Journal of Informetrics 15 (2021) 101188

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

System for evaluating the reliability and novelty of medical scientific papers

Isaac Martín de Diego, César González-Fernández*, Alberto Fernández-Isabel, Rubén R. Fernández, Javier Cabezas

Rey Juan Carlos University Data Science Laboratory (www.datasciencelab.es) c/ Tulipán, s/n, Móstoles, 28933, Spain

ARTICLE INFO

Keywords:
Reliability estimation
Novelty detection
Article evaluation
Knowledge graph
Medical articles

ABSTRACT

As society develops, the number of published research articles raises. Besides, the pressure to publish has been increased because the competitiveness between researchers working on similar topics. Although this increment is desirable, it leads to multiple issues. At a reader level, insurmountable barriers to keep up with the state of the art appear. From a publisher perspective, it is a very demanding task to determine which research articles are worth publishing. Automatizing these tasks appears as a core solution. In the case of readers, a previous evaluation of articles would simplify the filtering process. As for publishers, they could perform preliminary selections or estimate the reviewing effort. This paper presents *Medical Evaluator System for Scientific Interoperability* (*MESSI*) system to overcome all these issues. It is able to evaluate the novelty and reliability of health-related texts. The novelty calculation is based on previously acquired knowledge after processing more than 500,000 papers. The reliability estimation is based on the reputations of similar articles calculated based on previously defined metrics. Multiple experiments have been addressed to illustrate the viability of the proposal. The obtained results show a good performance that encourage to continue evolving the system.

1. Introduction

Over the last decades, the scientific community has experienced a continuous growth. The improvements in educational systems, the use of the Internet to search and share information, and the emergence of new actors with enormous scientific capacity are some factors that may explain this growth (Bornmann and Mutz, 2015; Naghizadeh and Naghizdeh, 2017). Consequently, the number of research articles published each year has increased two-fold in the last two decades (Guerrero-Bote and Moya-Anegón, 2012).

This vast number of research articles is pushing the edges of science, yet it poses problems for both scientific journals and readers. Regarding the former, research articles undergo a long process before publishing, whose main bottleneck is peer-reviews. These reviews usually comprehend checking novelty and trustworthiness, which are very demanding tasks (e.g., checking novelty involves keeping up with the current state of the art). With respect to the latter, readers face a similar problem. Selecting a manuscript to read involves skimming through several research articles that might be outdated (e.g., newer approaches improving the previous ones), or do not present significant contributions (e.g., approaches that are not relevant for the scientific community). Thus, the more research articles are considered, the harder these tasks become.

* Corresponding author.
E-mail addresses: isaac.martin@urjc.es (I. Martín de Diego), cesar.gonzalezf@urjc.es (C. González-Fernández), alberto.fernandez.isabel@urjc.es (A. Fernández-Isabel), ruben.rodriguez@urjc.es (R.R. Fernández), javier.cabezas@urjc.es (J. Cabezas).

https://doi.org/10.1016/j.joi.2021.101188
Received 19 November 2020; Received in revised form 17 February 2021; Accepted 18 June 2021
Available online 13 July 2021

In order to address this issue, many statistical analyses have appeared in the literature, which has led to the emergence of several bibliometric indicators (e.g., *h-index* or the *Journal Impact Factor*) Hicks and Melkers (2013). These indicators work at different levels according to the evaluated item (e.g., author, article or journal). In this way, users have available multiple approaches to evaluate quantitatively different concepts (e.g., the impact of specific journal or the scientific productivity of an author). Some of these metrics have been adopted by different academic databases (e.g., Google Scholar or Scopus), which simplify the process of searching.

This paper introduces *MESSI*, an original framework that evaluates two bibliometric indicators: novelty and reliability. The main innovation of this framework is the combination of both concepts to sort health-related texts without needed to know the authorship. This new way of sorting can facilitate the review of new research articles by experts. *MESSI* is developed as part of the Swarm Agent-Based Environment For Reputation in MEDicine (SABERMED) project funded by the Spanish Ministry of Economy and Competitiveness (MMG: MedLab Media Group, 2019). The main objective of SABERMED is to find a solution to the problem posed today by non-trusted digital content in the medical domain.

The novelty of a specific text is based on the previous knowledge that the system has about its topics. The reliability of a text is based on the reputation of previous works that address the same issue from a similar point of view. This framework allows comparing several papers at the same time by using graph-based techniques (Rajagopal et al., 2013). This feature enables readers to discard those manuscripts which are not considered as novel or reliable (saving time for readers) by the system. In the case of reviewers, the novelty and reliability help to estimate required efforts to review an article (e.g., novel papers should require more review effort as their claims may not have been intensively studied previously).

The medicine domain has been selected because three foundations. First, it is one of the most relevant application fields for mankind (More, 2016). Second, it is constantly upgraded by new scientific improvements, these tasks and automatize the process (Castiglioni, 2019). Finally, it is one of the most reconnoitered fields by the scientific community (Richta, 2018), easing the document gathering process.

Multiple experiments have been addressed with the purpose of validating the *MESSI* framework. In this regard, tests have been developed to illustrate the different use cases of the framework. The system has been trained and evaluated through a corpus of 500,000 articles.

The rest of the paper is organized as follows. Section 2 introduces relevant literature related to the proposal. Section 3 presents the *MESSI* framework detailing its different modules and components. Section 4 addresses the experiments to evaluate the performance of *MESSI*. Finally, Section 5 concludes and provides further research directions.

2. Background

This section introduces the foundations of the *MESSI* framework. It is decomposed into three main topics: metrics to evaluate research articles, knowledge graphs, and medical knowledge. The first one describes some relevant metrics for classifying research articles. The second one addresses the definition of knowledge graphs, their representations and functionalities. Finally, the third one is specifically focused on medical knowledge, showing different tools to process and storing it.

2.1. Metrics for scientific articles

Reputation systems (Resnick et al., 2000) are focused on rating items or individuals. This rating allows users or systems to evaluate how trustful or interesting an item or individual is. The combined use of these systems with bibliometric indicators has been shown useful for ranking articles, authors, and journals. For instance, the researchers can use their outcomes to find the most appropriate articles related to a specific approach, to identify potential research collaborators or to select a journal adequate for a given research paper (Okubo, 1997).

In this regard, multiples indices have been proposed, most of which are based on citations. Two well-known instances of these indices are the impact factor (Garfield, 2006) and the *h-index* (Hirsch, 2005). The former is focused on measuring the impact of a journal based on the number of citations of its articles. Likewise, the *h-index* use the citations to measure the productivity and impact of a researcher. Although citation-based metrics can be very useful, they present some limitations. For instance, they do not consider the differences between popularity and prestige (Franceschet, 2010). To avoid this issue, metrics like *f-Value* (Fragkiadaki et al., 2011) and *P-score* (Ribas et al., 2015) make use of more complex relationships between articles, citations, and authors to estimate the reputation of an article. At journal level, alternative metrics that do not use the citations directly have also been studied. For instance, a metric that calculates the impact factor of a journal based on the editorial team (Xie et al., 2019).

Another relevant instance is Unified Knowledge Compiler (UNIKO) (Fernández-Isabel et al., 2018). This system uses the Digital Object Identifier (DOI) of the research articles and the Open Researcher and Contributor ID (ORCID) (Haak et al., 2012) or the name of authors to calculate the reputation. It calculates the reputation based on the number of citations and the reputation, the seniority, and the previous publications of their authors.

Text novelty can also be used as a metric to analyze publications. It has been approached in different levels, namely sentence and document level (Ghosal et al., 2018). In both cases, these metrics measure how much information the sentence or document introduces with respect to the previous literature. Instances for the first case are based on: set of terms (Zhang et al., 2003), density of the unseen named entities (Gabrilovich et al., 2004), pattern-analysis and named entities (Li and Croft, 2005), and cosine similarity with respect to the sentences of the previous knowledge (Tsai and Zhang, 2011).

Regarding the document level approaches, they are mainly focused on comparing the topics in the documents (Yang et al., 2002). To achieve this task, they usually set metrics, geometric distance, and distributional similarity (Zhang et al., 2002), esti-

mating the novelty of the text as the average novelty of its sentences (Tsai and Zhang, 2011). Complementary, they could also include specific semantic-based methods (Kumar and Bhatia, 2020), counting the co-occurrences between specific concepts within the text (Hofstra et al., 2020), and Machine Learning (ML) methods discarding well-known hand-crafted features (e.g., bag-of-words) (Ghosal et al., 2018).

In the case of the *MESSI* framework, it estimates the reputation of scientific articles to determine the importance of the relationships between the entities extracted from their content. Regarding the novelty calculation, a new metric based on knowledge graphs is proposed at document level. This metric differs from previous approaches, as it is focused on relationships between concepts (i.e., not only unigrams). Notice that concepts are noun phrases representing a general notion or idea, in contrast to words that refer to a specific language unit. Thus, concepts address synonyms and make easier to focus on the general meaning of the sentence.

2.2. Knowledge graphs

A knowledge graph is a way to integrate information with the purpose of managing knowledge (Singhal, 2012). This knowledge is usually represented by an ontology or a knowledge base. Knowledge graphs were introduced to improve user-experience in Google searches (Ehrlinger and Wöß, 2016). There are other relevant well-known knowledge graphs, such as (Bergman, 2018) and (Bollacker et al., 2008), albeit they are significantly smaller.

Delving into the main functionalities of knowledge graphs, they have been widely used in text processing to relate concepts (Jiang et al., 2010). Thus, they can be used to establish relationships between concepts at lexical and semantic levels, or measure similarity (Zhong et al., 2002). For the first case, approaches based on text summarization have been the most typical in the domain. They produce a knowledge graph from a text to extract or organize the knowledge. Then, a simpler text embracing the main topics is generated. This text can be completely new (i.e., abstractive summarization (Liu et al., 2018)) or contain excerpts from the original (i.e., extractive summarization (Gupta and Lehal, 2010)). For the second, metrics that use the distance between nodes as a way to represent concepts are the most common approaches (Pedersen et al., 2004).

Regarding the construction of knowledge graphs, they are mostly built by combining structured and semi-structured sources (Bizer et al., 2009; Vrande and Krtzsch, 2014). For instance, information boxes, categorisation and links from *Wikipedia*. Although it is not so common, they have also been built from raw text (Hewett et al., 2002; Plake et al., 2006; Rotmensch et al., 2017) (e.g., using *PubMed* articles (Kamdar et al., 2017)).

Moreover, for the representation of knowledge graphs, they can usually be illustrated as a set of nodes with edges that establish relationships between them, or stated as sets of triplets. These triplets represent the type of relation between two concepts or individuals. For instance, the triplet (“mouse tail”, “is part of”, “mouse”), “mouse” and “mouse tail” are individuals, whereas “is part of” is the relationship between them.

On the other hand, an ontology is a specific knowledge graph type that imposes a determined structure (i.e., it is a formal definition). Thus, ontologies, in the computer science field, are defined as an organized set of concepts and individuals and the relationships between them (Dou et al., 2015). These concepts and individuals can contain attributes, describing properties and characteristics they might have, while the relationships express how they relate. Instances of well-known ontologies are: *DBpedia* (Bizer et al., 2009), a project that models *Wikipedia’s* knowledge (Leitch and Leitch, 2014) in a structured representation, *Dublin Core Ontology* (Kunze and Baker, 2007), an attribute set definition to describe digital resources, Friend of a Friend (FOAF) (Graves et al., 2007), a structure to characterize relationships between people on the Internet, and Yet Another Great Ontology (YAGO) (Suchanek et al., 2007), an ontology automatically extracted from *Wikipedia* and other sources such as *DBpedia* and *WordNet* (Miller, 1995).

In the case of the *MESSI* framework, it makes use of medical ontologies and knowledge graphs. For the first ones, they are used to detect diseases and their related medical concepts. These elements are processed and stored into a knowledge graph generating a knowledge base. For the latter, they are used to measure the reliability and novelty of the scientific articles according to graph-based metrics.

2.3. Medical knowledge

Knowledge bases have also been extensively studied in the medicine domain. *AliBaba* is a well-known approach that utilizes a knowledge base based on text gathered from *PubMed* articles (Plake et al., 2006). Concepts are extracted from existing databases (e.g., *UniProt* (Consortium et al., 2018), Medical Subject Headings (MeSH) (Baumann, 2016), National Center for Biotechnology Information (NCBI) Taxonomy (Sharma et al., 2018), *MedlinePlus* (Ahmed, 2019) and *PubMed* (Kamdar et al., 2017)) and their relations are calculated using co-occurrence. In contrast with the previous approaches, *PharmGKB* is created manually from literature (Hewett et al., 2002). This results in curated information, but limits the system ingest to human-resources. It is focused on understanding how genetic variations affect the manner in which people respond to drugs. Thus, relations between symptoms and diseases (i.e., concepts) are extracted from medical records. Then, ML models are used to determine their relation.

In the case of ontologies, one of the most common is Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) (Spackman et al., 1997). It is a collection of medical terms which includes their Concept Unique Identifier (CUI), synonyms and definitions. For instance, they include terms about diagnoses, procedures, and body parts among others. Additionally, terms are defined in different languages and translations can be easily achieved through CUIs and the structure of the ontology. Another ontology focused on medical knowledge is *Human Disease Ontology* (Kibbe et al., 2014). Its main goal is to organize human diseases based on their etiology, and it also includes related medical vocabulary.

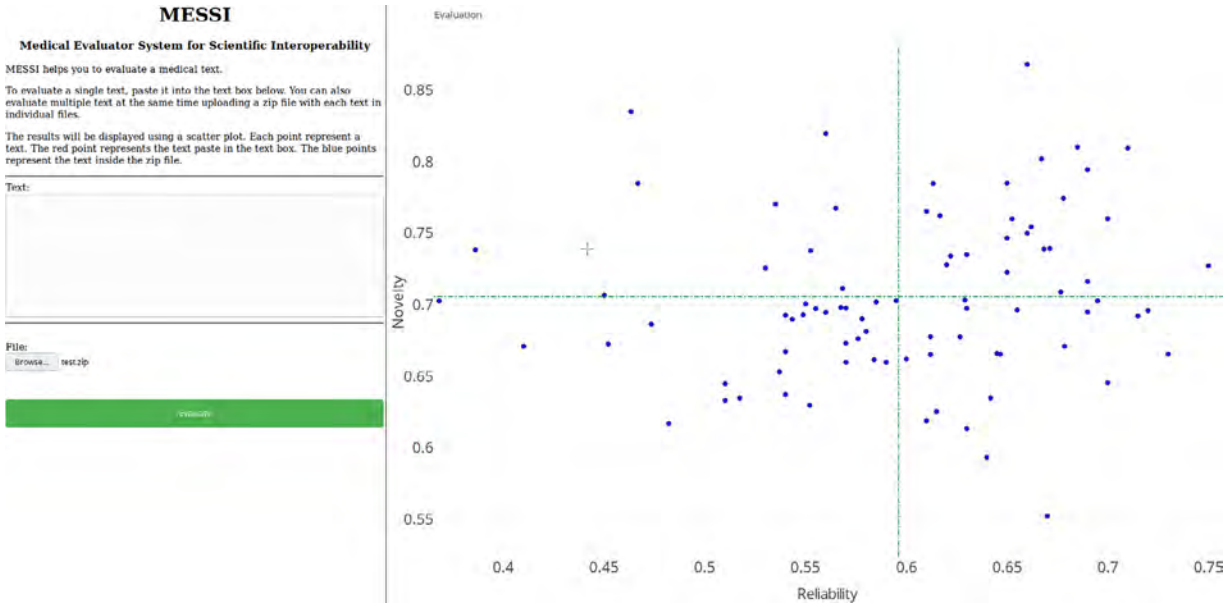

Fig. 1. The *MESSI* web portal.

Other projects like Unified Medical Language System (UMLS) (Bodenreider, 2004) group terms of ontologies such as SNOMED CT and *Human Disease Ontology*, with other medical vocabularies. Computer systems using different medical vocabularies can interoperate by providing a mapping between terms. Furthermore, UMLS has also been used to develop software to extract medical information from text. For instance, *MetaMap* (Aronson, 2001) gathers UMLS concepts from raw texts. *SemRep* (Rindflesch and Fiszman, 2003) makes use of *Metamap* to map UMLS concepts within a biomedical free text. In addition, *SemRep* supports the *ABGene* gene recognition system (Tanabe and Wilbur, 2002) that powers the identification of proteins and genes. *SemRep* also adds special functionalities such as the extraction of semantic relations (triples formed by a subject, a relation, and an object), coreferences resolution, among others. All of these characteristics make it possible to perform further analysis of the text and to find more accurate co-occurrences.

In the case of the *MESSI* framework, a knowledge base is used to store the medical knowledge. This knowledge base is developed using the entities extracted from *SemRep*. The development differs from previous approaches in the senses that all entities found by *SemRep* are included, not only the predicates (Kilicoglu et al., 2008). This decision is grounded by the fact that *MESSI* considers a relationship between concepts when they appear in the same text, and not just in the same sentence.

3. Proposed graph-based framework

This paper introduces the *MESSI* framework, a system to assist medical researchers in the scientific domain. The main purpose of *MESSI* is to evaluate medical research articles and to organize them according to their novelty and reliability. This arrangement is presented through a visualization tool that facilitates the interaction between users and the system (see Fig. 1). A release of the *MESSI* framework is available online for testing purposes.[2]

Concerning the architecture of *MESSI* (see Fig. 2), it is made up of two main modules: the *Processing* module and the *Evaluation* module. The *Processing* module is used to process research articles to populate the knowledge base. The *Evaluation* module processes new texts and estimates their novelty and reliability by using the information stored in the knowledge base.

The knowledge base consists of a knowledge graph that stores the knowledge of the system. This graph is formed by nodes that represent a medical concept or a semantic type, and edges between nodes representing those nodes that have appeared together in a research article. Further, these edges are independent for each research article, thus two nodes can be connected with more than one edge. The set of these edges is considered as the relationship between two nodes. Edges store the DOI of the research article and its estimated reputation using the UNIKO approach (Fernández-Isabel et al., 2018) as attributes. This reputation, among others, is used to estimate the reliability of the new articles that contain this same relationship between concepts, as shown in Section 3.2.3. Notice that while UNIKO estimates the reputation based on article metadata, the new proposed reliability takes into account only the concepts discussed in the text and the reputation of previous research articles that dealt with the same concepts. For instance, given two medical concepts (see Fig. 3): *Malignant Neoplams* and *colonoscopy*, they are related since they have been found together in two

[2] www.datasciencelab.es/research/projects/messi/

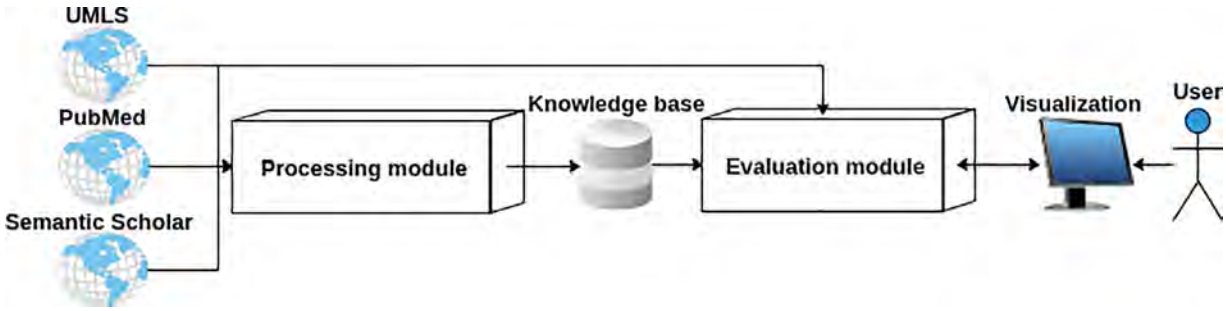

Fig. 2. Overview of *MESSI* architecture.

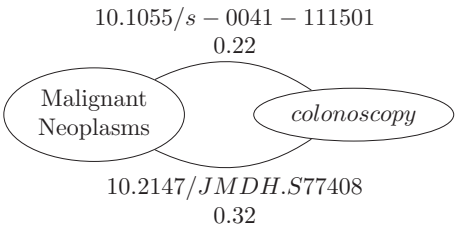

Fig. 3. Example of two medical concepts linked by two articles in the knowledge base.

different research articles with DOIs *10.1055/s-0041-111501* and *10.2147/JMDH.S77408*. The reputations of these articles are 0.32 and 0.22, respectively.

The following notation will be used throughout this paper. The knowledge base will be referred as KB. Let the KB represented as a graph defined as:

$$G(KB) = \{N^{KB}, E^{KB}, I_r^{KB}, I_d^{KB}\}, \tag{1}$$

where, N^{KB} is the set of nodes and E^{KB} is the set of edges between them, I_r^{KB} and $I_d^{KB}(i)$ are the edges and nodes intensities, respectively. Thus, $I_r^{KB}(i,j)$ is the number of edges between node i and node j, and $I_d^{KB}(i)$ is the maximum number of edges between i and one of its adjacent nodes. Let the text graph be defined as:

$$G(t) = \{N^t, E^t\}, \tag{2}$$

where, N^t is the set of nodes (i.e., medical concepts from the text) and E^t the set of edges between them. Let D be the set of diseases found in the intersection between N^{KB} and N^t, and d an instance from D. Let the subgraph of $G(KB)$ which contains the nodes in D and their neighbours (i.e., related medical concepts) defined as:

$$H(D) = \{N^{H(D)}, E^{H(D)}, I_r^{H(D)}, I_d^{H(D)}\}. \tag{3}$$

Finally, let m be a medical concept (e.g., diseases or treatments). Then

$$R(G(t), H(D)) = \{(d,m) | d \in D, m \in (N^t \cap N^{H(D)}) \wedge d \neq m\}, \tag{4}$$

is the set of relationships between the diseases in D and the intersection between N^t and N^{KB}.

The generation of the knowledge base using medical research articles using the *Processing* module is described in Section 3.1. Then, Section 3.2 details how the *Evaluation* module performs the evaluation of the text provided by the user. Finally, Section 3.3 introduces the web portal and explains the interpretation of the visualization.

3.1. Processing module

The main purpose of the *Processing* module is to create and to maintain the knowledge base. This module uses three web sources of information: *UMLS, PubMed* and *SemanticScholar* (Fricke, 2018). The first one is used by the system to extract medical concepts. The second one is consulted to obtain the abstract of research articles. Notice that the abstract usually contains the most important concepts while avoiding the noise other sections might introduce (e.g., the related work or background sections) (Tshitoyan et al., 2019). The last source of information is queried to estimate the reputation of the research articles.

Regarding the architecture of the module, it comprises four components (see Fig. 4): *Concept collector, Reputation calculator, Text graph builder* and *Knowledge integrator*.

The *Concept collector* component extracts medical concepts from abstracts by using *SemRep* (Rindflesch and Fiszman, 2003). These medical concepts are representations for a set of terms, which are different forms to refer to the same concept. For instance, *Cancer* and *Malignant Neoplasms* are terms grouped under the concept *Malignant Neoplasms*. Thus, when *SemRep* finds a term in the text, it

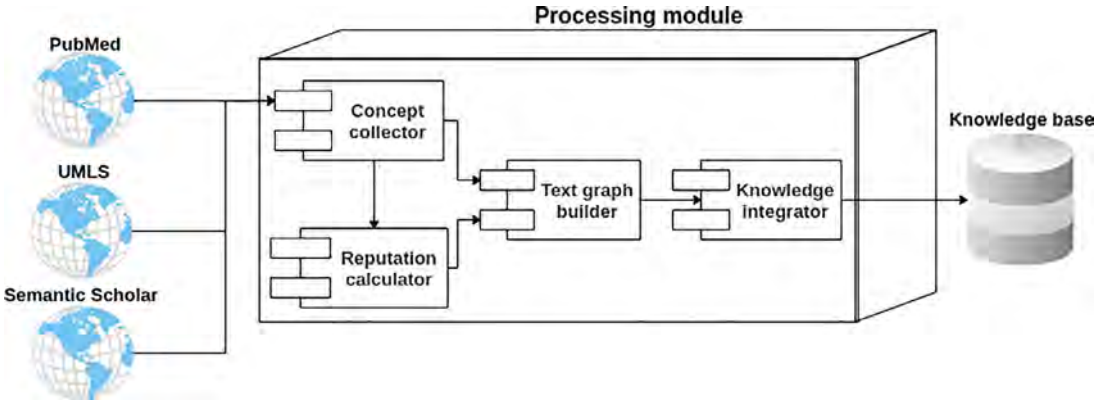

Fig. 4. Architecture of the *Processing* module.

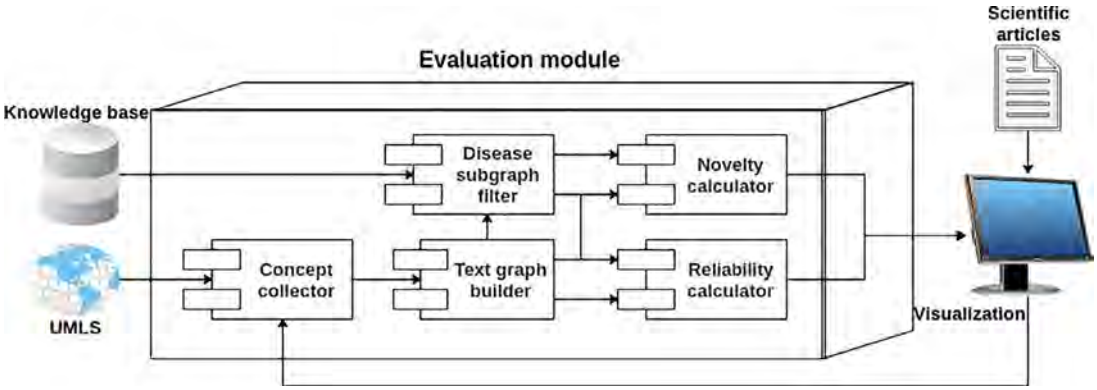

Fig. 5. Architecture of the *Evaluation* module.

returns the concept representing that term. This allows addressing problems related to the ambiguity of the language (Shen et al., 2014). The use of *SemRep* brings some benefits to the performance of this work. Tools such as *MetaMap, cTAKES* (Savova et al., 2010), *DNorm* (Leaman et al., 2013) or *MetaMap Lite* (Demner-Fushman et al., 2017) only identify medical concepts. In the case of *MetaMap* also implements a word-sense disambiguation server. On the other hand, *SemRep* makes use of *MetaMap* capabilities to identify medical concepts and extracts the relationship that binds them. These relationships are considered when the text graph is created. This consideration allows avoiding the creation of incorrect relations that affect the framework performance.

The *Reputation calculator* component estimates the reputation of the research articles. It uses the estimations made by the UNIKO framework, which implements a previously presented methodology to achieve this task (Fernández-Isabel et al., 2018). This methodology only requires the DOI of the research article. The reputation ranges from 0 to 1, where 0 means that no information about the authors was found, and bigger values imply a better reputation.

The *Text graph builder* component generates the text graph that represents the information gathered by the previous modules. This text graph is a complete graph (i.e., every pair of nodes has an edge between them) where the nodes are medical concepts, and the edges have the DOI and the calculated reputation as attributes.

Finally, the *Knowledge integrator* component updates the knowledge base with the text graph. This task consists of adding the nodes and the edges of the text graph to the knowledge base. The nodes are only added if they do not previously exist in the knowledge base. Regarding the edges, they are always added since they are defined for each research article. This process is repeated for each of the research articles.

3.2. Evaluation module

This module evaluates medical texts and calculates their novelty and reliability. The input text, in contrast with the *Processing* module, is not restricted to the abstract of research articles. This also allows processing free format texts (e.g., texts from blogs or user replies in question and answer websites).

Regarding the *Evaluation* module architecture, it is composed of five components (see Fig. 5): *Concept collector, Text graph builder, Reliability calculator, Disease subgraph filter* and *Novelty calculator*.

The *Concept collector* component is analogous to the component with the same name used by the *Processing* module.

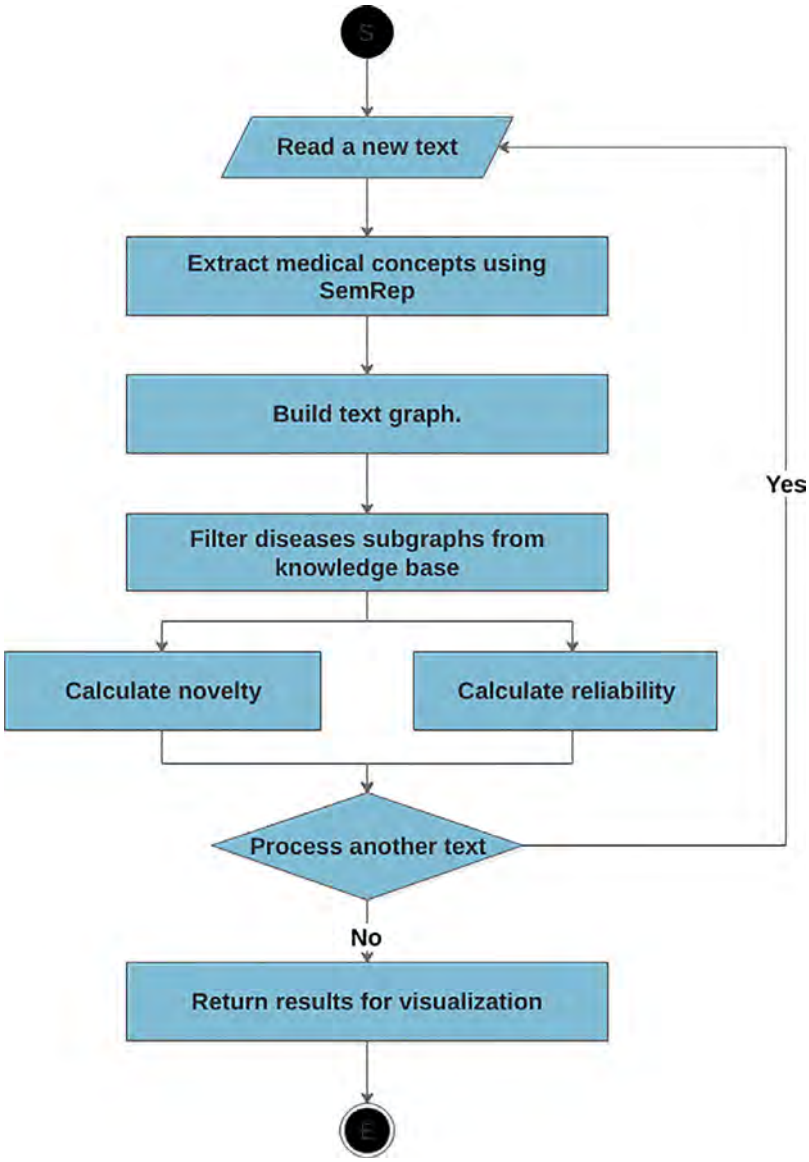

Fig. 6. Workflow of the *Evaluation* module.

The *Text graph builder* component produces a text graph by using the concepts gathered by the *Concept collector*. In this case, the text graph does not include any reputation information.

The *Disease subgraph filter* component isolates the concepts from the knowledge base that are related to the topics of the text. Thus, this component identifies the diseases that appear in the text graph and filters the subgraph associated with each disease of the knowledge base (this subgraph is formed by the disease node and its neighbors). This task enhances the evaluation of the text.

Finally, the *Reliability calculator* and *Novelty calculator* components use the concepts gathered from the text by the *Concept collector* component and the diseases subgraphs filtered by *Disease subgraph filter* component. Based on this information, these components estimate the novelty and the reliability values of the evaluated text, respectively. Novelty and reliability estimations are detailed in the next sections.

3.2.1. Evaluation module process

The workflow of the *Evaluation* module consists of seven steps (see Fig. 6), and one decision. Thus, given a set of texts, it starts reading a text to evaluate. Then, the medical concepts are extracted from it by using *SemRep*. These tasks are managed by the *Concept collector* component. After that, the text graph is generated with these concepts. The *Text graph builder* tackles this operation. Next, the subgraphs with the concepts related to the found diseases are obtained. The *Disease subgraph filter* component is in charge of carrying out this step. Finally, the novelty and the reliability of the text are estimated at the same time. The *Reliability calculator* and

Novelty calculator components produce these results respectively. If no more texts have to be evaluated, then the results are sent to the visualization module. Otherwise, the process starts again with the next text.

3.2.2. Novelty estimation

Following (Zhang et al., 2002), in the *MESSI* framework, the novelty of a text is defined as the opposite of the degree of previous knowledge about the text. This degree measures how much *MESSI* knows about the topics addressed in the text. This metric is based on the number of medical concepts in common between the disease subgraph and text graph (*common_nodes*), and the relevance of their common relationships (*relevance*). The former is calculated as the proportion of nodes in the text graph that are also present in the current disease subgraph as follows:

$$common\_nodes(G(t), H(D)) = \frac{|N^{H(D)} \cap N^t|}{|N^t|} \tag{5}$$

In the case of the relevance, it is defined as the average of the relative relevance (i.e., the importance of a medical concept in the disease subgraph) of the common relationships $R(G(t), H(D))$ between the text graph and the disease subgraphs as follows:

$$relevance(G(t), H(D)) = \frac{\sum_{(d,m)\in R(G(t),H(D))} e^{\log \frac{I_r^{H(D)}(d,m)}{I_d^{H(D)}(d)}}}{|R(G(t), H(D))|} \tag{6}$$

The relative relevance of a relationship (d, m) is calculated as the logarithm of edge intensity $I_r^{H(D)}(d, m)$, divided by the disease intensity $I_d^{H(D)}(d)$ (see notation in Section 3). The use of logarithm is aimed at smoothing the differences between the relative intensity of edge with greater intensity and the other edges.

A metric to measure the amount of information *MESSI* has about the topics in the text is proposed as a linear combination of the $relevance$ and the $common\_nodes$:

$$kd(G(t), KB) = \alpha \cdot common\_nodes(G(t), H(D)) + (1 - \alpha) \cdot relevance(G(t), H(D)) \tag{7}$$

Notice that the parameter α regulates the trade-off between the knowledge of the topic and its relevance, but ensures that the result is between 0 and 1.

Lastly, the novelty of the text $N_t(G(t), KB)$ is estimated as follows:

$$N_t(G(t), KB) = 1 - kd(G(t), KB)\,. \tag{8}$$

Thus, the novelty grows in inverse proportion to the number of nodes (i.e., *common_nodes*) and relationships (i.e., *relevance*) in common. The first metric conveys whether the concepts found in the new article have been considered together with the disease addressed in previous articles. The second metric compares the oddity of the relationships found with the oddity of the relationships stored in the knowledge base.

3.2.3. Reliability estimation

The reliability $R_t(G(t), KB)$ measures how trustworthy a text is. It is defined as the average over the reputations of the common relationships $R(G(t), H(D))$ between the disease subgraphs and the text graph as follows:

$$R_t(G(t), KB) = \frac{\sum_{(d,m)\in R(G(t),H(D))} R_{d,m}}{|R(G(t), H(D))|} \tag{9}$$

The relationship reputation between the disease d and the medical concept m ($R_{d,m}$) is calculated as the third quartile of the reputations of all the edges E which share the same nodes as follows:

$$R_{d,m} = Q_3(rep(E(d, m))) \tag{10}$$

The third quartile is selected over other aggregation methods (e.g., mean or median) because it is more robust to extreme values (Savoy, 1997; 2005), and it gives importance to higher values. Thus, the third quartile controls the left-skewed distribution of the implicated reputations, mitigating the high number of papers with low reputations, redundant, inconsequential, and outright poor research (Bauerlein et al., 2010).

Therefore, a high reliability is achieved when the relationships of the concepts in the evaluated text are detected in the disease subgraphs and these subgraphs present a high reputation value based on the reputation of the papers used to build the knowledge base.

3.3. Visualization

The web portal of the *MESSI* framework is able to evaluate individual texts providing their novelty and reliability results, or multiple texts organizing them into groups. In the second case, texts are shown in a scatter plot, where the *x* axis represents their reliability values and the *y* axis their novelty values. Then, a vertical line and a horizontal line that represent the average reliability and the average novelty values, respectively, are drawn. Thus, the texts could be organized into four different quadrants:

- *Not novel and not reliable (Bottom-Left)*: those texts that address broadly studied topics, however, the research articles that have addressed these topics before have a low reputation.

Table 1
Reputation and intensity of a subset of common relations between the text graph and the knowledge base. Rep. stands for Reputation, I_r and I_d for edge and node intensity, respectively.

Disease	Medical concept	Rep.	$\mathbf{I_r}$	$\mathbf{I_d}$
Malignant Neoplasm of lung	Therapeutic procedure	0.59	592	592
Malignant Neoplasm of lung	Diagnosis	0.74	448	592
Malignant Neoplasm of lung	Carcinoma	0.54	13	592
Malignant Neoplasms	Lung	0.64	850	4838
Malignant Neoplasms	Squamous cell carcinoma	0.55	222	4838
Malignant Neoplasms	Carcinoma, Large Cell	0.72	5	4838

- *Not novel and reliable (Bottom-Right)*: these texts contain widely-known topics, and the reputations of the research articles about these topics are high.
- *Novel and not reliable (Top-Left)*: texts in this group tackle topics that are not common, and the articles about these topics have a low reputation.
- *Novel and reliable (Top-Right)*: topics of the texts belonging to this group are not common, but the research articles that have addressed these topics before have a high reputation.

Notice that the assignment of texts into these groups is relative to the processed papers. For instance, a medical text placed in the Bottom-Left quadrant implies that the novelty is low and has a low reliability regarding the other considered medical texts.

4. Experiments

MESSI has been evaluated in three experiments that emphasize different functionalities. For these experiments, more than 500,000 medicine research articles from *PubMed* published between 2000 and 2019 have been used. The research articles are split into train and evaluation sets, being the former used to create the knowledge base using the *Processing* module and the latter used to test the *Evaluation* module. The research articles prior to 2018 have been used for training purposes, and the articles published from 2018 onward have been used as test[3]. By splitting the set according to the year of publication (i.e., using the elder ones for training and the more recent for testing), the influence of the temporary feature on the novelty and reliability can be simulated.

The first experiment focuses on how the system evaluates a specific text. The obtained result and the process followed by the system are analyzed. Moreover, a sample of the data used in the process is manually extracted from the knowledge base and interpreted for better understanding.

The second experiment compares the result provided by the system when it evaluates a real text and two synthetic texts. Synthetic texts are created by replacing the original disease in the real text with another disease. It is expected that these modifications affect the evaluation results provided by the system.

Finally, in the third experiment, multiple texts are used to test the visualization tool. Some of these texts and their result have been analyzed to give an explanation of their relative representation. This allows checking the performance of the system and whether the visualization organizes the text properly. In addition, a comparison has been made between *MESSI* and alternative bibliometric indicators.

To perform a sensitivity analysis, the parameter α in Eq. (7) has been set to three different values: 0, 0.5 and 1, in order to calculate the knowledge degree. Notice that, when $\alpha = 0$, the novelty estimation only considers the term of the relevance of the common relationships, whereas when $\alpha = 1$, the novelty estimation only considers the common nodes term. A value of $\alpha = 0.5$ provides the same importance to both terms in the calculation of the novelty. It is important to remark that this parameter does not affect the performance of the system. Instead, it is a trade-off parameter to configure it, following the user preferences.

4.1. Simple text evaluation

The main purpose of this experiment is to illustrate the performance of the system by evaluating a simple text. For this purpose, the abstract from the research article with DOI *10.1097/MD.0000000000012613* has been evaluated. The article is about the treatment of three patients of *lung cancer*.

The calculated reliability of this article is 0.64. This value is estimated as the average reputation of the relations of all common concepts between the text graph and the disease subgraphs (see Section 3.2.3). An excerpt of the reputations of the common relations is shown in Table 1. For instance, the relationship between *Malignant Neoplasms* and *Carcinoma, Large Cell* has a reputation of 0.72 based on the reputation of five research articles where they appear (see Table 2). In this case, the reputations of these five research articles are similar with the exception of the research article with DOI *10.1097/MD.0000000000007842* that is far lower. However, this exception does not affect the overall reputation due to the selection of the third quartile as the method to aggregate these reputations.

[3] https://github.com/sariogonfer/messi-experiments

Table 2
DOI and reputation of the processed articles treating *Malignant Neoplasms* and *Carcinoma, Large Cell*.

DOI	Reputation
10.4293/JSLS.2014.00062	0.53
10.1097/MD.0000000000007842	0.07
10.1038/sj.bjc.6600286	0.80
10.1038/sj.bjc.6600464	0.72
10.1007/s00432-011-0986-0	0.71

Table 3
Results obtained by the system for the original text and the synthetic cases 1 and 2.

Text	Novelty	Reliability
Original	0.77	0.65
Synthetic case 1	0.82	0.58
Synthetic case 2	1.00	0.00

Table 4
Intensity and reputation between tested diseases and the medical concept *Breast*. NA means that there is no relationship between those terms.

Disease	Reputation	Rel. Intensity	Medical Concept
Malignant neoplasm of breast (Breast Cancer)	0.65	0.75	Breast
Upper Respiratory Infections (Cold)	0.42	0.09	Breast
Plantar fasciitis	NA	NA	Breast

Regarding the calculation of the novelty, it is based on the knowledge degree, which is a linear combination of the common nodes and relevance metrics weighted by the parameter α. In the case of the common nodes metric, the calculated value is 0.34, which implies that 34% of the medical concepts are in common between the evaluated text and the disease subgraph. On the other hand, the relevance metric is 0.38, which means that, on average, the intensity of the relations between a disease and a medical concept is small with respect to the intensity of the disease. For instance (see Table 1), the ratio in the relationship between the disease intensity *Malignant Neoplasm of lung* and the medical concept *Therapeutic procedure* is 1. Thus, *Therapeutic procedure* is the medical concept which appears the most with *Malignant Neoplasm of lung* in the research articles used to populate the knowledge base. However, most of the other relations have low ratios, such as the relation between *Malignant Neoplasm of lung* and *Carcinoma* and the relation between *M. Neoplasms* and *Carcinoma, Large Cell*, with ratios 0.02 and 0.001 respectively. As a consequence, the relevance metric yields a low value. The novelty calculated using these values and setting the parameter $alpha$ to 0.5 (i.e., the common nodes and the relevance metrics have the same weight) is 0.64.

4.2. Original vs synthetic comparison

In this experiment, the ability of *MESSI* to detect unusual relationships in medical texts is evaluated. For this purpose, the system has considered three medical texts: an original text about *breast cancer* and two synthetic texts. The synthetic texts are based on the original case but replacing the occurrences of *breast cancer* with *common cold* (synthetic case 1) and *plantar fasciitis* (synthetic case 2). This introduces associations between medical concepts and diseases which are not common (i.e., *common cold*) or not related (i.e., *plantar fasciitis*).

Novelty and reliability values for the three texts are presented in Table 3. The synthetic case 1 obtained a lower reliability but higher novelty than the original case. This could be explained by the fact that the medical concepts that appear in the texts are more related to *breast cancer* (original case) than to *common cold* (synthetic case 1). Some of these medical terms are shared because the body parts affected by *breast cancer* and *common cold* could be closed (i.e., breast). However, their reputation and edge intensity are not the same. For instance, considering the relation of these diseases with the body part *breast*, the relative intensity between *breast* with *common cold* is much lower than with *breast cancer* (see Table 4). Since the novelty is defined as the opposite of the knowledge degree, it follows that fewer relations with less intensity increase the novelty. Consequently, the novelty of the text increases but the reliability decreases with respect to the original text because the research articles, where the introduced relationships appear, have lower reputations.

Regarding the synthetic case 2, common relationships between the text graph and the knowledge base cannot be found (see Table 5). In consequence, as there is no previous knowledge about those relations, the calculated reliability and novelty are 1 and 0 respectively.

Table 5
Medical concepts found in each text. Underlined concepts only appear in the original text.

	Medical Concepts
Original	{Health education, Breast, Wellness Programs, Intervention regimes, Knowledge acquisition, Level, Breast Self-Examination}
Synthetic case 1	{Health education, Breast, Intervention regimes, Paracentesis, Knowledge acquisition, Level}
Synthetic case 2	{}

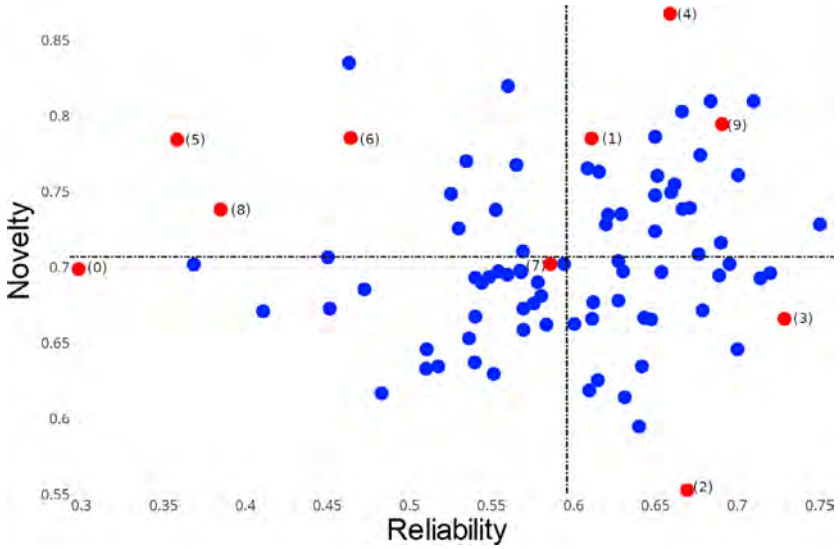

Fig. 7. Excerpt of the resulting scatter plot with the evaluated text.

Table 6
Evaluation results of studied texts.

Label	DOI	# nodes	# common nodes	cn
0	*10.4103/jfcm.JFCM_98_17*	63	2	0.03
1	*10.1186/s41747-018-0054-5*	71	9	0.13
2	*10.1016/j.idcr.2018.e00454*	21	6	0.29
3	*10.2147/TCRM.S166081*	79	13	0.16
4	*10.2147/NSS.S151085*	71	3	0.04
5	*10.1093/ofid/ofy210.2119*	77	6	0.08
6	*10.1159/000492745*	47	11	0.23
7	*10.2147/COPD.S186170*	99	16	0.16
8	*10.1097/MD.0000000000012376*	54	11	0.20
9	*10.1016/j.nicl.2019.101676*	70	10	0.14

4.3. Multiple texts evaluation

In this experiment, the capability of *MESSI* to evaluate and organize multiple texts is tested. A random subset of 200 abstracts from research articles published in 2018 has been selected. In this subset, *SemRep* only found medical concepts in 174 abstracts, discarding the others. The α parameter has been set to 0.5.

The novelty and reliability of these abstracts are represented in Fig. 7. The average reliability and the average novelty are used to define the four quadrants. As a result, the texts are distributed as follows: Top-Right (25), Top-Left (11), Bottom-Right (23) and Bottom-Left (28). There are 87 texts which are not represented in the visualization because they do not have common relations with the knowledge base (i.e., the system does not have information about their topics). Thus, for those papers, the system has estimated the novelty as 1 and the reliability as 0. Texts are mostly placed on the right quadrants, which implies that the reliability is negative-skewed (i.e., there are more values on the right side of the mean than on the left side). The novelty is positive-skewed, in contrast to the reliability. Thus, in the case that the skewness indicates that most of the points are placed on the left side (positive-skewed) or right side (negative-skewed) with respect to the mean, whereas the right side and left side respectively are more sparse. These distributions depend on the reliability and novelty values of the evaluated texts.

Table 7
Reliability estimated by *MESSI* and other bibliometric metrics for each selected manuscript.

Label	Reliability	UNIKO	CV	ICC	Citations	h-index
0	0.30	0.21	0	0	3	15
1	0.61	0.66	5	1	15	76
2	0.67	0.04	0	0	0	4
3	0.73	0.38	0	0	3	39
4	0.66	0.70	16	1	48	74
5	0.36	0.11	0	0	0	20
6	0.46	0.21	0	0	1	30
7	0.58	0.19	0	0	1	15
8	0.38	0.34	0	0	4	40
9	0.69	0.40	0	0	3	83

To deepen the understanding of the evaluation and organization, 10 texts distributed over all groups have been selected (red points in Fig. 7). Table 6 details information about these texts. Most of them have a rate of nodes in common with the knowledge base greater than 0.15.

Table 7 shows the reliability values calculated by *MESSI* for these papers. In addition, the reputation calculated with UNIKO and other metrics such as: Citation Velocity (CV), Influential Citation Count (ICC), the number of citations (Citations), and the authors highest h-index (h-index), are presented. It can be seen that CV and ICC, both estimated by Semantic Scholar, are not useful for comparison tasks. The number of citations is a biased reputation estimator because it strongly depends on factors such as the publication date and the journal. To compare *MESSI* reliability regarding the alternative methods, a test for association between paired samples, using Spearman's rank correlation, has been performed (Myles et al., 1973). The highest correlation values were achieved when *MESSI* was compared with UNIKO and h-index: 0.38 and 0.36, respectively. The alternative hypothesis for the test is that the true correlation is not equal to 0. In any of the cases, the correlations were not statistically different from zero (p-values higher than 0.1). Notice that UNIKO and h-index correlation is 0.89 (p-value < 0.001).

It is important to remark that *MESSI* only requires the text to estimate the reliability. In contrast, UNIKO needs that the text belongs to a previously published article, whereas in the case of the h-index metrics, it is necessary to know the authors. This is one of the most important strengths of *MESSI*, allowing it to be used in problems such as research paper filtering by journals.

The text 0 is focused on the knowledge of *Saudi medical students about alternative medicine* and it only has two relationships in common with the knowledge base (see Table 6). However, these relations have been incorrectly classified. It is caused by the incorrect classification of term *Statistical Package for the Social Sciences (SPSS)* (which is a statistical software) as *Stiff-Person Syndrome (SPS)* disease. Thus, it is the paper with the lowest reliability.

The texts 1 and 9 address *augmented reality in the medicine field* and the use of ML models to predict *stimulant dependence* respectively. The novelty of these texts is high because they use techniques that are not common in the medicine field. Besides, the reliability of these texts is high because the relationships in common with the knowledge base have a high reputation.

The text 2 has the lowest novelty among the selected texts. It only includes generic concepts, which results in more common nodes with the knowledge base. Further, the relevance of those relationships is high, which also penalizes the novelty.

The text 3 studies *the response to teduglutide for short bowel syndrome*. The *teduglutide* is a well-known treatment for short bowel syndrome. As a consequence, it produces a low novelty and a high associated reliability value.

The text 4 addresses the *Actigraphy-based sleep estimation*. This is the paper with the highest values in all the metrics but the *MESSI* system. The main reason is that the topic is not common, being treated by only 12 articles in 2018.

The text 5 is a study about the *Evolution of the Immunization on the Neonatal Intensive Care Unit at British Columbia Women's Hospital*. This type of studies, which addresses very specific topics, usually has a high level of novelty due to the fact that few articles cover the same topics.

Focusing on the texts 6 and 8, they have a large rate of nodes in common with the knowledge base, but they are evaluated as novel by *MESSI*. This evaluation is due to the low relevance (intensity) of their relations. Both texts also have a low reliability and a high novelty. Text 6 tackles a *new ocular implant* and the previous generation of this model of implant which has reported several complications in its use. In the text 8, a strange case of *impalement injuries in a construction worker* is reported. These kinds of injuries are not typical in medical articles (in 2018 only one medical article addresses it).

Finally, the text 7 addresses a study about *soluble receptor for advanced glycation end-products in the identification of Chronic Obstructive Pulmonary Disease (COPD) frequent exacerbator phenotype*. It is closed to the intersection of the quadrants in Fig. 7. Thus, this text presents average novelty and reliability values with respect to the other processed texts.

Notice that the distribution of the texts in the representation depends on the processed texts. Thus, the notions of high and low novelty and reliability are relative. From the perspective of the users (e.g., researchers or publishers), this relative classification into groups could support the selection of specific groups that they might consider most interesting. This leads to saving time and reducing the review effort.

5. Conclusions

This paper has presented a system for evaluating the novelty and the reliability of research articles through graph-based techniques. The novelty is a measure based on the knowledge previously stored in the system regarding the topics of the text. The reliability is a measure based on the reputation of the related research articles addressing similar topics. The main innovation of this work is the combination of both concepts to propose a novel way of sorting new health-related articles.

MESSI consists of two main modules: the *Processing* module and the *Evaluation* module. The system is completed with a web portal for visualization purposes and a knowledge base.

The *Processing* module manages the knowledge base built as a conceptual graph. This graph represents medical concepts as nodes, while edges indicate relationships between these concepts in each one of the research articles. The *Evaluation* module evaluates texts by using the knowledge stored in the knowledge base, and calculates their reliability and novelty.

Regarding the experiments, the framework has been evaluated through three different experiments. In the first experiment, a basic evaluation for a medical text was accomplished. In the second one, the ability to detect low reputed and novel medical test was evaluated. The third experiment illustrates how the system is able to provide support to users to discriminate between a set of manuscripts, organizing them according to their estimated reliability and novelty.

Some limitations have been mainly detected during the text extraction and graph building. For instance, for medical concepts extracted from the same text, the system assumed that they were related. However, these concepts might not be related despite appearing in the same text. Further, the system does not consider the type of the relation between two medical concepts (e.g., “Chemotherapy”, “treats”, “Malignant Neoplasm”), or if this relation is negated (e.g.,“Homeopathy”, “not treats”, “Malignant Neoplasm”).

In the future, the detected limitations of the system will be addressed. It will be also interesting to compare the performance of the system. From the best of our knowledge, no similar approaches have been found at the moment. Notice that most of the approaches in the literature (e.g., Google Scholar (Minasny et al., 2013) and Science Citation Index (De Bellis, 2009)) only use reputation and relevance measures usually based on citations to compare scientific articles. Moreover, *MESSI* is a prototype with complete functionality. Due to *MESSI* is based on the analysis of the text, further research in Natural Language Processing (NLP) and Natural Language Understanding (NLU) (Allen, 1995) techniques or combining multiple annotation tools (i.e., *SemRep* and *cTakes*) (Lin et al., 2017) will be very useful to enrich the knowledge graph. Other possible guidelines include alternative approaches for novelty estimation based on new metrics to evaluate the similarity between concepts. Moreover, sentiment analysis or affective computing techniques with the purpose of obtaining possible opinions of authors (positive or negative) about the research tackled in an article. Related to this issue, the evaluation of the type of relationships (positive, negative and neutral) between medical concepts could be a key functionality.

CRediT authorship contribution statement

Isaac Martín de Diego: Supervision, Project administration, Funding acquisition, Writing - review & editing. **César González-Fernández:** Conceptualization, Software, Formal analysis, Investigation, Data curation, Writing - original draft. **Alberto Fernández-Isabel:** Supervision, Writing - review & editing. **Rubén R. Fernández:** Software, Formal analysis, Investigation, Data curation, Writing - original draft. **Javier Cabezas:** Writing - original draft.

Acknowledgments

Research supported by grant from the Spanish Ministry of Economy and Competitiveness, under the Retos-Colaboración program: SABERMED (Ref: RTC-2017-6253-1), medical corpus provided by MMG and donation of the Titan V GPU by NVIDIA Corporation.

References

Ahmed, T. (2019). Medlineplus at 21: A website devoted to consumer health information. *Information Services & Use*, 1–10. (Preprint)

Allen, J. (1995). *Natural language understanding*. Pearson.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program.. In *Proceedings of the AMIA symposium* (p. 17). American Medical Informatics Association.

Bauerlein, M., Gad-el Hak, M., Grody, W., McKelvey, B., & Trimble, S. W. (2010). We must stop the avalanche of low-quality research. *The Chronicle of Higher Education, 13*.

Baumann, N. (2016). How to use the medical subject headings (MeSH).. *International Journal of Clinical Practice, 70*(2), 171–174.

Bergman, M. (2018). *The KBpedia resource* (pp. 409–419)).

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia – a crystallization point for the web of data.

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research, 32*(suppl_1), D267–D270.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on management of data* (pp. 1247–1250). AcM.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology, 66*(11), 2215–2222.

Castiglioni, A. (2019). *A history of medicine*. Routledge.

Consortium, U., et al. (2018). Uniprot: The universal protein knowledgebase. *Nucleic Acids Research, 46*(5), 2699.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Scarecrow press.

Demner-Fushman, D., Rogers, W. J., & Aronson, A. R. (2017). Metamap lite: An evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association, 24*(4), 841–844.

Dou, D., Wang, H., & Liu, H. (2015). Semantic data mining: A survey of ontology-based approaches. In *Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015)* (pp. 244–251). IEEE.

Ehrlinger, L., & Wöß, W. (2016). Towards a definition of knowledge graphs.. *SEMANTiCS (Posters, Demos, SuCCESS), 48*.
Fernández-Isabel, A., Prieto, J. C., Ortega, F., de Diego, I. M., Moguerza, J. M., Mena, J., ... Napalkova, L. (2018). A unified knowledge compiler to provide support the scientific community. *Knowledge-Based Systems, 161*, 157–171.
Fragkiadaki, E., Evangelidis, G., Samaras, N., & Dervos, D. A. (2011). f-value: Measuring an article's scientific impact. *Scientometrics, 86*(3), 671–686.
Franceschet, M. (2010). The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis. *Journal of Informetrics, 4*(1), 55–63.
Fricke, S. (2018). Semantic scholar. *Journal of the Medical Library Association: JMLA, 106*(1), 145.
Gabrilovich, E., Dumais, S., & Horvitz, E. (2004). Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th international conference on world wide web* (pp. 482–490). ACM.
Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA, 295*(1), 90–93.
Ghosal, T., Edithal, V., Ekbal, A., Bhattacharyya, P., Tsatsaronis, G., & Chivukula, S. S. S. K. (2018). Novelty goes deep. a deep neural solution to document level novelty detection. In *Proceedings of the 27th international conference on computational linguistics* (pp. 2802–2813).
Graves, M., Constabaris, A., & Brickley, D. (2007). FOAF: Connecting people on the semantic web. *Cataloging & Classification Quarterly, 43*(3–4), 191–202.
Guerrero-Bote, V. P., & Moya-Anegón, F. (2012). A further step forward in measuring journals scientific prestige: The SJR2 indicator. *Journal of Informetrics, 6*(4), 674–688.
Gupta, V., & Lehal, G. S. (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence, 2*(3), 258–268.
Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: A system to uniquely identify researchers. *Learned Publishing, 25*(4), 259–264.
Hewett, M., Oliver, D. E., Rubin, D. L., Easton, K. L., Stuart, J. M., Altman, R. B., & Klein, T. E. (2002). PharmGKB: The pharmacogenetics knowledge base. *Nucleic Acids Research, 30*(1), 163–165.
Hicks, D., & Melkers, J. (2013). Bibliometrics as a tool for research evaluation. *Handbook on the theory and practice of program evaluation*. Edward Elgar Publishing.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–16572.
Hofstra, B., Kulkarni, V. V., Galvez, S. M.-N., He, B., Jurafsky, D., & McFarland, D. A. (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences, 117*(17), 9284–9291.
Jiang, C., Coenen, F., Sanderson, R., & Zito, M. (2010). Text classification using graph mining-based feature extraction. In *Research and development in intelligent systems xxvi* (pp. 21–34). Springer.
Kamdar, B. B., Martin, J. L., Needham, D. M., & FCPA, M. (2017). Related articles from pubmed. *Ecology, 98*(5), 1290–1299.
Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., ... Vasant, D., et al. (2014). Disease ontology 2015 update: An expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research, 43*(D1), D1071–D1078.
Kilicoglu, H., Fiszman, M., Rodriguez, A., Shin, D., Ripple, A., & Rindflesch, T. C. (2008). Semantic medline: A web application for managing the results of pubmed searches. In *Proceedings of the third international symposium for semantic mining in biomedicine: vol. 2008* (pp. 69–76). Citeseer.
Kumar, S., & Bhatia, K. K. (2020). Semantic similarity and text summarization based novelty detection. *SN Applied Sciences, 2*(3), 332.
Kunze, J., & Baker, T. (2007). The Dublin core metadata element set. *Technical Report RFC 5013*. August
Leaman, R., Islamaj Doğan, R., & Lu, Z. (2013). DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics, 29*(22), 2909–2917.
Leitch, T., & Leitch, T. M. (2014). *Wikipedia U: Knowledge, authority, and liberal education in the digital age*. JHU Press.
Li, X., & Croft, W. B. (2005). Novelty detection based on sentence level patterns. In *Proceedings of the 14th ACM international conference on information and knowledge management* (pp. 744–751). ACM.
Lin, Y.-C., Christen, V., Groß, A., Cardoso, S. D., Pruski, C., Da Silveira, M., & Rahm, E. (2017). Evaluating and improving annotation tools for medical forms. In *International conference on data integration in the life sciences* (pp. 1–16). Springer.
Liu, F., Flanigan, J., Thomson, S., Sadeh, N., & Smith, N. A. (2018). Toward abstractive summarization using semantic representations. arXiv preprint arXiv:1805.10399.
Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM, 38*(11), 39–41.
Minasny, B., Hartemink, A. E., McBratney, A., & Jang, H.-J. (2013). Citations and the h index of soil researchers and journals in the web of science, scopus, and google scholar. *PeerJ, 1*, e183.
MMG: MedLab Media Group (2019). SABERMED. https://mmg-ai.com/es/el-ministerio-de-ciencia-elige-sabermed-de-mmg/. [Online: accessed 16-Feb-2021].
More, B. (2016). Overview of medicine- its importance and impact. *DJ International Journal Medical Research, 1*, 1–8.
Myles, H., Wolfe, D. A., & Chicken, E. (1973). *Nonparametric statistical methods* p. 503. New York, NY: Ed. John Wiley and Sons.
Naghizadeh, M., & Naghizdeh, R. (2017). Growth of scientific publications in Iran: Reasons, impacts, and trends. In *The development of science and technology in Iran* (pp. 75–86). Springer.
Okubo, Y. (1997). *Bibliometric indicators and analysis of research systems: Methods and examples*.
Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet:: Similarity: Measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004* (pp. 38–41). Association for Computational Linguistics.
Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J., & Leser, U. (2006). AliBaba: Pubmed as a graph. *Bioinformatics, 22*(19), 2444–2445.
Rajagopal, D., Cambria, E., Olsher, D., & Kwok, K. (2013). A graph-based approach to commonsense concept extraction and semantic similarity detection. In *Proceedings of the 22nd international conference on world wide web* (pp. 565–570). ACM.
Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM, 43*(12), 45–48.
Ribas, S., Ribeiro-Neto, B., de Souza e Silva, E., Ueda, A. H., & Ziviani, N. (2015). Using reference groups to assess academic productivity in computer science. In *Proceedings of the 24th international conference on world wide web* (pp. 603–608).
Richta, R. (2018). *Civilization at the crossroads: Social and human implications of the scientific and technological revolution (international arts and sciences press): Social and human implications of the scientific and technological revolution*. Routledge.
Rindflesch, T. C., & Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics, 36*(6), 462–477.
Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. *Scientific Reports, 7*(1), 5994.
Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association, 17*(5), 507–513.
Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management, 33*(4), 495–512.
Savoy, J. (2005). Bibliographic database access using free-text and controlled vocabulary: An evaluation. *Information Processing & Management, 41*(4), 873–890.
Sharma, S., Ciufo, S., Starchenko, E., Darji, D., Chlumsky, L., Karsch-Mizrachi, I., & Schoch, C. L. (2018). The NCBI biocollections database. *Database, 2018*.
Shen, W., Wang, J., & Han, J. (2014). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering, 27*(2), 443–460.
Singhal, A. (2012). Introducing the knowledge graph: Things, not strings. *Official google blog, 5*.
Spackman, K. A., Campbell, K. E., & Côté, R. A. (1997). SNOMED RT: A reference terminology for health care. In *Proceedings of the AMIA annual fall symposium* (p. 640). American Medical Informatics Association.
Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the 16th international conference on world wide web* (pp. 697–706). ACM.
Tanabe, L., & Wilbur, W. J. (2002). Tagging gene and protein names in biomedical text. *Bioinformatics, 18*(8), 1124–1132.
Tsai, F. S., & Zhang, Y. (2011). D2S: Document-to-sentence framework for novelty detection. *Knowledge and Information Systems, 29*(2), 419–433.
Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*(7763), 95.
Vrande, D., & Krtzsch, M. (2014). Wikidata: A free collaborative knowledge base. *Communications of the ACM, 57*, 78–85.

Xie, Y., Wu, Q., & Li, X. (2019). Editorial team scholarly index (ETSI): An alternative indicator for evaluating academic journal reputation. *Scientometrics, 120*(3), 1333–1349.

Yang, Y., Zhang, J., Carbonell, J., & Jin, C. (2002). Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 688–693). ACM.

Zhang, M., Lin, C., Liu, Y., Zhao, L., & Ma, S. (2003). THUIR at TREC 2003: Novelty, robust and web. In *Trec* (pp. 556–567). Citeseer.

Zhang, Y., Callan, J., Callan, J., & Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 81–88). ACM.

Zhong, J., Zhu, H., Li, J., & Yu, Y. (2002). Conceptual graph matching for semantic search. In *International conference on conceptual structures* (pp. 92–106). Springer.