw.*, 12(5):781–794, 2004.

[CH03] Kevin Crowston and James Howison. The social structure of open source software development teams. In *OASIS 2003 Workshop (IFIP 8.2 WG)*, 2003.

[Cha83] John M. Chambers. *Graphical Methods for Data Analysis (Statistics)*. Chapman & Hall/CRC, February 1983.

[Che06] Thomas Chesney. An empirical examination of wikipedia's credibility. *First Monday*, 11(11), November 2006.

[Chi07] E. Chi. Long tail of user participation in wikipedia. Technical report, May 2007.

[Cif03] Andrea Ciffolilli. Phantom authority, self–selective recruitment and retention of members in virtual communities: The case of wikipedia. *First Monday*, 8(12), December 2003.

[CLL+01a] K. Selçuk Candan, Wen-Syan Li, Qiong Luo, Wang-Pin Hsiung, and Divyakant Agrawal. Enabling dynamic content caching for database-driven web sites. *SIGMOD Rec.*, 30:532–543, May 2001.

[CLL+01b] K. Selçuk Candan, Wen-Syan Li, Qiong Luo, Wang-Pin Hsiung, and Divyakant Agrawal. Enabling dynamic content caching for database-driven web sites. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, SIGMOD '01, pages 532–543, New York, NY, USA, 2001. ACM.

[Cox72] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[Cra07] Michael J. Crawley. *The R Book*. Wiley, June 2007.

[Cro05] Michael J. Crowley. *Statistics: An Introduction using R.* John Wiley & Sons, 2005.

[CS+01] Carlos Castillo-Salgado et al. Measuring health inequalities: Gini coefficient and concentration index. *Epidemiological Bulletin*, 22(1), March 2001.

[CS07] Dianne Cook and Deborah F. Swayne. *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi (Use R).* Springer, 1 edition, December 2007.

[CSC+06] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: the case of wikipedia, Feb 2006.

[CSN07] Aaron Clauset, Cosma R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, Jun 2007.

[CVM09] Caterina Calefato, Fabiana Vernero, and Roberto Montanari. Wikipedia as an example of positive technology: How to promote knowledge sharing and collaboration with a persuasive tutorial. In *2009 2nd Conference on Human System Interactions*, pages 510–516. IEEE, May 2009.

[Dal08] Peter Dalgaard. *Introductory Statistics with R (Statistics and Computing)*, chapter 14. Springer, 2nd edition, August 2008.

[DB92] Paul Dourish and Victoria Bellotti. Awareness and coordination in shared workspaces. In *CSCW '92: Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pages 107–114, New York, NY, USA, 1992. ACM Press.

[DBWS06] Pierpaolo Dondio, Stephen Barrett, Stefan Weber, and Jean Seigneur. Extracting trust from domain analysis: A case study on the wikipedia project. pages 362–373. 2006.

[DG06] Ludovic Denoyer and Patrick Gallinari. The wikipedia xml corpus. *SIGIR Forum*, 40(1):64–69, June 2006.

[DGPS06] Alain Désilets, Lucas Gonzalez, Sébastien Paquet, and Marta Stojanovic. Translation the wiki way. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, pages 19–32, New York, NY, USA, 2006. ACM Press.

[DHPW05] Peter Denning, Jim Horning, David Parnas, and Lauren Weinstein. Wikipedia risks. *Commun. ACM*, 48(12):152–152, December 2005.

[Dor79] Robert Dorfman. A formula for the gini coefficient. *The Review of Economics and Statistics*, 61:146–149, March 1979.

[DSC05] Chris Dibona, Mark Stone, and Danese Cooper. *Open Sources 2.0 : The Continuing Evolution.* O'Reilly Media, Inc., October 2005.

[EG04] Anja Ebersbach and Markus Glaser. Towards emancipatory use of a medium: The wiki. *International Journal of Information Ethics*, 11, 2004.

[EGH05] Anja Ebersbach, Markus Glaser, and Richard Heigl. *Wiki : Web Collaboration.* Springer, November 2005.

[EH05] W. Emigh and S. C. Herring. Collaborative authoring on the web: A genre analysis of online encyclopedias. In *FALTA*, pages 99a–99a, 2005.

[ELB08] K. Ehmann, A. Large, and J. Beheshti. Collaboration in context: Comparing article evolution among subject disciplines in wikipedia. *First Monday*, 13(10), October 2008.

[FAL] FALTA. Wikibibliographie encyclen. consulted on April 13th, 2008.

[Fal08] Don Fallis. Toward an epistemology of wikipedia. *J. Am. Soc. Inf. Sci. Technol.*, 59(10):1582–1597, 2008.

[FB06] Andrea Forte and Amy Bruckman. From wikipedia to the classroom: exploring online publication and learning. In *ICLS '06: Proceedings of the 7th international conference on Learning sciences*, pages 182–188. International Society of the Learning Sciences, 2006.

[FdR06a] Fissaha and Maarten de Rijke. Exploratory search in wikipedia. In *Proceedings SIGIR 2006 workshop on Evaluating Exploratory Search Systems (EESS)*, 2006.

[FdR06b] Fissaha and Maarten de Rijke. Finding similar sentences across multiple languages in wikipedia. In *EACL 2006 Workshop on New Text.Wikis and Blogs and Other Dynamic Text Sources*, 2006.

[FGR+06] Jochen Fischer, Zeno Gantner, Steffen Rendle, Manuel Stritt, and Lars S. Thieme. Ideas and improvements for semantic wikis. *Lecture Notes in Computer Science*, 4011 / 2006, 2006.

[FPH+95] V. Franco, R. Piirto, H. Y. Hu, B. V. Lewenstein, R. Underwood, and N. K. Vidal. Anatomy of a flame: conflict and community building on the internet. *Technology and Society Magazine, IEEE*, 14(2):12–21, 1995.

[Fry07] Ben Fry. *Visualizing Data*. O'Reilly Media, Inc., December 2007.

[GCXZ05] Lei Guo, Songqing Chen, Zhen Xiao, and Xiaodong Zhang. Analysis of multimedia workloads with implications for internet streaming. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 519–528, New York, NY, USA, 2005. ACM.

[GH06] S. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 30(4):378–391, 2006.

[Gil04] Dan Gillmor. *We the Media*. O'Reilly, August 2004.

[Gil05] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005.

[Gin36] Conrando Gini. On the measure of concentration with especial reference to income and wealth. In *Cowless Comission*, 1936.

[GJL09] Y. Ganjisaffar, S. Javanmardi, and C. Lopes. Review-based ranking of wikipedia articles. In *Computational Aspects of Social Networks, 2009. CASON '09. International Conference on*, pages 98 –104, june 2009.

[GM06] Evgeniy Gabrilovich and Shaul Markovitch. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06*, pages 1301–1306, 2006.

[GM07] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007.

[GMY04a] Michel L. Goldstein, Steven A. Morris, and Gary G. Yen. Fitting to the power-law distribution. *FALTA*, Aug 2004.

[GMY04b] Michel L. Goldstein, Steven A. Morris, and Gary G. Yen. Problems with fitting to the power-law distribution, August 2004.

[HBB05] Todd Holloway, Miran Bozicevic, and Katy Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors, Dec 2005.

[HCL05] Jeffrey Heer, Stuart K. Card, and James A. Landay. prefuse: a toolkit for interactive information visualization. In *CHI '05: Proceeding of the SIGCHI conference on Human factors in computing systems*, pages 421–430, New York, NY, USA, 2005. ACM Press.

[HE10] Alison Head and Michael Eisenberg. How today's college students use Wikipedia for course-related research. *First Monday*, 15(3), 2010.

[Her08] Israel Herraiz. *A statistical examination of the evolution and properties of libre software*. PhD thesis, Universidad Rey Juan Carlos, 2008. http://purl.org/net/who/iht/phd.

[Hin06] Dion Hinchcliffe. The state of web 2.0, April 2006.

[HKVV06] Heiko Haller, Markus Krötzsch, Max Völkel, and Denny Vrandecic. Semantic wikipedia. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, pages 137–138, New York, NY, USA, 2006. ACM Press.

[HL08] Alexander Halavais and Derek Lackaff. An analysis of topical coverage of wikipedia. *Journal of Computer-Mediated Communication*, 13(2):429–440, 2008.

[HLM08] David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis: Regression Modeling of Time to Event Data (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2 edition, March 2008.

[HLMH0] Yair A. Hamburger, Naama Lamdan, Rinat Madiel, and Tsahi Hayat. Personality characteristics of wikipedia members. *CyberPsychology & Behavior*, 0(0):1–3, 0.

[HLS+07a] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady W. Lauw, and Ba-Quy Vuong. Measuring article quality in wikipedia: models and evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 243–252, New York, NY, USA, 2007. ACM.

[HLS+07b] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady W. Lauw, and Ba-Quy Vuong. On improving wikipedia search using article quality. In *WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management*, pages 145–152, New York, NY, USA, 2007. ACM.

[JGLB09] S. Javanmardi, Y. Ganjisaffar, C. Lopes, and P. Baldi. User contribution and trust in wikipedia. In *Collaborative Computing: Networking, Applications and Worksharing, 2009 CollaborateCom 2009. 5th International Conference on*, pages 1 –6, nov. 2009.

[JHG06] Frank T. Johnsen, Trude Hafsoe, and Carsten Griwodz. Analysis of server workload and client interactions in a news-on-demand streaming system. In *Proceedings of the Eighth IEEE International Symposium on Multimedia*, ISM '06, pages 724–727, Washington, DC, USA, 2006. IEEE Computer Society.

[KB04] Aaron Krowne and Anil Bazaz. Authority models for collaborative authoring. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, 2004.

[KI01] A. Khunkitti and W. Intraha. Web object life measurement using squid log file. pages 100–103, 2001.

[KK05] David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text (Statistics for Biology and Health)*. Springer, 2nd edition, August 2005.

[KKT06] K. Kawamoto, Y. Kitamura, and Y. Tijerino. Kawawiki: A semantic wiki based on rdf templates. In *FALTA*, pages 425–432, 2006.

[KM05] Josef Kolbitsch and Hermann Maurer. Community building around encyclopaedic knowledge. *Journal of Computing and Information Technology*, 13, 2005.

[KM07] Josef Kolbitsch and Hermann Maurer. The growing importance of e-communities on the web. In Joaquim Filipe, José Cordeiro, and Vitor Pedrosa, editors, *Web Information Systems and Technologies*, volume 1 of *Lecture Notes in Business Information Processing*, chapter 3, pages 19–37. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[KNNL04] Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 5th edition, August 2004.

[KNP+06] Korfiatis, Nikolaos, Poulos, Marios, Bokos, and George. Evaluating authoritative sources using social networks: an insight from wikipedia. *Online Information Review*, 30(3):252–262, May 2006.

[Kon] Piotr Konieczny. Wikis and wikipedia as a teaching tool. *International Journal of Instructional Technology & Distance Learning*, (1), January.

[Kos10] Vasilis Kostakis. Peer governance and Wikipedia: Identifying and understanding the problems of Wikipedia's governance. *First Monday*, 15(3), 2010.

[KPSM07] Aniket Kittur, Bryan A. Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In

Proceedings of the 25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007). ACM, April 2007.

[Kri06a] K. Krishnamoorthy. *Handbook of Statistical Distributions with Applications (Statistics: a Series of Textbooks and Monographs)*. Chapman & Hall/CRC, June 2006.

[Kri06b] A. Krizhanovsky. Synonym search in wikipedia: Synarcher, Jun 2006.

[KSPC07] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: conflict and coordination in wikipedia. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462, New York, NY, USA, 2007. ACM Press.

[Kuz06] Stacey Kuznetsov. Motivations of contributors to wikipedia. *SIGCAS Comput. Soc.*, 36(2), June 2006.

[KVV] Markus Krötzsch, Denny Vrandecic, and Max Völkel. Wikipedia and the semantic web - the missing links.

[KVV06] Markus Krötzsch, Denny Vrandecić, and Max Völkel. *Semantic MediaWiki*. 2006.

[LC01] Bo Leuf and Ward Cunningham. *The Wiki Way: Collaboration and Sharing on the Internet*. Addison-Wesley Professional, April 2001.

[LD07] Miro Lehtonen and Antoine Doucet. Extirp: Baseline retrieval from wikipedia. In *MISSING*, pages 115–120. MISSING, 2007.

[Lih04] Andrew Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *5th International Symposium on Online Journalism*. University of Texas at Austin, April 2004.

[Lin10] David Lindsey. Evaluating quality control of Wikipedia's feature articles. *First Monday*, 15(4), 2010.

[LKSY07] Brendan Luyt, Wee Kwek, Ju Sim, and Peng York. Evaluating the comprehensiveness of wikipedia: The case of biochemistry. pages 512–513. 2007.

[LLXX10] Alexandros Labrinidis, Qiong Luo, Jie Xu, and Wenwei Xue. Caching and materialization for web databases. *Found. Trends databases*, 2:169–266, March 2010.

[LNX08] Qiong Luo, Jeffrey F. Naughton, and Wenwei Xue. Form-based proxy caching for database-backed web sites: keywords and functions. *The VLDB Journal*, 17:489–513, May 2008.

[Lor06] Michael Lorenzen. Vandals, administrators, and sockpuppets, oh my! an ethnographic study of wikipedia's handling of problem behavior. *MLA forum*, 5(2), December 2006.

[Lot26] Alfred J. Lotka. The frequency distribution of scientific productivity. In *Journal of the Washington Academy of Sciences*, volume 12, pages 317–324, 1926.

[LS05] Lipczynska and Sonya. Power to the people: the case for wikipedia. *Reference Reviews incorporating ASLIB Book Guide*, 19(2):6–7, February 2005.

[LS10] Teun Lucassen and Jan Maarten PRIMERO Schraagen. Trust in wikipedia: how users trust information from an unknown source. In *WICOW '10: Proceedings of the 4th workshop on Information credibility*, pages 19–26, New York, NY, USA, 2010. ACM.

[LWZ04] Ming-Kuan Liu, Fei-Yue Wang, and Daniel Dajun Zeng. Web caching: a way to improve web qos. *J. Comput. Sci. Technol.*, 19:113–127, March 2004.

[M+97] Jeffrey A. Mills et al. Statistical inference via bootstrapping for measures of inequality. *Epidemiological Bulletin*, 22(12), March/April 1997.

[Ma05] Cathy Ma. The social, cultural, economical implications of the wikipedia. *Computers and Writing Online 2005*, 2005.

[MB06] John Maindonald and John Braun. *Data Analysis and Graphics Using R: An Example-based Approach (Cambridge Series in Statistical and Probabilistic Mathematics)*, chapter 8, pages 275–284. Cambridge University Press, December 2006.

[Mcc05] R. Mccool. Rethinking the semantic web, part 1. *Internet Computing, IEEE*, 9(6):88–87, 2005.

[Mcc06] Rob Mccool. Rethinking the semantic web, part 2. *Internet Computing, IEEE*, 10(1):96–95, 2006.

[McG09] John McGrady. Gaming against the greater good. *First Monday*, 14(2), February 2009.

[Mck05] Gerry Mckiernan. Wikimediaworlds. *Library Hi Tech News incorporating Online and CD Notes*, 22(8):46–54, August 2005.

[Mil05] Nora Miller. Wikipedia and the disappearing "author". *ETC: A Review of General Semantics*, 62(1):37–40, January 2005.

[Mit] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions.

[Mit07] Domas Mituzas. Wikipedia: Site internals, configuration, code examples and management issues. Technical report, Wikimedia Foundation, April 2007.

[MMB08] Claudia Müller, Benedikt Meuthrath, and Anne Baumgrass. Analyzing wiki-based networks to improve knowledge processes in organizations. In *MISS*, volume 14, pages 526–545. MISS, February 2008.

[MMLW09] Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. Mining meaning from wikipedia. May 2009.

[MMW06] David Milne, Olena Medelyan, and Ian H. Witten. Mining domain-specific thesauri from wikipedia: A case study. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 442–448, Washington, DC, USA, 2006. IEEE Computer Society.

[Mor07] Joseph C. Morris. Distriwiki:: a distributed peer-to-peer wiki network. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 69–74, New York, NY, USA, 2007. ACM.

[Mur05] Paul Murrell. *R Graphics (Computer Science and Data Analysis)*. Chapman & Hall/CRC, July 2005.

[MZdS+06] Deborah L. Mcguinness, Honglei Zeng, Paulo P. da Silva, Li Ding, Dhyanesh Narayanan, and Mayukh Bhaowal. Investigations into trust for collaborative information repositories: A wikipedia case study. In *Proceedings of the Workshop on Models of Trust for the Web*, 2006.

[Nej02] W. Nejdl. Semantic web and peer-to-peer technologies for distributed learning repositories, 2002.

[New04] M. E. J. Newman. Power laws, pareto distributions and zipf's law, December 2004.

[NGJ03] G. B. Newby, J. Greenberg, and P. Jones. Open source software development and lotkas law: Bibliometric patterns in programming. *JASIST (Journal of the American Society for Information Science and Technology)*, 54(2):1169–1178, 2003.

[Nie07] Finn A. Nielsen. Scientific citations in wikipedia, May 2007.

[NKCM90] Christine M. Neuwirth, David S. Kaufer, Ravinder Chandhok, and James H. Morris. Issues in the design of computer support for co-authoring and commenting. In *CSCW '90: Proceedings of the 1990 ACM conference on Computer-supported cooperative work*, pages 183–195, New York, NY, USA, 1990. ACM Press.

[Nov07] Oded Nov. What motivates wikipedians? *Commun. ACM*, 50(11):60–64, November 2007.

[NR03] S. Noël and J. M. Robert. How the web is used to support collaborative writing. *Behaviour and Information Technology*, 22(4):245–262, July 2003.

[obKHpbRGa] S original by Kenneth Hess and R port by R. Gentleman. *muhaz: Hazard Function Estimation in Survival Analysis*. R package version 1.2.3.

[obKHpbRGb] S original by Kenneth Hess and R port by R. Gentleman. *muhaz: Hazard Function Estimation in Survival Analysis*. R package version 1.2.3.

[OC09] Felipe Ortega and Kevin Crowston. Introduction to open movements: Floss, open content and open communities minitrack. In *Proceedings of the 42nd Hawaiian International Conference on System Sciences (HICSS 2009)*, January 2009.

[OGB07] Felipe Ortega and Jesus M. Gonzalez-Barahona. Quantitative analysis of the wikipedia community of users. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 75–86, New York, NY, USA, 2007. ACM.

[OGB08] Felipe Ortega and Jesus M. Gonzalez-Barahona. On the inequality of contributions to wikipedia. In *Proceedings of the 41st Hawaiian International Conference on System Sciences (HICSS 2008)*, January 2008.

[OGBR07] Felipe Ortega, Jesus M. Gonzalez-Barahona, and Gregorio Robles. The top ten wikipedias: A quantitative analysis using wikixray. In *Proceedings of the 2nd International Conference on Software and Data Technologies (ICSOFT 2007)*. INSTICC, Springer-Verlag, July 2007.

[Oll08] F.X. Olleros. Learning to trust the crowd: Some lessons from wikipedia. In *e-Technologies, 2008 International MCETECH Conference on*, pages 212 –216, jan. 2008.

[ON08] Shaul Oreg and Oded Nov. Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values. *Computers in Human Behavior*, 24:2055–2073, September 2008.

[O'R05] Tim O'Reilly. O'reilly – what is web 2.0, 2005.

[Ort09] Felipe Ortega. *Wikipedia: A quantitative analysis*. PhD thesis, Universidad Rey Juan Carlos, 2009. http://libresoft.es/Members/jfelipe/phd-thesis.

[OS07] Yann Ollivier and Pierre Senellart. Finding related pages using green measures: An illustration with wikipedia. In *Conference on Artificial Intelligence (AAAI 2007)*. Association for the Advancement of Artificial Intelligence, 2007.

[Paq03] Sébastien Paquet. Seb's "wikis and knowledge sharing" survey: Results, 2003.

[PCS+07] Reid Priedhorsky, Jilin Chen, Shyong, Kathering Panciera, Loren Terveen, and John. Creating, destroying, and restoring value in wikipedia. *MISSING*, November 2007.

[Pot07] Martin Potthast. Wikipedia in the pocket: indexing technology for near-duplicate detection and high similarity search. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 909–909, New York, NY, USA, 2007. ACM.

[PY06] Barbara Poblete and Ricardo B. Yates. A content and structure website mining model. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 957–958, New York, NY, USA, 2006. ACM.

[PY08] Barbara Poblete and Ricardo B. Yates. Query-sets: using implicit feedback and query patterns to organize web documents. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 41–50, New York, NY, USA, 2008. ACM.

[PZA06] U. Pfeil, P. Zaphiris, and C. S. Ang. Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113, January 2006.

[Qui06] John Quiggin. Blogs, wikis and creative innovation. *International Journal of Cultural Studies*, 9(4):481–496, December 2006.

[R D08] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[R D09] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[RAGBH05] G. Robles, J. J. Amor, J. M. Gonzalez-Barahona, and I. Herraiz. Evolution and growth in large libre software projects. In *MISSING*, pages 165–174, 2005.

[RAH06] Sheizaf Rafaeli, Yaron Ariel, and Tsah Hayat. Wikipedians sense of (virtual) community. In *eighth International Conference General Online Research (GOR06)*, Bielefeld, Germany, 2006.

[Ray01] Eric S. Raymond. *The Cathedral & the Bazaar*. O'Reilly, January 2001.

[RCAC05] Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In *MISSING*, pages 380–386. MISSING, 2005.

[Rea05] Joseph M. Reagle. A case of mutual aid: Wikipedia, politeness, and perspective taking. In *Proceedings of Wikimania 2005.The First International Wikimedia Conference*, Frankfurt, Germany, 2005.

[Rea07] Joseph M. Reagle. Do as i do: authorial leadership in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 143–156, New York, NY, USA, 2007. ACM.

[RF07] Casey Reas and Ben Fry. *Processing: A Programming Handbook for Visual Designers and Artists*. The MIT Press, September 2007.

[RGB06] Gregorio Robles and Jesus Gonzalez-Barahona. Contributor turnover in libre software projects. In *MISSING*, pages 273–286. MISSING, 2006.

[RGBOR08] Antonio J. Reinoso, Jesus M. Gonzalez Barahona, Felipe Ortega, and Greogrio Robles. Quantitative analysis and characterization of Wikipedia requests. In *Proceedings of the 4th International Symposium on Wikis*, WikiSym '08, New York, NY, USA, 2008. ACM.

[RGBRO09] Antonio J. Reinoso, Jesús M. González Barahona, Gregorio Robles, and Felipe Ortega. A quantitative approach to the use of the wikipedia. In *ISCC*, pages 56–61. IEEE, July 2009.

[RH08] Rector and Lucy Holman. Comparison of itwikipedia/it and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review*, 36(1):7–22, 2008.

[RHA05a] Sheizaf Rafaeli, Tsahi Hayat, and Yaron Ariel. Wikipedia participants and "ba": Knowledge building and motivations. In *Cyberculture 3rd Global Conference. Prague, Czech Republic*, 2005.

[RHA05b] Sheizaf Rafaeli, Tsahi Hayat, and Yaron Ariel. Wikipedians' sense of community, motivations, and knowledge building. In *Proceedings of Wikimania 2005 - The First International Wikimedia Conference*, Frankfurt, Germany, 2005.

[Rhe00] Howard Rheingold. *The Virtual Community: Homesteading on the Electronic Frontier*. The MIT Press, rev sub edition, November 2000.

[Rie06] Dirk Riehle. How and why wikipedia works: an interview with angela beesley, elisabeth bauer, and kizu naoko. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, pages 3–8, New York, NY, USA, 2006. ACM Press.

[Rob06] Gregorio Robles. *Software Engineering Research on Libre Software: Data Sources, Methodologies and Results*. PhD in Computer Science, Escuela Superior de Ciencias Experimentales y Tencologia, Universidad Rey Juan Carlos, 2006.

[ROGBH10] Antonio J. Reinoso, Felipe Ortega, Jesus M. Gonzalez-Barahona, and Israel Herraiz. A statistical approach to the impact of featured articles in wikipedia. In *International Conference on Knowledge Engineering and Ontology Development*, Valencia, Spain, 2010.

[RSJD04] Bendel R.B., Higgins S.S., Teberg J.E., and Pyke D.A. Comparison of skewness coefficient, coefficient of variation, and gini coefficient as inequality measures within populations. *Oecologia*, 78(3):394–400, Mar. 2004.

[Sar08a] Deepayan Sarkar. *lattice: Lattice Graphics*, 2008. R package version 0.17-15.

[Sar08b] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R (Use R)*. Springer, March 2008.

[SBB+06] Sebastian Schaffert, Diana Bischof, Tobias Bürger, Andreas Gruber, Wolf Hilzensauer, and Sandra Schaffert. Learning with semantic wikis. In *First Workshop "SemWiki2006 - From Wiki to Semantics", co-located with the 3rd Annual European Semantic Web Conference (ESWC)*, June 2006.

[SCCP09] Bongwon Suh, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. The singularity is not near: slowing growth of wikipedia. In *WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, pages 1–10, New York, NY, USA, 2009. ACM.

[Sch06a] Sebastian Schaffert. Ikewiki: A semantic wiki for collaborative knowledge management. In *1st International Workshop on Semantic Technologies in Collaborative Applications STICA 06*, June 2006.

[Sch06b] Peter Schonhofen. Identifying document topics using the wikipedia category network. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 456–462, Washington, DC, USA, 2006. IEEE Computer Society.

[Sch08] N. J. Schweitzer. Wikipedia and psychology: Coverage of concepts and its use by undergraduate students. *Teaching of Psychology*, 35(2):81–85, 2008.

[SCPK07] Bongwon Suh, Ed H. Chi, Bryan A. Pendleton, and Aniket Kittur. Us vs. them: Understanding social dynamics in wikipedia with revert graph visualizations. In *2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 163–170. IEEE, October 2007.

[SdMD06] Mareike Schoop, Aldo de Moor, and Jan L. G. Dietz. The pragmatic web: a manifesto. *Commun. ACM*, 49(5):75–76, May 2006.

[Sel08] Steve Selvin. *Survival Analysis for Epidemiologic and Medical Research (Practical Guides to Biostatistics and Epidemiology)*. Cambridge University Press, 1 edition, March 2008.

[SG08] Besiki Stvilia and Les Gasser. An activity theoretic model for information quality change. *First Monday*, 13(8), April 2008.

[SH02] Felix Stalder and Jesse Hirsh. Open source intelligence. *First Monday*, 7(6), 2002.

[SH07] Klaus Stein and Claudia Hess. Does it matter who contributes: a study on featured articles in the german wikipedia. In *HT '07: Proceedings of the 18th conference on Hypertext and hypermedia*, pages 171–174, New York, NY, USA, 2007. ACM.

[SH09] Joachim Schroer and Guido Hertel. Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it. *Media Psychology*, 12(1):96–120, 2009.

[Shi08] Daniel Shiffman. *Learning Processing: A Beginner's Guide to Programming Images, Animation, and Interaction (Morgan Kaufmann Series in Computer Graphics)*. Morgan Kaufmann, illustrated edition edition, August 2008.

[SL08] Diomidis Spinellis and Panagiotis Louridas. The collaborative organization of knowledge. *Commun. ACM*, 51(8):68–73, August 2008.

[SMHM99] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[SML07] Patrick A. S. Sinclair, Kirk Martinez, and Paul H. Lewis. Dynamic link service 2.0: using wikipedia as a linkbase. In *HT '07: Proceedings of the 18th conference on Hypertext and hypermedia*, pages 161–162, New York, NY, USA, 2007. ACM.

[SMZ04] Kunwadee Sripanidkulchai, Bruce Maggs, and Hui Zhang. An analysis of live streaming workloads on the internet. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, IMC '04, pages 41–54, New York, NY, USA, 2004. ACM.

[Sou05] A. Souzis. Building a semantic wiki. *Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications]*, 20(5):87–91, 2005.

[SP06] Michael Strube and Simone P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI'06: proceedings of the 21st national conference on Artificial intelligence*, pages 1419–1424. AAAI Press, 2006.

[SPH06] Sander Spek, Eric Postma, and Jaap Herik. Wikipedia: organisation from a bottom-up approach, Nov 2006.

[Spo07a] A. Spoerri. What is popular on wikipedia and why? *First Monday*, 12, April 2007.

[Spo07b] Anselm Spoerri. What is popular on wikipedia and why? *First Monday*, 12(4), 2007.

[SSB07] Christoph Sauer, Chuck Smith, and Tomas Benz. Wikicreole: a common wiki markup. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 131–142, New York, NY, USA, 2007. ACM.

[STGS05] B. Stvilia, M. B. Twidale, L. Gasser, and L. C. Smith. Information quality in a community-based encyclopedia. In S. Hawamdeh, editor, *Knowledge Management: Nurturing Culture, Innovation, and Technology - Proceedings of the 2005 International Conference on Knowledge Management*, pages 101–113, Charlotte, NC, 2005. World Scientific Publishing Company.

[STS05] Besiki Stvilia, Michael B. Twidale, and Linda C. Smith. Information quality: Discussions in wikipedia. *MISSING*, 2005.

[STSG05] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality - ICIQ 2005*, pages 442–454, 2005.

[Stv08] Besiki FALTA Stvilia. Information quality work organization in wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001, 2008.

[Sur04] James Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, May 2004.

[Sur05] James Surowiecki. *The Wisdom of Crowds*. Anchor, August 2005.

[Swa06] Aaron Swartz. Who writes wikipedia, September 2006.

[TL08] Terry Therneau and Thomas Lumley. *survival: Survival analysis, including penalised likelihood.*, 2008. R package version 2.34-1.

[TR09] Shyong Tony and John Riedl. Is wikipedia growing a longer tail? In *GROUP '09: Proceedings of the ACM 2009 international conference on Supporting group work*, pages 105–114, New York, NY, USA, 2009. ACM.

[TW08] Don Tapscott and Anthony D. Williams. *Wikinomics: How Mass Collaboration Changes Everything*. Portfolio Hardcover, April 2008.

[TWF01] Vinod Thomas, Yan Wang, and Xibo Fan. Measuring education inequality - gini coefficients of education. Policy Research Working Paper Series 2525, The World Bank, January 2001.

[UPvS07a] Guido Urdaneta, Guillaume Pierre, and Maarten van Steen. A decentralized wiki enginge for collaborative wikipedia hosting. In *Proceedings of the 3rd International Conference on Web Information Systems and Technologies*, pages 156–163, March 2007.

[UPvS07b] Guido Urdaneta, Guillaume Pierre, and Maarten van Steen. Wikipedia workload analysis. *MISSING*, September 2007.

[VD04] J. Voss and P. Danowski. Bibliothek, information und dokumentation in der wikipedia. *Information : Wissenschaft und Praxis*, 8:457–462, 2004.

[Vie07] Fernanda B. Viegas. The visual side of wikipedia. In *Proceedings of Hawaiian International Conference of Systems Sciences*, January 2007.

[VKV+06] Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic wikipedia. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 585–594, New York, NY, USA, 2006. ACM Press.

[Vos05] Jakob Voss. Measuring wikipedia. In *International Conference of the International Society for Scientometrics and Informetrics : 10th*. ISSI, July 2005.

[Vos06] Jakob Voss. Collaborative thesaurus tagging the wikipedia way, Apr 2006.

[VR03] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*, chapter 13. Springer, September 2003.

[VWD04] F. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of SIGCHI*, pages 575–582, Vienna, Austria, 2004.

[VWKvH07] Fernanda B. Viégas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type: Coordination in wikipedia. In *MISSING*, pages 78–78, 2007.

[VWM07] Fernanda B. Viégas, Martin Wattenberg, and Matthew Mckeon. The hidden order of wikipedia. In *MISSING*, pages 445–454. MISSING, 2007.

[W+91] A Wagstaff et al. On the measurements of inequalities in health. *Soc. Sci. Med.*, 33(5):545–577, 1991.

[Wag04] Christian Wagner. Wiki: A technology for conversational knowledge management and group collaboration. *Communications of the AIS*, 13:256–289, 2004.

[Wag05] Christian Wagner. Breaking the knowledge acquisition bottleneck through conversational knowledge management. *Information Resources Management Journa*, 2005.

[Wal05] Jimmy Wales. Wikipedia in the free culture revolution. In *OOPSLA '05: Companion to the 20th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications*, pages 5–5, New York, NY, USA, 2005. ACM.

[Wat07] Neil L. Waters. Why you can't cite wikipedia in my class. *Commun. ACM*, 50(9):15–17, September 2007.

[WB05] Christian Wagner and Narasimha Bolloju. Supporting knowledge management in organizations with conversational technologies: Discussion forums, weblogs, and wikis. *Journal of Database Management*, 16(2):i–viii, 2005.

[WCIB06] Christian Wagner, Karen Cheung, Rachael Ip, and Stefan Bottcher. Building semantic webs for e-government with wiki technology. *Electronic Government*, 3(1):36–55, 2006.

[WH07a] Dennis M. Wilkinson and Bernardo A. Huberman. Assessing the value of coooperation in wikipedia, April 2007.

[WH07b] Dennis M. Wilkinson and Bernardo A. Huberman. Cooperation and quality in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 157–164, New York, NY, USA, 2007. ACM.

[wik] Wiki research bibliography. consulted on April 13th, 2008.

[Wil07] John Willinsky. What open access research can do for wikipedia. *First Monday*, 12(3), March 2007.

[Wil08] Chris Wilson. The wisdom of the chaperones. digg, wikipedia, and the myth of web 2.0 democracy. Technical report, February 2008.

[WSC06] Gabriel Weaver, Barbara Strickland, and Gregory Crane. Quantifying the accuracy of relational statements in wikipedia: a methodology. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 358–358, New York, NY, USA, 2006. ACM Press.

[WVH07] Martin Wattenberg, Fernanda Viégas, and Katherine Hollenbach. Visualizing activity on wikipedia with chromograms. In *MISSING*, pages 272–287. MISSING, 2007.

[WWZ09] Dietmar Wolfram, Peiling Wang, and Jin Zhang. Identifying Web search session patterns using cluster analysis: A comparison of three search environments. *Journal of the American Society for Information Science and Technology*, 9999(9999):NA+, 2009.

[WWZY07] Yang Wang, Haofen Wang, Haiping Zhu, and Yong Yu. Exploit semantic information for category annotation recommendation in wikipedia. In *Natural Language Processing and Information Systems*, pages 48–60. MISSING, 2007.

[WZM06] Harris Wu, Mohammad Zubair, and Kurt Maly. Harvesting social knowledge from folksonomies. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 111–114, New York, NY, USA, 2006. ACM Press.

[XCY07] Wenpeng Xiao, Changyan Chi, and Min Yang. On-line collaborative software development via wiki. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 177–183, New York, NY, USA, 2007. ACM.

[ZAD+06] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. Mcguinness. Computing trust from revision history. *MISSING*, November 2006.

[ZBvD06] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74(1), 2006.

[Zei07] Achim Zeileis. *ineq: Measuring inequality, concentration and poverty*, 2007. R package version 0.2-8.

[ZG07] Torsten Zesch and Iryna Gurevych. Analysis of the wikipedia category graph for nlp applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, 2007.

[ZGM06] Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. Analyzing and accessing wikipedia as a lexical semantic resource. In *Biannual Conference of the Society for Computational Linguistics and Language Technology*, pages 213–221, 2006.

[ZGM07] Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. Comparing wikipedia and german wordnet by evaluating semantic relatedness on multiple datasets. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. Association for Computational Linguistics, 2007.

[Zha06] Yuejiao Zhang. Wiki means more: hyperreading in wikipedia. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 23–26, New York, NY, USA, 2006. ACM Press.

Appendix A

Validation tables

Tables A.1 and A.2 contain the result of the comparison of the number of pageviews reported from Mituzas's log files (Rows indicated with 'Mituzas') with the number of pageviews obtained after our analysis (Rows heading by 'Ours'). The quotient (Rows with 'Quotient') between the two measures is also presented to evaluate its closeness to the sampling factor (1/100). This close match means that we are disregarding very few log lines, if any, considered of interest. The difference with Mituzas' figures may be also affected by articles in namespaces not considered in this thesis. As a result, we can consider our filtering process as rightly driven and trust enough.

Tables A.3 and A.4 present the comparison between the number of edits from Zachte's site corresponding to articles in the considered Wikipedias (Rows indicated with 'Zachte') and the number of edits after our own results (Rows heading by 'Ours'). The quotient (Rows with 'Quotient') between the two measures is also presented to assess that its closeness to the sampling factor (1/100). Again, the general quotient of 0.01 means that the our feed consists on the 1/100 sample of all the requests and, again, the filtering process is not overlooking any request asking for edit operations.

Tables A.5 and A.6 present the comparison between the number of edit operations after our analysis and after the *WikiXRay* tool used by Ortega in [Ort09]

Lang.	Jan.	Feb.	Mar.	Apr.	May.	Jun.
DE (Ours)	10,821,625	6,833,171	8,034,636	6,945,878	7,612,949	7,249,244
DE (Mituzas)	1,271 M	982 M	978 M	817 M	875 M	909 M
Quotient	0.009	0.007	0.008	0.009	0.009	0.008
EN (Ours)	47,369,841	43,136,627	51,845,199	48,242,580	48,085,156	43,950,168
EN (Mituzas)	5,615 M	5,944 M	6,092 M	5,989 M	6,066 M	5,819 M
Quotient	0.0084	0.0073	0.0085	0.0081	0.0079	0.0076
ES (Ours)	4,411,173	4,752,977	6,057,891	5,438,380	6,079,028	5,419,625
ES (Mituzas)	526 M	665 M	709 M	623 M	713 M	689 M
Quotient	0.0084	0.0071	0.0085	0.0087	0.0085	0.0079
FR (Ours)	3,945,670	3,433,034	4,133,455	4,025,746	4,195,556	3,604,704
FR (Mituzas)	489 M	490 M	511 M	513 M	518 M	479 M
Quotient	0.0081	0.0070	0.0081	0.0078	0.0081	0.0075
IT (Ours)	2,815,854	2,491,855	2,926,519	2,836,434	2,941,568	2,857,848
IT (Mituzas)	324M	331M	334M	321M	325M	339M
Quotient	0.0087	0.0075	0.0088	0.0088	0.0091	0.0084
JA (Ours)	9,202,652	8,022,811	8,835,897	8,508,914	9,488,843	8,816,399
JA (Mituzas)	1,020 M	1,016 M	966 M	936 M	1,054 M	1,076 M
Quotient	0.0090	0.0079	0.0091	0.0091	0.0090	0.0082
NL (Ours)	1,301,279	1,085,099	1,349,849	1,166,997	1,269,936	1,161,305
NL (Mituzas)	154 M	147 M	158 M	133 M	143 M	142 M
Quotient	0.0084	0.0074	0.0085	0.0088	0.0089	0.0082
PL (Ours)	3,359,914	2,654,506	3,387,327	2,800,633	3,052,641	2,370,672
PL (Mituzas)	379 M	348 M	378 M	309 M	333 M	278 M
Quotient	0.0089	0.0076	0.0090	0.0091	0.0092	0.0085
PT (Ours)	1,468,445	1,414,783	2,163,905	2,016,947	2,183,219	2,056,801
PT (Mituzas)	174 M	196 M	251 M	226 M	249 M	252 M
Quotient	0.0084	0.0072	0.0086	0.0089	0.0088	0.0082
RU (Ours)	1,990,244	1,841,822	2,335,899	2,354,768	2,497,543	2,306,491
RU (Mituzas)	244 M	261 M	285 M	276 M	285 M	287 M
Quotient	0.0082	0.0071	0.0082	0.0085	0.0088	0.0080

Table A.1: Comparison of the number of pageviews for the whole set of Wikipedia editions during the first semester of 2009. M stands for Million.

Lang	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
DE (Ours)	6,626,701	6,942,208	7,404,872	7,223,746	7,615,539	7,102,197
DE (Mituzas)	819 M	813 M	889 M	885 M	904 M	760 M
Quotient	0.008	0.009	0.008	0.008	0.008	0.009
EN (Ours)	44,451,649	48,426,122	49,713,090	49,392,482	49,738,157	47,687,869
EN (Mituzas)	5,614 M	5,604 M	5,938 M	6,041 M	5,842 M	5,259 M
Quotient	0.0079	0.0086	0.0084	0.0082	0.0085	0.0091
ES (Ours)	4,632,767	6,058,239	6,955,212	6,603,739	6,507,704	4,467,558
ES (Mituzas)	569 M	670 M	805 M	793 M	750 M	500 M
Quotient	0.0081	0.0090	0.0086	0.0083	0.0087	0.0089
FR (Ours)	3,056,991	3,319,903	3,854,688	4,058,351	4,207,051	3,738,801
FR (Mituzas)	402 M	405 M	468 M	518 M	518 M	457 M
Quotient	0.0076	0.0082	0.0082	0.0078	0.0081	0.0082
IT (Ours)	2,568,739	2,545,767	3,051,185	2,899,914	2,936,762	2,723,087
IT (Mituzas)	302 M	281 M	349 M	340 M	335 M	293 M
Quotient	0.0085	0.0091	0.0087	0.0085	0.0088	0.0093
JA (Ours)	9,093,702	9,710,101	9,224,460	9,072,514	9,178,759	9,043,711
JA (Mituzas)	1,072 M	1,057 M	1,036 M	1,042 M	1,019 M	948 M
Quotient	0.0085	0.0092	0.0089	0.0087	0.0090	0.0095
NL (Ours)	954,441	1,043,484	1,206,443	1,269,412	1,289,915	1,174,796
NL (Mituzas)	116 M	118 M	140 M	149 M	149 M	128 M
Quotient	0.0082	0.0088	0.0086	0.0085	0.0087	0.0092
PL (Ours)	2,013,671	2197485	2,696,572	2,704,090	2,854,847	2,689,520
PL (Mituzas)	237 M	240 M	300 M	309 M	317 M	280 M
Quotient	0.0085	0.0092	0.0090	0.0088	0.0090	0.0096
PT (Ours)	1,714,607	2,215,491	2,534,121	2,286,352	2,416,963	1,797,790
PT (Mituzas)	205 M	239 M	285 M	265 M	271 M	193 M
Quotient	0.0084	0.0093	0.0089	0.0086	0.0089	0.0093
RU (Ours)	2,043,838	2,301,908	2,578,112	2,826,355	3,021,851	3,106,244
RU (Mituzas)	250 M	263 M	305 M	336 M	351 M	342 M
Quotient	0.0082	0.0088	0.0085	0.0084	0.0086	0.0091

Table A.2: Comparison of the number of pageviews for the whole set of Wikipedia editions and from July till December 2009. M stands for Million.

Lang.	Jan.	Feb.	Mar.	Apr.	May.	Jun.
DE (Ours)	11,041	9,457	10,341	8,361	8,052	7,754
DE (Zachte)	876 K	752 K	802 K	655 K	684 K	701 K
DE (Quotient)	0.0126	0.0126	0.0129	0.0128	0.0118	0.0111
EN (Ours)	53,121	46,778	54,564	47,921	47,692	42,282
EN (Zachte)	4,300 K	4,200 K	4,400 K	4,000 K	4,300 K	4,000 K
EN (Quotient)	0.0124	0.0111	0.0124	0.0120	0.0111	0.0106
ES (Ours)	6,513	6,487	6,383	5,534	5,480	5,051
ES (Zachte)	563 K	573 K	559 K	536 K	614 K	628 K
ES (Quotient)	0.0116	0.0113	0.0114	0.0103	0.0089	0.0080
FR (Ours)	8,146	7,280	7,549	6,403	6,630	5,989
FR (Zachte)	672 K	638 K	633 K	621 K	771 K	676 K
FR (Quotient)	0.0121	0.0114	0.0119	0.0103	0.0086	0.0089
IT (Ours)	7,345	5,696	5,685	5,322	5,113	4,393
IT (Zachte)	522 K	443 K	446 K	468 K	543 K	494 K
IT (Quotient)	0.0141	0.0129	0.0127	0.0114	0.0094	0.0089
JA (Ours)	4,506	4,083	4,606	4,193	4,253	3,694
JA (Zachte)	420 K	381 K	430 K	414 K	451 K	417 K
JA (Quotient)	0.0107	0.0107	0.0107	0.0101	0.0094	0.0089
NL (Ours)	3,126	3,155	3,995	2,779	2,815	2,130
NL (Zachte)	253 K	279 K	334 K	285 K	311 K	264 K
NL (Quotient)	0.0124	0.0113	0.0120	0.0098	0.0091	0.0081
PL (Ours)	3,686	3,086	4,317	2,636	2,458	2,222
PL (Zachte)	308 K	275 K	291 K	260 K	285 K	266 K
PL (Quotient)	0.0120	0.0112	0.0148	0.0101	0.0086	0.0084
PT (Ours)	3,045	2,781	2,793	2,397	2,433	2,186
PT (Zachte)	259 K	247 K	240 K	245 K	266 K	259 K
PT (Quotient)	0.0118	0.0113	0.0116	0.0098	0.0091	0.0082
RU (Ours)	5,511	4,516	5,576	5,068	4,842	4,614
RU (Zachte)	458 K	393 K	467 K	452 K	474 K	479 K
RU (Quotient)	0.0120	0.0115	0.0119	0.0112	0.0102	0.0173

Table A.3: Comparison of the edit operations reported by Zachte's site for the whole set of Wikipedia editions and for the first semester of 2009 with the results of our analysis. K stands for thousands. M stands for Million.

Lang	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
DE (Ours)	7,688	8,393	8,111	7,968	7,942	7,581
DE (Zachte)	688 K	729 K	680 K	714 K	716 K	714 K
DE (Quotient)	0.0112	0.0115	0.0119	0.0112	0.0111	0.0106
EN (Ours)	41,087	45,492	43,969	38,631	37,641	36,568
EN (Zachte)	3,800 K	3,900 K	4,000 K	4,000 K	3,900 K	4,400 K
EN (Quotient)	0.0108	0.0117	0.0110	0.0097	0.0097	0.0083
ES (Ours)	5,263	5,735	5,769	5,100	4,938	4,529
ES (Zachte)	635 K	574 K	603 K	586 K	563 K	532 K
ES (Quotient)	0.0083	0.0100	0.0096	0.0087	0.0088	0.0085
FR (Ours)	5,558	6,183	5,815	5,712	5,851	5,527
FR (Zachte)	622 K	681 K	633 K	671 K	660 K	661 K
FR (Quotient)	0.0089	0.0091	0.0092	0.0085	0.0089	0.0084
IT (Ours)	4,279	4,110	4,486	3,761	3,739	3,684
IT (Zachte)	498 K	465 K	498 K	469 K	462 K	458 K
IT (Quotient)	0.0086	0.0088	0.0090	0.0080	0.0081	0.0080
JA (Ours)	3,653	3,926	3,862	3,680	3,716	3,484
JA (Zachte)	421 K	461 K	447 K	429 K	428 K	406 K
JA (Quotient)	0.0087	0.0085	0.0086	0.0086	0.0087	0.0086
NL (Ours)	1,984	2,134	2,093	1,965	1,849	1,858
NL (Zachte)	236 K	249 K	265 K	291 K	267 K	244 K
NL (Quotient)	0.0084	0.0086	0.0079	0.0068	0.0069	0.0076
PL (Ours)	2,234	2,339	2,281	1,997	2,017	2,041
PL (Zachte)	285 K	266 K	260 K	293 K	282 K	290 K
PL (Quotient)	0.0078	0.0088	0.0088	0.0068	0.0072	0.0070
PT (Ours)	2,255	2,706	2,320	1,948	1,790	1,925
PT (Zachte)	258 K	286 K	264 K	310 K	256 K	277 K
PT (Quotient)	0.0087	0.0095	0.0088	0.0063	0.0070	0.0069
RU (Ours)	4,549	6,425	6,163	4,429	4,497	4,445
RU (Zachte)	472 K	481 K	472 K	525 K	528 K	508 K
RU (Quotient)	0.0096	0.0134	0.0131	0.0084	0.0085	0.0088

Table A.4: Comparison between the number of edits from Zachte's site corresponding to the whole set of Wikipedias and from July till December. K stands for thousands. M stands for Million.

Lang.	Jan.	Feb.	Mar.	Apr.	May.	Jun.
DE (Ours)	11,041	9,457	10,341	8,361	8,052	7,754
DE (Ortega)	1,227,017	1,069,725	1,148,209	962,561	987,244	1,013,734
DE (Quotient)	0.0090	0.0088	0.0090	0.0087	0.0082	0.0076
EN (Ours)	53,121	46,778	54,564	47,921	47,692	42,282
EN (Ortega)	6,195,518	5,926,109	6,614,845	5,876,645	6,166,014	5,702,894
EN (Quotient)	0.0086	0.0079	0.0082	0.0082	0.0077	0.0074
ES (Ours)	6,513	6,487	6,383	5,534	5,480	5,051
ES (Ortega)	703,823	710,674	719,996	683,336	778,404	783,012
ES (Quotient)	0.0093	0.0091	0.0089	0.0081	0.0070	0.0065
FR (Ours)	8,146	7,280	7,549	6,403	6,630	5,989
FR (Ortega)	931,125	890,550	949,120	885,512	1,077,889	1,010,830
FR (Quotient)	0.0087	0.0082	0.0080	0.0072	0.62	0.0059
IT (Ours)	7,345	5,696	5,685	5,322	5,113	4,393
IT (Ortega)	673,821	583,216	583,689	613,025	674,298	622,251
IT (Quotient)	0.0109	0.0098	0.0097	0.0087	0.0076	0.0071
JA (Ours)	4,506	4,083	4,606	4,193	4,253	3,694
JA (Ortega)	489,815	448,522	511,996	478,603	529,484	491,352
JA (Quotient) 0.0092	0.0091	0.0090	0.0088	0.0080	0.0075	
NL (Ours)	3,126	3,155	3,995	2,779	2,815	2,130
NL (Ortega)	333,345	347,098	415,458	362,097	388,637	359,057
NL (Quotient)	0.0094	0.0091	0.0096	0.0077	0.0072	0.0059
PL (Ours)	3,686	3,086	4,317	2,636	2,458	2,222
PL (Ortega)	385,127	348,300	359,269	326,777	354,200	330,687
PL (Quotient)	0.0096	0.0089	0.0120	0.0081	0.0069	0.0067
PT (Ours)	3,045	2,781	2,793	2,397	2,433	2,186
PT (Ortega)	355,209	345,603	346,850	329,893	364,971	350,702
PT (Quotient)	0.0086	0.0080	0.0081	0.0073	0.0067	0.0062
RU (Ours)	5,511	4,516	5,576	5,068	4,842	4,614
RU (Ortega)	622,510	529,972	649,664	606,935	631,921	636,549
RU (Quotient)	0.0089	0.0085	0.0086	0.0084	0.0077	0.0072

Table A.5: Comparison between the number of edits on articles of all the considered Wikipedias obtained from our results (Rows heading by 'Ours') for January till June 2009 and the same number of operations reported by Ortegas's tool *WikiXRay* (Rows indicated with 'Ortega') for the same period. Both data correspond to articles in the main namespace. Rows headed by 'Quotient' correspond to the quotient between the two measures.

Lang	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
DE (Ours)	7,688	8,393	8,111	7,968	7,942	7,581
DE (Ortega)	993,866	1,048,137	975,990	1,056,171	1,091,001	1,073,048
DE (Quotient)	0.0077	0.0080	0.0083	0.0075	0.0073	0.0071
EN (Ours)	41,087	45,492	43,969	38,631	37,641	36,568
EN (Ortega)	5,492,827	5,557,041	5,762,412	5,747,647	5,497,166	6,060,027
EN (Quotient)	0.0075	0.0082	0.0076	0.0067	0.0068	0.0060
ES (Ours)	5,263	5,735	5,769	5,100	4,938	4,529
ES (Ortega)	790,497	728,937	780,566	760,488	722,453	683,143
ES (Quotient)	0.0067	0.0079	0.0074	0.0067	0.0068	0.0066
FR (Ours)	5,558	6,183	5,815	5,712	5,851	5,527
FR (Ortega)	831,180	927,447	885,531	979,869	926,301	976,643
FR (Quotient)	0.0067	0.0067	0.0066	0.0058	0.0063	0.0057
IT (Ours)	4,279	4,110	4,486	3,761	3,739	3,684
IT (Ortega)	625,344	596,566	695,965	608,687	594,970	584,376
IT (Quotient)	0.0068	0.0069	0.0064	0.0062	0.0063	0.0063
JA (Ours)	3,653	3,926	3,862	3,680	3,716	3,484
JA (Ortega)	485,637	530,283	514,313	504,413	504,767	171,329
JA (Quotient)	0.0075	0.0074	0.0075	0.0073	0.0074	0.0203
NL (Ours)	1,984	2,134	2,093	1,965	1,849	1,858
NL (Ortega)	300,051	319718	334913	365767	340977	314050
NL (Quotient)	0.0066	0.0067	0.0062	0.0054	0.0054	0.0059
PL (Ours)	2,234	2,339	2,281	1,997	2,017	2,041
PL (Ortega)	349,181	329,800	319,368	359,047	344,707	359,795
PL (Quotient)	0.0064	0.0071	0.0071	0.0056	0.0059	0.0057
PT (Ours)	2,255	2,706	2,320	1,948	1,790	1,925
PT (Ortega)	342,490	379,966	355,995	390,200	337,589	164,836
PT (Quotient)	0.0066	0.0071	0.0065	0.0050	0.0053	0.0117
RU (Ours)	4,549	6,425	6,163	4,429	4,497	4,445
RU (Ortega)	625,538	652,256	625,536	686,849	700,084	677,512
RU (Quotient)	0.0073	0.0099	0.0099	0.0064	0.0064	0.0066

Table A.6: Comparison between the number of edits on articles corresponding to all the considered Wikipedias obtained from our results (Rows heading by 'Ours') for July till December 2009 and the same number of operations reported by Ortegas's tool *WikiXRay* (Rows indicated with 'Ortega') for the same period. Both data correspond to articles in the main namespace. Rows headed by 'Quotient' correspond to the quotient between the two measures.

Appendix B

Glossary

Article : Every entry in a wiki-based platform containing encyclopaedic information about a particular subject, event, person, date, etc. Articles may contain text, formulae and media content such as images, music or videos. Moreover, articles may link to other related ones in the platform or, even, to external pages and resources. Articles are stored in a basic markup language called *wikitext* and they are rendered to common HTML pages when they are requested. Article's titles consist of two parts separated by a colon :, a prefix or *namespace* of the article and the article's title properly said.

NameSpace : Articles are grouped under different namespaces which are used to organize them according to their content, functionality or purpose. Unless the *main namespace* which have no prefix the rest of namespace add their name as a prefix to the article's title (like *Wikipedia:NameSpace*).

Main NameSpace : Visited articles requested when browsing the Wikipedia are usually in the *main namespace* that is the namespace in which articles are created by default.

Talk NameSpaces : Every article in any namespace has a page intended to receive the discussion issues about the article's content. In this way discussion pages of a given namespace add the "*_Talk*" clause to the namespace name (such as *User_Talk:* whereas discussion pages corresponding to articles in the *main namespace* just add the prefix "*_Talk*" to the article's title.

User NameSpace : Every registered user is provided with a page to publish personal information and for message exchanging with other users. The *user namespace* put together all these pages.

Special NameSpace : Common article correspond to static content stored in a database. However there are articles whose content is dynamically created as a result of users' demand. All these articles are grouped in the *Special namespace* and include search operations, articles linking to a given one, etc...

Visit or pageview : Request for the content of a certain article. Although it may refer to an article in any namespace, when browsing Wikipedia, users usually request articles in the *main namespace*.

Edit or save request : Contribution or modification performed over the contents of an article and which result in an write operation issued to the database server.

Edit request : Request for modifying the contents of an article. It is issued by following the "edit" tab in an article's page and, as a result, the users receives the content of the article inside a basic editor that allow to perform the desired corrections.

Submit request : Request for previewing the result of the changes introduced after an edit request or to highlight the changes introduced in comparison with the current version of the article. In any case, a submit request does not involve a write operation into the database but just the web server to render the HTML code.

History request : A request to obtain the list chronologically ordered with all the edidions performed over a given article.

Search request : Request to the Wikipedia's own search engine to look for the articles containing in their titles or in their contents the a certain topic.

Featured Articles (FA) : Article considered as the best quality ones all over the Wikipedia. Features articles must meet a set of demanding criteria to deserve the promotion to this state. Prior to their nomination as candidates for featured status, articles are encouraged to pass a peer reviewing process to improve their quality. Once they have been nominated, editor and reviewers must reach a consensus about the promotion of the article to the featured status. Otherwise, the nomination will be archived. After being considered as featured, articles may lose their status if quality lack or featured criteria mismatch is observed. A two-step process is then started and, again, a consensus about the demotion of the demotion of the article has to be reached. If not, the article will remain considered as featured.

FLOSS (Free, Libre, Open Source Software) : Term to refer to *free software* according to the Free Software Foundation definition as well as to the Open Source Inititiave manifest about *open source software*.

GNU R : Statistical software package released under the GNU GPL license which offers a large number of functionalities for statistical analysis (available at [1]).

[1] `http://lib.stat.cmu.edu/R/CRAN/`

Appendix C

Resumen en español

C.1 Introducción

El enfoque basado en la colaboración y cooperación de una comunidad de miembros ha demostrado ser eficaz y altamente eficiente cuando se ha aplicado a la consecución de objetivos concretos o a la resolución de un problemas. Así, su aplicación en áreas específicas, como el desarrollo de aplicaciones software, ha proporcionado notables avances y ha permitido obtener resultados de gran calidad y aceptación por parte de sus destinatarios finales. En relación con la gestión del conocimiento, este nuevo paradigma ha supuesto una absoluta revolución tanto en la producción del mismo como en su divulgación y transmisión. El esquema tradicional donde el conocimiento emana de un conjunto muy concreto de fuentes de autoridad reconocida se ve ahora alterado por un nuevo modelo que persigue involucrar a cualquier usuario en la construcción y revisión de dicho conocimiento. Más aún, se promueve y invita continuamente a toda la comunidad a contribuir al proyecto con sus aportaciones. Ello sin considerar en ningún momento la pertenencia de sus miembros a instituciones o esferas tradicionalmente relacionadas con el saber o con alguna rama de este en particular. Además, se espera que los miembros contribuyan de forma completamente voluntaria y desinteresada, lo que supone un extraordinario aliciente a la hora de observar y examinar el resultado final de una obra construida bajo tales preceptos.

Sin duda, este nuevo esquema de producción del conocimiento se ha visto ampliamente respaldado por el soporte ofrecido por las herramientas y servicios desarrollados en el ámbito de las tecnologías de la información y las comunicaciones. Así pues, los nuevos métodos de acceso y gestión de la información se han implementado bajo novedosas formas de interacción entre los usuarios y los sistemas desplegados para recoger y poner el conocimiento a disposición de toda la comunidad. Es en este punto donde herramientas como *blogs*, *wikis* y otras comúnmente relacionadas con el término Web 2.0 incorporan su funcionalidad al escenario actual de la gestación y divulgación del saber. En concreto el enfoque “wiki” de producción intelectual, además de perseguir que los usuarios se involucren en la generación de los contenidos ofrecidos, promueve la facilidad y sencillez de los mecanismos de acceso y contribución que normalmente se articulan en torno al concepto de plataforma. De esta manera, el usuario final sólo precisará, para toda interacción con la plataforma de contenidos, de un navegador web común. Por otro lado, cualquier compendio de conocimiento debe consistir en un conjunto estructurado de unidades básicas de información. Las unidades estructurales de las plataformas “wiki” son los denominados *artículos* que se relacionan y enlazan entre sí a través de vínculos que imitan los hiper-enlaces característicos del lenguaje HTML. También existen otros

elementos organizativos como *namespaces* y categorías que permiten agrupar a los artículos en base a su naturaleza, su funcionalidad o el área correspondiente a los temas tratados.

Wikipedia es, en la actualidad, la plataforma más importante basada en un motor *wiki* y sirve como herramienta eficaz para la creación y difusión del conocimiento en cualquiera de sus áreas dado su carácter enciclopédico. Wikipedia es mantenida, junto con otros proyectos también basados en el esquema ”wiki“, por una organización con fines no lucrativos denominada *Fundación Wikimedia* y consta de más de 250 ediciones cada una correspondiente a un idioma determinado. Wikipedia ofrece recursos de información en gran cantidad de formatos con el fin de poner a su disposición de sus usuarios una herramienta de referencia más rica y diversa. Wikipedia utiliza el concepto de edición para agrupar a los distintos artículos escritos en cada idioma. La facilidad en el acceso a la información presentada y el extraordinario compromiso de su comunidad de usuarios por la calidad de la misma han hecho que Wikipedia adquiera la dimensión y el éxito del que actualmente goza. El crecimiento de Wikipedia jamás se ha detenido desde sus comienzos, al igual que su popularidad que ha situado su portal dentro de las diez páginas más visitadas en Internet. Un éxito de esta magnitud ha propiciado que Wikipedia transcienda rápidamente de entornos típicamente académicos y adquiera la categoría de fenómeno de masas.

Sin embargo también hay lugar para la controversia. El carácter abierto de la Enciclopedia on-line, la ausencia del respaldo de algún tipo de autoridad en materia de conocimiento que actúe como garante de la información presentada y la posibilidad de opiniones sesgadas o, más aún, auténtico vandalismo informativos constituyen las principales amenazas y también los principales argumentos esgrimidos por los detractores de Wikipedia para desaconsejar la consideración de sus contenidos.

Quizá una de las cuestiones más interesantes relacionadas con Wikipedia es su contribución a la difusión del paradigma “Wiki” como mecanismo de utilidad para la compartición e intercambios de información. De hecho, un gran número de organizaciones, tanto institucionales como corporativas, y de comunidades en general lo han adoptado y han puesto en marcha portales *wiki* destinados a la publicación y gestión de sus activos de información.

C.2 Antecedentes

Debido a la dimensión adquirida de fenómeno de masas y a la extraordinaria importancia derivada de su uso masivo como herramienta de consulta, Wikipedia se ha revelado como un tema de gran interés para la comunidad científica. Sin embargo, la mayor parte de la investigación realizada hasta la fecha se ha centrado en aspectos relacionados con la calidad y fiabilidad de los contenidos ofrecidos y en el grado de reputación y confiabilidad de sus autores y colaboradores. Además, la cuestión relativa a su crecimiento y tendencia evolutiva ha atraído a un buen número de investigadores. Por el contrario, nuestro interés se aleja de estos esquemas y pretende centrarse en la forma en que los usuarios hacen uso de Wikipedia.

Los resultados de la aplicación de enfoques basados en la cooperación de comunidades de individuos en proyectos e iniciativas consideradas de interés general han sido ya ampliamente tratados y discutidos por muchos investigadores y desde un número considerable de perspectivas ([NKCM90], [DB92], [CH03] or [Sur04]). En relación a la gestión del conocimiento, el nuevo modelo de producción distribuida de la información no contempla el respaldo de fuentes centralizadas de reconocida autoridad sino, más bien, la participación colectiva de toda la comunidad ([Ben06]). En este sentido, esta concepción descentralizada de la génesis del conocimiento supuso una ruptura con el esquema tradicional y constituyó una auténtica revolución en la esfera de la producción intelectual y en el acceso a las fuentes de información. Diversos autores aplicaron al nuevo e incipiente paradigma

el, tan luego recurrido término, de *Inteligencia abierta o colectiva* [SH02].

El enfoque descrito de producción distribuida del conocimiento requería de herramientas eficaces para su implementación y encontró en las nuevas tecnologías de la información y las comunicaciones el soporte ideal para su articulación. Conceptos como *blogs* y *wikis* aparecieron como instrumentalizaciones concretas del concepto más amplio de Web 2.0 [O'R05] que pretendía otorgar a los usuarios un papel mucho más activo en la construcción de los contenidos ofrecidos en los portales asociados. Así apareció Wikipedia como uno de estos portales destinados a recoger y ofrecer las aportaciones colectivas recibidas. Su posterior expansión y vertiginoso crecimiento convirtieron sus datos en objeto de interés por parte de los investigadores que demandaban información sobre sus distintos parámetros. Así, aparecieron diversas iniciativas, tanto en el ámbito académico como fuera de él, dirigidas a proporcionar información, eminentemente cuantitativa, sobre aspectos como el número de accesos, usuarios o artículos, el número de ediciones realizadas o el tamaño de las contribuciones aportadas. Muchas de estas iniciativas continúan activas y algunas resultan especialmente interesantes como las basadas en datos suministrados por la propia Wikimedia Foundation. Los datos que proporcionan estas fuentes pueden considerarse de confianza y constituyen un elemento fundamental para realizar comparaciones que permitan validar los resultados obtenidos por cualquier análisis. Desafortunadamente, muchas de estas iniciativas se encuentran, en el momento de escribir esta tesis, desactualizadas y sin mantenimiento alguno.

Los portales *wiki* en general y Wikipedia, en particular, han sido objeto de numerosos estudios, like [DBWS06], especialmente preocupados por establecer el nivel de calidad de sus contenidos. Las técnicas utilizadas para este fin incluyen desde medidas de centralidad entre artículos [KNP+06] hasta comparación de contenidos con enciclopedias tradicionales [Gil05] pasando por métricas basadas en el número de errores [LKSY07], de contribuciones [WH07b] o de referencias constrastadas [Nie07]. Otros aspectos recurrentes en la investigación previa sobre Wikipedia incluyen la determinación de la reputación de los autores [AdA07] y el estudio de las tendencias de evolución tanto de la Enciclopedia en su conjunto como de sus distintas ediciones [CSC+06] [ZBvD06]. La relación entre Wikipedia y otras iniciativas relacionadas con la recuperación y categorización de la información, como la *Web semántica*, también han sido objeto de estudio por parte de los investigadores, [SP06] and [GM07].

Considerando que el principal objetivo de esta tesis es el de determinar patrones temporales y de comportamiento que ayuden a describir el uso que las distintas comunidades de usuarios hacen de Wikipedia, se han revisado los estudios y experiencias realizadas en la misma línea. En este sentido cabe destacarse que la mayoría han consistido en encuestas realizadas sobre grupos con poblaciones muy concretas y normalmente pertenecientes al ámbito académico, [Kon], [Sch08], [Wat07] or [Wil07]. Nuestro enfoque, sin embargo, se aleja radicalmente de este tipo de análisis tanto en la población objeto de estudio como en la metodología de realización del mismo. Así, esta tesis se basa en el análisis de las peticiones que los usuarios envían a Wikipedia a través de la caracterización del tráfico dirigido a sus servidores de soporte. Esta línea de trabajo hasta ahora apenas si ha sido desarrollada por lo que son muy escasos los trabajos relacionados que pueden citarse. Sí existen, en cambio, múltiples estudios basados en el análisis de peticiones y solicitudes de usuarios, normalmente registradas en archivos de bitácora especiales, que tienen por objeto determinar la adecuación de los contenidos y servicios ofrecidos desde determinados sistemas.

C.3 Objetivos

El principal objetivo de esta tesis es el estudio de patrones temporales y de comportamiento en la interacción habitual entre Wikipedia y sus usuarios. Así pues, se persigue analizar, tanto

cuantitativamente como cualitativamente, aspectos relacionados con el uso dado a la Enciclopedia por parte de sus usuarios.

El enfoque utilizado resulta novedoso tanto por los datos en los que se basa como por los resultados que permite obtener y consiste, básicamente, en la caracterización del tráfico formado por las peticiones que los usuarios envían a Wikipedia. De esta forma, el primer objetivo perseguido es la validación del propio enfoque como metodología de análisis para lo que se han comparado y contrastado algunos de los resultados obtenidos con los proporcionados por fuentes consideradas de confianza.

Por otro lado, los resultados de un análisis como el descrito pueden ayudar a conocer la naturaleza de las peticiones a las que los sistemas de soporte de Wikipedia tienen que dar respuesta y pueden resultar en mejoras para aumentar el rendimiento, escalabilidad y capacidad operativa de los mismos.

En concreto se pretende ofrecer una respuesta adecuada a diversas preguntas de investigación que se explican a continuación. En primer lugar, el análisis macroscópico del tráfico a Wikipedia persigue caracterizar las distintas peticiones que forman parte de él y sus respectivas proporciones. En este sentido, el objetivo perseguido es claramente la determinación de la composición del tráfico dirigido a Wikipedia. Específicamente las preguntas relacionadas con este aspecto serían:

1. **¿Es posible caracterizar las peticiones que forman el tráfico dirigido a las distintas ediciones de Wikipedia ?**
   Para responder a esta cuestión se ha analizado el tráfico dirigido a cada edición de Wikipedia utilizando expresiones regulares. De esta forma se ha podido determinar la proporción de las distintas peticiones y, en particular, de aquellas que consisten en visitas o ediciones a los correspondientes artículos. Además, también se han cuantificado las que solicitan algún tipo de acción sobre los artículos o se remiten como parte de una operación de búsqueda. Finalmente, las peticiones que involucran elementos de personalización y visualización, como *"skins"* y estilos css también han sido tenidas en cuenta.

2. **¿Existe una relación de proporción entre el número de artículos de cada edición de Wikipedia y el tráfico que recibe?**
   La respuesta a esta pregunta incluye la comparación del tamaño de cada edición, expresado en número de artículos, con la cantidad de tráfico dirigido a ella. Además, se ha analizado la evolución de ambas medidas, tamaño y tráfico, durante todo el año.

   A continuación, basaremos nuestro examen en las peticiones ya filtradas por nuestra propia aplicación. Estas peticiones se refieren a elementos de información específicos (fundamentalmente determinados namespaces) y a acciones cuya cuantificación y análisis entra dentro de nuestros intereses. Nuestro estudio, aquí, se centra en aspectos temporales y de comportamiento que puedan extraerse del tráfico y que resulten de utilidad en la descripción de la interacción entre Wikipedia y sus usuarios. En concreto, las preguntas propuestas serían:

3. **¿Es posible identificar patrones repetidos en el tiempo que impliquen determinados tipos de peticiones a Wikipedia?**
   Para ofrecer una respuesta adecuada a esta pregunta, se analizarán las peticiones realizadas a Wikipedia durante diferentes unidades de tiempo. Esto permitirá obtener distintas perspectivas correspondientes a los diferentes períodos considerados. Para obtener una mayor precisión, se analiza separadamente cada tipo de peticiones con el fin de evitar efector colaterales derivadas de las diferencias en escala. Por la misma razón, las peticiones correspondientes a cada edición de Wikipedia se tratarán por separado.

4. **¿Están las visitas a los contenidos de Wikipedia relacionadas con las ediciones y los otros tipos de peticiones de alguna manera?**
   Esta pregunta se responderá poniendo en relación el número de peticiones de cada tipo lanzadas en períodos de tiempo similares de manera que puedan observarse correlaciones entre ellas. Las relaciones entre algunos tipos de peticiones pondrían de manifiesto hábitos concretos de conducta por parte de los usuarios cuando interactúan con Wikipedia. Además, este tipo de comparaciones puede servir para distribuir las contribuciones enviadas a las distintas ediciones entre sus respectivos usuarios y también conducir a la determinación del grado de participación correspondiente a las distintas comunidades de usuarios.

   Finalmente, nos centramos en el tráfico dirigido a contenidos concretos y muy particulares. Wikipedia establece distintos mecanismos para promover y presentar contenidos considerados de una calidad excepcional y nosotros evaluamos su efectividad en relación con el tráfico que consiguen atraer. Por otro lado, nos interesa conocer que tipo de artículos reciben un mayor número de visitas y si son los mismos en las distintas ediciones de Wikipedia. Además, Wikipedia también ofrece un motor de búsqueda integrado que nos interesa desde el punto de vista del estudio de los tipos de contenidos correspondientes a las operaciones de búsqueda solicitadas por los usuarios. Las siguientes cuestiones reflejan estas inquietudes de investigación:

5. **¿Cómo afecta la consideración de artículos como contenido destacado en el número de visitas que reciben?**
   Esta cuestión se considera desde una doble perspectiva. Por un lado, se analiza el impacto, en términos del número de visitas que atraen, de los artículos destacados que se presentan en las páginas principales de las distintas ediciones de Wikipedia como ejemplos de contenidos de calidad. Además, se analiza también el número de visitas que atraen los artículos candidatos a contenido destacado durante su proceso de promoción. Las visitas a estos artículos pueden servir para interpretar la dinámica que sigue cada comunidad de usuario durante la búsqueda del consenso necesario para otorgar a los artículos la consideración de contenido de calidad. Un número elevado de visitas a artículos destacados puede ser un indicio del interés de una determinada comunidad de usuarios por artículos de gran calidad y, por tanto, su relación de uso con Wikipedia no respondería a la forma de mera consulta o búsqueda de información. Las visitas a artículos destacados mostrados en la página principal de alguna edición de Wikipedia implican la visita previa a estas páginas y, por tanto, tienen una probabilidad considerablemente menor de ser el resultado de una operación de búsqueda realizada desde un motor externo o del propio motor de Wikipedia. Por tanto, la visita a estos artículos con toda probabilidad es el resultado de cautivar la atención del usuario al paso de éste por la página principal. Por supuesto, se considera de un interés especial el poder determinar si la inclusión de artículos destacado en sus páginas principales tiene la misma repercusión en todas las ediciones de Wikipedia.

6. **¿Qué tipo de contenidos son los más visitados en Wikipedia?**
   Esta pregunta no tiene un carácter marcadamente cuantitativo como las anteriores sino más bien cualitativo y pretende determinar que artículos de cada edición de Wikipedia atraen más la atención de sus usuarios en función del tipo de contenido desarrollado. Así mismo, también se analizará el tipo de artículos que recibe mayores tasas de contribución. Los dos resultados pueden servir como indicadores del tipo de uso que las diferentes comunidades de usuarios hacen de Wikipedia. Las categorías de artículos consideradas para dar respuesta a esta pregunta se basan en las presentadas en el estudio conducido por Spoerry en [Spo07b].

7. **¿Influyen las operaciones de búsqueda sobre determinados temas en las visitas a los artículos relacionados con dichos temas?**
   Esta pregunta es, nuevamente, de naturaleza cualitativa y pretende determinar y categorizar, en primer lugar, las categorías de artículos sobre las que se realiza un mayor número de operaciones de búsqueda. Para ello se empleará la misma categorización utilizada para resolver la pregunta anterior. Para determinar la influencia de las operaciones de búsqueda en las subsiguientes visitas a los correspondientes artículos se correlarán los dos tipos de peticiones.

C.4 Metodología

El análisis descrito a lo largo de esta tesis consiste, básicamente, en la caracterización de las solicitudes que los usuarios de Wikipedia envían a ésta. Para ello contamos con una muestra consistente en el 1% de todas las peticiones servidas por los sistemas Squid que la Fundación Wikimedia ha dispuesto con el fin de actuar como caché de las páginas más solicitadas y aliviar, así, la carga de trabajo de los servidores web y de bases de datos situados detrás de ellos. Por cada petición que sirven, los servidores Squid registran distintos datos relacionados con ella. La información relativa a cada petición queda finalmente reflejada en una línea de *log* cuyos campos se establecen con arreglo al formato de registro utilizado por la Fundación Wikimedia. Estas líneas, una vez despojadas de cualquier información susceptible de ser utilizada para practicar alguna forma de identificación de los usuarios que las originaron, son puestas en paquetes y enviadas hasta nuestros sistemas donde quedan almacenadas para su posterior análisis. A partir de esta información se procede a la caracterización de las peticiones mediante un proceso que consiste en la obtención y filtrado de los diversos elementos de información contenidos en los distintos campos de cada línea de log y, particularmente, en el relativo a la URL enviada a la Wikipedia. El proceso de filtrado es necesario debido al ingente volumen de información a procesar y se lleva a cabo con el fin de obtener únicamente aquellos elementos de interés considerados de interés para el análisis. En nuestro caso, tales elementos consistirán en las peticiones enviadas al proyecto Wikipedia (los servidores Squid registran peticiones enviadas a todos los proyectos de la Fundación Wikimedia) y dentro de éstas, aquellas dirigidas a sus ediciones más importantes en volumen tanto de artículos como de tráfico. Además, se considerarán sólo las que involucren a los *namespaces* y acciones más comunes. La información de todas estas peticiones quedará almacenada en una base de datos disponible para un posterior análisis estadístico. Aunque sólo la información de las peticiones consideradas de interés quede almacenada en la base de datos, todo el tráfico general es caracterizado de manera que podemos obtener una apreciación muy exacta de su composición. Todas estas actividades relacionadas con el proceso de la información recibida de la Fundación Wikimedia son llevadas a cabo por parte de la aplicación *WikiSquilter* diseãda y desarrollada ex-profeso para esta misión. La ingeniería de software utilizada para su proceso de desarrollo otorga una gran importancia tanto a las cuestiones relativas al rendimiento como a la modularización y ausencia de dependencias entre sus partes. Además, se ha prestado una especial atención a la flexibilidad y extensibilidad que permiten la adición de nuevos servicios y funcionalidades de manera sencilla y eficiente. Finalmente cabe destacar las facilidades que introduce para la configuración y especificación de los elementos información considerados de interés para cada análisis. Esto la convierte en una herramienta de gran versatilidad fácilmente adaptable para analizar información de log procedente de cualquier plataforma basada en un motor *wiki*, en general, y de los proyectos actualmente soportados por la Fundación Wikimedia en particular.

Después de analizar el tráfico correspondiente a un aõ completo, el presente estudio muestra diversos patrones correspondientes a la distribución temporal de las peticiones enviadas a Wikipedia

por sus usuarios. Además, este estudio también presenta patrones que describen la manera en la que los usuarios interactúan con Wikipedia y el tipo y frecuencia de las acciones que le solicitan. Además, esta tesis analiza la relación entre el número de visitas y las operaciones de edición sobre artículos de distintas ediciones de Wikipedia con el fin de determinar el grado de participación y comportamiento colaborativo exhibido por sus usuarios. Se analiza, además, la influencia de las características de los artículos en el número y tipo de visitas que reciben y en las acciones de que son objeto. En este sentido se considera, por ejemplo, la distribución de visitas y ediciones a los artículos en función del espacio organizativo (*namespace*) al que pertenecen o la distribución de las distintas acciones en torno a estos espacios. La influencia de la calidad de los contenidos de Wikipedia en las visitas y ediciones recibidas también es tenida en cuenta. Así, se estudia el impacto de la promoción de artículos a la consideración de *destacados* en su posterior número y tipo de accesos. Otra cuestión de gran interés tratada en esta tesis es la categorización de los artículos más solicitados en las distintas ediciones Wikipedias. Este aspecto, sin duda, ofrecerá una visión cualitativa del tipo de contenido más solicitado por los usuarios y, por tanto, contribuirá a establecer un perfil del uso que se hace de Wikipedia. En relación con esta cuestión, este trabajo es el primero en considerar el uso de Wikipedia como motor de búsqueda de forma que, además, de una clasificación cualitativa de los elementos buscados, se analiza su influencia sobre las visitas a los contenidos.

La consideración cuantitativa de los datos presentados en esta tesis puede contribuir a la estimación de la carga de proceso impuesta a los servidores que soportan tanto el proyecto Wikipedia como el resto de proyectos mantenidos por la Fundación Wikimedia, así como ser de utilidad en la evaluación de la escalabilidad y rendimiento de la arquitectura de soporte en su conjunto. Por tanto, este tipo de análisis puede dar lugar a diversas mejoras en aspectos relacionados con sistemas tanto software como hardware.

Hasta el momento, y que nosotros conozcamos, no se ha realizado ningún otro análisis tan pormenorizado sobre el uso de Wikipedia ni que considere los elementos de información utilizados en el que se presenta aquí. Esperamos que nuestros esfuerzos y resultados sirvan como contribución en el estudio de las dinámicas de uso e interacción entre usuarios y plataformas relacionadas con la gestión colaborativa del conocimiento como Wikipedia.

C.5 Conclusiones

El desarrollo de la presente tesis ha permitido obtener un conjunto de conclusiones relacionadas con los objetivos y preguntas de investigación planteados que se exponen a continuación:

- En primer lugar se han validado los resultados obtenidos a partir del estudio desarrollado como parte de esta tesis y que se basa en el análisis de las peticiones realizadas a Wikipedia por sus usuarios. Este análisis resulta novedoso tanto por la naturaleza de la muestra de datos utilizada como por los resultados que permite obtener. La validación ha resultado posible gracias a la disponibilidad de fuentes de datos de fidedignas y sus resultados han mostrado la fiabilidad del análisis tanto en ámbitos marcadamente generalistas como los relativos a ediciones o a contenidos como en los mayor nivel de detalle relativos a artículos o acciones concretos.

- Atendiendo a los resultados del proceso de validación, es posible, además, concluir que la mayor parte de las visitas a artículos de la Wikipedia corresponden a los *namespaces* considerados en esta tesis: *Main*, *Talk*, *User*, *User_talk* and *Special*. En el caso de las operaciones de edición, los correspondientes resultados permiten asegurar aún con más seguridad que tales operaciones sólo involucran a los *namespaces* mencionados.

- Como resultado del proceso de caracterización del tráfico de peticiones a Wikipedia se ha determinado que las visitas a artículos constituyen aproximadamente la cuarta parte de todo el tráfico a Wikipedia. Las peticiones que solicitan realizar algún tipo de acción alcanzan otro 25% y destaca la baja proporción de operaciones de edición. Por el contrario, las operaciones de búsqueda son las más demandadas con una tasa cercana al 10%. Destaca el número de peticiones relacionadas con opciones de presentacián y visualización de los contenidos que suponen aproximadamente un 35% de todo el tráfico.

- Tras comprar el tamaño de las distintas ediciones de Wikipedia con el tráfico que atraen, podemos concluir que mayores volúmenes de artículos no significan necesariamente mayores volúmenes de tráfico. Esto significa que los recursos relacionados con el almacenamiento y el servicio de contenidos escalan de forma completamente distinta.

- El estudio de los patrones temporales ha revelado, en primer lugar, que el tráfico consistente en las peticiones filtradas para el análisis realizado en esta tesis puede servir de modelo del tráfico general a Wikipedia. Las peticiones filtradas son aquellas que involucran los *namespaces* anteriores en peticiones de visita, edición, búsqueda, solicitud de edición, consulta de histórico y visualización de cambios introducidos. Además, se ha comprobado que sólo visitas y operaciones de búsqueda siguen patrones regulares en el tiempo mientras que el resto de peticiones tiene una naturaleza mucho más espúrea.

- En relación con el comportamiento de los usuarios, se ha podido comprobar que un gran número de solicitudes de edición no terminan con la correspondiente operación de escritura en la base de datos. Esto significa que los usuarios en algún momento deciden abandonar el proceso de edición iniciado con la correspondiente solicitud. En este sentido, hemos obtenido una clasificación con las tasas de abandono de operaciones de edición en las distintas Wikipedias. Por el contrario, se ha comprobado que en la mayoría de las ediciones, las peticiones de visualización de cambios y edición son muy similares en número lo que índica un uso generalizado de la primera antes de realizar la segunda.

- La correlación de ediciones y visitas ha mostrado que éstas sólo se relacionan positivamente en algunas Wikipedias. Las mismas que tienen una correlación positiva de solicitudes de edición y ediciones finalizadas (con escritura en la base de datos). Estas ediciones son la alemana, inglesa, española, italiana y rusa. La correlación entre visitas y las distintas acciones es positiva en todos los casos y para todas las Wikipedias.

- La evaluación del impacto de los contenidos destacados ha permitido determinar que los artículos presentados durante períodos concretos de tiempo en las páginas principales de las distintas ediciones, como ejemplos de contenidos de calidad, atraen de forma segura la atención de los visitantes en dicho período sólo en el caso de la Wikipedia inglesa. Por otro lado, el análisis de las visitas a los artículos que reciben la consideración de destacados ha puesto de manifiesto las distintas dinámicas empleadas por las respectivas comunidades en la búsqueda de consenso para la promoción de los artículos.

- Se han asignado categorías a los artículos más visitados y editados en las distintas ediciones de Wikipedia. Como resultado, en la Wikipedia inglesa la categoría más visitada corresponde a artículos relacionados con el entretenimiento y el ocio mientras que en la española corresponden a Ciencia y Humanidades. También se han categorizado los temas relacionados con las operaciones de búsqueda remitidas a Wikipedia. Destaca la abundante de cantidad de búsquedas

relacionadas con contenidos de ocio, sobre todo en edición inglesa donde predominan. La edición española, sin embargo, realiza más búsquedas de temas geográficos que de ningún otro. Cuando se ha realizado la correlación entre las búsquedas sobre determinados temas y el número de visitas a artículos de los mismos temas, se ha encontrado que sólo es positiva en el caso de la Wikipedia inglesa y alemana.

Appendix D

License Creative Commons Attribution-ShareAlike 3.0

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE (”CCPL” OR ”LICENSE”). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

   (a) ”Adaptation” means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image (”synching”) will be considered an Adaptation for the purpose of this License.

   (b) ”Collection” means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are

assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined below) for the purposes of this License.

(c) ”Creative Commons Compatible License” means a license that is listed at http://creativecommons.org/compatiblelicenses that has been approved by Creative Commons as being essentially equivalent to this License, including, at a minimum, because that license: (i) contains terms that have the same purpose, meaning and effect as the License Elements of this License; and, (ii) explicitly permits the relicensing of adaptations of works made available under that license under this License or a Creative Commons jurisdiction license with the same License Elements as this License.

(d) ”Distribute” means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.

(e) ”License Elements” means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, ShareAlike.

(f) ”Licensor” means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

(g) ”Original Author” means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

(h) ”Work” means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

(i) ”You” means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

(j) ”Publicly Perform” means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place

individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

(k) ”Reproduce” means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

   (a) to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;

   (b) to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked ”The original work was translated from English to Spanish,” or a modification could indicate ”The original work has been modified.”;

   (c) to Distribute and Publicly Perform the Work including as incorporated in Collections; and,

   (d) to Distribute and Publicly Perform Adaptations.

   (e) For the avoidance of doubt:

      i. Non-waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

      ii. Waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor waives the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; and,

      iii. Voluntary License Schemes. The Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License.

   The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved.

4. Restrictions. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

(a) You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(c), as requested.

(b) You may Distribute or Publicly Perform an Adaptation only under the terms of: (i) this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License (e.g., Attribution-ShareAlike 3.0 US)); (iv) a Creative Commons Compatible License. If you license the Adaptation under one of the licenses mentioned in (iv), you must comply with the terms of that license. If you license the Adaptation under the terms of any of the licenses mentioned in (i), (ii) or (iii) (the ”Applicable License”), you must comply with the terms of the Applicable License generally and the following provisions: (I) You must include a copy of, or the URI for, the Applicable License with every copy of each Adaptation You Distribute or Publicly Perform; (II) You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License; (III) You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform; (IV) when You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License.

(c) If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution (”Attribution Parties”) in Licensor’s copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable,

the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and (iv) , consistent with Ssection 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., ”French translation of the Work by Original Author,” or ”Screenplay based on original Work by Original Author”). The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

(d) Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author’s honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author’s honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

5. Representations, Warranties and Disclaimer

   UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

(a) This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

(b) Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

(a) Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

(b) Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.

(c) If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

(d) No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

(e) This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

(f) The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any

damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensor hereunder, it shall have all rights and obligations of Licensor.

Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, Creative Commons does not authorize the use by either party of the trademark ”Creative Commons” or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons’ then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time. For the avoidance of doubt, this trademark restriction does not form part of the License.

Creative Commons may be contacted at `http://creativecommons.org/`.