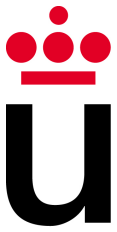

Universidad
Rey Juan Carlos

Tesis Doctoral

Métricas de reputación para la construcción de Sistemas Basados en el Conocimiento

Autor:

Juan Carlos Prieto Hernández

Directores:

Isaac Martín de Diego

Alberto Fernández-Isabel

Programa de Doctorado en Tecnologías de la Información y las Comunicaciones

Escuela Internacional de Doctorado

Septiembre de 2023

A mis nietos,
que probablemente me verán presentar la tesis.

Agradecimientos

Es curioso que, después de llevar años desarrollando y escribiendo la tesis, éstas vayan a ser las líneas de mayor alcance. O no, que se lo pregunten a Raymond Radiguet.

Lo primero –bueno, no ha sido lo primero–, quiero agradecer a Isaac y a Alberto la paciencia infinita que han tenido con el ritmo que he impuesto. Verme avanzar con los tres artículos publicados en revistas, la conferencia en el Imperial College de Londres y la escritura de la tesis, ha tenido que ser para ellos como una tarde de Tour de Francia. Quiero agradecerles también su capacidad de adaptación para ayudarme de madrugada y en fines de semana. Han empatizado al máximo con mi situación. Compaginar el desarrollo del doctorado con el lanzamiento de una empresa de software desde cero pone las cosas difíciles. Difíciles y lentas. Por suerte –porque es su mérito, no el mío–, estoy a punto de terminar mi tesis. Y, entretanto, hemos desarrollado un SaaS *top* –gracias a David y *devs*– y hemos plantado pica ya en 7 países –gracias a todo el equipo–. *Not bad.* Además, emprender como padre en mitad de todo esto tampoco ha ayudado a ir más rápido, pero sí lo ha hecho mucho más divertido. Tanto que he emprendido dos veces.

También quiero agradecer a mis padres todo lo que han hecho por mí desde que nací. Antes no tengo nada que agradecerles, la verdad, sólo pensaban en ellos. Sin embargo, mi nacimiento supuso un punto de inflexión en nuestra relación. Desde entonces se volcaron en darme todo su apoyo y ánimo. Y en proporcionarme todas las herramientas que he necesitado para realizar mi doctorado. Especialmente la capacidad de respirar, eso me ha alentado.

Por supuesto, quiero agradecer a Patri todo lo que ha hecho para que yo haya podido terminar el doctorado. No se me ocurre una manera más bonita de agradecérselo que pidiéndole que nos casemos. Pero de momento sólo le voy a dar las gracias.

También quiero agradecer a mi abuela y a mi tía Almu su interés en mis avances académicos durante estos años. Sobre todo a mi abuela, que el interés de mi tía no bajaba de 60 minutos al teléfono. En cualquier caso, han sido un par de empujoncitos que me han ayudado a avanzar, uno de ellos desde abajo.

Por último, quiero agradecer a todas las personas que han contribuido a mi felicidad. Y también a mi familia y a mis amigos. A Bilborito, que sabe qué color es el amarillo.

A Jesús, que le gustan las carrozas. A Almu, *no como a mi madre*. A Willy Wonka. A Winston Churchill. A Karim Benzema, a Simon Neil, a Mike Tomlin y a todos los demás. Go Steelers!

Resumen

La reputación es un concepto tratado de manera tradicional como una medida afectada por agentes subjetivos. Además, los esfuerzos realizados para proponer una métrica calculada con parámetros objetivos se han visto comprometidos por aspectos que no se han tenido en cuenta y reducen la calidad de la métrica. En la comunidad científica, por ejemplo, no existe una métrica única, objetiva y precisa para clasificar el trabajo de los investigadores en función de su mérito científico. La mayoría de las métricas existentes se basan en la cantidad de publicaciones asociadas con un autor junto con la cantidad de citas recibidas por esas publicaciones. Para mitigar alguna de estas limitaciones se presenta el framework *Unified Knowledge Compiler* (UNIKO). Este framework incluye la evaluación de puntuaciones de reputación de artículos y autores, y el análisis de sentimiento de los textos de los artículos. Además, también incluye la recomendación de artículos y autores relacionados con un campo de aplicación específico para facilitar el proceso de investigación a la comunidad científica. UNIKO está construido como un framework híbrido basado en sistemas basados en el conocimiento y sistemas de recomendación basados en el contenido. Con el fin de evaluar el rendimiento del sistema, se han realizado varios experimentos. El primer experimento se desarrolla para ilustrar la tarea de puntuación de reputación. El segundo aborda el cálculo de puntuaciones de sentimiento basado en un léxico que es compatible con una red neuronal convolucional. El último experimento muestra las tareas de recomendación basadas en medidas específicas de similaridad y aprendizaje no supervisado. También se ha propuesto el *Framework for Reputation Estimation of Scientific Authors* (FRESA), centrándose en la calidad de la métrica de reputación. Es un sistema capaz de estimar la reputación de un investigador haciendo uso de los conceptos de relevancia y novedad en el dominio científico. El sistema es capaz de representar las trayectorias científicas de los investigadores a través de los índices propuestos para ilustrar su evolución en el tiempo. FRESA utiliza fuentes de información web y aplica medidas de similaridad, técnicas de minería de texto y algoritmos de clustering para clasificar y agrupar a los investigadores. Por último, también se presenta el framework *Domains Classifier based on Risky Websites* (DOCRIW), que estudia el cálculo de la reputación de sitios web. Se basa en dos componentes principales. El primer componente es una base de conocimiento construida previamente que contiene información de sitios web riesgosos. El

segundo complementa el sistema con un clasificador binario capaz de etiquetar un sitio web (como riesgoso o no) en base a su reputación considerando solo su dominio web. El sistema hace uso de fuentes de información web e incluye variables basadas en host. También aplica medidas de similaridad, algoritmos de aprendizaje supervisado y métodos de optimización para mejorar su rendimiento. El trabajo presentado es experimental y ofrece resultados prometedores.

Abstract

Reputation is a concept traditionally treated as a measure affected by subjective agents. In addition, the efforts made to propose a metric calculated with objective parameters have been compromised by aspects that have not been taken into account and reduce the quality of the metric. In the scientific community, for example, there is no single, objective and accurate metric to classify the work of researchers based on its scientific merit. Most existing metrics are based on the number of publications associated with an author along with the number of citations received by those publications. To reduce some of these limitations, the UNIKO framework is introduced. This framework includes article and author reputation assessment, and the sentiment analysis of the texts of the articles. In addition, it also includes the recommendation of articles and authors related to a specific field of application in order to ease the research process to the scientific community. UNIKO is built as a hybrid framework based on Knowledge-Based Systems and Content-Based Recommendation Systems. In order to evaluate the performance of the system, several experiments have been done. The first experiment is developed to illustrate the reputation scoring task. The second one addresses the sentiment scores calculation based on a lexicon which is supported by a Convolutional Neural Network. The last experiment shows the recommendation tasks based on specific similarity measures and unsupervised learning. The FRESA has also been proposed, focusing on the quality of the reputation metric. It is a system able to estimate the reputation of a researcher using the concepts of relevance and novelty in the scientific domain. The system is able to depict the scientific trajectories of the researchers through the proposed indexes to illustrate their evolution over time. FRESA uses web information sources and applies similarity measures, text mining techniques, and clustering algorithms to also rank and group the researchers. Finally, the framework DOCRIW is also presented, which studies the calculation of domains *Uniform Resource Locator* (URL) reputation. It is based on two main components. The first component is a previously built knowledge base containing information from risky websites. The second one complements the system with a binary classifier able to label a website (as risky or not) based on the reputation score considering just its URL domain. The system makes use of web information sources and includes host-based variables. It also applies similarity measures, supervised learning algorithms and optimization methods

to enhance its performance. The presented work is experimental, rendering promising outcomes.

Índice general

Índice de figuras

Índice de tablas

Listado de acrónimos

AB *Adaboost*.

ABS *Agent-Based Systems*.

AI *Artificial Intelligence*.

API *Application Programming Interface*.

BNLFT *Biased Negative Latent Factorization Tensor*.

CNN *Convolutional Neural Network*.

CORE *Computing Research and Education Association*.

DL *Deep Learning*.

DNM *Dendritic Neural Model*.

DNS *Domains Name System*.

DOCRIW *Domains Classifier based on Risky Websites*.

DOI *Digital Object Identifier*.

DT *Decision Tree*.

DTW *Dynamic Time Warping*.

ERT *Extremely Randomized Trees*.

FLS *Fuzzy Logic Systems*.

FN *False Negative*.

FP *False Positive*.

FRESA *Framework for Reputation Estimation of Scientific Authors*.

GB *Gradient Boosting*.

ICIMP *International Conference on Information Management and Processing*.

IRC *Information Retrieval Controller*.

ISI *Institute of Science Information*.

JCR *Journal Citation Reports*.

KBS *Knowledge-Based Systems*.

KMC *Knowledge Management Controller*.

KMS *Knowledge Management Systems*.

kNN *k-Nearest Neighbour*.

LDA *Linear Discriminant Analysis*.

LR *Logistic Regression*.

LSA *Latent Semantic Analysis*.

MIT *Massachusetts Institute of Technology*.

ML *Machine Learning*.

NB *Naïve Bayes*.

NLF *Negative Latent Factorization*.

NLP *Natural Language Processing*.

NMF *Non-Negative Matrix Factorization*.

NNS *Neural Networks Systems*.

OEDI *Observatorio Español de Delitos Informáticos*.

OSI *Observatorio Español de Seguridad de la Información*.

Q1 *Quartile 1*.

Q2 *Quartile 2*.

Q3 *Quartile 3*.

Q4 *Quartile 4*.

QoS *Quality of Service*.

RF *Random Forest*.

RS *Recommender Systems*.

SVM *Support Vector Machine*.

TN *True Negative*.

TP *True Positive*.

TRS *Trust and Reputation Systems*.

UNIKO *Unified Knowledge Compiler*.

URL *Uniform Resource Locator*.

WGI *Weighted Gini Index*.

Capítulo 1

Introducción

La reputación es originalmente un constructo social que permite determinar el grado de fiabilidad o ética en una persona o entidad en comparación al resto de agentes de la sociedad (Barnett *et al.*, 2006). La reputación ha sido tradicionalmente sometida en la mayoría de los casos a una evaluación expuesta a componentes subjetivos y a un desconocimiento de los datos que la definen. Con el paso de los años se ha convertido en un tema de interés científico y se han intentado establecer fundamentos objetivos que permitan implantar rigor en su cálculo. Parece que establecer un estándar que mida la reputación de manera precisa a nivel global es del dominio de lo imposible, pero se han conseguido imponer algunas métricas de reputación por distintas áreas de aplicación gracias a los esfuerzos de los investigadores (por ejemplo, el PageRank (Page *et al.*, 1999)). Hoy en día, existen mecanismos y tecnologías que permiten definir la reputación como una métrica calculada a partir de fórmulas matemáticas y de modelos de aprendizaje máquina. La reputación tiene un impacto, en muchos casos inconsciente, en nuestra forma de relacionarnos y de actuar. Nuestro pensamiento se va formando por la información que consumimos, y esas fuentes de información son seleccionadas en gran medida por el nivel reputacional de las mismas. Por tanto, la reputación de las personas, empresas, medios de comunicación o personalidades públicas, por citar algunos ejemplos, tienen una influencia enorme en nuestra forma de vida.

En esta tesis analizamos la reputación como un componente global y entramos en detalle proponiendo un reputation score en el ámbito científico/académico y en Internet. En el área científica, ofrecemos una mejora a las propuestas anteriores introduciendo en el cálculo de la reputación de los autores y/o artículos compenentes de novedad y despenalizando la calidad de las aportaciones (dado que las métricas propuestas hasta ahora premian mucho la productividad), para ello se han utilizado sistemas basados en el conocimiento. En cuanto al cálculo de la reputación de los sitios web, hemos utilizado modelos de aprendizaje supervisado y métodos de optimización para evitar tener que acceder al contenido de las webs, simplificando el proceso y la carga de cómputo, además del riesgo

de acceso a sitios web con contenido malicioso. En el resto del capítulo, definiremos el término reputación y presentaremos la reputación en distintos campos de aplicación.

1.1. Motivación

La evaluación de la reputación es un aspecto fundamental en diversos dominios, desde la academia hasta el mundo en línea, donde la calidad y la objetividad son esenciales. Sin embargo, a lo largo de los años, hemos enfrentado el desafío de medir la reputación de manera precisa y justa, ya que tradicionalmente se ha considerado un concepto subjetivo, susceptible a la influencia de factores no tenidos en cuenta. Este desafío se ha vuelto aún más evidente en el ámbito científico, donde la evaluación del mérito científico de investigadores y publicaciones es esencial para el avance del conocimiento. La falta de métricas de reputación sólidas y objetivas ha llevado a la necesidad de desarrollar enfoques innovadores y sistemas capaces de abordar estas limitaciones.

La motivación detrás de esta tesis reside en la búsqueda de soluciones efectivas para mejorar la medición y evaluación de la reputación en diversos contextos. A lo largo de este trabajo, exploraremos la creación de sistemas y frameworks que incorporan el poder del conocimiento, la minería de texto, el aprendizaje automático y otras técnicas avanzadas para ofrecer métricas de reputación más precisas y útiles. Estos sistemas están diseñados para abordar no solo la cantidad, sino también la calidad de las métricas de reputación, considerando factores como la relevancia, la novedad y el sentimiento.

En última instancia, esta tesis busca contribuir a la comunidad científica y a otros contextos donde la evaluación de la reputación es crucial, proporcionando herramientas y enfoques prometedores que pueden tener un impacto significativo en la mejora de la toma de decisiones, la calidad de la investigación y la confiabilidad en la evaluación de la reputación. A través de una serie de experimentos y resultados prometedores, nuestro objetivo es avanzar en la comprensión y aplicación de métricas de reputación más sólidas y objetivas en diversos campos, allanando el camino hacia una evaluación más justa y precisa en el mundo académico y más allá.

1.1.1. Definición de reputación

El término reputación viene del latín *reputatio* y la Real Academia Española de la Lengua lo define como la "opinión o consideranción en que se tiene a alguien o algo" o el "prestigio o estima en que son tenidos alguien o algo" (RAE, s.f.-b). El Compact Oxford English Dictionary lo define como "the beliefs or opinions that are generally held about someone or something" y como "a widespread belief that someone or something has a particular characteristic" (RAE, s.f.-a). En resumen, la reputación denota la percepción en que un individuo, conjunto de individuos u organización es vista en términos de prestigio.

Es por ello que hablamos de la reputación de un autor, de una revista, de un centro universitario o de una empresa, entre otras, como el grado de credibilidad, competencia y/o nivel ético.

No hay que confundir reputación con relevancia, que la Real Academia Española de la Lengua define como "cualidad o condición de relevante, importancia, significación."(RAE, s.f.-b). La relevancia incluye el aspecto social, la exposición o el alcance de algo o alguien. No obstante, algo o alguien con cierta relevancia puede tener tanto buena como mala reputación, no hay una correlación de 1 entre ambos términos.

La reputación se ha convertido en una métrica utilizada en todas las áreas. En el ámbito científico se ha buscado establecer una reputación para los autores, artículos o revistas científicas, por ejemplo (Over, 1982). Pero la reputación no es una métrica exclusiva del ámbito científico, también se mide la reputación de las empresas, de los sitios web, de los usuarios en redes sociales o de personalidades públicas, entre otras (Madden y Smith, 2010).

Pese a que existe un acuerdo en la definición de reputación, no se ha conseguido establecer un estándar para medir la reputación de manera precisa. Hay dos problemas principales que tienen mayor o menor impacto en el cálculo de la reputación en función del ámbito donde se calcule: la falta de datos y la introducción de valores subjetivos. En los frameworks presentados en la Sección 3 de este documento, se discuten estos problemas y se proponen soluciones para aliviar el efecto negativo de los mismos.

1.1.2. Reputación de autor

La reputación de un autor se ha convertido en un tema de conversación recurrente dentro de la comunidad científica debido a que no hay un mecanismo único y automático capaz de calificar a los científicos. Y, en cambio, sí existe la necesidad de emplear dicha reputación en procesos de evaluación y/o selección del profesorado. De manera tradicional la reputación de un autor ha estado ligada a dos factores. El primero es la producción científica, que tiene en cuenta únicamente el número de artículos publicados, ignorando aspectos básicos relacionados con la calidad de las publicaciones y de la aportación real del autor en la investigación. El segundo es el impacto que ha podido tener alguno de sus artículos, considerando más reputado un autor con un artículo con mucha relevancia que uno con publicaciones de alta calidad pero con menor impacto. Esto quiere decir que autores enfocados en temas de menor interés común son penalizados pese a la calidad de sus aportaciones científicas. Se han producido varios intentos de definir métricas de reputación basadas en las publicaciones como h-index (Hirsch y Buela-Casal, 2014), g-index (Egghe, 2006) o el índice i10 de Google (Dhamdhere, 2018). Estas métricas serán definidas en la Sección 2. Sin embargo, ninguna métrica ha alcanzado una relevancia especial para convertirse en un estándar (Yu *et al.*, 2016) debido a que son demasiado simples para ser

justos con todos los perfiles de autores existentes, ya que priman la productividad del autor sobre la calidad o la novedad de los enfoques encarados en sus artículos. Por tanto, los autores más reputados son generalmente los que tienen mayor número de artículos publicados; lo cual provoca que los autores con más años de experiencia se vean premiados en términos de reputación. El no haber encontrado un estándar que permita ranquear a los investigadores de manera objetiva y única (Meho y Rogers, 2008) dificulta utilizar criterios objetivos, en la medida de lo posible, para resolver situaciones actuales de manera más justa como asignar plazas públicas en universidades o para designar premios por méritos científicos. Algo que ha dificultado encontrar una métrica precisa es que no hay un repositorio único donde se recopilen todas las publicaciones. No obstante, existen repositorios bien conocidos que intentan dar una solución aceptable al asunto. Ejemplos famosos de ellos son Google Scholar (López-Cózar *et al.*, 2019) y Semantic Scholar (Fricke, 2018). Ambos son motores de búsqueda especializados en contenido científico-académico. El segundo está respaldado por un sistema de inteligencia artificial desarrollado por el Allen Institute for Artificial Intelligence, y provee mayor información. El principal problema en el uso de estos repositorios es la capacidad de recopilación de información de los autores (es decir, las publicaciones realizadas por ellos o las citas asociadas). Por lo tanto, elaborar una métrica estándar para puntuar a los investigadores es una tarea exigente que podría influir en la comunidad científica para alimentar estos repositorios transformándolos en bases de conocimiento más completas.

1.1.3. Reputación de artículo

La reputación de un artículo científico siempre ha precedido a la del autor, de modo que la de éste siempre se ha calculado en base al número de artículos y la reputación de los mismos. La reputación de un artículo normalmente ha estado basada en el número de citas recibidas, lo cual penaliza a los artículos que tratan sobre temas muy específicos y/o novedosos. Con los artículos no se ha hecho un esfuerzo para proponer métricas que tengan en cuenta la novedad del artículo o el grado de complejidad del caso resuelto. Probablemente no se ha hecho este ejercicio porque estos dos aspectos tienen el peligro potencial de introducir muchos valores subjetivos en el cálculo. No obstante, medir la reputación únicamente por el número de citas que tiene el mismo, se antoja un cálculo incompleto. La extracción de palabras clave que tiene el artículo o el cálculo del valor de sentimiento del mismo podrían ayudar a completar el análisis de reputación (Fernández-Isabel *et al.*, 2018).

1.1.4. Reputación de revistas científicas

Las revistas científicas son calificadas en función de su calidad mediante distintas métricas. Una métrica es el índice de la revista determinado por el cuartil en el que se encuentra basado en su factor de impacto (Lariviere y Sugimoto, 2019). Las revistas in-

dexadas en el *Journal Citation Reports* (JCR) (JCR, s.f.) pueden pertenecer a distintos cuartiles que definen la reputación de la revista. Estos cuartiles se dividen en función del factor de impacto de la revista. El 25 % de las revistas con mayor factor de impacto pertenecen al *Quartile 1* (Q1) (por ejemplo: Computing Surveys); las que se encuentran entre el 25 % y el 50 % pertenecen al *Quartile 2* (Q2) (por ejemplo: Online Information Review); entre el 50 % y el 75 % pertenecen al *Quartile 3* (Q3) (por ejemplo: Journal of Applied Logic) y las últimas pertenecen al *Quartile 4* (Q4) (por ejemplo: Theoretical Informatics and Applications). Hay revistas que no están indexadas en el JCR y su reputación es menor.

El factor de impacto de la revista es el número promedio de veces que los artículos de la revista publicados en los últimos dos años han sido citados en el año JCR. El factor de impacto se calcula dividiendo el número de citas en el año JCR por el número total de artículos publicados en los dos años anteriores. Un factor de impacto de 1,0 significa que, en promedio, los artículos publicados hace uno o dos años han sido citados una vez. Un factor de impacto de 2,5 significa que, de media, los artículos publicados hace uno o dos años han sido citados dos veces y media. Los artículos citados pueden ser de la misma revista, pese a que la mayoría de los artículos que citan son de diferentes revistas. El número del factor de impacto no significa mucho por sí mismo, aporta valor en comparación con el factor de impacto con el resto de revistas y su ranking.

Además del factor de impacto, JCR también incluye las siguientes métricas para una revista: el Article Influence Score, el Eigenfactor Score, el Five-Year Impact Factor y el Immediacy Index (of Florida, s.f.). La puntuación de la influencia del artículo (Article Influence Score) determina la influencia promedio de los artículos de una revista durante los primeros cinco años después de la publicación. Se calcula dividiendo la puntuación de factor propio de una revista por el número de artículos en la revista, normalizado como una fracción de todos los artículos en todas las publicaciones. El cálculo de la puntuación de factor propio (Eigenfactor Score) se basa en la cantidad de veces que los artículos de la revista publicados en los últimos cinco años han sido citados en el año JCR, pero también considera qué revistas han contribuido con estas citas para que las revistas altamente citadas influyan en la red más que revistas menos citadas. Las referencias de un artículo de una revista a otro artículo de la misma revista se eliminan, de modo que las puntuaciones de los factores propios no se vean influenciadas por las autocitas de la revista. El factor de impacto a 5 años (Five-Year Impact Factor) se calcula como el factor de impacto pero usando 5 años en lugar de 2 años. El índice de inmediatez (Immediacy Index) es el promedio de veces que se cita un artículo en el año de su publicación.

Además, numerosos artículos se publican en conferencias, que también son calificadas y rankeadas. La clasificación de conferencias *Computing Research and Education Association* (CORE) (CORE, s.f.) proporciona evaluaciones de las principales conferencias en las disciplinas informáticas. Las clasificaciones son administradas por el Comité Ejecutivo

de CORE, con rondas periódicas para la presentación de solicitudes de adición o reclasificación de conferencias. Las decisiones son tomadas por comités académicos con base en datos objetivos solicitados como parte del proceso de presentación. Las conferencias (portal, s.f.) se asignan a una de las siguientes categorías:

- A*: determina que la conferencia es líder en un área de disciplina (por ejemplo: Computer Aided Verification).
- A: define la conferencia como excelente y muy respetada en un área de disciplina (por ejemplo: Symposia on VLSI Technology and Circuits).
- B: define la conferencia de buena a muy buena y bien considerada en un área de disciplina (por ejemplo: Conference on Information Science, Technology and Management).
- C: define las conferencias como que cumplen con los estándares básicos (por ejemplo: Asilomar Conference on Signals, Systems and Computing).
- Australasia: define la conferencia como la cual cuya audiencia es principalmente australiana y neozelandesa (por ejemplo: Australasian Database Conference).
- Sin clasificar: es la categoría para las conferencias para las que no se ha tomado una decisión de clasificación (por ejemplo: Engineering Interactive Computing Systems).
- Nacional: define una conferencia como local en un país y que no es lo suficientemente conocida como para ser clasificada (por ejemplo: IEEE Conference on Cognitive and Computational Aspects of Situation Management).
- Regional: es similar a Nacional pero puede cubrir una región que cruza fronteras nacionales (por ejemplo: International Baltic Conference on Databases and Information Systems).

Las clasificaciones de la conferencia están determinadas por una combinación de indicadores, que incluyen las tasas de citación, las tasas de envío y aceptación de trabajos, y la visibilidad y el historial de investigación de las personas clave que organizan la conferencia y administran su programa técnico.

1.1.5. Reputación de centros universitarios

Existen numerosos listados que califican las universidades por relevancia o reputación. Estos listados son emitidos anualmente por distintas entidades de todo el mundo (Volkwein y Sweitzer, 2006). Estos rankings se construyen por norma general utilizando una metodología bibliométrica, de acuerdo a criterios objetivos que son medibles (Delgado-Márquez *et al.*, 2012). La mayoría de ellos son globales, basados en más de un criterio,

para determinar la calidad de la universidad a nivel general, pero también existen rankings específicos que se centran en una característica única (K.-h. Chen y Liao, 2012). También se pueden encontrar listados basados en criterios objetivos no bibliométricos. Además, existen rankings basados en criterios subjetivos, calificados por expertos o a través de sondeos de opinión de profesores y alumnos. Este tipo de ranking suele tener menor valor. Los criterios objetivos bibliométricos no son proporcionados por las propias universidades, por ello son fiables y los análisis son reproducibles y rigurosos. Algunos de los criterios objetivos bibliométricos son (Bordons *et al.*, 1999):

- número de artículos publicados en revistas indexadas y cálculos del factor de impacto de esas revistas,
- número total de citas de los artículos publicados por los académicos de la universidad,
- número de artículos publicados en revistas de alto factor de impacto (por ejemplo, Science o Nature),
- y número de académicos o de ex-alumnos galardonados con premios de gran relevancia internacional (por ejemplo, Premio Nobel o Medalla Fields).

Los criterios objetivos no bibliométricos son proporcionados por las propias universidades, de modo que pueden estar sujetos a manipulación dado que no son siempre verificables. Estas medidas, por tanto, suelen proporcionar información sobre la infraestructura y poder económico de una universidad más que la calidad de la misma. Algunos criterios objetivos no bibliométricos (Rousseau, 2001) son:

- ratio del número de estudiantes graduados sobre el número de estudiantes matriculados,
- número de académicos con doctorado,
- número y tipo de grados impartidos,
- y número de posgrados registrados en patrones de calidad.

Independientemente de la institución que realice el ranking, hay universidades que aparecen habitualmente en el top 10. Algunas de ellas son las americanas de Harvard, Stanford, *Massachusetts Institute of Technology* (MIT), Berkley o Princeton. También son recurrentes las universidades británicas de Oxford, Cambridge y Londres.

1.1.6. Reputación de empresas

La reputación es el activo corporativo más relevante para una empresa. Se adquiere poco a poco pero puede perderse fácilmente y de un día para otro si no se gestiona de manera adecuada. Es un desafío comprender y explicar qué valor específico está asociado con el logro de una buena reputación para una empresa. La evidencia sobre el valor de la reputación o su poder para alcanzar las metas corporativas es un tanto difusa. Por ello, la reputación debe gestionarse si tiene un valor explícito para la empresa, "stakeholders" y la sociedad en general, y siempre supone un esfuerzo y tiene una complejidad (Brown *et al.*, 2022).

Demostrar y proporcionar evidencia del valor financiero y no financiero de la reputación es tema de investigación. Algunos autores consideran que la reputación es indispensable en cualquier proceso de intercambio en los mercados porque los interesados generalmente celebran un contrato con una empresa en función de su reputación (Carmeli y Freund, 2002); por tanto, la reputación debe considerarse como una condición previa para que las personas estén dispuestas a hacer negocios con una empresa (Ettenson y Knowles, 2008). Desde la perspectiva de las partes interesadas, la reputación corporativa indica la contribución de la empresa al bienestar de las partes interesadas y al bienestar social en general. La reputación corporativa es crucial para que las partes interesadas determinen su propio apoyo a la empresa. Desde un punto de vista utilitario, vincular la reputación con el beneficio corporativo puede servir como prueba de la relevancia de la reputación para la empresa. La reputación puede afectar el resultado final (Abdel-Kader y Mentzeniot, 2007). Sin embargo, ni el valor concreto de la reputación corporativa ni su impacto en el éxito financiero de las empresas se pueden corroborar empíricamente de manera satisfactoria (Herbig y Milewicz, 1995). Por tanto, sigue siendo cierto que a pesar de su evidente valor, el valor en dólares de la reputación de una empresa resulta difícil de cuantificar (Fombrun y Van Riel, 1997). No obstante, esto no significa que medir la reputación no sea posible, sólo es difícil establecer una relación entre la reputación y su valor económico. Además, en las empresas, la reputación también puede basarse en percepciones de los principios morales que parecen guiar la conducta de una empresa. Hemos de considerar que la reputación es una construcción social que puede basarse en la observación de las consecuencias de las acciones. Es decir, la sociedad recompensa las reputaciones que se basan en motivos morales más de lo que valora las reputaciones que se basan en recibir o distribuir beneficios (Mitnick y Mahon, 2007). La reputación alinea el comportamiento corporativo y determina el papel que las empresas pueden desempeñar en la sociedad.

Para los gerentes, la reputación corporativa es un activo, una ventaja competitiva, un recurso, un valor; un impulsor del desempeño económico que, en el mejor de los casos, debería ser medible. La reputación y su mejora son también un componente ético del trabajo diario de los directivos. Se enfrentan a la tarea de tomar decisiones éticamente aceptables al mismo tiempo que garantizan la eficiencia de sus empresas y accionistas,

lo que a su vez da forma a la reputación. Por lo tanto, los gerentes apuntan al aspecto utilitario y deontológico de la reputación (Helm *et al.*, 2011).

1.1.7. Reputación en Internet

Los economistas han estudiado durante mucho tiempo el fenómeno de la reputación, definida ampliamente como lo que los agentes (por ejemplo, los compradores online) creen o esperan de otros agentes (por ejemplo, los vendedores online). La reputación en Internet es un fenómeno separado por derecho propio, y se han realizado numerosos estudios sobre ello (M. Chen y Singh, 2001). En primer lugar, el crecimiento de Internet ha estado acompañado por el crecimiento de los sistemas formales y sistemáticos de revisión y retroalimentación (tanto de los agentes del mercado online como de los agentes del mercado offline que son calificados online). Por el contrario, la investigación tradicional sobre la reputación solía estar motivada por problemas del mundo real en los que solo estaba presente información suave (por ejemplo, la reputación de un monopolio local titular como American Airlines por ser duro en el trato con los competidores que entran en su mercado). La segunda razón por la que la reputación en Internet forma un área de investigación separada es que los sistemas formales de reputación online generan una gran cantidad de información que hasta ahora no estaba disponible para el investigador. Como resultado, el análisis económico de la reputación online es principalmente empírico, mientras que la literatura anterior sobre reputación era principalmente de naturaleza teórica. Por último, la necesidad del estudio de reputación online viene muy ligada a la importancia de los mercados online y del aumento en la actividad realizada en Internet (Tadelis, 2016).

Una de las características distintivas de la reputación online es la existencia de mecanismos formales de retroalimentación y reputación. Por ejemplo, el sistema de comentarios de eBay es particularmente importante, tanto por el valor en dólares que representa como por la cantidad de investigación que ha inducido. En función de los comentarios proporcionados por los agentes (compradores y vendedores), eBay muestra varios agregados correspondientes a la reputación de un vendedor, incluida la diferencia entre la cantidad de calificaciones de comentarios positivos y negativos; el porcentaje de calificaciones de retroalimentación positiva; la fecha en que el vendedor se registró en eBay; y un resumen de los comentarios más recientes recibidos por el vendedor. Finalmente, eBay proporciona un registro completo de los comentarios recibidos por cada vendedor, comenzando por los más recientes. Toda la información sobre cada vendedor está disponible públicamente; en particular, está disponible para cualquier comprador potencial. eBay no es el único sistema de retroalimentación y reputación en línea. Amazon.com ofrece un sistema de reseñas de clientes mediante el cual los compradores pueden calificar tanto el producto en sí y, si se compra a un vendedor que no sea Amazon, al vendedor. Aunque el sistema de revisión del vendedor de Amazon es bastante similar al de eBay, su sistema de revisión del producto es algo más complejo, ya que los revisores pueden calificar otras revisiones. El hecho de que

exista un sistema de reputación de retroalimentación en un mercado en línea determinado no implica que dicho sistema importe, es decir, que tenga algún significado práctico. En principio, es posible que los agentes (compradores y vendedores) ignoren los niveles de reputación que genera el sistema. Si ese fuera el caso, habría pocos incentivos para proporcionar retroalimentación, lo que a su vez justificaría que los agentes no se preocuparan por el sistema en primer lugar. En términos más generales, muchos, si no la mayoría, de los juegos de transmisión de información (como los sistemas de retroalimentación y calificación) admiten equilibrios en los que los agentes brindan retroalimentación y calificaciones de forma aleatoria (o simplemente no brindan retroalimentación) y, consistentemente, los agentes ignoran la información generada por el sistema (Cabral, 2012).

En Internet también tiene interés conocer la reputación de un sitio web para consumir información o entretenimiento. El algoritmo de reputación más relevante y utilizado también para otros casos de uso es el PageRank propuesto por Larry Page (Page *et al.*, 1999). Originalmente el PageRank fue un algoritmo usado por Google Search para ranquear las páginas web en los resultados de su motor de búsquedas. Actualmente, PageRank no es el único algoritmo que utiliza Google para ordenar los resultados de búsqueda, pero es el primer algoritmo que utilizó la empresa, y es el más conocido. PageRank funciona contando el número y la calidad de los enlaces a una página para estimar la importancia del sitio web (ver Fig. 3.1). La suposición subyacente es que es probable que los sitios web más importantes reciban más enlaces de otros sitios web. PageRank es un algoritmo de análisis de enlaces y asigna una ponderación numérica a cada elemento de un conjunto de documentos con hipervínculos, como la World Wide Web, con el fin de medir su importancia relativa dentro del conjunto. El algoritmo se puede aplicar a cualquier colección de entidades con citas y referencias recíprocas. El valor de ranking indica la importancia de una página en particular. Un hipervínculo a una página cuenta como un voto de apoyo. El PageRank de una página se define de forma recursiva y depende del número y la métrica de PageRank de todas las páginas que enlazan con ella (enlaces entrantes). Una página que está vinculada a muchas páginas con alto PageRank recibe un alto PageRank.

1.1.8. Reputación en redes sociales

El número de personas utilizando las redes sociales para compartir información, buscar información y para comunicarse con otras personas ha crecido de manera abrupta en los últimos años. Esta situación ha puesto el foco en el análisis de la reputación también en las redes sociales (Sabater y Sierra, 2002). Una red social es una estructura social entre actores, individuos u organizaciones. Las personas se conectan con otras personas más allá de lo geográfico y barreras de tiempo, disminuyendo los límites físicos y temporales creando nuevos lazos. Estos lazos pueden caracterizar cualquier tipo de relación, amistad, autoría, etc. Los servicios de microblogging se han desarrollado rápidamente como servicio emergente debido a su puntualidad y conveniencia. Este entorno social singular ha recibido

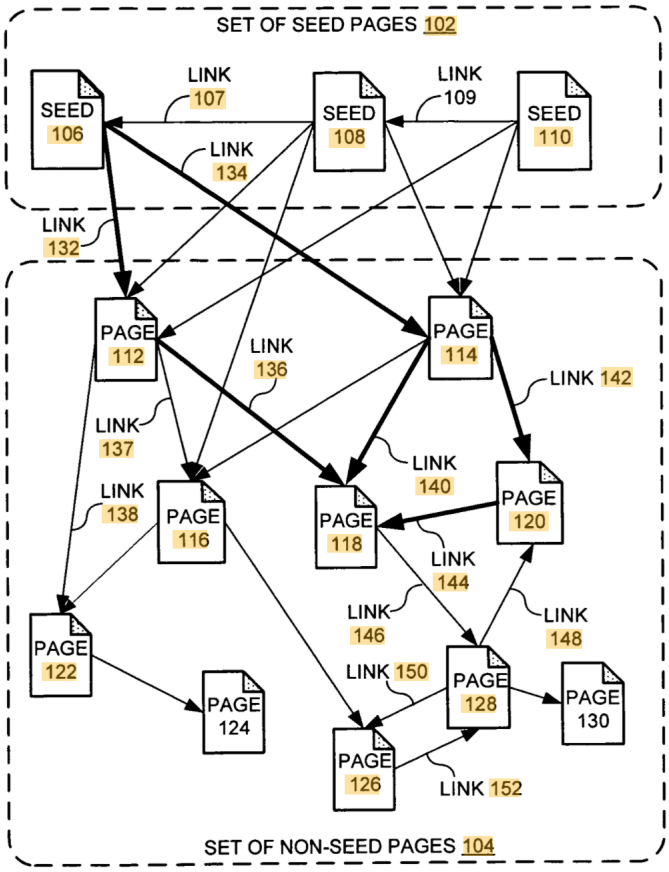

Figura 1.1: *Ejemplo de PageRank*

una atención considerable por parte de los investigadores académicos, ya que representa un nuevo medio para la búsqueda y difusión de información. El problema principal que plantean las redes sociales es que cualquiera puede escribir y publicar lo que quiera, de modo que la distribución de la calidad del contenido tiene una gran variación, desde elementos de muy alta calidad hasta artículos de baja calidad. Esto dificulta que los usuarios determinen la calidad de la información y la reputación de los otros usuarios, especialmente cuando no son personalidades o instituciones reconocidas. Por ejemplo, los usuarios de Twitter publican tweets y, si a otros usuarios les gusta o encuentran su contenido interesante, lo re-publican o retuitean. Retuitear es un mecanismo clave para la difusión de información en microblogging. Al permitir que los usuarios de Twitter reenvíen información que consideren interesante, importante o entretenida, el proceso de retweet se comporta como un sistema de recomendación informal. Por tanto, se considera que el usuario que ha sido retuiteado obtiene un reconocimiento y su reputación se ve aumentada (Weitzel *et al.*, 2013).

En los últimos años, la detección de "spam" en los sitios de redes sociales ha atraído la atención de los investigadores. La detección de "spam" es una tarea difícil para mantener la

seguridad de las redes sociales. Es esencial reconocer los mensajes no deseados en las redes sociales para proteger a los usuarios de varios tipos de ataques maliciosos y para preservar su seguridad y privacidad. Estas peligrosas maniobras adoptadas por los "spammers" causan una destrucción masiva de la comunidad en el mundo real. Los "spammers" de Twitter, por ejemplo, tienen varios objetivos, como difundir información no válida, noticias falsas o rumores. Los "spammers" logran sus objetivos maliciosos a través de anuncios y varios otros medios en los que admiten diferentes listas de correo y, posteriormente, envían mensajes de "spam" al azar para difundir sus intereses. Estas actividades causan molestias a los usuarios originales que se conocen como no "spammers". Además, también disminuye la reputación de las propias redes sociales (Masood *et al.*, 2019).

1.2. Objetivo

El objetivo de esta tesis es investigar y desarrollar métricas de reputación efectivas y confiables para la construcción de sistemas basados en el conocimiento. Para ello, se ha realizado una revisión de la literatura que investiga la reputación para distintas entidades dentro de la comunidad científica como la reputación del autor, del artículo, de las revistas científicas y de los centros universitarios; pero también en ámbitos externos a la comunidad científica como la reputación en empresas, en Internet o en redes sociales. También se ha realizado una revisión detallada del estado del arte sobre las métricas de reputación científicas (por ejemplo, índice h (Hirsch y Buela-Casal, 2014), índice i10 (Dhamdhere, 2018) o índice g (Egghe, 2006)) y se ha investigado sobre la relevancia y novedad, que tienen un impacto en el cálculo de la reputación. Por último, se ha realizado una revisión de sistema basados en el conocimiento (Akerkar y Sajja, 2010), sistemas de gestión del conocimiento (Alavi y Leidner, 2001), sistemas de recomendación (Resnick *et al.*, 2000), sistemas de análisis de los sentimientos (Scherer, 1984) y de minería de opinión (Pang *et al.*, 2008), sistemas de reputación y confianza (Jøsang y Golbeck, 2009) y sistemas de detección de riesgo (Abraham y Chengalur-Smith, 2010). Con este trabajo, se han identificado las limitaciones de las evaluaciones actuales de la reputación y se han definido y propuesto nuevas fórmulas para el cálculo de la reputación.

El alcance de este trabajo es experimental, por ello se han desarrollado tres frameworks para demostrar de forma empírica la eficacia y eficiencia de las nuevas métricas de reputación presentadas para sistemas basados en el conocimiento. Los dos primeros frameworks presentados se han diseñado dentro del ámbito de la investigación científica. El tercer framework se ha construido para el cálculo de la reputación en Internet, con la motivación de presentar conclusiones que apliquen a la reputación en sistemas basados en el conocimiento independientemente del ámbito de aplicación. Se han presentado diferentes experimentos en cada uno de los frameworks para cumplir con este objetivo. En estos frameworks se han utilizado distintas técnicas y métodos como la heurística,

modelos de aprendizaje máquina, sistemas de recomendación, procesamiento natural del lenguaje, análisis del sentimiento o clustering, que han permitido evaluar la capacidad de las nuevas métricas para mejorar la precisión, confiabilidad y escalabilidad de los sistemas. Además, se han comparado los resultados con modelos teóricos existentes para compartir conclusiones sobre la efectividad y viabilidad de las nuevas métricas propuestas.

1.3. Estructura del documento

El documento se ha organizado de la siguiente forma. El Capítulo 1 introduce el concepto de reputación y lo presenta en distintos ámbitos. El Capítulo 2 establece el contexto y describe los trabajos relacionados dentro del estado del arte. El Capítulo 3 presenta la arquitectura de los frameworks *Unified Knowledge Compiler (UNIKO)*, *Framework for Reputation Estimation of Scientific Authors (FRESA)* y *DOmains Classifier based on RIsky Websites (DOCRIW)*. El Capítulo 4 ilustra la viabilidad de la propuesta mostrando los experimentos realizados. El Capítulo 5 presenta las conclusiones generales e identifica las futuras líneas de investigación.

Capítulo 2

Estado del arte

La reputación como concepto ha sido ampliamente utilizado en temas comerciales. Por ejemplo, la colaboración en línea, los sitios web de comercio electrónico y las transacciones son algunos de los dominios más comunes. En este tipo de comercio, la reputación ayuda a reducir la incertidumbre entre pares, ya que un miembro de alta reputación será, en consecuencia, más confiable y seguro (Rufener, 2006). Los sistemas de reputación proporcionan información esencial para la confianza computacional como predicciones sobre el comportamiento futuro basadas en las acciones pasadas de un compañero (Baxter y Vogt, 2002). En este mundo cada vez más interconectado, la necesidad de reputación se vuelve más importante a medida que aumenta el número de personas y servicios que interactúan en línea. La reputación es una herramienta para facilitar la confianza entre entidades, ya que aumenta la eficiencia y eficacia de los servicios y comunidades online. Como la mayoría de las entidades no tendrán ninguna experiencia directa con otras entidades, deben confiar cada vez más en los sistemas de reputación. Dichos sistemas permiten estimar quién es probable que sea confiable en función de los comentarios de transacciones pasadas (Lanubile *et al.*, 2010).

En Online Rating Systems (Rosenberg, 2001), la reputación del usuario es fundamental para detectar a los "spammers" y eliminar sus calificaciones. El efecto de dichos "spammers" puede mitigarse mediante un sistema de reputación que pueda estimar la credibilidad de los usuarios en función de su comportamiento. Algunos sistemas de reputación detectan a los "spammers" en función de la desviación entre sus calificaciones y la calificación real del elemento. Esta desviación puede ser generalmente estimada utilizando una función de penalización (Zhou *et al.*, 2005).

Por otro lado, medir la influencia y la reputación de los investigadores académicos ha sido un tema desafiante que ha llamado mucho la atención de la comunidad científica. En este sentido, se han propuesto y se siguen proponiendo muchos indicadores bibliométricos, la mayoría basados en citas de artículos y sus variaciones. En el resto de este capítulo explicaremos las medidas de reputación científicas utilizadas hasta ahora, así co-

mo los sistemas utilizados en base a la reputación como detección de riesgo, de análisis de sentimiento, de recomendación o de conocimiento.

2.1. Medidas de reputación científicas

La reputación científica puede definirse como el grado de confianza y reconocimiento otorgados a un investigador científico por su trabajo. Este reconocimiento debe reducirse a un número determinado para poder compararlo con otros y construir una métrica válida para evaluar a los investigadores. A lo largo de los años, se han utilizado diferentes métricas de calidad en la comunidad científica para evaluar a los investigadores científicos. Las principales métricas son:

- Número total de citas: es un indicador bibliométrico que refleja la relevancia de un artículo. Es útil para detectar autores con una carrera larga y relevante. No obstante, penaliza a los autores jóvenes, ya que un autor con pocos artículos de gran relevancia puede tener un número de citas muy inferior al de un autor con muchos artículos irrelevantes.

- Factor de impacto (Garfield *et al.*, 1994): mide la importancia de una publicación científica. Es calculado cada año por el *Institute of Science Information* (ISI) para aquellas publicaciones del JCR. El factor de impacto de un año se calcula como la relación de A a B, donde A es el número de citas en el año en curso de cualquier publicación en una determinada revista de los dos años anteriores; y B denota el número de artículos publicados en la revista en los dos años anteriores. Es una métrica muy útil para medir la relevancia de los últimos trabajos de un autor, sin embargo, presenta dos problemas que limitan su uso: su cálculo se basa en el número de citas, por lo que no mide realmente la calidad de la publicación, y el periodo de cálculo base es muy corto, ya que los artículos clásicos se citan con frecuencia incluso después de décadas, de modo que no sirve para medir la reputación de los autores comparando vidas científicas.

- Índice h (Hirsch y Buela-Casal, 2014): mide la calidad profesional de los investigadores, considerando tanto el número de publicaciones del investigador como el número total de citas de esos trabajos. Dado un conjunto de publicaciones científicas de un investigador clasificadas en orden decreciente de número de citas, el índice h es el número en el que el número ordinal coincide con el número de citas. Es la métrica más utilizada y popular porque el cálculo es muy sencillo y ofrece una estimación de la reputación del autor. El principal inconveniente del índice h es que penaliza a los investigadores con una carrera corta pero brillante, ya que al estar limitados por el número de publicaciones no tendrán un índice h alto.

- Índice i10 (Dhamdhere, 2018): es una métrica de Google disponible desde 2011 que indica el número de trabajos académicos con al menos 10 citas que ha publicado un autor. Es una métrica que pretende mejorar el índice h evitando recompensar la productividad dejando fuera del cálculo todos los artículos con muy poca relevancia. No obstante, como el índice h y el factor de impacto, sigue valorando la productividad por encima de la calidad del trabajo.

- Índice g (Egghe, 2006): es una métrica presentada para mejorar el índice h. Se calcula de la siguiente manera: dado un conjunto de artículos ordenados en orden decreciente de número de citas, el índice g es el número más alto que cumple que estos g artículos han recibido en total al menos g^2 citas. Esta métrica evita recompensar la productividad ya que a los artículos con más citas se les da más valor.

Finalmente, en los últimos años se han sugerido nuevas medidas que contemplan el impacto y la difusión de la obra en Internet. Se denominan genéricamente "altmetrics" (Priem *et al.*, 2011). Una medida bien conocida de este tipo es un indicador bibliométrico novedoso llamado "citing authors index" (Cappelletti-Montano *et al.*, 2021).

En esta tesis se proponen dos índices de reputación nuevos, el índice *Unified Knowledge Compiler* (UNIKO) y el índice *Framework for Reputation Estimation of Scientific Authors* (FRESA), que serán presentados en el Capítulo 3. UNIKO pretende compensar los problemas que introducen métricas como el índice h, el índice g o el índice i10, mitigando la influencia de la productividad del autor. Esta métrica se calcula como una suma ponderada del número de citas influyentes, el número de citas totales, la antigüedad y el número de artículos publicados. Sin embargo, esta métrica no impide que se penalice a autores de gran relevancia y con una corta trayectoria. FRESA pretende compensar los problemas que no alcanza a resolver el índice UNIKO, introduciendo el concepto de novedad de las publicaciones y haciendo un cálculo de trayectorias de reputación de los autores teniendo en cuenta la relevancia para que la longitud de la trayectoria científica del autor no tenga influencia en su reputación.

2.2. Relevancia

El concepto de relevancia fue reconocido por primera vez a finales de la década de 1950 para dar solución a un concepto que se venía tratando desde el siglo XVII con la publicación de las primeras revistas científicas (Mizzaro, 1997). En el caso del contenido textual sin procesar, uno de los métodos más importantes para medir la relevancia es el algoritmo PageRank (Page *et al.*, 1999). Es capaz de medir la importancia de un texto generando a partir de él una red conceptual y luego contando el número y calidad de los enlaces (relaciones) de un nodo (generalmente un concepto).

En el contexto científico, la relevancia se evalúa utilizando medidas cuantitativas. Existen varios indicadores para medir la relevancia científica de un autor, un artículo o una revista. La idea básica es que los artículos muy relevantes se citarán con más frecuencia que los artículos menos relevantes. El “acoplamiento de bibliografía"(Kessler, 1963) es una medida bien conocida para evaluar la similitud entre los documentos y el número de citas recientes, considerando este número de citas recientes de un artículo científico en un cierto periodo de tiempo. En (Guerrero-Sosas *et al.*, 2019) se propone un sistema de recomendación de relevancia científica para investigadores utilizando un modelo ontológico descrito previamente (Sosa *et al.*, 2019). Además, para medir la relevancia de los investigadores, se han propuesto varios indicadores. Así, las métricas tradicionales como el Scimago Journal Ranking (Falagas *et al.*, 2008) y el número de citas de la publicación suelen combinarse con otras características dependientes del contexto, como el tipo de publicación o la ciudad y el país de la institución de investigación.

2.3. Novedad

La detección de novedades es la tarea de identificar información novedosa dado un conjunto de antecedentes ya acumulados. Las aplicaciones potenciales de los sistemas de detección de novedades son abundantes, dada la sobrecarga de información en todos los dominios (Gamon, 2006). Se han propuesto varios métodos para determinar información novedosa a partir de documentos de texto sin procesar como la minería de texto (He *et al.*, 2008). Para establecer si la información dentro de una oración se ha visto en el material leído previamente, algunos enfoques integran la información sobre el contexto de la oración con palabras nuevas y entidades nombradas dentro de la oración y usan un algoritmo de aprendizaje especializado para ajustar los parámetros del sistema (Schiffman y McKeown, 2005). El reconocimiento de entidades nombradas y el etiquetado de partes del discurso han logrado excelentes resultados comparando varios tipos de entidades dentro de oraciones (Ng *et al.*, 2007). En el caso de la comunidad científica, los científicos están bajo una presión cada vez mayor para producir investigaciones novedosas. Se han realizado estudios de revisión de la sociología de la ciencia para anticipar cómo el énfasis en la novedad podría afectar la estructura y función de la comunidad científica (B. A. Cohen, 2017). Se concluye que es necesario encontrar un equilibrio entre la novedad y la reproducibilidad (Holding, 2019). Además, el impacto tecnológico no siempre se define por la novedad de la investigación (Veugelers y Wang, 2019).

En la literatura se han propuesto diferentes indicadores para detectar, analizar y evaluar la novedad de la investigación científica. Ejemplos de este tipo de enfoques son (Wang *et al.*, 2017) y (Carayol *et al.*, 2019). El primero desarrolló un indicador de novedad para evaluar publicaciones científicas enunciando que la novedad de un artículo se mide como el número de nuevos pares de revistas en sus referencias ponderadas por la similitud

del coseno entre las revistas recién emparejadas. El segundo proponía un indicador de novedad basado en la frecuencia de combinaciones por pares de palabras clave de autor aplicadas a todas las publicaciones de investigación durante diferentes años. Por lo tanto, dado un año y un campo de investigación, los artículos pueden clasificarse en función de su puntuación de novedad, donde la propuesta con la combinación de palabras clave por pares más infrecuente obtiene puntuaciones altas capturando su originalidad.

2.4. Sistemas de gestión del conocimiento

Los Sistemas de gestión del conocimiento o *Knowledge Management Systems* (KMS) (Alavi y Leidner, 2001) son sistemas diseñados específicamente para recopilar, usar y aplicar el conocimiento organizacional. Este conocimiento generalmente se almacena en una fuente de información (por ejemplo, bases de datos o administradores de archivos). *Los KMS tienen como objetivo principal proporcionar conocimiento e información almacenada previamente para ganar influencia en las actividades presentes.* Así, pueden producir niveles de eficacia organizacional que se incrementan de acuerdo con la adquisición de nuevos conocimientos. En cuanto a los campos de aplicación de los KMS, se podrían organizar en Intranet Infrastructures (Swan *et al.*, 1999), Document and Content Management Systems (Rufener, 2006), Collaborative Software (también llamado Groupware (Lanubile *et al.*, 2010)) y E-Learning Systems (Rosenberg, 2001).

Las Intranet Infrastructures pueden adoptarse para ser la plataforma principal de un grupo colaborativo donde el conocimiento sensible debe mantenerse y compartirse (actuando como un repositorio) solo para los miembros. Conduce a mejorar la colaboración y la socialización entre los usuarios.

Los Document and Content Management Systems se centran en hacer que el contenido (en particular, los documentos) esté disponible para los usuarios y aborde una amplia gama de temas relacionados. La suposición principal es que el usuario sabe qué contenido relacionado con el tema desea recuperar. Por lo general, se basa en una gran cantidad de metadatos para organizar y evaluar el contenido. Por lo tanto, se logra una búsqueda y recuperación de contenido más eficiente.

El Collaborative Software o Groupware permite colaborar entre múltiples grupos de desarrollo generando entornos de trabajo virtuales útiles para diferentes tareas. Suelen ser sistemas de propósito general. Profundizando en sus funcionalidades, gestionan las posibles inconsistencias y versiones del desarrollo durante las distintas etapas del proceso. También proporcionan un canal de comunicación seguro entre los usuarios. Ejemplos de estos sistemas están relacionados con la detección de intrusos (Zhou *et al.*, 2005), la biomedicina (Gómez *et al.*, 1998) o la gestión de emergencias (Qin *et al.*, 2012). Mención especial a Blockchain, que es un paradigma que engloba un conjunto de tecnologías (por

ejemplo, Peer to Peer (P2P) (Balakrishnan *et al.*, 2003) o métodos criptográficos (Mao, 2003)), y registros descentralizados y sincronizados en lugar de bases de datos.

Los E-Learning Systems están enfocados principalmente en ámbitos formativos y educativos. Son capaces de proporcionar conocimiento estructurado recopilado de diferentes fuentes de información. Este conocimiento se presenta de forma simplificada para estimular y facilitar el proceso de aprendizaje. El objetivo principal de estos sistemas es apoyar a los usuarios para realizar mejoras y avances en habilidades específicas. Ejemplos de estos sistemas son los sistemas de recomendación orientados a un área determinada (Zaíane, 2002) o los juegos educativos (Torrente *et al.*, 2010).

El framework UNIKO puede clasificarse como un Software Colaborativo donde los usuarios pueden establecer relaciones entre ellos a través de un canal de comunicación (es decir, un foro), y como un Sistema de Gestión de Contenidos. Este último está directamente relacionado con el conocimiento organizado recopilado por el sistema a partir de fuentes virtuales de información (por ejemplo, web scraping (Haddaway, 2015)).

2.5. Sistemas basados en el conocimiento

Los Sistemas basados en el conocimiento o *Knowledge-Based Systems* (KBS) (Akerkar y Sajja, 2010) son sistemas que utilizan *Artificial Intelligence* (AI) para resolver tareas generales. La arquitectura de estos sistemas suele comprender un componente de almacenamiento (por ejemplo, una base de datos) para facilitar la recuperación del conocimiento en respuesta a posibles solicitudes. Esta arquitectura se completa con diferentes módulos para atender las necesidades de los usuarios o para optimizar el sistema. Ejemplos de estos módulos son la interfaz de visualización y el conjunto de técnicas de *Machine Learning* (ML) (Witten *et al.*, 2016) presentadas en el Capítulo 3. Un KBS podría ser considerado como un sistema informático que genera conocimiento a partir de datos, información o el propio conocimiento. Estos sistemas son capaces de comprender la información tomando decisiones en consecuencia. Los enfoques KBS más importantes podrían clasificarse en:

- Sistemas de lógica difusa (Yager y Zadeh, 2012), también conocidos como *Fuzzy Logic Systems* (FLS). Son una extensión de los sistemas lógicos clásicos que proporcionan soluciones de representación del conocimiento en entornos donde la información es imprecisa o incierta. Para lograr esta representación, los sistemas de lógica difusa utilizan la semántica de puntaje de prueba, lo que les permite manejar entidades imprecisas como conjuntos de restricciones elásticas. En el caso de operaciones de razonamiento, estos sistemas propagan estas restricciones elásticas para generar comportamientos específicos. Por lo tanto, los sistemas de lógica difusa suelen incluir un conjunto de reglas que actúan como guías para generar comportamientos específicos. Los sistemas de lógica difusa se han aplicado ampliamente

en dominios como energía, combustibles (Singh y Markeset, 2009) y simulaciones de comportamientos complejos (Gokulan y Srinivasan, 2010). En estos campos, los FLS permiten modelar y controlar sistemas en los que las variables de entrada y salida son difíciles de medir o se expresan en términos lingüísticos. Por ejemplo, un FLS puede ser utilizado para controlar la temperatura en un horno industrial, basándose en variables como la "sensación de calor.º la ïntensidad del fuego.en lugar de medidas precisas de temperatura. Esta capacidad para manejar la imprecisión y la incertidumbre hace que los sistemas de lógica difusa sean una herramienta valiosa en una amplia gama de aplicaciones.

- Sistemas basados en agentes (Jennings, 2001) o *Agent-Based Systems* (ABS). Estos utilizan agentes inteligentes para recopilar, evaluar y procesar el conocimiento. Estos agentes pueden definirse como abstracciones de entidades autónomas. Son capaces de observar el entorno circundante e interactuar con él y con los demás individuos (es decir, agentes). Los agentes generalmente persiguen un objetivo principal que es su motivación para lograr tareas en el sistema. Para facilitar el procesamiento de una gran cantidad de conocimiento, los ABS se enfocan en arquitecturas distribuidas. Por esta razón, los ABS se suelen utilizar para resolver tareas complejas que requieren múltiples operaciones que pueden descomponerse en otras más simples (por ejemplo, inteligencia de enjambre (Luna y Stefansson, 2012)) o para simular comportamientos complejos de individuos (Wilensky y Rand, 2015). Estos sistemas están enfocados en medicina (Foster *et al.*, 2005), simulaciones de tránsito (Fernández-Isabel y Fuentes-Fernández, 2017) y simulaciones económicas (Luna y Stefansson, 2012).

- Sistemas de redes neuronales (LeCun *et al.*, 2015) o *Neural Networks Systems* (NNS). Estos incluyen técnicas de ML (generalmente aprendizaje supervisado (Hastie *et al.*, 2009)), que pueden identificar e interpretar características relevantes de los datos. Una rama actual de estos sistemas es *Deep Learning* (DL) (por ejemplo, (Schmidhuber, 2015)). DL utiliza redes neuronales para construir estructuras neuronales complejas que se organizan de manera similar al sistema nervioso de los mamíferos. Por lo tanto, estos sistemas DL pueden incluir partes específicas de la red global que están especializadas en la detección de características específicas de datos ocultos. La Visión Artificial (Ciregan *et al.*, 2012) y el *Natural Language Processing* (NLP) (Socher *et al.*, 2011) son unos de los alcances más extendidos de estos sistemas (Socher, 2014).

- Algoritmos genéticos (Freitas, 2003). Los sistemas que incluyen algoritmos genéticos para procesar el conocimiento utilizan una función de ajuste que proporciona la solución perfecta. Generan múltiples operaciones inspiradas en la reproducción biológica (dos individuos producen una descendencia que comparte su código genético, y por tanto características específicas de los ancestros) de los seres vivos y las mutaciones que pueden ocurrir durante este proceso. Luego, solo las entidades

que son más eficientes para abordar la tarea deseada que sus entidades antepasadas son seleccionadas para nuevas tareas de reproducción (es decir, selección natural (Forrest *et al.*, 1993)). El principal inconveniente de este tipo de sistemas es el elevado tiempo de cómputo necesario para completar la tarea selectiva (es decir, hasta obtener una entidad suficientemente factible). Los sistemas de algoritmos genéticos se utilizan generalmente en campos como la minería de datos y el descubrimiento del conocimiento (Freitas, 2003), y también para cuestiones de seguridad (Li, 2004), entre otros.

- Sistemas de soft computing (Pal *et al.*, 2002). Estos sistemas hacen uso de resultados imprecisos para tareas computacionales difíciles (por ejemplo, problemas NP-completos (Rosenkrantz y Stearns, 2003)) donde no es posible lograr una solución exacta en tiempo polinomial. Estos sistemas se diferencian del software convencional (duro) en que se adaptan bien a la imprecisión, la incertidumbre y la verdad parcial. Algunos de los campos de aplicación más importantes de estos sistemas son las cuestiones predictivas (por ejemplo, las predicciones de rendimiento bancario (Ravi *et al.*, 2008)) y los sistemas de control inteligente (por ejemplo, el ahorro de energía (Dounis *et al.*, 2011)).

Los tres frameworks presentados en esta tesis son sistemas híbridos que incluyen algunas de las características presentadas anteriormente. Todos ellos utilizan agentes inteligentes (arañas) para recopilar y procesar información de los sitios web. En el framework UNIKO y en *Domains Classifier based on Risky Websites* (DOCRIW), se han incluido múltiples técnicas de ML. Además, en UNIKO se ha integrado un componente DL en el módulo de análisis de sentimiento y para seleccionar los autores más adecuados cuando sus nombres no coinciden completamente con investigadores de renombre, se han propuesto algunas técnicas asociadas al Soft Computing.

2.6. Sistemas de reputación y sistemas de recomendación

Los sistemas de reputación (Resnick *et al.*, 2000) están orientados a permitir a los usuarios a calificar un elemento o persona en concreto. Esto lleva a crear comunidades en línea que construyen un nivel de confianza asociado al individuo o elemento que se está evaluando. El objetivo principal de los sistemas de reputación es proporcionar información sobre temas específicos a los usuarios para generar la confianza suficiente en ellos. Así, se construyen sistemas como el de Gestión de la Reputación en Línea (Dolle, 2014), que contiene productos de mayor calidad. La Gestión de la Reputación On-line está enfocada a la gestión de los resultados de búsqueda de productos y servicios en Internet. Es una herramienta muy popular en los mercados digitales y las comunidades en línea, ya que

puede controlar de manera efectiva los nodos que pueden minimizar la amenaza y proteger el sistema de posibles usos indebidos y abusos por parte de los usuarios.

Los sistemas de recomendación o *Recommender Systems* (RS) (Trewin, 2000) son un tipo de sistema de información que utiliza técnicas de inteligencia artificial y aprendizaje automático para analizar los datos de usuarios y ofrecer recomendaciones personalizadas y relevantes para ellos. Los RS pueden aprovechar los sistemas de reputación o ser agnósticos a las calificaciones del usuario. El propósito principal del RS es guiar a los usuarios aplicando un proceso personalizado para filtrar los objetos más interesantes en un conjunto de opciones posibles. Un tipo específico de sistemas RS son los sistemas de recomendación basados en contenido (Lops *et al.*, 2011). Presentan habilidades de memoria para recomendar elementos que comparten características específicas con otros previamente aprobados por el usuario. El proceso de almacenamiento de las características en memoria consiste en hacer coincidir los atributos que presenta un perfil de usuario. Así, el sistema se centra especialmente en las preferencias, intereses y también en la reputación para elaborar una plantilla de cada usuario. Esta plantilla se actualiza de acuerdo a los nuevos movimientos y nivel de confianza adquirido por el usuario a través del tiempo. Estos sistemas son ampliamente utilizados en plataformas de comercio electrónico, streaming de contenido, redes sociales y aplicaciones de viajes, entre otros. Además, se han convertido en una herramienta valiosa para las empresas, ya que pueden mejorar la experiencia del usuario, aumentar las ventas y mejorar la fidelidad del cliente.

Los tres frameworks presentados el Trabajo de Tesis definen la reputación de artículos, autores y sitios web, respectivamente. En base a estas reputaciones se puede ofrecer un sistema de recomendación según el nivel de confianza deseado. El framework UNIKO incluye un sistema de recomendaciones (con un nivel de certeza) que recomienda a los usuarios sobre los autores o artículos más apropiados.

2.7. Sistemas de análisis de sentimiento y de minería de opinión

Los sentimientos son puntos de vista u opiniones que los humanos expresan comúnmente en lenguaje natural. Los sentimientos se pueden clasificar según las emociones básicas (Scherer, 1984): ira, miedo, alegría, repulsión, tristeza y sorpresa. Esta organización tiene sus carencias en cuanto a posibles dificultades para procesar y clasificar los datos de entrada. Por este motivo, se han adoptado ampliamente representaciones basadas en la polaridad (positiva, negativa o neutra). Estas representaciones también pueden incluir una escala numérica (discreta o continua) para mayor precisión (Pang *et al.*, 2008). Los sistemas de análisis de sentimiento y minería de opinión podrían clasificarse de diferentes maneras. Centrándonos en los enfoques más extendidos, se podrían organizar en:

- Sistemas de diccionarios o léxicos. Estos utilizan una bolsa de palabras preconstruida

con una polaridad asociada (escala continua de sentimientos) para cada palabra. SentiWordNet (Baccianella *et al.*, 2010) es un ejemplo de diccionario. SentiWordNet es una base de datos que asocia a cada palabra un valor numérico que indica su grado de positividad, negatividad o neutralidad. Este recurso léxico es útil para procesamiento de lenguaje natural y análisis de sentimiento, ya que permite a los algoritmos entender el tono emocional de un texto. SentiWordNet se basa en el WordNet, un diccionario de inglés que organiza las palabras en grupos de sinónimos llamados "synsets". Cada synset contiene una lista de palabras relacionadas que tienen un significado similar. SentiWordNet asigna un valor de sentimiento a cada synset, basado en la intensidad de las emociones asociadas con las palabras en ese synset. Para obtener la polaridad de un texto completo, es necesario calcular el promedio de polaridades de las palabras en el texto.

- Enfoques de aprendizaje automático ML. En estos enfoques se crea un modelo utilizando instancias de entrenamiento previamente seleccionadas. Luego, el modelo se utiliza para clasificar las palabras. Los sistemas DL también se incluyen en este enfoque. Ejemplos de ellos son (Chikersal *et al.*, 2015), que es un enfoque tradicional de ML, y (Tang *et al.*, 2014) donde se han incluido las redes neuronales.

- Sistemas semánticos (Baldoni *et al.*, 2012). Se centran en calcular la similitud entre palabras. Sus fundamentos se basan en la premisa de que dos palabras con significados semánticos similares deben presentar un valor de polaridad equivalente. Esto permite predecir los valores de sentimiento de palabras mal clasificadas, identificando cualquiera de sus sinónimos. Ejemplos de estos sistemas son los enfoques basados en ontologías (Baldoni *et al.*, 2012).

- Enfoques estadísticos (He *et al.*, 2008). Obtienen la polaridad de sentimiento de una palabra analizando las otras palabras que aparecen cerca de ella en el texto. Si algunas de estas otras palabras se han identificado previamente como positivas o negativas, entonces la nueva palabra podría clasificarse correctamente. El principal inconveniente de estos enfoques consiste en esta dependencia, pudiendo cometer errores asociativos debido a un contexto específico. Sin embargo, son enfoques muy útiles para obtener nuevos valores basados en el contexto para palabras mal clasificadas. Un ejemplo de estos sistemas es (He *et al.*, 2008).

En particular, el framework UNIKO presenta un módulo híbrido para el análisis de sentimientos. El módulo se basa en el flujo de trabajo presentado en (Cambria, 2016). Por lo tanto, utiliza un diccionario (SentiWordNet) para obtener valores de sentimiento de las palabras coincidentes. Posteriormente, incluye un componente DL que proporciona un modelo previamente entrenado para predecir el valor de sentimiento de palabras no identificadas.

2.8. Sistemas de confianza y reputación

Los sistemas de confianza y reputación o *Trust and Reputation Systems* (TRS) representan una clase importante de herramientas de apoyo a la toma de decisiones que pueden ayudar a reducir el riesgo al participar en transacciones e interacciones en Internet. Desde el punto de vista de la parte de confianza individual, un TRS puede ayudar a reducir el riesgo asociado con cualquier interacción en particular. Desde el punto de vista del proveedor de servicios, representa una herramienta de marketing. Desde el punto de vista comunitario, representa un mecanismo para moderación y control social, así como un método para mejorar la calidad de los mercados en línea y comunidades.

Los TRS en línea también utilizan los mismos principios básicos para la creación y propagación de la confianza y la reputación en las comunidades tradicionales. La principal diferencia es que la confianza y la formación de la reputación en las comunidades tradicionales suele ser relativamente ineficiente y depende en la comunicación física (por ejemplo, a través del boca a boca), mientras que los TRS en línea están respaldados por redes y sistemas informáticos extremadamente eficientes. En teoría, es posible diseñar una gestión muy eficaz de la confianza y la reputación en las comunidades en línea, pero la fiabilidad de las puntuaciones calculadas de confianza y reputación y, por lo tanto, la utilidad del propio TRS, también depende de la robustez del TRS en cuestión.

Los intentos de tergiversar la confiabilidad y manipular la reputación son comunes en las comunidades tradicionales. Los estafadores emplean métodos para parecer dignos de confianza (por ejemplo, a través de la fabricación y presentación de credenciales falsas). Tipos similares de ataques también se aplicarían a las comunidades en línea. En caso de que se utilice alguna forma de TRS para moderar una comunidad o mercado en línea, las vulnerabilidades en el propio TRS pueden abrir vectores de ataque adicionales. Por lo tanto, es crucial que los TRS sean robustos contra los ataques que podrían conducir a puntuaciones engañosas de confianza y reputación. En el peor de los casos, un TRS vulnerable podría ser dado la vuelta y utilizado como una herramienta de ataque para manipular maliciosamente el cómputo y difusión de las puntuaciones. La consecuencia de esto podría ser una pérdida total de la confianza de la comunidad causada por la incapacidad de sancionar y evitar servicios engañosos y de baja calidad.

Cuando se producen ataques contra un TRS, normalmente no significa que un servidor que aloja funciones TRS está siendo pirateado. Los ataques a los TRS generalmente consisten en desempeñar el papel de confiar partes y/o entidades de servicio, y de manipular el TRS a través de comportamientos específicos que es contrario a un supuesto comportamiento fiel. Por ejemplo, una parte que confía o que es idéntica a la entidad de servicio podría proporcionar calificaciones positivas falsas o injustas al TRS con el propósito de inflar el puntaje de la entidad de servicio, lo que a su vez aumentaría la probabilidad de que esa entidad de servicio sea seleccionada por otras partes que confían.

Se pueden imaginar muchos otros escenarios de ataque que, de tener éxito, darían lugar a ventajas para los atacantes. Todos estos ataques tienen en común que resultan en la erosión de la confianza de la comunidad que, a su vez, sería perjudicial para los servicios y las aplicaciones en el mercado o la comunidad afectada. Por lo tanto, la robustez de TRS puede ser crucial para el mercado o comunidad donde se está aplicando la TRS (Jøsang y Golbeck, 2009).

2.9. Sistemas de detección de riesgo

Los organismos públicos que persiguen sitios web fraudulentos y maliciosos dedican una cantidad significativa de tiempo y recursos a detectar estafas y malware en Internet (OSI, 2011). La mayor parte de este trabajo suele ser manual, lo que se traduce en esfuerzos duros e ineficientes. Por esta razón, se ha vuelto fundamental desarrollar sistemas capaces de automatizar la clasificación de sitios web en potencialmente riesgosos o no riesgosos según las características de estos sitios. En este contexto, un sitio web de riesgo es aquel con contenido malicioso, inseguro o fraudulento con intenciones peligrosas contra sus visitantes (Abraham y Chengalur-Smith, 2010).

El estudio del *Observatorio Español de Seguridad de la Información* (OSI) recoge la magnitud de los problemas de las webs de riesgo en España (OEDI, 2018). Entre los principales resultados y conclusiones del estudio, cabe destacar que un 53,1 % de los internautas españoles afirma haber sido víctima de un intento (no necesariamente consumado) de fraude en los últimos tres meses. Del análisis de situaciones potencialmente fraudulentas ocurridas a los usuarios durante la navegación por Internet destaca la recepción de invitaciones para visitar alguna web sospechosa (34,4 %). En el periodo analizado, el 95,2 % de los internautas españoles comparte que no ha sufrido perjuicio económico en los últimos tres meses como consecuencia de un fraude a través de Internet, mientras que el 4,8 % ha sufrido pérdidas. Además, el análisis empírico de los equipos muestra que el 39,8 % de los equipos alberga algún tipo de troyano, el 6,8 % alberga troyanos bancarios (fragmentos de código maliciosos destinados a interceptar credenciales de banca electrónica de entidades concretas) y un 5,8 % sufre una infección de rogue-software (o antivirus falso). Además, el 81,8 % de los internautas que han sufrido un incidente de este tipo no han modificado sus hábitos de navegación, frente al 5 % que ha abandonado esta actividad y el 13,2 % que ha reducido el uso de Internet. El *Observatorio Español de Delitos Informáticos* (OEDI) informó de 110.613 ciberdelitos en España en 2018, el 74 % de ellos siendo fraude (Akerkar y Sajja, 2010).

Los sitios web de riesgo son propensos a distribuir diferentes tipos de malware, técnicas de fraude y phishing, y otras formas de actos de ciberdelincuencia (Abraham y Chengalur-Smith, 2010). El malware y el fraude han sido explotados como un problema muy común en la sociedad actual provocando grandes perjuicios económicos a particulares y empresas.

Las heurísticas han sido la forma tradicional de luchar contra estas prácticas nocivas. Sin embargo, este tipo de análisis ya no se considera eficaz porque los casos de fraude pueden ser similares en apariencia y contenido, pero por lo general no son idénticos. El fraude es un delito adaptativo, por lo que necesita métodos especiales de análisis inteligente de datos para detectarlo y prevenirlo. Por lo tanto, se han desarrollado métodos automatizados para detectar estas amenazas. La solución más típica para la detección de malware se basa en el comportamiento. El análisis incluye reproducir el malware en un entorno emulado para generar informes de comportamiento (Ma *et al.*, 2009). Los métodos importantes para detectar y prevenir el fraude se basan en la red. Los módulos de detección de phishing detectan ataques de fraude al determinar que un dominio es similar a un dominio de phishing conocido, o que una dirección del recurso basado en la red desde el cual se recibe el contenido tiene propiedades de red sospechosas (Chiba *et al.*, 2012).

Sin embargo, estas soluciones tienen importantes inconvenientes. Téngase en cuenta que es necesario reproducir el malware en un entorno virtual o mostrar el contenido de una *Uniform Resource Locator* (URL). Por lo tanto, lograr buenos resultados implica altos costos tanto en tiempo como en recursos. El enfoque presentado en esta tesis en el framework DOCRIW, ampliamente presentado en el Capítulo 3 en la Sección 3, enfrenta este problema al simplificar todo el proceso. La única entrada que exigirá el sistema propuesto para determinar si un sitio web es potencialmente riesgoso es el nombre de dominio y sus características relacionadas. En esta línea, existen estudios similares que utilizan el dominio para detectar sitios web maliciosos. Sin embargo, estos estudios generalmente usan características textuales (Forbes *et al.*, 2011) o usan un enfoque de dirección IP (Devaki *et al.*, 2014). Otras alternativas utilizan las características del *Domains Name System* (DNS) (Ngai *et al.*, 2011) y Whois (Hilas, 2009), que se parecen más a la propuesta presentada. La principal diferencia radica en la tarea de clasificación, ya que se lleva a cabo utilizando tanto variables léxicas como basadas en el host. Adicionalmente, otra contribución diferencial es el uso de medidas de similaridad y métodos de ensamblaje para la clasificación óptima. DOCRIW utiliza técnicas de ML para mejorar el rendimiento de los clasificadores binarios de acuerdo con diferentes métricas de evaluación. Para la detección de sitios web de riesgo se han utilizado tradicionalmente técnicas de análisis estadístico de datos. Ejemplos de estas técnicas de análisis de datos estadísticos son: cálculo de parámetros estadísticos tales como distribución de probabilidad y cuantiles (Forbes *et al.*, 2011), análisis de series de tiempo (Devaki *et al.*, 2014), agrupamiento para encontrar patrones entre conjuntos de datos, coincidencia de datos utilizada para comparar dos conjuntos de datos o análisis de regresión para detectar relaciones entre variables de interés (Ngai *et al.*, 2011). Sin embargo, recientemente han aparecido técnicas de AI más avanzadas: sistemas expertos para detectar fraudes en forma de reglas (Hilas, 2009), reconocimiento de patrones para aproximar clases o patrones de comportamiento sospechoso (Bolton y Hand, 2002), ML para detectar automáticamente características de riesgo (Shabtai *et al.*, 2009), redes neuronales que pueden aprender patrones sospechosos a partir de los datos

(Aleskerov *et al.*, 1997), optimización de máquinas de aprendizaje extremas ponderadas para la clasificación desequilibrada en la detección de fraudes con tarjetas de crédito (Zhu *et al.*, 2010), detección de fraudes de transacciones basada en la relación total de pedidos y diversidad de comportamientos (M. Zhang *et al.*, 2008), modelos de detección de fallas en línea y estrategias basadas en nubes (X. Zhang *et al.*, 2017), y aprendizaje profundo de representación con pérdida total del centro para la detección de fraudes con tarjetas de crédito (Li, 2004). Hay varias técnicas de ML utilizadas en el contexto de la detección de sitios web riesgosos. En (Zhou *et al.*, 2005) se presenta una encuesta exhaustiva y una comprensión estructural de las técnicas de detección de URL maliciosas utilizando ML. Entre las técnicas más comunes en este campo se encuentran *Support Vector Machine* (SVM) (Cortes y Vapnik, 1995), *Logistic Regression* (LR) (Ma *et al.*, 2009), *Naïve Bayes* (NB) y *Decision Tree* (DT) (Ma *et al.*, 2009). En (Singh y Markeset, 2009) se evaluó un conjunto de modelos ML para clasificar sitios web maliciosos dada su URL como entrada. Además, se ha propuesto un método ML basado en SVM para clasificar sitios web maliciosos utilizando solo nombres de dominio (Dhamdhere, 2018).

En los últimos años, otros trabajos relevantes en torno a los algoritmos de clasificación han propuesto nuevos caminos para evitar los problemas introducidos por los métodos tradicionales de predicción. En el campo de las redes neuronales artificiales, se ha desarrollado un nuevo Modelo de Neurona Dendrítica (*Dendritic Neural Model* (DNM)) para una mejor comprensión de un sistema neuronal biológico y para proporcionar un método más útil para resolver problemas prácticos al considerar la no linealidad de las sinapsis. Las predicciones confiables para la calidad de servicio (*Quality of Service* (QoS)) también han sido un tema de investigación importante en el dominio de la computación de servicios. Dos líneas interesantes para hacer una predicción altamente precisa de los datos de QoS que faltan son construir un conjunto de modelos de factor latente no negativo (*Negative Latent Factorization* (NLF)) (Zhou *et al.*, 2005) y presentar un modelo de factorización latente no negativa sesgada de tensores (*Biased Negative Latent Factorization Tensor* (BNLFT)) para temporal predicción de QoS basada en patrones (Lu, 2009). En cuanto al procesamiento de matrices dispersas y de alta dimensión y datos desequilibrados, los modelos de factorización de matrices no negativas (*Non-Negative Matrix Factorization* (NMF)) han demostrado ser altamente efectivos debido a su fina representatividad de los datos no negativos y el embebido. El método de selección de características que utiliza el índice de Gini ponderado (*Weighted Gini Index* (WGI)) ha mejorado la precisión (Li, 2004).

2.10. Conclusiones

El estudio de distintos sistemas para el cálculo de la reputación puede llevar a varias conclusiones importantes. No hay un sistema único y universal para el cálculo de la repu-

tación. Cada sistema tiene sus singularidades con sus ventajas y desventajas, y la elección del sistema adecuado depende del contexto en el que se utilice. Los sistemas de cálculo de la reputación se basan en diferentes criterios y es importante comprender los criterios en los que se basa cada sistema y no ignorar criterios fundamentales como, por ejemplo, la calidad del trabajo de los investigadores científicos en el caso de calcular la reputación del autor. Los sistemas de cálculo de la reputación son alterados por factores externos y es importante tener en cuenta estas variables al analizar los resultados de un sistema de reputación. El estudio de distintos sistemas para el cálculo de la reputación permite una mejor comprensión de cómo funciona la reputación en diferentes contextos, y cómo se puede utilizar para mejorar la confianza y la calidad en diversos ámbitos.

En resumen, existe la oportunidad de mejorar las métricas que aparecen en la literatura científica para el cálculo de la reputación en distintos ámbitos y simplicar su cálculo. En este Trabajo de Tesis se aportan métodos nuevos con la intención de aportar en esa mejora.

Capítulo 3

Método

En esta tesis ahondamos en el cálculo de la reputación utlizando sistemas basados en el conocimiento y aprendizaje máquina. Además, se introducen componentes como el sentimiento, la novedad y se usan métodos de optimización para proponer métricas que eliminan la dependencia de la producción (por ejemplo, entre los autores) y simplifican el cálculo (por ejemplo, evitando reproducir el malware en un entorno virtual). En el método, proponemos tres distintos cálculos de reputación orientados a distintos dominios de aplicación, los dos primeros tienen que ver con el ámbito científico y el último con Internet. *Unified Knowledge Compiler* (UNIKO) es una plataforma que calcula, entre otras cosas, la reputación de los artículos y autores. *Framework for Reputation Estimation of Scientific Authors* (FRESA) es un framework que se centra en el cálculo de la reputación de los autores introduciendo parámetros como la relevancia y la novedad. Por último, *Domains Classifier based on Risky Websites* (DOCRIW) es un sistema que calcula la reputación de los sitios web.

3.1. Reputación de artículos y autores: UNIKO

UNIKO es una plataforma innovadora diseñada para proveer soporte sobre las búsquedas científicas que incluye un portal (ver Fig. 3.1) usado para la visualización de las consultas y la interacción con los usuarios.

UNIKO permite buscar y procesar la información relacionada con un campo de investigación específico. Tiene la capacidad de proveer conocimiento sobre autores y artículos, y feedback sobre otros usuarios de la plataforma. El conocimiento puede medirse aplicando diferentes métricas (por ejemplo, reputación de los atículos o autores procesados), y organizado a través de algoritmos de clustering, ilustrando posibles relaciones entre elementos. Para alcanzar estas tareas, UNIKO considera diferentes fuentes de información. Algunas de estas fuentes son páginas web scrapeadas para extraer metadata relacionada

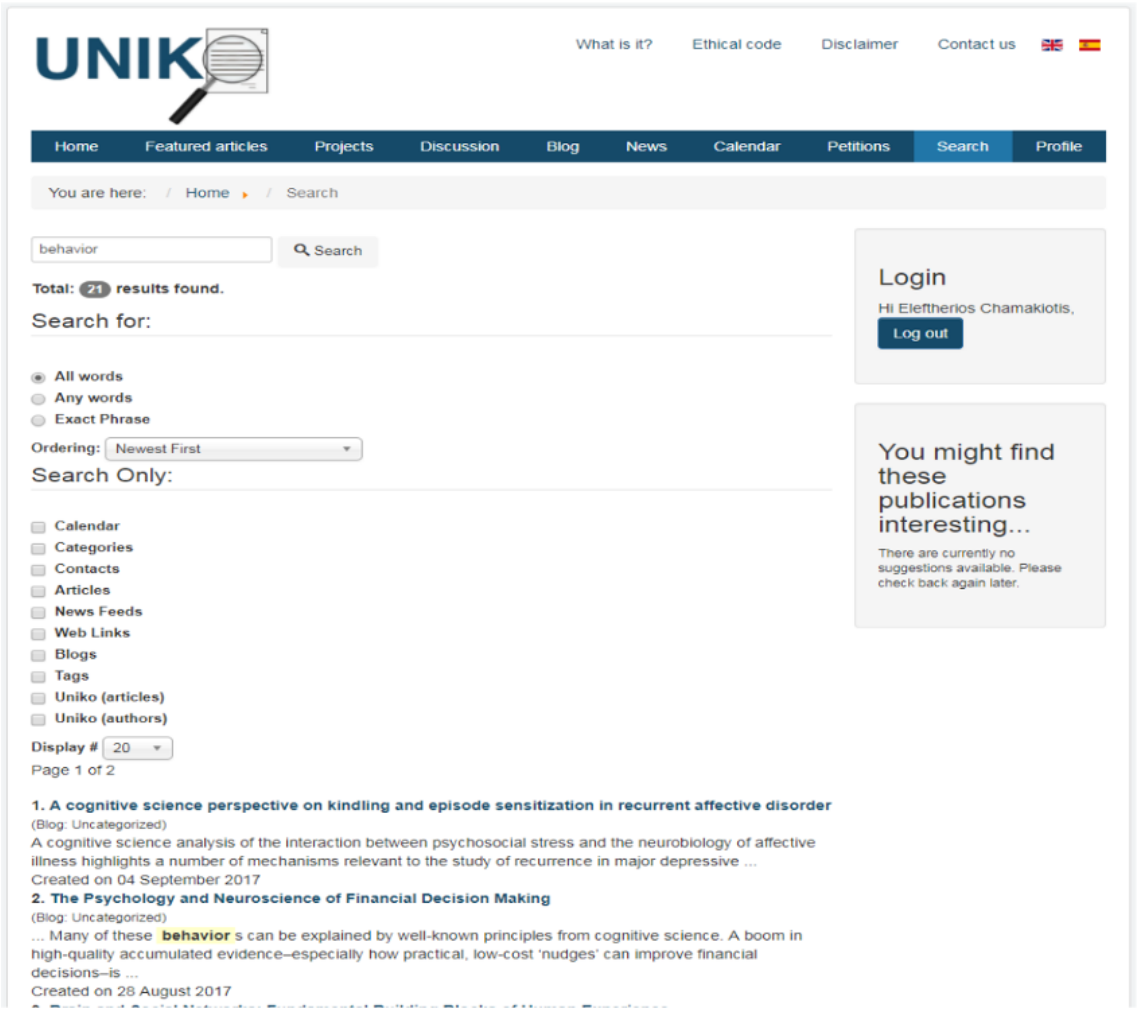

Figura 3.1: *Extracto del portal* UNIKO

con los artículos (por ejemplo, Elsevier (Group, 2018) o Springer (Springer International Publishing AG, 2018)), o servicios web API REST (Richardson y Ruby, 2008) de proveedores externos para los autores (por ejemplo, *Semantic Scholar* (Allen Institute for Artificial Intelligence and Semantic Scholar, 2018)). Esta información puede ser combinada y filtrada para generar conocimiento específico disponible para los usuarios.

Respecto a la arquitectura de UNIKO, está organizada en dos sistemas principales: el *Controlador de Extracción de Información o Information Retrieval Controller (IRC)* y el *Controlador de Gestión del Conocimiento o Knowledge Management Controller (KMC)* (ver Fig. 3.2). El sistema IRC contiene cuatro módulos: *Procesador de Información de Artículos (Articles Information Processor)*, *Procesador de Información de Autores (Authors Information Processor)*, *Calculadora de Reputación (Reputation Calculator)* y *Calculadora de Sentimiento (Sentiment Calculator)*. El sistema KMC contiene dos módulos: *Gestor de Operaciones de Usuarios (Users Operations Manager)* y *Evaluador de Similaridad y Clusturing (Similarity and Clustering Evaluator)*.

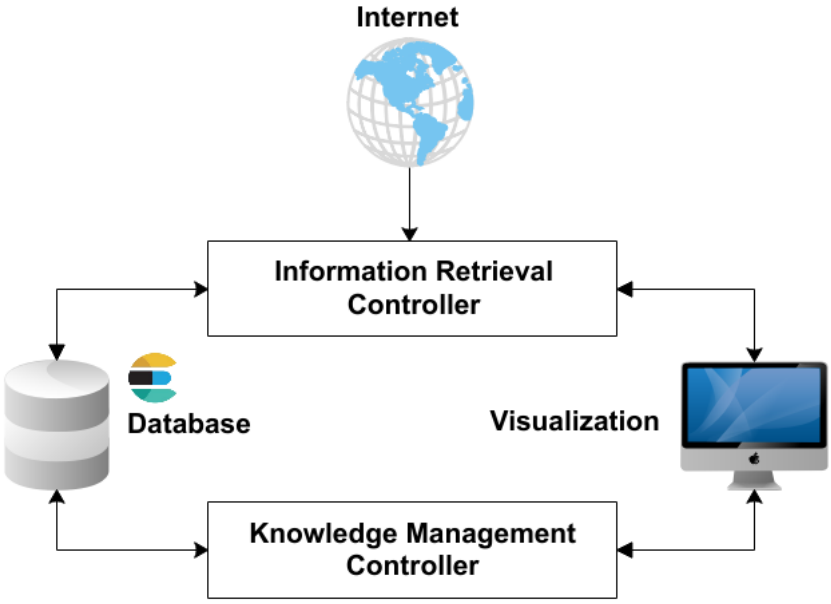

Figura 3.2: *Extracto de la arquitectura general de* UNIKO.

Además, se incluye un módulo para la consolidación de la información basado en una base de datos Elasticsearch (Gormley y Tong, 2015). Esta base de datos está conectada con los sistemas IRC y KMC para adquirir nueva información y proporcionársela a los usuarios. Por último, el framework UNIKO se completa con el módulo de visualización.

El principal objetivo del sistema IRC es recoger, procesar y almacenar el conocimiento extraído de sus módulos. Por otro lado, el sistema KMC se encarga de la interacción con los usuarios, proporcionando relaciones entre el conocimiento almacenado, las métricas y las sugerencias. Los módulos de ambos sistemas están conectados entre ellos usando una API REST.

La sección 3.1.1 presenta el sistema IRC y sus módulos asociados en detalle. La sección 3.1.2 describe el sistema KMC y sus componentes asociados para la interacción con los usuarios.

3.1.1. Controlador de extracción de información o Information Retrieval Controller (IRC)

El IRC es el sistema que recoge la información de las fuentes web externas. Esta información está principalmente relacionada con artículos y autores procedentes de revistas científicas de editoriales conocidas como Elsevier o Springer. Usa cuatro módulos distribuidos (ver Fig. 3.3). El módulo principal es el *Article Information Processor* que interactúa con los otros tres módulos a través de la API REST. Estos módulos (*Reputation Calculator*, *Authors Information Processor* y *Sentiment Calculator*) proporcionan información específica (extraída o calculada) relacionada con los autores o artículos.

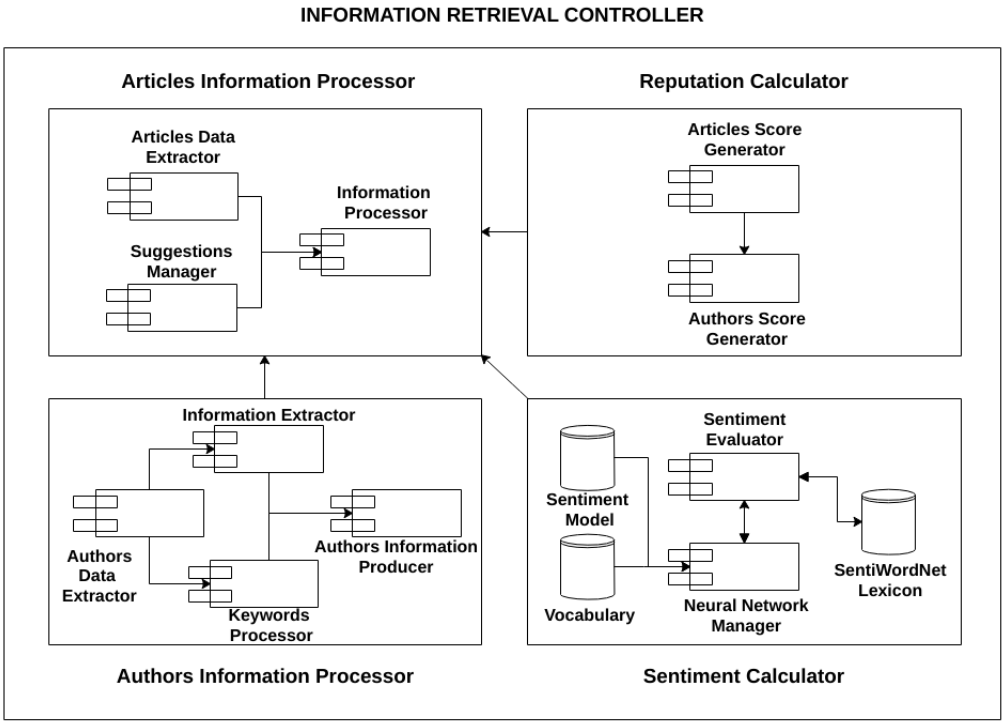

Figura 3.3: *Extracto de la arquitectura general del* Information Retrieval Controller.

Las siguientes cuatro subsecciones describen estos módulos en detalle.

Procesador de información de artículos (Articles Information Processor)

Este módulo recoge artículos de investigación publicados en distintas revistas científicas (por ejemplo, Elsevier (Group, 2018) and Springer (Springer International Publishing AG, 2018)). Esto se realiza gracias al uso de múltiples crawlers web (Garg *et al.*, 2017) cuyo objetivo es consultar y extraer información de las *Application Programming Interface* (API) de código abierto.

Los artículos que tienen que ser recolectados son solicitados por los usuarios el día anterior a través del portal. Esta acción se realiza una vez al día actualizándose los índices correspondientes de la base de datos con la nueva información. Este módulo usa una cola Apache Kafka (Narkhede *et al.*, 2016) y Apache Spark (Zaharia *et al.*, 2016) como principal tecnología. Apache Kafka se utiliza para ordenar las solicitudes de los usuarios (previamente almacenadas en base de datos). Apache Spark se encarga de las operaciones de extracción de datos.

Los artículos se extraen de diferentes editoriales, por ello los crawlers funcionan de manera ligeramente distinta. Por este motivo es necesario normalizar las entradas para mantener un formato consistente. Además, debido a la ausencia de detalles relvantes sobre los artículos que no se proveen por las API, otras páginas web (por ejemplo, ScienceDirect (RELX Group, 2018)) son scrapeadas.

Respecto a la arquitectura del módulo, contiene tres compenentes: el *Extractor de Información de Artículos (Articles Data Extractor)*, el *Gestor de Sugerencias (Suggestions Manager)* y el *Procesador de Información (Information Processor)*. El primer componente utiliza los crawlers web para recopilar y normalizar la información sobre un artículo solicitado. Esta información consiste principalmente en palabras clave, texto completo (si está disponible, si no se recoge el abstract), autores y referencias. El segundo componente genera las etiquetas del artículo a través de operaciones semánticas para hacer sugerencias a los usuarios. El tercer componente genera la estructura de la información recopilada y la almacena en la base de datos cuando el resto de módulos de IRC han proporcionado su información correspondiente (por ejemplo, *Sentiment Calculator* devuelve el valor de sentimiento del abstract y *Reputation Calculator* genera la reputación de un artículo y sus autores).

El *Suggestions Manager* genera una lista de etiquetas que representan con precisión el contenido de un artículo. Además, este componente mide la similaridad entre las etiquetas para sugerir los artículos más adecuados a los usuarios. Para producir la lista de etiquetas, el *Suggestions Manager* realiza un conjunto de operaciones basadas en técnicas de procesamiento de lenguaje natural (*Natural Language Processing* (NLP)) (Sekine y Ranchhod, 2009) para procesar el texto completo o el abstract de un artículo (cuando el cuerpo del artículo no está disponible). Primero, se logra el etiquetado de Penn TreeBank (Marcus *et al.*, 1993). Luego, se eliminan las stopwords y los signos de puntuación, y finalmente se lematizan los tokens restantes. Estas tareas se basan en cadenas léxicas (T. Erekhinskaya and D. Moldovan, 2013). Las cadenas léxicas son secuencias de palabras relacionadas semánticamente que están interconectadas a través de relaciones semánticas. En UNIKO, estas cadenas se construyen utilizando WordNet (Miller, 1995). WordNet contiene conceptos y relaciones entre conceptos. Así, este recurso puede verse como un gran grafo semántico donde encontrar caminos entre conceptos corresponde a encontrar cadenas léxicas.

Para medir la similitud entre las etiquetas, el componente *Suggestions Manager* utiliza métricas como *Resnik*, *Hirst-StOnge*, *Lin*, *Lesk*, *Jiang- Conrath*, *Wu-Palmer*, *Leacock-Chodorow* y *Path* (ver (X. Zhang *et al.*, 2017) para una descripción completa). Se aplica un procedimiento de ponderación para obtener la métrica final. Una vez realizada esta tarea, UNIKO está lista para sugerir los artículos que tienen la mayor similaridad agregada con un usuario específico.

Procesador de información de autores (Authors Information Processor)

El objetivo principal de este módulo es recopilar la información de los autores de un artículo determinado. Utiliza la API de *Semantic Scholar* (Allen Institute for Artificial Intelligence and Semantic Scholar, 2018) para obtener la mayoría de los datos relevantes. El módulo usa el *Digital Object Identifier* (DOI) del artículo para identificarlo (Langston

y Tyler, 2004).

En cuanto a la arquitectura del módulo, comprende cuatro componentes: el *Extractor de Información de Autores (Authors Data Extractor)*, el *Extractor de Información (Information Extractor)*, el *Procesador de Palabras Clave (Keywords Processor)* y el *Productor de Información de Autores (Authors Information Producer)*. La información resultante sobre los autores de un artículo es enviada al módulo *Articles Information Processor* para integrarla y consolidarla en la base de datos.

El componente *Authors Data Extractor* es responsable de ejecutar consultas en ambos endpoints de la API de Semantic Scholar (es decir, datos de autores y artículos). Recopila y almacena los metadatos relacionados sin procesar.

El *Information Extractor* recibe los metadatos sin procesar del componente anterior y procesa los metadatos para extraer información relevante: *nombre completo de los autores*, *seniority* de los autores, sus *citas influyentes* (es decir, una publicación citada tiene un impacto significativo sobre otras del dominio relacionado), sus *citas* (en general), el número de publicaciones y sus *títulos*. Así, para cada autor, el resultado consiste en la suma de las citas influyentes, la suma de las citas, la antigüedad (calculada como el año de la última publicación menos el año de la primera publicación) y el número total de artículos publicados.

El objetivo principal del componente *Keywords Processor* es generar un conjunto de palabras clave para identificar a los autores de acuerdo con su dominio de investigación. Recibe los *títulos* de las publicaciones recopiladas previamente por el *Information Extractor*. Luego, procesa estos títulos extrayendo sus sintagmas nominales a través de técnicas de NLP. Estos sintagmas nominales se lematizan y almacenan para tener en cuenta su aparición. Una vez que se ha evaluado cada título, las palabras clave para un autor son las frases nominales más comunes. De forma predeterminada, el componente devuelve un máximo de seis palabras clave.

El *Authors Information Producer* organiza la información recopilada por los módulos anteriores uniendo sus respectivos resultados. El resultado final producido es útil para el módulo *Reputation Calculator* (*nombre del autor*, *citas influyentes*, *citas*, *seniority* y *número de publicaciones*) y para el módulo *Similarity and Clustering Evaluator* (palabras clave que identifican el dominio de investigación de cada autor).

Calculadora de reputación (Reputation Calculator)

Este módulo calcula la reputación de autores y artículos en función de un conjunto de características que son recopiladas por el módulo *Authors Information Processor* presentado anteriormente.

En cuanto a la arquitectura del módulo, contiene dos componentes: *Generador de Puntuación de Autores (Authors Score Generator)* y *Generador de Puntuación de Ar-*

tículos (Article Score Generator). El primero se centra en la reputación de los autores. El segundo evalúa la reputación de los artículos. Dado que se necesita la reputación de los autores para calcular la reputación del artículo, el *Article Score Generator* incluye el *Authors Score Generator*.

El componente *Author Score Generator* utiliza cuatro características relacionadas con un autor específico para establecer su valor de reputación. Estas características son: número de citas recibidas en artículos publicados, número de citas influyentes en artículos publicados, seniority y número de artículos publicados. Sea $rep_i \in [0, 1]$ la reputación del autor i, calculada de la siguiente manera:

$$rep_i = \omega_1 * inf\_citations + \omega_2 * citations + \omega_3 * seniority + \omega_4 * papers\,, \tag{3.1}$$

donde $\sum_{i=1}^{4} \omega_i = 1$. Estos parámetros permiten fijar la importancia relativa de las características medidas modificando el resultado final según las más relevantes. Hay que observar que las características están normalizadas para estar en el rango entre 0 y 1. Para ello, se define un umbral superior para cada característica relacionada con la máxima calidad de la característica (por ejemplo, en el caso de las citas, los autores con al menos 500 deben tener una influencia aceptable en su dominio de investigación).

El componente *Article Score Generator* considera dos reputaciones parciales para obtener la reputación completa de un artículo. La primera es la reputación promedio de los autores del artículo, mientras que la segunda tiene en cuenta el número de citas del artículo (es decir, el impacto en su dominio científico).

Sea $citations_p$ el valor normalizado de las citas del artículo p, y sea $rep\_authors_p$ la reputación media de los n autores del artículo p:

$$rep\_authors_p = \sum_{i=1}^{n} rep_i/n\,. \tag{3.2}$$

Por tanto, la reputación completa del artículo p, rep_p, se calcula como:

$$rep_p = \alpha * rep\_authors_p + (1 - \alpha) * citations_p\,, \tag{3.3}$$

donde α es un parámetro de compensación para gestionar la importancia relativa de la reputación de los autores sobre las citas del artículo.

Calculadora de sentimiento (Sentiment Calculator)

El propósito principal de este módulo es producir un valor de sentimiento en un rango entre -1 y 1 (siendo 0 neutral). Este valor de sentimiento se calcula a partir de los textos de los artículos y se almacena en la base de datos. Se debe tener en cuenta que hay algunas revistas que no proporcionan los textos. En tal caso, se utiliza en su lugar el resumen del artículo. Este valor de sentimiento es útil para realizar una tarea de clasificación en el conjunto de artículos.

En cuanto a la arquitectura del módulo, consta de dos componentes: el *Gestor de Redes Neuronales (Neural Network Manager)* y el *Evaluador de Sentimiento (Sentiment Evaluator)*. El *Neural Network Manager* está a cargo de predecir los valores de sentimiento para palabras específicas. Utiliza un modelo previamente entrenado y un vocabulario producido por una Red Neural Convolucional (*Convolutional Neural Network* (CNN)) (Poria *et al.*, 2015). El *Sentiment Evaluator* lematiza el texto de entrada a través de técnicas de NLP (y también aplica el etiquetado de Penn TreeBank (Marcus *et al.*, 1993), elimina stopwords y signos de puntuación). Luego, utiliza el léxico SentiWordNet (Baccianella *et al.*, 2010) para obtener los valores de sentimiento de las palabras. Cuando una palabra no se encuentra en el léxico, el *Sentiment Evaluator* solicita predicciones al *Neural Network Manager*.

La arquitectura CNN se basa en un enfoque bien conocido. Se ha conservado la estructura secuencial de la original provista por (Bhavsar *et al.*, 2017) (ver Tabla 3.1). Además, se incluyen mejoras relacionadas con la optimización del proceso. Tanto en los pasos de entrenamiento como de validación se han incluido técnicas de NLP para procesar el texto. Esto lleva a producir una lista de palabras lematizadas como entrada en lugar de texto sin formato. Por lo tanto, se producen más coincidencias entre palabras cuando la CNN necesita predecir el valor de sentimiento de una palabra (solo si las palabras proporcionadas también están lematizadas).

Layers
1. Embedding input_dim 5000 output_dim 50
2. Dropout rate 0.2
3. Conv1D 250 filters of 3 with stride 1
4. Pool1D (max) with stride 1
5. Dense units 250
6. Dropout rate 0.2
7. Relu
8. Dense units 1
9. Sigmoid

Tabla 3.1: *Arquitectura de Red Neuronal.*

La CNN ha sido modificada para producir dos resultados: un vocabulario con las palabras aprendidas y el modelo con los pesos específicos. Ambos han sido configurados para ser utilizados con una palabra, oraciones o textos completos. Este problema facilita su integración en el componente *Sentiment Evaluator*. Este componente necesita palabras separadas para predecir su valor de sentimiento cuando el léxico de SentiWordNet no las proporciona (es decir, el léxico no contiene la palabra o el valor de sentimiento asociado de una palabra es neutral).

3.1.2. Controlador de gestión de conocimiento o Knowledge Management Controller (KMC)

El KMC es el segundo sistema principal del framework UNIKO (ver Fig. 3.2). El KMC gestiona los usuarios de UNIKO y controla los eventos creados por ellos (ver Fig. 3.4). Utiliza la información recopilada por el IRC para proporcionar sugerencias y organizaciones específicas de los elementos de interés (es decir, artículos y autores). El KMC muestra el conocimiento adquirido en el módulo *Visualization* (es decir, el portal) del framework UNIKO.

En cuanto a la arquitectura, el KMC contiene dos módulos: el *Gestor de operaciones de Usuario (User Operations Manager)* y el *Evaluador de Similaridad y Clustering (Similarity and Clustering Evaluator)*. El primero gestiona los eventos producidos por los usuarios (por ejemplo, creación y modificación de nuevos usuarios, o incorporación de nuevos artículos a la base de datos de ElasticSearch). El segundo módulo adquiere información y produce conocimiento específico relacionado con artículos y autores.

El KMC comunica al usuario aquellos eventos relacionados con la recuperación de información de fuentes web (por ejemplo, información de un artículo específico requerido y sus autores) a través de la modificación de la información almacenada en un índice determinado en la base de datos. El IRC no logra estos eventos en tiempo de usuario. Así, se recuperan y resuelven en un momento predefinido del día, generando nueva información

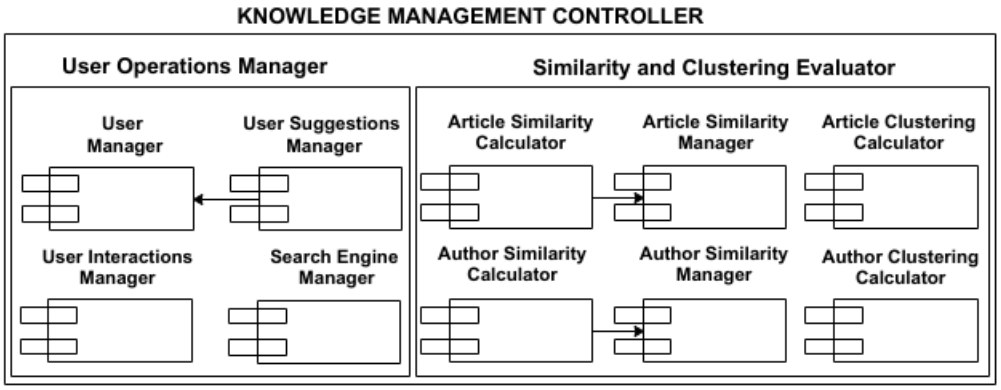

Figura 3.4: *Extracto de la arquitectura general del* Knowledge Manager Controller.

estable de los índices relacionados con autores y artículos. La misma política se cumple en el caso de nuevos usuarios y las sugerencias de tareas de actualización.

La sección 3.1.2 presenta el *User Operations Manager* y sus componentes. La sección 3.1.2 describe en detalle el módulo *Similarity and Clustering Evaluator* y los diferentes componentes encargados de proporcionar autores y artículos similares, y sus posibles organizaciones.

Gestor de Operaciones de Usuario (User Operations Manager)

Este módulo es un gestor que engloba las operaciones asociadas a los usuarios del framework UNIKO. Estas operaciones consisten en dar de alta y modificar usuarios, proporcionar sugerencias relacionadas con autores similares y artículos de interés, buscar por palabra clave autores o artículos y almacenar las solicitudes realizadas para recopilar información sobre artículos y autores no considerados anteriormente.

Respecto a la arquitectura del módulo (ver Fig. 3.4), comprende cuatro componentes: el *Gestor de Usuario (User Manager)*, el *Gestor de Sugerencias de Usuario (User Suggestions Manager)*, el *Gestor de Motor de Búsqueda (Search Engine Manager)* y el *Gestor de Interacciones de Usuario (User Interactions Manager)*. El primer componente gestiona la elaboración de perfiles de usuario y su modificación. El segundo componente selecciona las posibles sugerencias sobre artículos o autores que podrían ser de interés para los usuarios (por ejemplo, autores y usuarios con dominios de investigación similares). Estas sugerencias se configuran cuando se crea un usuario y se actualizan cuando se incluye en la base de datos información relacionada con nuevos artículos o autores. El tercer componente logra las búsquedas en la base de datos filtrando artículos o autores por sus palabras claves, y devolviéndolos ordenados por su similaridad. El cuarto componente se centra en las interacciones de los usuarios. Almacena los comentarios en foros y las solicitudes logradas para recopilar información sobre un artículo específico (y sus autores relacionados).

Evaluador de similaridad y clustering (Similarity and Clustering Evaluator)

Este módulo genera conocimiento a partir de la información de los artículos y autores. Este conocimiento está relacionado con la identificación de medidas adecuadas de similaridad enfocadas a organizar la información. Los resultados producidos se muestran a los usuarios para proporcionar información visual. Esto facilita el proceso de búsqueda de elementos de investigación similares (es decir, autores y artículos) relacionados con un tema específico. Así, brinda pautas a los usuarios indicando qué elementos (autores o artículos) tienen más relación con otro previamente seleccionado.

La arquitectura del módulo comprende (ver Fig. 3.4) seis componentes: el *Calculadora de Similaridad de Artículo (Article Similarity Calculator)*, el *Calculadora de Similaridad de Autor (Author Similarity Calculator)*, el *Gestor de Similaridad de Artículo (Article*

Similarity Manager), el *Gestor de Similaridad de Autor (Author Similarity Manager)*, el *Calculadora de Clustering de Artículo (Article Clustering Calculator)* y el *Calculadora de Clustering de Autor (Author Clustering Calculator).*

El *Article Similarity Calculator* y el *Author Similarity Calculator* calculan las matrices de similaridad para los artículos y para los autores respectivamente. Spark (Zaharia *et al.*, 2016) ha sido la tecnología seleccionada para realizar los cálculos ya que se ha previsto gran cantidad de información.

Respecto a la metodología utilizada, es la misma en ambos componentes. En primer lugar, las palabras clave de los artículos y autores se recopilan de los índices correspondientes de la base de datos. Luego, estas palabras clave se usan para construir las matrices de similaridad usando el conocido algoritmo *tf-idf* y la métrica *cosine similarity* (Pazzani y Billsus, 2007). *Tf-idf* caracteriza los conjuntos de palabras clave en función de la frecuencia de las palabras en cada conjunto (tanto en artículos como en autores) en comparación con el resto de conjuntos de palabras clave. La métrica *similitud del coseno* calcula el coseno del ángulo entre las palabras clave. Así, no solo se considera la magnitud de cada palabra clave, sino el ángulo entre ellas. Estas matrices de similaridad se actualizan diariamente y se almacenan en la base de datos para incluir la información de nuevos artículos y autores que pueda ser solicitada por los usuarios.

Los componentes *Article Similarity Manager* y *Author Similarity Manager* reciben como entrada un artículo o un autor respectivamente. Estos son proporcionados por un evento de usuario en la interfaz gráfica de UNIKO. Usando estas entradas, las matrices creadas por los dos componentes anteriores se evalúan para filtrar solo los elementos más similares (artículos o autores). La cantidad de elementos recuperados está controlada por un umbral específico (5 por defecto).

El *Article Clustering Calculator* y el *Author Clustering Calculator* se han desarrollado para agrupar por separado conjuntos homogéneos de artículos y autores. Dado un conjunto de elementos, estos módulos obtienen sus palabras clave y aplican las métricas *tf-idf* y *cosine similarity* para producir las matrices de similaridad. Luego, se calcula un *Hierarchical Clustering* (Steinbach *et al.*, 2000) para construir una jerarquía de clusters.

3.1.3. Proceso del controlador de extracción de información (Information Retrieval Controller Process)

Este proceso proporciona pautas generales que describen el pipeline y las interacciones entre los módulos del sistema IRC (ver Sección 3.1.1). Comprende ocho pasos (ver Fig. 3.5): *Evento de nuevo artículo (New article event)*, *Extracción de información (Data retrieval)*, *Normalización de información (Data normalization)*, *Etiquetado (Tagging)*, *Extracción de información de autores (Authors information retrieval)*, *Reputación de autores y artículos (Authors and article reputation)*, *Análisis de sentimiento (Sentiment analysis)*

and *Gestión de sugerencias (Suggestion management)*. El sistema se ejecuta una vez al día actualizando los índices correspondientes de la base de datos.

El paso *New article event* inicia el flujo de trabajo recopilando, desde la base de datos, las solicitudes realizadas por los usuarios el día anterior. Esto se realiza en el módulo *Articles Information Processor*. Para cada artículo, se recopila el DOI (Langston y Tyler, 2004). Luego, si la información del artículo está disponible, los scrapers buscan información sobre el mismo en el paso *Data retrieval*. El *Articles Data Extractor* es el componente encargado de lograr estas tareas. Una vez completada la información, se normaliza a través del componente *Information Processor* en el paso *Data normalization*. A continuación, en el paso de *Tagging*, el *Suggestions Manager* elabora las etiquetas relacionadas con el artículo procesado. El paso *Authors information retrieval* selecciona la información sobre los autores. Este proceso lo lleva a cabo el componente *Authors Information Processor*. Establece comunicación con el componente *Articles Information Processor*. Posteriormente, se calcula la reputación de los autores y del artículo. Esto se logra con el paso *Authors and article reputation* y se produce con el módulo *Reputation Calculator*. En primer lugar, se calcula la reputación de los autores. Nótese que, si la reputación de un autor ha sido previamente evaluada (debido a una autoría en un artículo previamente procesado), se toma directamente de la base de datos. A continuación, se produce la reputación del artículo. El paso *Sentiment analysis* evalúa la polaridad del artículo. Se lleva a cabo en el módulo *Sentiment Calculator*. Finalmente, el paso *Suggestion manager* actualiza las sugerencias a los usuarios de acuerdo con las etiquetas producidas para el artículo actual. Esta tarea es implementada por el componente *Information Processor* en el módulo *Articles Information Processor*. Este componente también concluye el flujo de trabajo almacenando la información correspondiente generada en la base de datos.

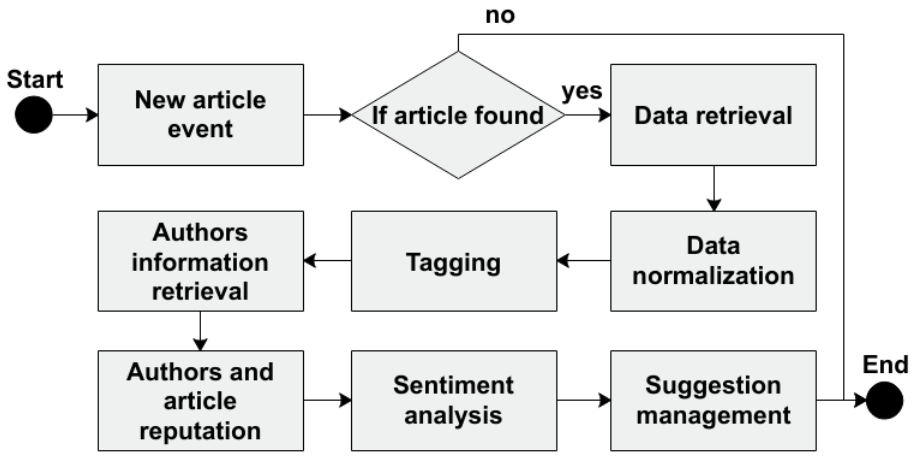

Figura 3.5: *Extracto del flujo de trabajo del* IRC.

Proceso de análsis de sentimiento (Sentiment Analysis Process)

El *Sentiment Analysis Process* es un flujo de trabajo interno del *Information Retrieval Controller Process* (consulte la Sección 3.1.3). Se realiza mediante el módulo *Sentiment Calculator* (ver Sección 3.1.1 y Fig. 3.3) para proporcionar un valor de sentimiento entre -1 y 1 (siendo 0 neutral) a un texto específico (texto completo o abstract de un artículo).

El flujo de trabajo (ver Fig. 3.6) comienza evaluando si el texto completo del artículo está disponible (pasos *Obtener texto completo (Obtain complete text)* u *Obtener texto del abstract (Obtain abstract text)*). Se utilizan técnicas de NLP sobre el texto (paso *Limpiar y tokenizar texto*). Luego, se etiqueta el texto (usando el etiquetado de Penn Treebank (Marcus *et al.*, 1993)) para evaluar su sintaxis. Las stopwords y los signos de puntuación se descartan y se extraen los tokens. A continuación, en el paso *Lematizar tokens (Lemmatize tokens)*, los tokens se lematizan usando las etiquetas generadas. Los tokens se evalúan para obtener su valor de sentimiento individual. Si el token se encuentra en el léxico de SentiWordNet (Baccianella *et al.*, 2010) y su valor de sentimiento no es neutral (es decir, 0), el valor se acumula. Por otro lado, si no se encuentra el token o su valor de sentimiento es neutral, la CNN se utiliza para predecir su valor de sentimiento. Este valor de sentimiento también se acumula. Estas operaciones se abordan mediante el paso *Obtener valor de SentiWordNet (Obtain value from SentiWordNet)* y el paso *Predecir*

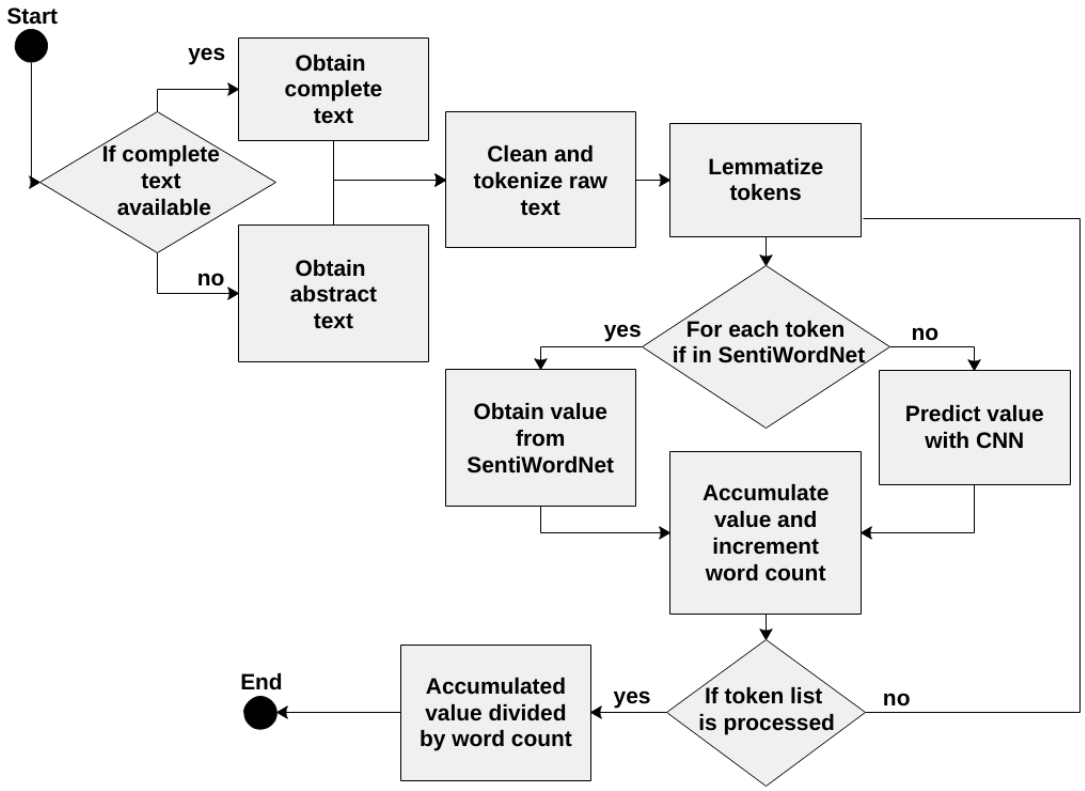

Figura 3.6: *Extracto del flujo de trabajo del Sentiment Analysis.*

valor con CNN (Predict value with CNN), respectivamente, hasta que se obtiene el valor de sentimiento de cada token. Finalmente, el valor de sentimiento promedio del texto se calcula en el paso *Dividir el valor acumulado entre el número de palabras (Accumulated value divided by word count)*. Divide el valor de sentimiento resultante acumulado para todos los tokens procesados por la cantidad total de tokens.

3.1.4. Proceso del controlador de gestión del conocimiento (Knowledge Management Controller Process)

Este proceso describe las interacciones de los usuarios con el portal UNIKO (gestión de usuarios, sugerencias y etiquetado y comentarios en los foros previstos). Está enfocado en los módulos presentados en el sistema KMC (ver Sección 3.1.2 y Fig. 3.4). Hay ocho pasos independientes (ver Fig. 3.7): *Clustering de la lista de artículos (Article list clustering)*, *Artículos similares (Similar articles)*, *Clustering de la lista de autores (Author list clustering)*, *Autores similares (Similar authors)*, *Creación o modificación de usuarios (User creation or modification)*, *Publicación de comentarios (Comments publication)*, *Filtrado de artículos (Filtering articles)*, y *Filtrado de autores (Filtering authors)*. Nótese que no tienen un pipeline común (es decir, no son secuenciales), sino que son flujos de trabajo que logran operaciones específicas de acuerdo a las solicitudes realizadas por los usuarios del portal.

El paso *Article list clustering* está relacionado con las operaciones realizadas en la *Article Clustering Calculator*. Dado un conjunto de artículos de interés, se agrupan según sus palabras clave.

El paso *Similar articles* es ejecutado por el *Article Similarity Manager*. Dado un artículo de interés, se obtiene una lista con los cinco artículos más similares de la base de datos.

El paso *Author list clustering* abarca las tareas realizadas por el *Author Clustering Calculator*. Dado un conjunto de autores de interés, se agrupan según sus palabras clave.

El paso *Similar authors* es ejecutado por el *Author Similarity Manager*. Dado un autor de interés, se obtiene una lista con los cinco autores más similares en la base de datos.

El paso *User creation or modification* incluye las operaciones relacionadas con el registro y modificación de los usuarios en el portal de UNIKO. Estas tareas también incluyen la evaluación de las etiquetas de interés seleccionadas por los usuarios y la asociación de etiquetas existentes de los artículos ya almacenados en la base de datos. Los componentes asociados a ellos son el *User Manager* y el *User Suggestions Manager* del módulo *User Operations Manager*.

El paso *Comments publication* cumple con las funciones relacionadas con la publicación de mensajes en los foros del portal. Estos comentarios también pueden ser modificados

por los usuarios. El componente responsable es el *User Interactions Manager*.

Los pasos *Filtering articles* y *Filtering authors* se encargan de recuperar artículos y autores relevantes según una palabra clave específica proporcionada por el evento de usuario correspondiente. Estos elementos se clasifican según su similaridad utilizando sus etiquetas. Por lo tanto, el framework actúa como un sistema de recomendación.

3.2. Reputación de autores: FRESA

Framework for Reputation Estimation of Scientific Authors (FRESA) es una plataforma innovadora enfocada en proporcionar una nueva métrica para clasificar a los autores por su reputación. Por lo tanto, FRESA es capaz de calcular y representar una trayectoria temporal de reputación basada en la relevancia y novedad de las obras del autor. Para ello, recopila información de fuentes web externas, obtiene conocimiento relevante y estima las

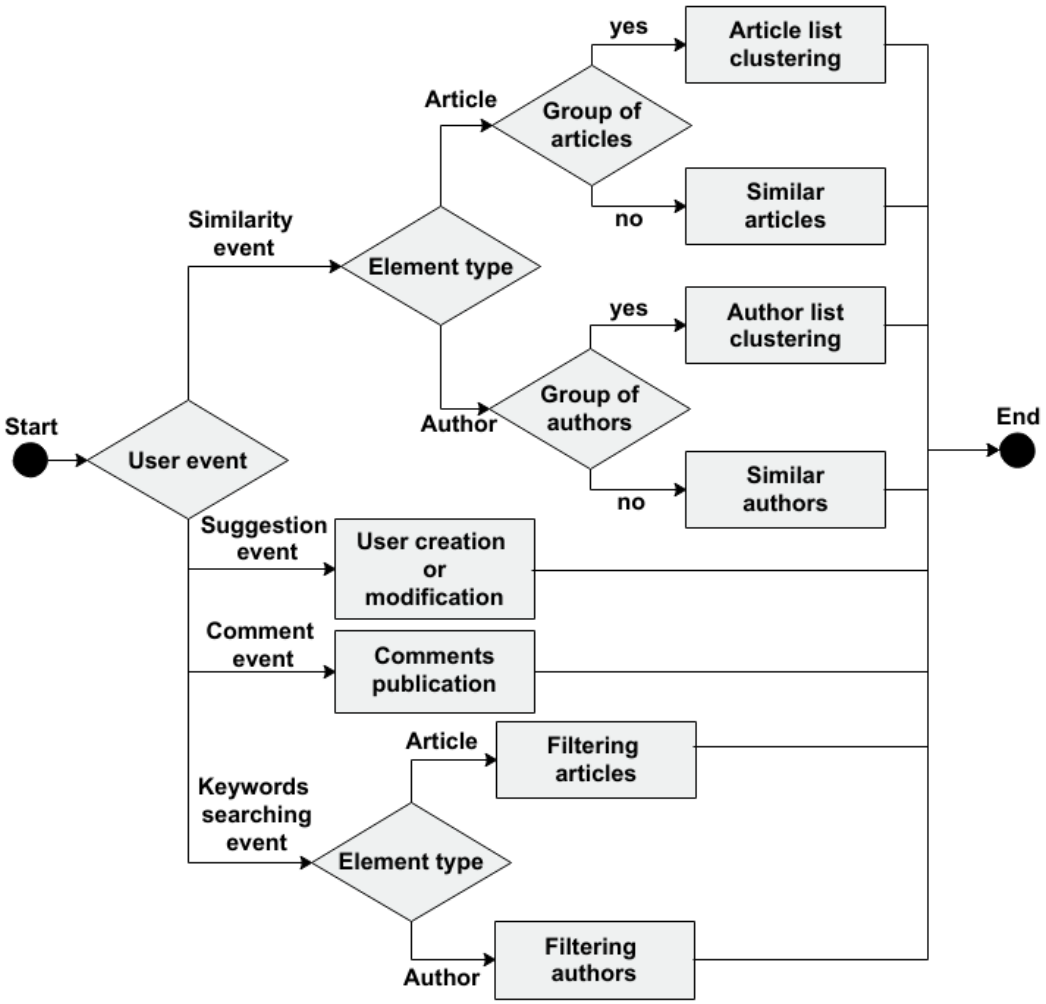

Figura 3.7: *Extracto del flujo de trabajo del Knowledge Management Controller.*

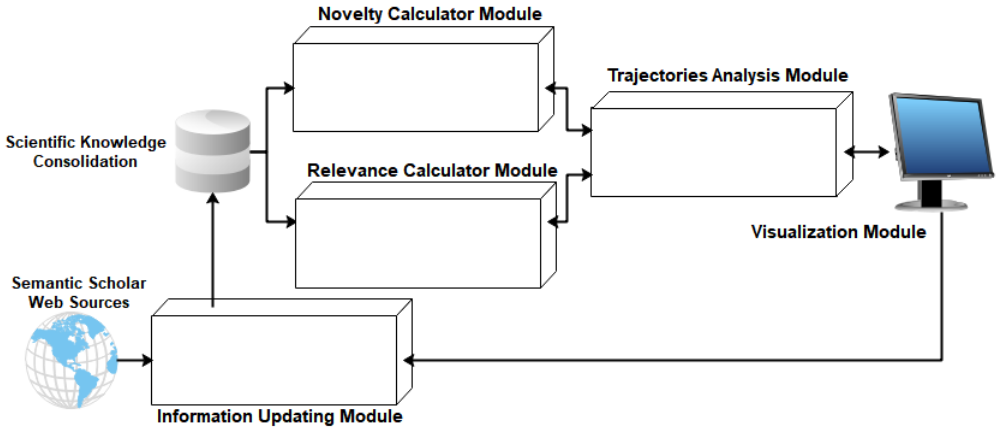

Figura 3.8: *Extracto de la arquitectura general de FRESA.*

trayectorias.

Respecto a la arquitectura del sistema, presenta cuatro módulos principales que se detallan en las siguientes secciones (ver Fig. 3.8): el módulo *Actualización de Información (Information Updating)*, el módulo *Calculadora de Relevancia (Relevance Calculator)*, el módulo *Calculadora de Novedad (Novelty Calculator)* y el módulo *Análisis de Trayectorias (Trajectories Analysis)*. Estos cuatro módulos recopilan y actualizan la información de fuentes web utilizando la API de Semantic Scholar. Es un motor de búsqueda respaldado por *Artificial Intelligence* (AI) para publicaciones académicas (Gormley y Tong, 2015) diseñado para resaltar los artículos más importantes e influyentes. A continuación, estiman los índices de relevancia y novedad para cada autor y año y calculan la puntuación final de reputación. Por último, construyen las trayectorias de reputación de cada autor que representan toda la obra científica del autor. Además, el sistema también contiene un módulo de *Visualización (Visualization)* y el módulo de *Consolidación de Conocimiento Científico (Scientific Knowledge Consolidation)* que actúa como una base de conocimiento. El último contiene la información sobre las revistas *Journal Citation Reports* (JCR) y las conferencias *Computing Research and Education Association* (CORE) (*The Computing Research and Education Association of Australasia*, s.f.), como el nombre de la revista, el factor de impacto, el cuartil, el índice CORE, el tema de la revista o el acrónimo de la conferencia. También se actualiza con la información de los autores y artículos.

3.2.1. Módulo de actualización de información (Information Updating module)

Este módulo recopila y actualiza la información de las fuentes web y procesa estos datos para enriquecer el módulo *Scientific Knowledge Consolidation*. Presenta dos componentes (ver Fig. 3.9): el *Extractor de Información de Autores (Author's Information Collector)* y el *Procesador de Información de Autores (Author's Information Processor)*. El primero utiliza la API Semantic Scholar para recopilar la información relativa a los autores y sus

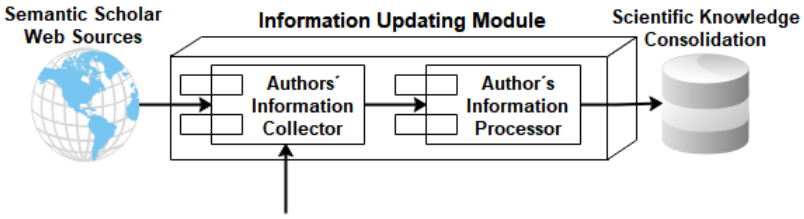

Figura 3.9: *Extracto de la arquitectura del módulo* Information Updating.

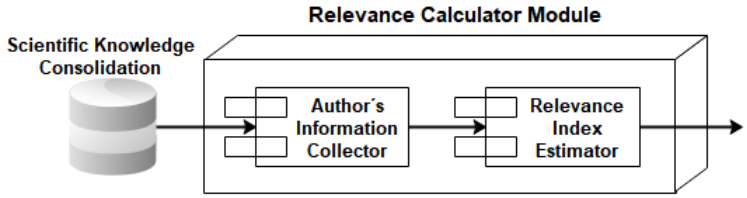

Figura 3.10: *Extracto de la arquitectura del módulo* Relevance Calculator.

publicaciones. La información recopilada de cada artículo publicado es la siguiente: título, abstract, autores, citas, campos de estudio, doi, temas, nombre de la revista y año de publicación. Debido a la falta de cuartiles e información de índices CORE en Semantic Scholar, se necesita el *Author's Information Processor*. Este módulo recibe la información del *Author's Information Collector* y la información relacionada con las revistas JCR y congresos CORE del módulo *Scientific Knowledge Consolidation*: nombre de la revista, factor de impacto, cuartil y categoría para revistas JCR; y nombre de la conferencia, acrónimo e índice CORE para conferencias CORE. Este componente consolida los datos desagregados en base a una estimación de similaridad TF-IDF entre el nombre de la revista extraído de Semantic Scholar y el nombre de la revista del módulo *Scientific Knowledge Consolidation*. Se debe tener en cuenta que los nombres de las revistas proporcionados por Semantic Scholar no siempre coinciden con los nombres oficiales proporcionados por JCR.

3.2.2. Módulo de cálculo de relevancia (Relevance Calculator module)

Este módulo calcula el índice de relevancia para cada autor y año. Presenta dos componentes (ver Fig. 3.10): el *Extractor de Información de Autores (Author's Information Collector)* y el *Estimador del Índice de Relevancia (Relevance Index Estimator)*. El primero extrae información del módulo *Scientific Knowledge Consolidation* para alimentar el *Relevance Index Estimator*. Este último calcula el índice de relevancia del autor para cada año. Se utiliza una media móvil de tres años para suavizar el impacto de años con pocas publicaciones o de baja calidad. Esto se aplica para evitar que la relevancia de un autor reputado sea de 0 en un año en el que no se realizan publicaciones. Del mismo modo, no se utiliza una media móvil de cinco años porque aplanaría demasiado la trayectoria, penalizando en exceso los años buenos y premiando los malos.

A continuación, se especifica la métrica utilizada para estimar el índice de relevancia. En primer lugar, se presentan las definiciones básicas con el fin de aclarar la explicación de la métrica propuesta:

- *Citations*: es el número total de citas de los artículos de un autor en ese año.
- J_P: es el número de artículos publicados en revistas JCR con un cuartil P, donde P es *Quartile 1* (Q1), *Quartile 2* (Q2), *Quartile 3* (Q3) o *Quartile 4* (Q4).
- C_P: es el número de artículos publicados en congresos CORE con índice P, donde P es A^*, A, B o C. Nótese que se ha considerado la equivalencia en importancia entre Q3 y Q4 con A^* y A respectivamente. Por lo tanto, los artículos CORE B y CORE C se consideran de importancia similar.
- A_P: representa el número de autores que han publicado artículos con un cuartil P en revistas JCR o un índice P en congresos CORE, donde P es Q1, Q2, Q3, Q4, A^*, A, B o C.
- w_j: se refiere al elemento de peso, con $w_j \geq 0$ y $\sum_{j=1}^{6} w_j = 1$.
- *Seniority*: se refiere al tiempo transcurrido, en años, entre el último trabajo publicado del autor y el primero.
- *Papers*: es el número total de artículos publicados de los autores durante su carrera.
- *CurrentSeniority*: se refiere al tiempo transcurrido entre el año para el cual se está calculando la relevancia y el año de publicación del primer artículo del autor.
- *CurrentYear*: se refiere al número de año para el cual se calcula la relevancia. Por ejemplo, para un autor cuyo primer año de publicación es 2020, el CurrentYear de 2020 será 1, el de 2021 será 2, y así sucesivamente.

Así, el índice de relevancia se calcula a través de la ecuación (3.4):

$$RelevanceIndex = \frac{Citations * WeightedCitationsValue}{YearsPenalty}, \tag{3.4}$$

donde:

$$WeightedCitationsValue = \frac{(S1 + S2 + S3 + S4 + S5)}{ProductivityPenalty}, \tag{3.5}$$

$$S1 = \frac{J_{Q1}}{A_{Q1}} * w_1, \tag{3.6}$$

$$S2 = \frac{J_{Q2}}{A_{Q2}} * w_2\,, \tag{3.7}$$

$$S3 = \frac{J_{Q3} + C_{A^*}}{A_{Q3} + A_{A^*}} * w_3\,, \tag{3.8}$$

$$S4 = \frac{J_{Q4} + C_A}{A_{Q4} + A_A} * w_4\,, \tag{3.9}$$

$$S5 = \frac{C_B + C_C}{A_B + AC} * w_5\,, \tag{3.10}$$

$$ProductivityPenalty = \frac{Papers}{(CurrentSeniority * w_6)}\,, \tag{3.11}$$

y:

$$YearsPenalty = ((Seniority - CurrentYear) * w_6) + 1 \tag{3.12}$$

Nótese que el número de citas ha sido ponderado por el valor de las citas. Este valor se basa en el cuartil de revistas donde se han publicado los artículos citados y el número de autores por publicación. De esta forma, se evita que un autor con $1,000$ citas en un artículo publicado en una revista de cuartil de Q4 tenga mayor relevancia que un autor con 500 citas en un artículo publicado en una revista de cuartil de Q1.

Además, para evitar premiar la productividad por encima de la calidad de los artículos, el valor ponderado de las citas se divide por el número de artículos publicados. Y se tiene en cuenta el tiempo transcurrido hasta el año en que se calcula la relevancia para evitar que se penalice a autores con una vida científica muy corta pero muy influyente.

3.2.3. Módulo de cálculo de novedad (Novelty Calculator module)

Este módulo calcula el índice de novedad para cada autor y año. Presenta dos componentes (ver Fig. 3.11): el *Extractor de Información de Autores (Author's Information Collector)* y el *Estimador del Índice de Novedad (Novelty Index Estimator)*. El primero extrae la información necesaria del módulo *Scientific Knowledge Consolidation* para alimentar el *Novelty Index Estimator*. Este último calcula el índice de novedad del autor para cada año. El índice de novedad ofrece una puntuación anual a cada autor en función de la originalidad de los artículos publicados en ese año. Se calcula comparando los temas de los artículos. Para ello se utilizan el título, palabras clave y abstract de los artículos científicos. Primero, se aplican algunas técnicas de NLP para recopilar los tipos de palabras relevantes (en este caso, solo se seleccionan los nombres propios y comunes). Esta

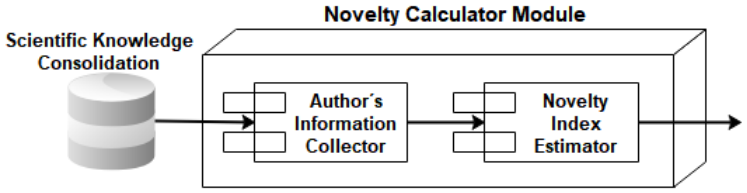

Figura 3.11: *Extracto de la arquitectura del módulo* Novelty Calculator.

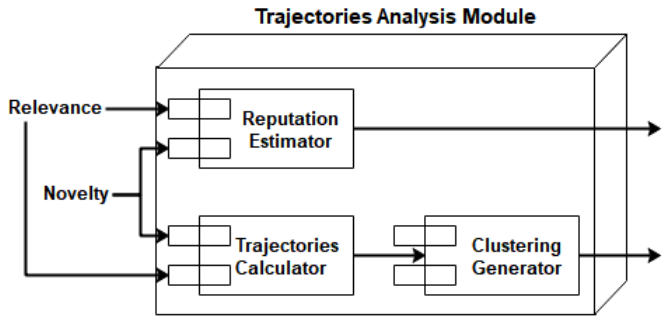

Figura 3.12: *Extracto de la arquitectura del módulo* Trajectories Analysis.

tarea mejora los resultados obtenidos a través de la medida de similaridad. Se aplican las técnicas TF-IDF y Cosine Similarity para calcular la medida de similaridad global entre los artículos.

El índice de novedad del autor en un año se estima a través de la media de los índices de similaridad calculados para cada artículo en ese año. La similaridad de los artículos se calcula de la siguiente manera: el primer artículo tiene similaridad igual a 1 ya que es totalmente original, el segundo artículo tiene similaridad igual a 1 menos la similaridad entre el primer y el segundo artículo, el tercer artículo tiene similaridad igual a 1 menos el máximo de las similaridades entre los artículos 1 y 3, y 2 y 3, y así sucesivamente.

3.2.4. Módulo de análisis de trayectorias (Trajectories Analysis module)

Este módulo realiza tres tareas diferentes. La primera consiste en estimar la puntuación de reputación final. La segunda construye las trayectorias de reputación de cada autor representando todo el trabajo científico del autor. La tercera agrupa a los autores en tres grupos en función de su reputación. Este módulo presenta tres componentes asociados con estas tareas (ver Fig. 3.12): el *Estimador de Reputación (Reputation Estimator)*, la *Calculadora de Trayectorias (Trajectories Calculator)* y el *Generador de Clusters (Clustering Generator)*, respectivamente.

El módulo recibe los índices de relevancia y novedad para cada autor de los módulos *Relevance Calculator* y *Novelty Calculator*.

Profundizando en el cálculo de las tareas, la puntuación de reputación se calcula como la media geométrica entre la ***relevancia*** y la ***novedad***. Nótese que el índice de novedad

está acotado por 1, pero el índice de relevancia no está acotado en absoluto. Así, se ha elegido la media geométrica ya que da más peso a los elementos pequeños que la media aritmética.

En el caso de las trayectorias de reputación, consisten en una secuencia de puntos. Cada punto tiene asociada información espacial dada por los índices de relevancia y novedad e información temporal dada por el año de dichos índices. Estos puntos se estiman como la combinación del índice de relevancia y el índice de novedad en ese año. Las trayectorias de reputación se trazan en un gráfico cuyo eje X es el índice de relevancia y el eje Y es el índice de novedad.

Finalmente, para los grupos de autores basados en las trayectorias de reputación, se realiza un grid análisis. La cuadrícula se divide en 25 celdas para convertir las trayectorias en matrices unidimensionales. Se utilizan 25 celdas debido al compromiso entre la complejidad que se introduce al insertar más celdas y la granularidad necesaria para obtener buenos resultados. Una vez convertidas todas las trayectorias, se aplica el algoritmo *Dynamic Time Warping* (DTW) (Senin, 2008) para calcular las distancias. Este algoritmo permite medir la similaridad entre secuencias temporales, obteniendo un buen ajuste incluso con desfase temporal. Esto se consigue buscando el punto más cercano entre cada punto de las dos trayectorias permitiendo discernir formas similares que pueden estar desfasadas. Para ello, se genera una matriz de similaridad que compara las dos trayectorias y calcula las distancias entre todos los puntos de cada trayectoria; y, en base a esta matriz, se selecciona la trayectoria más óptima. Finalmente, se utiliza el algoritmo Hierarchical Clustering (Johnson, 1967) para agrupar a los diferentes autores.

3.3. Reputación de dominios web: DOCRIW

Domains Classifier based on Risky Websites (DOCRIW) es una plataforma innovadora centrada en la detección de sitios web potencialmente peligrosos. Por lo tanto, puede clasificar los sitios web en sitios con riesgo (*risky*) o sitios sin riesgo (*non-risky*) en base a su reputación. Para ello, extrae conocimiento de fuentes de información web externas y realiza predicciones cuando no hay información disponible. Para hacer estas predicciones, DOCRIW crea medidas de similaridad para entrenar algoritmos de *Machine Learning* (ML) y utiliza métodos de optimización para seleccionar el mejor modelo y los parámetros adecuados. Nótese que el enfoque propuesto hace foco en el acceso directo a un sitio web por parte de los usuarios (es decir, cuando los usuarios intentan ser engañados).

Respecto a la arquitectura general del sistema, presenta cuatro módulos principales (ver Fig. 3.13): el módulo *Extracción y Validación de Dominios (Domains Extraction and Validation)*, el módulo *Extracción de Variables basadas en Host (Host-based Variables Extraction)*, el módulo de *Clasificación (Classification)* y el módulo de *Actualización de*

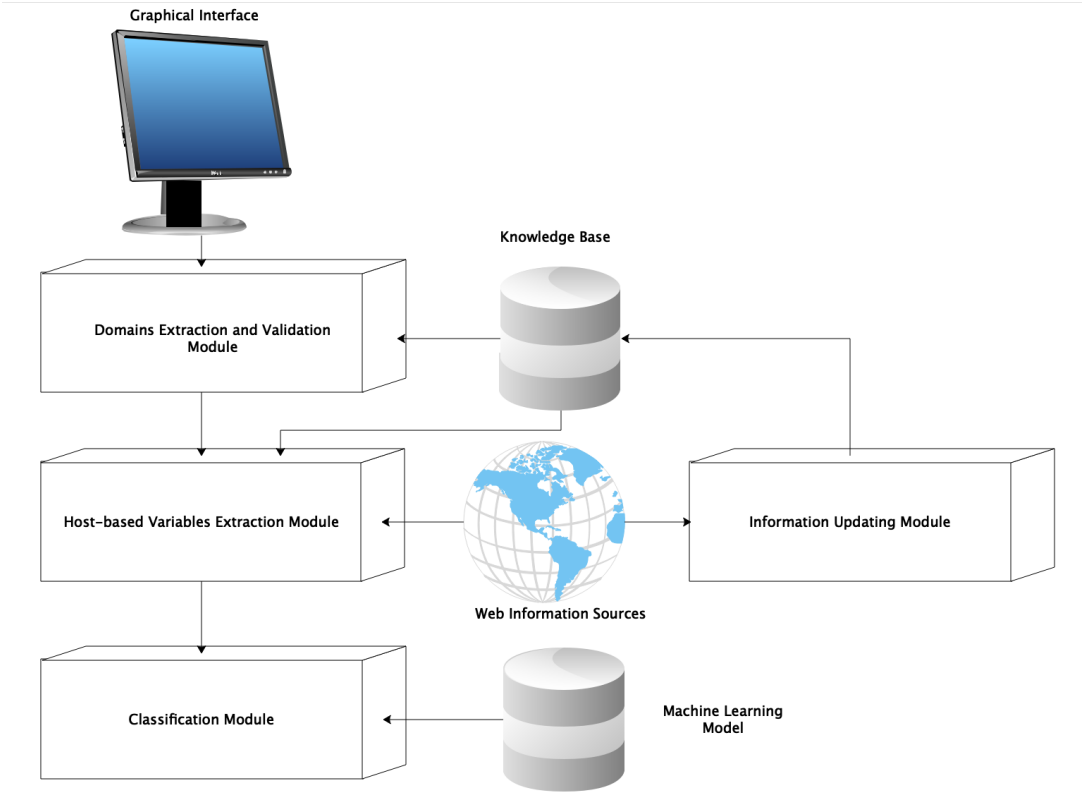

Figura 3.13: *Extracto de la arquitectura de* DOCRIW.

Información (Information Updating). Además, el sistema también contiene una *Interfaz Gráfica (Graphical Interface)* y dos bases de datos: la *Base de Conocimiento (Knowledge Base)* y el *Modelo de Aprendizaje Máquina (Machine Learning Model)*. La *Graphical Interface* se encarga de la interacción con los usuarios. La *Knowledge Base* es una base de datos de ElasticSearch (Gormley y Tong, 2015) que organiza el conocimiento recopilado de las *Fuentes de Información Web (Web Information Sources)*. El *Machine Learning Model* incluye un clasificador previamente entrenado. A continuación, se describen el resto de módulos.

3.3.1. Módulo de extracción y validación de dominios (Domains Extraction and Validation module)

Este módulo procesa *Uniform Resource Locator* (URL) extrayendo sus dominios correspondientes y analizándolos. Para lograr estas tareas, utiliza el módulo *Knowledge Base* para obtener los dominios *risky* previamente etiquetados.

El módulo presenta dos componentes: el *Analizador de URLs (URL Analyzer)* y el *Evaluador de Dominios (Domains Evaluator)* (ver Fig. 3.14). El primero recibe información de la *Graphical Interface* y actúa en respuesta a las solicitudes realizadas por los usuarios. La información proporcionada por la *Graphical Interface* puede ser URLs com-

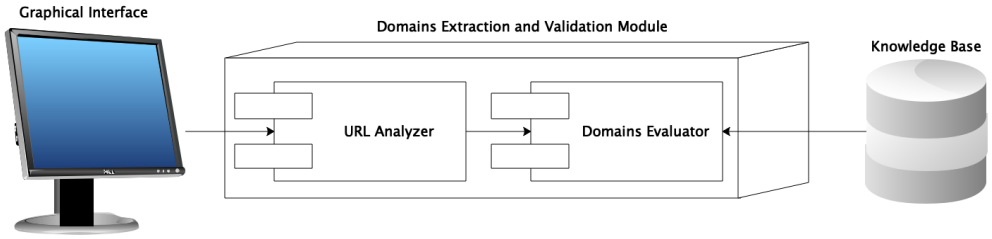

Figura 3.14: *Extracto de la arquitectura del módulo* Domains Extraction and Validation.

pletas o dominios previamente procesados. El *URL Analyzer* evalúa el dominio propuesto en ambas situaciones. Por lo tanto, verifica si el dominio es correcto (es decir, el código de estado es igual a 200) y detecta posibles redirecciones a páginas de destino. En este caso se incluyen todas las landing pages para ser analizadas, extrayendo los dominios asociados. El componente *Domains Evaluator* compara los dominios obtenidos y los dominios almacenados en la base de datos. Cuando se encuentran coincidencias, el dominio se etiqueta como *risky*. Cuando ninguno de los dominios coincide, el módulo envía el dominio original al módulo *Host-based Variables Extraction*.

3.3.2. Módulo de extracción de variables basadas en host (Host-based Variables Extraction module)

El módulo *Host-based Variables Extraction* recopila nuevas variables del host a través de la API REST de Whois (Whois.com, 2019), que forma parte de las *Web Information Sources*. Ese módulo proporciona información sobre la ciudad, país, fecha de creación, fecha de vencimiento y correo electrónico. Así, este módulo caracteriza el dominio analizado.

Respecto a la arquitectura del módulo (ver Fig. 3.15), consta de dos componentes: el *Extractor de Variables basadas en Host (Host-based Variables Collector)* y el *Limpieza de Datos (Data Cleaning)*. El primero gestiona la información proporcionada por la API de Whois, construyendo un conjunto de datos como salida. El segundo aborda la tarea de limpieza unificando los resultados. Por ejemplo, las abreviaturas de los países se adaptan de acuerdo con el código ISO (International Organization for Standardization, 2019) y se normalizan las posibles discrepancias entre los valores (por ejemplo, un nombre de ciudad con acentos y el mismo nombre de ciudad sin ellos).

3.3.3. Módulo de clasificación (Classification module)

Este módulo clasifica los dominios con las etiquetas *risky* o *non-risky* cuando no se encuentran en la *Knowledge Base*. Utiliza las variables generadas por el módulo *Host-based Variables Extraction* para alimentar el *Machine Learning Model* con el fin de obtener un

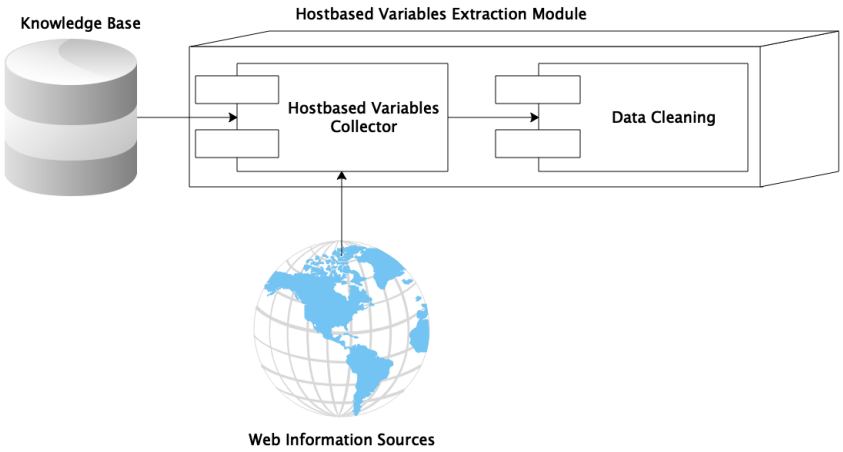

Figura 3.15: *Extracto de la arquitectura del módulo* Host-based Variables Extraction.

valor predicho para los dominios. El *Machine Learning Model* ha sido seleccionado en base a resultados empíricos. Más adelante se explicará el estudio completo para seleccionar los elementos relacionados con este modelo. El modelo incluye una definición de la similaridad entre dominios, un algoritmo *Logistic Regression* (LR), un umbral para la probabilidad proporcionada por el algoritmo y un conjunto de dominios de referencia.

Respecto a la arquitectura del módulo (ver Fig. 3.16), consta de dos componentes: el *Calculadora de Similaridad (Similarity Creator)* y el *Clasificador (Classifier)*. El primero calcula similaridades entre el nuevo dominio y cualquiera de los dominios del conjunto de referencia, para cada variable (es decir, nombre de dominio, ciudad, país, fecha de creación, fecha de vencimiento y correo electrónico). La similaridad basada en el nombre de dominio se calcula utilizando la distancia *Levenshtein* (Sarkar, 2016). Las otras cinco similaridades (correspondientes a las variables basadas en host) evalúan si dos dominios tienen el mismo valor para la variable correspondiente o no. Por ejemplo, para la variable país, la similaridad es 0 cuando los dos dominios están alojados en dos países diferentes y es 1 cuando los dos dominios están alojados en el mismo país. A continuación, se calcula una similaridad global entre el nuevo dominio y cualquiera de los dominios del conjunto de referencia como un promedio ponderado de las similaridades anteriores. Estos pesos son proporcionados por el *Machine Learning Model*.

El segundo componente del módulo *Classification* es el *Classifier*. El algoritmo LR proporcionado por el *Machine Learning Model* se alimenta con el vector de similaridades globales previamente calculado para obtener una predicción (entre 0 y 1) del riesgo. Dado un umbral de probabilidad de corte predefinido que maximiza el rendimiento general, el dominio se etiqueta como *risky* o *non-risky*.

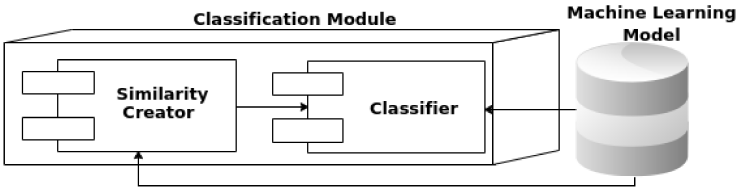

Figura 3.16: *Extracto de la arquitectura del módulo* Classification.

3.3.4. Módulo de actualización de información (Information Updating module)

Este módulo recopila datos de las *Web Information Sources* para actualizar la información almacenada en la *Knowledge Base*. Por tanto, obtiene nuevos dominios maliciosos de *AA419* (db.aa419.org, 2019) y *MalwareURL.com* (MalwareURL.com, 2019), dos sitios web públicos que identifican dominios de riesgo y ponen estos datos a disposición como un servicio público.

Esta tarea se ejecuta periódicamente actualizando el registro anterior con la nueva información recopilada. Este módulo almacena la información en el índice *Risky Domains* del módulo *Knowledge Base*.

3.3.5. Proceso de etiquetado de dominios (Labeling Domains Process)

El framework DOCRIW aborda el proceso de etiquetar un dominio como *risky* o *non-risky*. Este proceso describe las interacciones entre los módulos de DOCRIW para etiquetar un dominio como *risky* o *non-risky* según su naturaleza maliciosa o fraudulenta. Implica siete pasos secuenciales y una decisión.

El flujo de trabajo comienza procesando la URL proporcionada por los usuarios en el paso *URL processing* (ver Fig. 3.17). Este input podría ser un nombre de dominio o una URL específica. En el segundo caso, se procesa la URL para extraer el nombre de dominio adecuado. Luego, este dominio se compara con los dominios almacenados en el módulo *Knowledge Base* en el paso *Domain found in Risky Domains DB*. Si se encuentra el dominio, el próximo paso es *Risky domain labeling*. Allí se proporciona la respuesta del sistema. Esta respuesta es el dominio etiquetado como *risky* con probabilidad igual a 1. Todas estas tareas se logran en el módulo *Domains Extraction and Validation*. Por el contrario, si no se encuentra el nombre de dominio, el siguiente paso es la *Host-based variables collection*. Allí, el sistema recopila información sobre el dominio de acuerdo con las cinco características seleccionadas basadas en host (es decir, ciudad, país, fecha de creación, fecha de vencimiento y correo electrónico). Esta información se normaliza en el paso *Normalización de Información (Data normalization)*. Estas operaciones se logran mediante el módulo *Host-based Variables Extraction*. Una vez completada esta parte, se calculan las similaridades globales correspondientes, utilizando la información del dominio

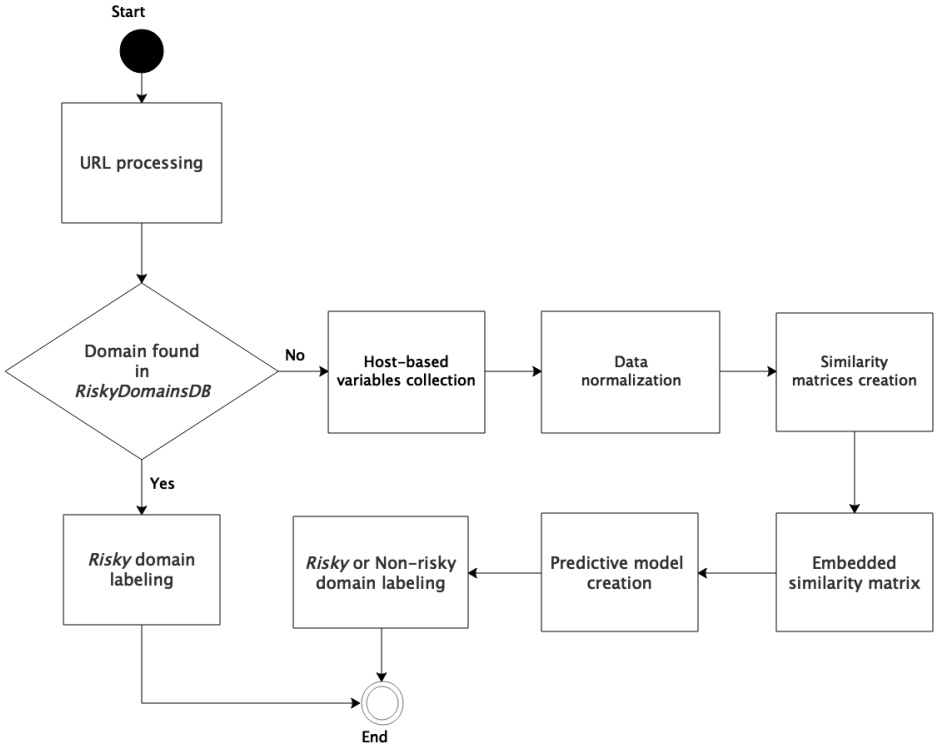

Figura 3.17: *Extracto del flujo de trabajo del* Domain Labeling.

actual y los dominios utilizados para entrenar el modelo ML. Finalmente, el proceso termina de hacer una predicción en el paso de *Cálculos del Modelo Predictivo (Predictive model creation)* y el etiquetado final se logra en el paso de *Etiquetado del dominio como risky o non-risky (Risky or non-risky domain labeling)*. En la arquitectura DOCRIW, el módulo *Classification* es responsable de estas tareas.

Además, el framework actualiza la información del sistema utilizando las fuentes de información web correspondientes para adquirir nuevos dominios etiquetados como *risky*. Esto es ejecutado por el módulo *Information Updating* una vez al día.

Capítulo 4

Experimentos

4.1. Unified Knowledge Compiler (UNIKO)

Esta sección aborda un conjunto de experimentos para ilustrar la viabilidad del framework UNIKO y los principales módulos asociados a los sistemas: IRC y KMC. Por cuestiones de reproductibilidad, se ha utilizado en los experimentos un conjunto de diez artículos fijos seleccionados al azar de la base de datos. Los DOI de los artículos seleccionados se presentan en la Tabla 4.1 para fines de identificación. Los autores asociados se presentan en la Tabla 4.2.

Respecto a las funcionalidades probadas, se han considerado aquellas que están asociadas a la recuperación de información de artículos específicos y sus operaciones relacionadas (por ejemplo, cálculo de reputación tanto de artículos como de sus correspondientes autores, cálculo de análisis de sentimiento, tareas de similaridad y clustering). También se incluyen las funcionalidades de similaridad entre artículos y autores, y las tareas de generación de clustering.

Article	DOI
1	10.3758/s13420-017-0270-5
2	10.3758/s13420-017-0263-4
3	10.3758/s13420-017-0294-x
4	10.3758/s13420-017-0288-8
5	10.3758/s13420-017-0258-1
6	10.3758/s13420-016-0222-5
7	10.3758/s13420-017-0290-1
8	10.3758/s13420-017-0306-x
9	10.3758/s13420-017-0277-y
10	10.3758/s13420-017-0259-0

Tabla 4.1: *Artículos con su DOI seleccionados para realizar los experimentos.*

Article	Authors
1	T. R. Smith, M. J. Beran, M. E. Young
2	W. A. Roberts, H. MacDonald, L. Brown, K. Macpherson
3	L. Fields
4	K. Andrews
5	B. Alonso-Álvarez, L. A. Pérez-González
6	J. Vonk
7	J. Fagot, R. Malassis, T. Medam
8	G. Miguez, B. McConnell, C. W. Polack, R. R. Miller
9	M. Uengoer, J. M. Pearce, H. Lachnit, S. Koenig
10	J. F. Briggs, B. P. Olson

Tabla 4.2: *Autores de los artículos seleccionados.*

Para calcular la reputación de los autores se ha utilizado la ecuación (3.1) (ver Apartado 3.1.1). Los parámetros de reputación se han corregido para dar más importancia a los artículos ($\omega_4 = 0{,}4$) y las citas ($\omega_2 = 0{,}3$) sobre el seniority y las citas influyentes ($\omega_1 = \omega_3 = 0{,}1$).

Los umbrales utilizados para normalizar las puntuaciones de reputación se seleccionaron ad-hoc de acuerdo con restricciones moderadas. En el caso de los autores, se ha configurado un umbral de citas influyentes a 2 citas, se ha asignado el umbral de número de citas a 150 citas, se ha fijado el umbral de seniority a 15 años y se ha fijado el número de artículos publicados a 50. El umbral para las citas de artículos se ha fijado en 100.

Finalmente, para calcular la reputación de un artículo, el parámetro de compensación α se ha fijado en $0,5$. Así, se da la misma importancia al número de citas del artículo y a la reputación de los autores.

La sección 4.1.1 presenta los experimentos relacionados con la reputación. La sección 4.1.2 describe el experimento relacionado con el análisis de sentimientos del texto de los artículos. La sección 4.1.3 concluye los experimentos que abordan las tareas de similaridad y clustering.

4.1.1. Evaluación de la reputación de autores y artículos

Este experimento está relacionado con el proceso de cálculo de la puntuación de reputación. Se ha implementado tanto para autores como para artículos. Como se describió anteriormente, el proceso para calcular la puntuación de reputación de los artículos depende de la puntuación de reputación de los autores. Por lo tanto, el experimento comienza recogiendo la información de los autores asociados a los artículos seleccionados.

Se recopilan los valores de las características consideradas (es decir, *citas influyentes*, *citas*, *artículos* y *seniority*). Luego se obtienen las puntuaciones de reputación. Se han

detectado algunos problemas inciertos una vez que se completa la tarea.

Hay cuatro autores con valores superiores en alguna de las características a los umbrales prefijados que les han sido asignados. Estos desbordamientos conducen a producir puntajes de reputación superiores. Ejemplos de ellos son *M. E. Young*, *J. Fagot*, *R. R. Miller* y *H. Lachnit* (ver Tabla 4.2). De hecho, todos ellos son investigadores experimentados durante muchos años con varias publicaciones en sus dominios. Por lo tanto, se puede decir que los umbrales moderados predefinidos son razonables y producen puntuaciones de reputación aceptables.

Hay autores con puntuaciones de reputación igual a cero. Ejemplos de ellos son: *M. J. Beran*, *W. A. Roberts*, *H. MacDonald*, *B. McConnell*, *J. M. Pearce*, *S. Koenig* y *B. P. Olson*. Analizando los valores de las características consideradas se puede concluir que ninguna de ellas tiene información relevante. Este percance es causado por problemas durante el proceso de recuperación de información. Estos problemas probablemente estén relacionados con desajustes en los nombres de los autores (aunque se han incluido algunas técnicas basadas en soft computing). Esto lleva a considerar este tema como un punto a analizar profundamente en el trabajo futuro con el fin de mejorar el rendimiento del framework.

Hay un caso especial. El autor *J. F. Briggs* presenta un valor de *seniority* de 72 años. Esto parece extraño a simple vista. Profundizando en este caso, se puede concluir que es posible que se superponga información correspondiente a diferentes autores con las mismas firmas, siendo esta una situación no deseable. Nótese que se utilizan recursos como ORCID (ORCID, 2018) y ResearcherID (Reuters, 2018) para mitigar este problema.

Una vez obtenidas las puntuaciones de reputación de los autores, se calcula la reputación de los artículos. En este caso las características consideradas son el promedio de las reputaciones de los autores asociados y el número de citas recibidas por el artículo.

En este experimento, dos artículos (es decir, *3* y *6*) presentan puntuaciones superiores a 0,5. Esto está alineado con una cantidad aceptable de citas y una excelente reputación de sus autores (ver Tabla 4.3).

Nótese que los artículos *2* y *9* cuya autoría incluye algunos de los autores que no pudieron ser identificados para extraer sus características, son penalizados (ver Tabla 4.3 y Tabla 4.4). Ambos tienen una cantidad aceptable de citas pero una reputación promedio mediocre.

4.1.2. Evaluación del análisis de sentimiento

Este experimento evalúa la influencia de la CNN en el proceso de análisis de sentimiento (ver Sección 3.1.3). La premisa principal detrás de este experimento es que SentiWordNet es apropiado para obtener la puntuación de sentimientos de los artículos científicos

Author	Infl.	Citations	Papers	Sens.	Rep.
T. R. Smith	0	2	3	9	**0,15**
M. J. Beran	NA				**0**
M. E. Young	111	843	100	40	**1**
W. A. Roberts	NA				**0**
H. MacDonald	NA				**0**
L. Brown	2	35	3	18	**0,39**
K. Macpherson	13	106	11	11	**0,55**
L. Fields	21	137	39	41	**0,89**
K. Andrews	10	124	12	17	**0,64**
B. Alonso-Álvarez	1	6	6	9	**0,23**
L. A. Pérez-González	3	39	19	23	**0,53**
J. Vonk	68	596	49	15	**0,99**
J. Fagot	116	898	85	30	**1**
R. Malassis	0	0	1	0	**0,01**
T. Medam	0	0	3	1	**0,04**
G. Miguez	1	31	17	6	**0,33**
B. McConnell	NA				**0**
C. W. Polack	2	21	15	5	**0,33**
R. R. Miller	144	1557	224	46	**1**
M. Uengoer	5	64	25	5	**0,49**
J. M. Pearce	NA				**0**
H. Lachnit	68	568	87	37	**1**
S. Koenig	NA				**0**
J. F. Briggs	6	51	56	72	**0,8**
B. P. Olson	NA				**0**

Tabla 4.3: *Influence citeps (Infl.), Citations, Papers, Seniority (Sens.) and Reputation (Rep.) of authors associated to the articles selected.*

(Sendhilkumar *et al.*, 2013). Utiliza el módulo *Sentiment Calculator* del sistema IRC (ver Sección 3.1.1).

Por tanto, el experimento consta de dos análisis de sentimiento diferentes. El primer análisis utiliza el módulo completo que comprende principalmente los componentes *Sentiment Evaluator* y *Neural Network Manager*. Incluye el léxico de SentiWordNet y predice los valores de sentimiento para las palabras que son neutrales o que no coinciden en el léxico con la CNN ya entrenada (consulte la Sección 3.1.3). El segundo análisis solo usa el componente *Sentiment Evaluator* y el conocido léxico para producir el valor de sentimiento final del texto.

En este caso todos los artículos seleccionados presentan su respectivo abstract, pero no los textos completos. La falta de textos completos es una situación común que dificulta la evaluación del sentimiento de los artículos. No obstante, los abstracts suelen incluir

Article	Citations	Authors Rep.	Reputation
1	0,57	0,38	**0,48**
2	0,44	0,23	**0,34**
3	0,46	0,89	**0,68**
4	0,50	0,64	**0,35**
5	0,31	0,38	**0,34**
6	0,05	0,99	**0,52**
7	0,34	0,35	**0,35**
8	0,40	0,41	**0,41**
9	0,25	0,37	**0,31**
10	0,48	0,40	**0,44**

Tabla 4.4: *Reputación de los artículos basada en la reputación de los autores y las citas normalizadas.*

suficientes palabras para hacer una estimación aceptable. Dado que el abstract podría limitarse a un número máximo de caracteres, no se producirían suficientes coincidencias con el léxico.

Evaluando el rendimiento, la CNN no modifica significativamente el resultado final. Sin embargo, hay casos específicos (*8* y *9* en la Tabla 4.5) donde la polaridad del sentimiento se ha modificado de negativa a positiva cuando solo se ha utilizado el léxico. Los resultados se asocian con una palabra (o un conjunto de palabras) cuya polaridad es predicha por la CNN como claramente negativa. El artículo *8* muestra una modificación más significativa, probablemente asociada a una palabra (o conjunto de palabras) predichas por la CNN que se identifican claramente con polaridad negativa. El artículo *9* es similar pero las palabras predichas están más cerca de la polaridad neutra.

Respecto al rendimiento de la CNN para hacer predicciones, se puede decir que la CNN enriquece el resultado final. Se pueden encontrar casos específicos donde la polaridad del sentimiento ha cambiado de positiva (cuando solo se usa el léxico) a negativa (cuando se agrega la CNN).

Nótese que como se comentó anteriormente, los abstracts no podían proporcionar suficiente número de palabras para apreciar cambios sustanciales en la evaluación. Este problema también se considerará en el trabajo futuro que intenta recopilar textos completos junto con la posibilidad de volver a entrenar a la CNN. Este último permitirá incrementar su vocabulario y producir adaptaciones específicas a los temas o dominios de los artículos evaluados.

Article	Sent. lexicon + CNN	Sent. lexicon
1	0,12	0,13
2	0,08	0,07
3	0,11	0,13
4	0,03	0,1
5	0,1	0,1
6	0,11	0,16
7	0,09	0,1
8	-0,03	0,08
9	-0,05	0,1
10	0,07	0,09

Tabla 4.5: *Valores de sentimiento de los abstracts utilizando el lexicon más la CNN y sólo el lexicon.*

4.1.3. Evaluación de la similaridad y clustering

Este experimento evalúa el rendimiento del módulo *Similarity and Clustering Evaluator* incluido en el sistema KMC (ver Sección 3.1.2). Se divide en dos experimentos independientes: el primero evalúa el *Article Similarity Manager* y el *Author Similarity Manager*, y los resultados (matrices de similaridad) producidos por los componentes *Article Similarity Calculator* y *Author Similarity Calculator*. El segundo experimento considera los componentes *Artículo Clustering Calculator* y *Author Clustering Calculator*, los cuales son responsables de generar clusterings de acuerdo a un conjunto de artículos o autores.

El primer experimento comienza usando los componentes *Article Similarity Calculator* y *Author Similarity Calculator* para producir las matrices de similaridad correspondientes. Estas matrices se almacenan en un índice específico de la base de datos. Luego, el *Article Similarity Manager* y el *Author Similarity Manager* filtran los artículos seleccionados y sus autores relacionados. Esto genera un vector por artículo (diez para este experimento) y un vector por autor (veinticinco para este experimento) con valores de similaridad y longitudes iguales al número de artículos y autores, respectivamente, almacenados en la base de datos. Estos vectores se ordenan de mayor a menor valor conservando el número de posición original para encontrar su correspondencia con cada uno de los artículos. Finalmente, los cinco primeros elementos se recogen de los vectores por ser los más similares.

Evaluando en detalle el experimento, se puede concluir que produce bastantes similaridades satisfactorias. Por ejemplo, el artículo identificado como *1* en este documento (consulte la tabla 4.1) tiene estas palabras clave: *gambling*, *monkeys*, *risk preference*, *risk sensitivity* y *signaled reinforcement* (ver Tabla 4.6). Sus artículos más similares tienen palabras clave como: *rhesus monkeys*, *place preference*, *subjective reinforcement* o *gaze*

Article	Article keywords	Similar articles
1	gambling monkeys risk preference risk sensitivity signaled reinforcement	10.3758/s13420-017-0310-1 10.3758/s13420-015-0185-y 10.3758/s13420-013-0118-6 10.3758/s13420-017-0295-9 10.3758/s13420-015-0204-z
2	radial maze rats proactive interference spatial memory	10.3758/s13420-014-0163-9 10.3758/s13420-017-0306-x 10.3758/s13420-015-0175-0 10.3758/s13420-015-0186-x 10.3758/s13420-013-0107-9
3	blocking errorless learning reinforcement contiguity stimulus fading	10.3758/s13420-013-0102-1 10.3758/s13420-014-0166-6 10.3758/s13420-017-0295-9 10.3758/s13420-015-0183-0 10.3758/s13420-015-0191-0
4	apes social cognition theory of mind	10.3758/s13420-017-0268-z 10.3758/s13420-012-0093-3 10.3758/s13420-014-0165-7 10.3758/s13420-017-0258-1 10.3758/s13420-016-0220-7

Tabla 4.6: *Extracto de cuatro de los artículos seleccionados con sus keywords y los artículos más similares.*

sensitivity. Estas palabras clave comparten algunos de los términos que están presentes

en las palabras clave del artículo *1*.

En el caso de los autores, el comportamiento es muy similar. Por ejemplo, el autor *M. E. Young* tiene como palabras claves: *stimulus*, *delay*, *causation*, *discrimination* y *pigeon* (ver Tabla 4.7). Los autores más similares son: *E. Wasserman*, *B. Ploog*, *W. A. Bowditch*, *B. Williams* y *M. Elcoro*. Tienen palabras clave como: *pigeon function*, *discrimination reversal* o *delay reinforcement*. Estas palabras clave comparten términos similares a las palabras clave originales asignadas a *M. E. Young*.

Author	**Author keywords**	**Similar authors**
M. E. Young	stimulus delay causation discrimination pigeon	E. Wasserman B. Ploog W. A. Bowditch B. Williams M. Elcoro
Lyn Brown	release interference rat memory precise judgment	H. Macdonald R. R. Miller M. S. Matell M. E. Bouton J. D. Crystal
L. Fields	class function stimulus effect class formation	E. Arntzen R. Nartey L. A. Pérez-González I. T. Tyndall C. Baizán
J. Vonk	chimpanzee lowland gorilla bear ursus group member observation	R. P. Gazes S. F. Brosnan F. Schnöller D. Beecham C. C. Luhmann

Tabla 4.7: *Extracto de cuatro de los autores seleccionados con sus keywords y los autores más similares.*

El segundo experimento trata de hacer grupos en base a las palabras clave de los elementos (los artículos y los autores). Así, dados todos los artículos del experimento, se obtienen sus palabras clave. Luego, se calculan el *tf-idf* y la similaridad del coseno. El método *Elbow* (Bertin y Atanassova, 2017) se usa para estimar la distancia de corte óptima. Este corte permite obtener el número óptimo de grupos a través de un *Hierarchical Clustering*. En el caso de los artículos, la distancia de corte es de 0,86 (ver Fig. 4.1), mientras que en el caso de los autores la distancia de corte es de 0,64 (ver Fig. 4.2).

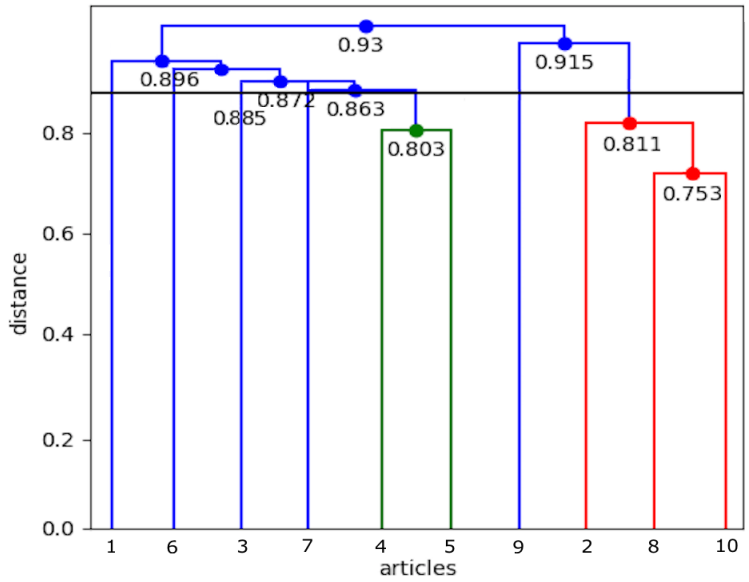

Figura 4.1: *Dendograma de los artículos seleccionados incluyendo la distance de corte.*

Ejemplos de los grupos de artículos obtenidos son los correspondientes a artículos identificados como *2*, *8* y *10* que comparten palabras clave como: *latent inhibition*, *memory*, *spatial working memory* y *conditioned inhibition* entre otros. Otro grupo lo forman los artículos *4* y *5*. En este caso estos artículos comparten la palabra clave *theory of mind*.

En el caso de los autores se obtienen 22 clusters (de 25 autores). Es decir, las similaridades entre autores en los artículos propuestos no son, en general, muy altas. Esto lleva a pensar que tanto los artículos como los autores presentan temas muy diferentes.

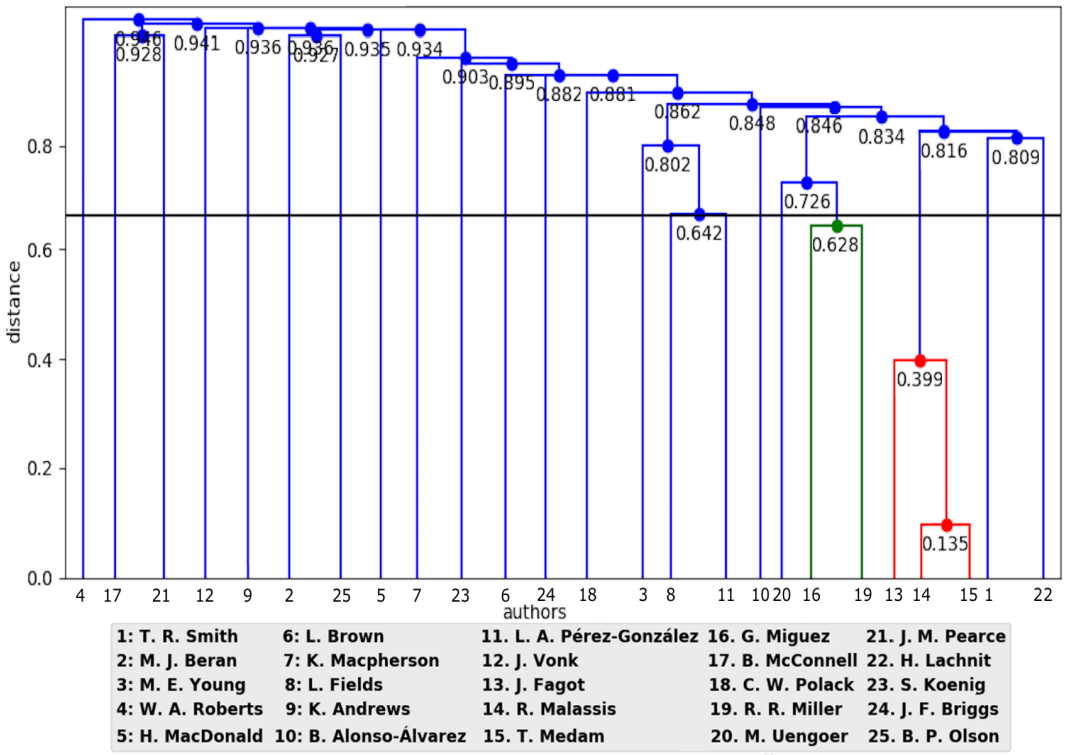

Figura 4.2: *Dendograma de los autores seleccionados incluyendo la distance de corte.*

4.2. Framework for Reputation Estimation of Scientific Authors (FRESA)

Esta sección aborda un conjunto de experimentos que evalúan las métricas propuestas de reputación, relevancia y novedad y validan el rendimiento general del framework FRESA . Para la puesta en marcha del sistema, el valor de las citas se pondera con los siguientes pesos: $w_1 = 0{,}4$, $w_2 = 0{,}25$, $w_3 = 0{,}15$, $w_4 = 0{,}1$, $w_5 = 0{,}05$ y $w_6 = 0{,}05$. Así, la importancia relativa otorgada a los cuartiles de las revistas en las que se publica un artículo disminuye a medida que aumenta el número de cuartiles. Además, al seniority se le otorga una relevancia moderada siguiendo la literatura previa en el dominio (Fernández-Isabel *et al.*, 2018).

Profundizando en los experimentos, el primero, descrito en la Sección 4.3.1, comparte una explicación completa sobre la trayectoria de reputación calculada para un autor reconocido. Esta trayectoria de reputación se interpreta heurísticamente en base a las publicaciones del autor para confirmar que las bases de datos utilizadas y las métricas propuestas son válidas.

El segundo experimento, descrito en la Sección 4.3.2, aborda tres problemas diferentes. En el primero se calculan el índice de relevancia, el índice de novedad y la puntuación

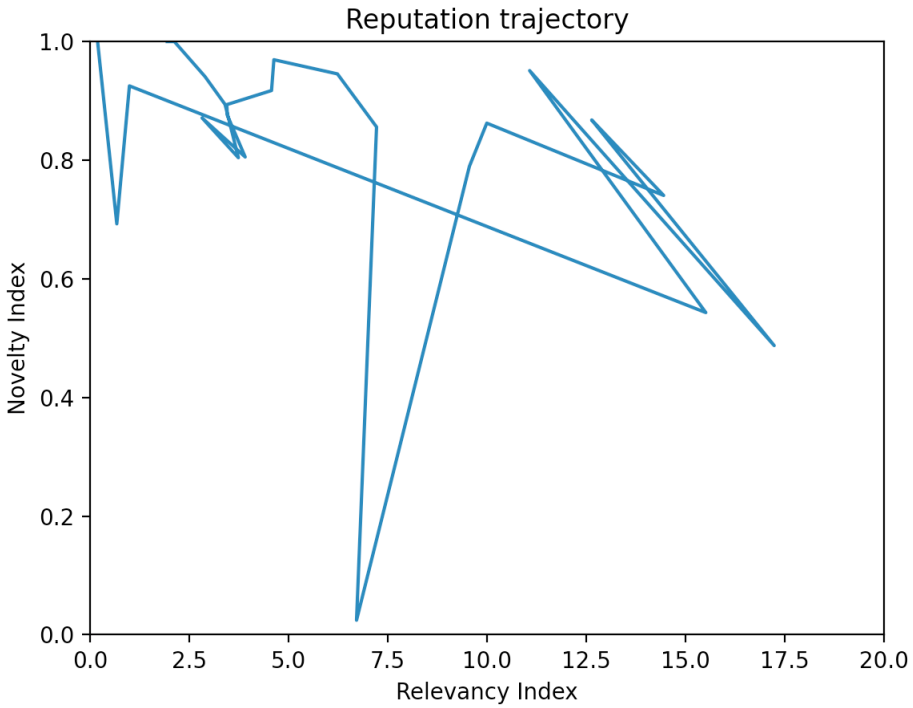

Figura 4.3: *Trayectoria de reputación de David Tax.*

de reputación y se comparan con el índice h y el índice i10. En el segundo, se hace una comparación para probar las limitaciones que presentan el índice h y el i10 cuando el autor presenta una carrera científica corta pero de gran impacto. Finalmente, se presenta un escenario hipotético para comparar el desempeño utilizando la métrica propuesta, el índice h, y el i10 cuando se comete un fraude.

El tercer experimento, descrito en la Sección 4.3.3 está enfocado en elaborar perfiles de acuerdo al desempeño científico de los investigadores en base a sus trayectorias de reputación. Ejecuta un algoritmo de clustering para lograr esta tarea para 45 autores.

4.2.1. Trayectoria de reputación de un autor reconocido

Este experimento se utiliza para validar heurísticamente el propósito de este artículo. Para ello, se calcula e interpreta la trayectoria de reputación de tres autores concretos a partir de sus publicaciones. Estos autores son David Tax, Terrance E. Boult e Isaac M. de Diego (véanse Fig. 4.3, Fig. 4.4 y Fig. 4.5). La tabla 4.8 muestra las puntuaciones de reputación de David Tax para cada año desde 1997 hasta 2020, así como los índices de relevancia y novedad de este autor. La tabla 4.9 muestra las puntuaciones de reputación de Terrance E. Boult para cada año desde 1985 hasta 2020, así como los índices de relevancia y novedad para este autor. Y la Tabla 4.10 muestra las puntuaciones de reputación de Isaac M. de Diego para cada año desde 1997 hasta 2022, así como los índices de relevancia y novedad para este autor.

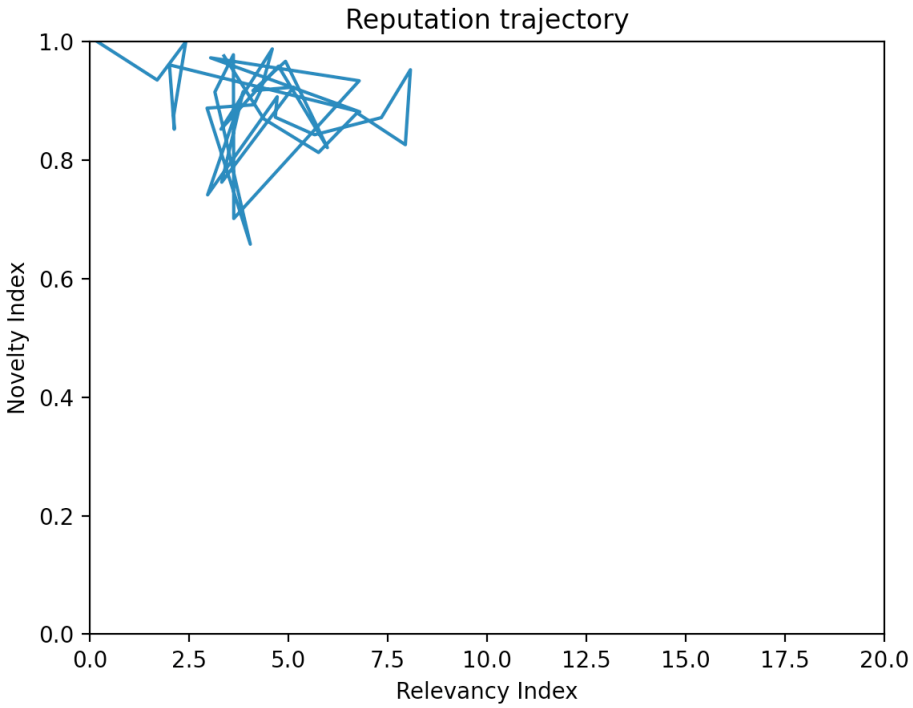

Figura 4.4: *Trayectoria de reputación de Terrance M. Boult.*

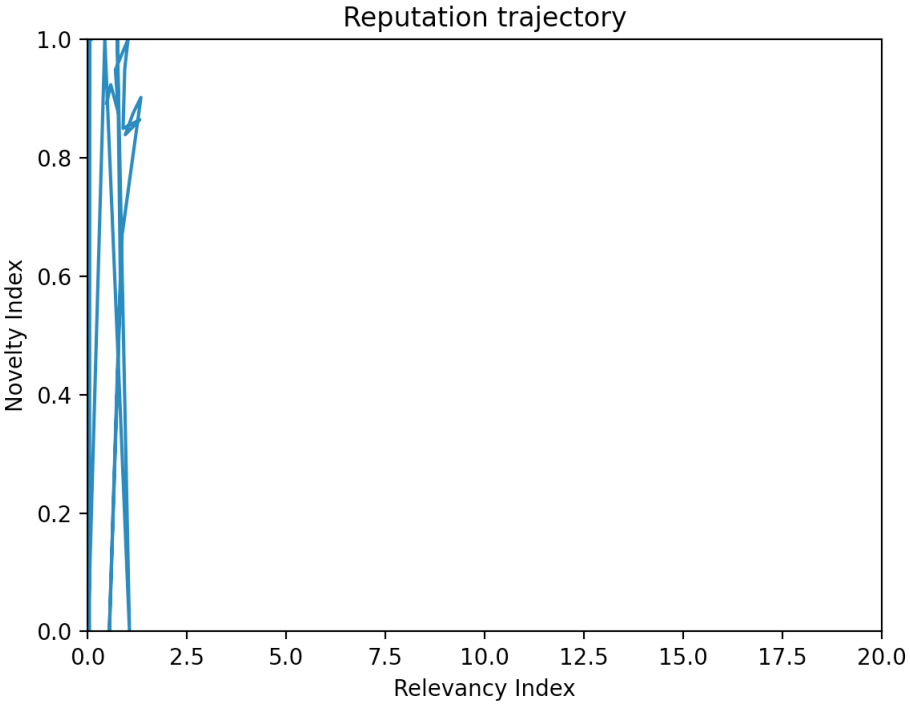

Figura 4.5: *Trayectoria de reputación de Isaac M. de Diego.*

En el primer año de publicación científica del autor, el índice de relevancia es de $0,19$ y el índice de novedad es de 1. Estos resultados son lógicos, ya que el trabajo realizado aún no ha tenido un impacto en la comunidad. Además, los artículos son novedosos ya

que son los primeros. En 1998, el autor publica 5 artículos. Dos de estos artículos están publicados en revistas de alto cuartil y son citados más de 130 veces, por lo que parece natural que su índice de relevancia aumente a $0,72$. Dado que los temas tratados este año son similares a los del año anterior, también es normal que el índice de novedad baje a $0,92$. En 1999, el autor publicó 7 artículos. Varios de ellos obtienen numerosas citas, alcanzando $1,803$ citas en un artículo publicado en una revista Q1. Por lo tanto, el índice de relevancia aumentó a 1,05. Como este índice se calcula en ciclos de tres años, (y el primer año penaliza) la subida del índice de relevancia no es muy pronunciada. El índice de novedad cae por la misma razón a 0,54. En 2000, el autor publicó 5 artículos nuevos. Uno de ellos alcanzó $8,136$ citas. Además, el primer año (1997) ya no se incluye en el cálculo del índice de relevancia para su cuarto año, lo que contribuye a un aumento muy pronunciado del índice de relevancia, alcanzando los $16,3$. En cambio, el índice de novedad aumenta nuevamente a 0,95, ya que los temas son novedosos en comparación con las publicaciones de años anteriores. En 2001, el autor publicó 4 artículos sin demasiadas citas. Además, tres de ellos se publican en congresos. De esta forma, el índice de relevancia tiene una ligera caída hasta los $11,76$, aunque sigue siendo elevado debido a la trayectoria de los últimos años. El índice de novedad en este año baja a 0,48 porque los temas de los artículos son similares. En 2002, el autor publicó 6 artículos. Tres de ellos se publican como primer autor, siendo el único autor y los otros dos compartiéndolos únicamente con otro autor. Además, alcanzó un elevado número de citas. El índice de relevancia de este año es el más alto de su carrera científica, $18,03$. El índice de novedad aumenta a 0,86. En 2003, el autor publicó 5 artículos con muy poco impacto. Las revistas donde se publicaron estos artículos tienen un cuartil bajo y los artículos apenas obtienen citas. El índice de relevancia baja a $13,26$, aunque sigue siendo alto debido a los dos años anteriores. En 2004, el autor publicó 14 artículos. Uno de ellos está publicado en una revista Q1. Este artículo está escrito por solo dos autores y David Tax es el primer autor. Además, obtiene $2,838$ citas. Sin embargo, el resto de artículos no tienen mucha repercusión. Por tanto, el índice de relevancia se mantiene casi constante, $13,21$. Sin embargo, debido a la falta de nuevos temas en estos artículos, el índice de novedad se reduce a $0,74$. En los años siguientes, la tendencia es la misma y los índices calculados explican la relevancia y novedad de las publicaciones (ver Tabla 4.8). En 2007, hay una caída significativa en el índice de novedad a $0,02$. Esto se produce debido a la publicación de solo dos artículos muy similares. En 2008, el índice de relevancia también disminuyó a $6,92$. En los años siguientes, el índice de relevancia comenzó a descender paulatinamente. Las publicaciones fueron teniendo cada vez menos repercusión, no superando las 45 citas en ninguna de ellas. Además, David Tax deja de publicar como primer autor y los artículos tienen un mayor número de autores.

Hay un patrón y una tendencia en la trayectoria del autor. El patrón muestra que el autor intercala años en los que publica artículos con temas novedosos y de menor impacto, con años en los que publica artículos de alto impacto sobre temas en los que el autor es experto. En los primeros, el autor obtiene altos índices de novedad y bajos índices de

Year	Relevance index	Novelty index	Reputation score
1997	0,19	1,00	0,19
1998	0,72	0,92	0,66
1999	1,05	0,54	0,57
2000	16,32	0,95	15,50
2001	11,76	0,48	5,64
2002	18,03	0,86	15,51
2003	13,26	0,86	11,40
2004	13,21	0,74	9,78
2005	14,86	0,86	12,78
2006	10,35	0,78	8,07
2007	9,82	0,02	0,20
2008	6,92	0,85	5,88
2009	7,37	0,94	6,93
2010	6,37	0,96	6,12
2011	4,73	0,91	4,30
2012	4,65	0,89	4,14
2013	3,46	0,80	2,77
2014	3,78	0,87	3,29
2015	2,86	0,80	2,29
2016	3,94	0,87	3,43
2017	3,50	0,88	3,08
2018	3,48	0,94	3,27
2019	2,96	1,00	2,96
2020	2,15	1,00	2,15

Tabla 4.8: *Relevance index, novelty index y reputation score por años de David Tax.*

relevancia. Y en los últimos, el autor obtiene bajos índices de novedad y altos índices de relevancia. La tendencia muestra que el impacto del trabajo científico es ascendente hasta 2005 y descendente desde 2006 en adelante.

En los primeros años de Terrance E. Boult, la relevancia es baja porque los artículos no reciben muchas citas. Sin embargo, su puntuación de relevancia es superior a la de Tax en los tres primeros años. Terrance publica artículos solo o con muy pocos autores en revistas de cuartil superior, lo que explica este hecho. La novedad de sus artículos también es mayor que la de Tax porque no se centra en los mismos temas. En 1991 y 1993, Bault obtiene una relevancia superior a $6,00$ porque supera las 200 y 500 citas respectivamente. Sin embargo, Boult no alcanza una relevancia tan alta como Tax porque publica muchos artículos durante estos años en revistas de cuartil inferior que no tienen citas. La novedad sigue siendo alta porque los temas que trata siguen siendo variados. Durante las décadas de 1990 y 2000, el índice de relevancia se mantiene constante en torno a $4,00$, lejos de los índices de relevancia alcanzados por Tax en la década de 2000 porque Bault sigue

Year	Relevance index	Novelty index	Reputation score
1985	0,19	1,00	0,44
1986	1,69	0,93	1,44
1987	2,41	1,00	1,80
1988	2,11	0,87	1,50
1989	2,12	0,85	2,05
1990	2,00	0,96	3,31
1991	6,79	0,88	2,44
1992	3,04	0,97	1,48
1993	6,78	0,93	2,39
1994	3,62	0,70	1,25
1995	3,61	0,97	2,49
1996	3,15	0,91	1,40
1997	4,04	0,65	2,22
1998	2,95	0,88	4,48
1999	4,13	0,89	0,44
2000	4,30	0,91	1,44
2001	4,59	0,98	1,80
2002	3,30	0,85	1,50
2003	4,74	0,95	2,05
2004	5,98	0,82	3,31
2005	4,92	0,96	2,44
2006	4,10	0,91	1,48
2007	5,12	0,92	2,39
2008	3,32	0,76	1,25
2009	3,86	0,91	2,49
2010	2,96	0,74	1,40
2011	4,71	0,90	2,22
2012	4,66	0,87	4,48
2013	5,66	0,84	2,36
2014	7,34	0,87	1,27
2015	8,07	0,95	3,61
2016	7,94	0,82	1,84
2017	6,72	0,87	3,49
2018	5,75	0,81	3,61
2019	4,37	0,86	1,84
2020	3,37	0,97	3,49

Tabla 4.9: *Relevance index, novelty index, and reputation score of Terrance E. Boult.*

publicando en revistas de menor calidad y no obtiene tantas citas como Tax. Además, Boult comparte artículos con varios autores y Tax publica en solitario o solo con otro autor. En 2014, 2015 y 2016 se observa un aumento del índice de relevancia de Bault. En 2012 y 2016 se publicaron sus dos artículos con mayor número de citas. Nótese que el

Year	Relevance index	Novelty index	Reputation score
1997	0,09	1,00	0,27
1998	0,05	1,00	0,24
1999	0,04	0,00	0,03
2000	0,03	0,00	0,02
2001	0,03	1,00	0,22
2002	0,02	0,00	0,02
2003	0,02	0,00	0,01
2004	0,01	0,00	0,01
2005	0,43	1,00	0,54
2006	1,05	0,00	0,84
2007	0,75	1,00	0,80
2008	0,82	0,59	0,78
2009	0,82	0,59	0,77
2010	0,55	0,00	0,44
2011	0,87	0,66	0,83
2012	1,34	0,90	1,25
2013	1,14	0,87	1,09
2014	0,95	0,83	0,93
2015	1,35	0,86	1,22
2016	0,89	0,84	0,88
2017	0,03	0,95	0,93
2018	1,01	1,00	1,01
2019	0,70	0,94	0,75
2020	0,78	0,87	0,80
2021	0,58	0,92	0,65
2022	0,48	0,89	0,56

Tabla 4.10: *Relevance index, novelty index and reputation score of Isaac M. de Diego.*

índice de relevancia se calcula en ciclos de tres años, por lo que el año 2014 se benefició de los artículos de 2012. Además, en 2013, 2014 y 2015 la mayoría de los artículos están escritos por él mismo con otro autor. La novedad se mantiene bastante constante a lo largo de los años. Ninguno de los artículos de Bault alcanza las citas generadas por los tres artículos de Tax con más citas.

Por último, Isaac tiene una relevancia mucho menor que los dos autores anteriores. En sus primeros años, alcanza un índice de relevancia extremadamente bajo, con una media de $0,04$, aunque 1997 es el año en que publica el artículo con mayor número de citas. Durante los primeros años, sus publicaciones son compartidas con muchos autores en revistas de cuartil bajo. A partir de 2005 su índice de relevancia sube, aunque está lejos de los índices de relevancia de los otros dos autores en los mismos años, provocado por la publicación con menos autores y la consecución de más citas. En 2012 y 2013 se obtienen el segundo y

tercer índice de relevancia más altos de su carrera gracias a artículos publicados en revistas de cuartil superior y con menos autores. Además, su segundo artículo con más citas fue publicado en 2010. Nótese que el cálculo de la relevancia se realiza en ciclos de tres años. En cuanto a la novedad, Isaac obtuvo altibajos durante los 10 primeros años. Hay años en los que se trata el mismo tema y años en los que el tema cambia radicalmente. Esto explica por qué Isaac pasa de $1,00$ a $0,00$ durante los diez primeros años. La novedad se hace más constante y elevada en los últimos diez años.

La interpretación de la trayectoria de la reputación de estos tres autores muestra que la métrica propuesta proporciona resultados interesantes, teniendo en cuenta factores ignorados por otras métricas. Por lo tanto, la reputación de un autor no debería medirse únicamente en función de la relevancia del trabajo producido. Además, la relevancia debería recoger detalles como el número de autores que participan en el artículo publicado y debería penalizarse si el trabajo del autor empieza a perder calidad con el paso de los años. Métricas como el índice h o el i10 no disminuyen aunque el investigador empiece a publicar artículos de muy baja calidad. Eso explica por qué Bault tiene un índice h más alto a pesar de no haber conseguido escribir nunca un artículo que iguale el número de citas de los tres artículos con más citas en Tax, por ejemplo.

4.2.2. Comparación entre la métrica propuesta y las alternativas

El propósito del segundo experimento es validar el desempeño de las métricas propuestas. Para ello se han realizado tres validaciones diferentes para comparar las métricas propuestas con el índice h e i10. La primera calcula el índice de relevancia, el índice de novedad y la reputación final para 45 autores diferentes. La segunda compara el índice h e i10 con la reputación final de un autor con una carrera científica muy corta pero de gran impacto. La tercera introduce un escenario hipotético para detectar un posible fraude.

Correlación entre la reputación propuesta y los índices h e i10

Esta primera validación calcula la correlación entre la reputación propuesta y los índices h e i10 para 45 autores diferentes. Estos autores se seleccionan en función de su reputación pública debido al índice h: 25 autores de reputación media, 18 autores de alta reputación y 2 autores de máxima reputación. Los autores de cada grupo tienen un número similar de artículos, citas y seniority, para evitar que el índice h y el i10 introduzcan errores por las limitaciones que presentan. El objetivo de este experimento es comparar la reputación propuesta con el índice h e i10 de cada autor. Para ello, se ha calculado la reputación final de todos los autores. Las tablas 4.11 y 4.12 muestran la información relevante para cada autor. La correlación de Pearson entre la reputación propuesta y el índice h es de $0,75$, entre la reputación propuesta y el índice i10 es de $0,54$, y entre el índice h y el índice i10 es de $0,93$. La alta correlación entre el índice h y el índice i10 se debe a que ambas métricas se calculan de manera similar. La reputación

Author name	Reputation score	H-index	i10
Andrew N. Meltzoff	4,08	111	255
D.E. Walling	2,31	110	470
Xavier Marie	0,74	57	195
Prat N.	1,04	55	186
Terrance E. Boult	1,90	55	156
Sefaattin Tongay	1,46	54	125
Douglas Irwin	1,81	54	107
Sebastien Marcel	1,85	52	132
Timothy J. Hatton	1,60	52	119
Jan Luiten Van Zanden	0,87	51	151
Salvador García	2,03	51	106
Cindy Morris	1,54	49	96
Joao Carlos Setubal	1,91	48	91
Tom Heskes	2,11	47	131
David Tax	2,12	46	114
Francesc Gallart	1,03	46	105
Víctor Izquierdo-Roca	1,11	46	102
Jean Greenberg	2,81	46	64
David Studholme	2,29	45	106
Giorgio Giacinto	1,94	45	91
Polemio M	0,55	23	60
Marco Petitta	0,56	23	50
Zanoni Dias	0,83	23	41

Tabla 4.11: *Reputation score, índice h e i*10 *para* 45 *autores. (Parte 1)*

propuesta considera la calidad y la novedad de los trabajos, por lo que es normal que la correlación entre esta reputación y las otras métricas sea menor. Como la correlación con el índice h es mayor que con el índice i10, se considera la correlación de Pearson entre el índice h y los índices de relevancia y novedad. De esta forma se puede comprobar el peso de la calidad y la novedad de las obras en el índice h. El primero es 0,72 y el segundo es 0,49. Observe que la correlación entre el índice de relevancia y el índice h no es tan alta como entre el índice h y el índice i10. Las variables incluidas en la reputación propuesta que consideran la calidad de los artículos provocan este descenso. Además, la correlación entre la novedad y el índice h es aún menor. Estos resultados se deben a que el índice h no considera la novedad de los trabajos de los autores. Por lo tanto, la reputación propuesta en este artículo parece más completa que el índice h y el índice i10.

Author name	Reputation score	H-index	i10
Katja Fiehler	1,28	23	40
Rainer Fremdling	0,78	23	35
Tine de Moor	0,98	23	35
Antonio Alonso Ayuso	0,93	23	32
David S. Jacks	1,14	23	27
Matthieu Boisgontier	0,86	21	34
Teodoro Estrela	0,69	21	42
Nerantzis Kazakis	1,12	21	29
Laura Ding	0,54	21	28
Olga Petrucci	0,63	20	42
Bas Van Leeuwen	0,40	20	39
Giovanni Pardini	0,69	20	33
Bert de Vries	1,20	20	31
Eduardo M. García Roger	1,04	20	26
Ángela Aurora Pasqua	0,57	19	30
Thorsten Rissom	0,36	19	29
Tjeerd Dijkstra	0,56	19	29
Syantan Ghosal	2,11	19	27
Ioannis Kougias	1,05	19	23
Kelvin S. Oie	1,00	18	18
Isaac M. de Diego	0,56	18	26
Bishnupriya Gupta	0,76	17	25

Tabla 4.12: *Reputation score, índice h e i*10 *para* 45 *autores. (Parte 2)*

Limitaciones de los índices h e i10 cuando el autor tiene una carrera corta pero de gran impacto

La segunda validación comprueba las limitaciones de los índices h e i10 cuando el autor tiene una carrera corta pero de gran impacto. La falta de información sobre la calidad y novedad de los trabajos al calcular el índice h y el índice i10, provoca que estos índices sean inútiles para perfilar la reputación de un autor de corta trayectoria científica pero de gran impacto. Hay bastantes autores con carreras poco prolíficas que han tenido una gran influencia en la evolución de importantes asuntos científicos. Uno de ellos es Evariste Galois (Galois y Neumann, 2011), cuyo caso es muy conocido en la comunidad científica y siempre se utiliza como ejemplo de una de las limitaciones que presentan métricas como el h-index o i10. La reputación propuesta para Galois no se puede calcular porque no se conocen los valores de las variables necesarias para calcular la métrica. En este experimento se calcula la reputación final de un autor hipotético que podría ser el propio Galois. Por tanto, se plantea que un investigador científico publicó solo dos artículos en toda su vida y lo hizo en el mismo año. Ambos artículos fueron publicados en revistas

de Q1 como autor único y recibieron 15.550 citas y 33.450 citas respectivamente. La reputación propuesta de este autor es de 490. En cambio, este autor tiene un índice h e i10 de 2. Es obvio que la reputación de este autor es alta debido a la influencia de sus dos artículos. De este modo, se puede concluir que es importante considerar la calidad y la novedad de los trabajos al calcular una métrica de reputación.

Detección de fraude

El último problema abordado en el segundo experimento es la detección de fraude. Se presenta un escenario hipotético para probar que métricas como h-index e i10 no detectan un posible fraude, mientras que la medida de reputación propuesta sí lo detecta. El escenario fraudulento hipotético es el siguiente: un grupo de 10 autores logra un índice h de 30 en dos años utilizando técnicas fraudulentas. Para ello, publican 30 artículos entre los 10 autores. Cada autor escribe solo 3 artículos en estos dos años y todos los autores están incluidos en todos los artículos. Para facilitar el trabajo de publicación, todos los artículos se envían a las publicaciones de la conferencia CORE C indexadas por Google Scholar. Y para lograr las 30 citas por artículo necesarias para alcanzar el índice h igual a 30, todos los artículos se citan en artículos posteriores, así como en artículos de científicos amigos. De esta forma, estos 10 autores logran un índice h de 30 en 2 años, a pesar de haber publicado en publicaciones de baja reputación, compartir estas publicaciones con otros 9 autores y obtener citas no legítimas. El índice h no considera factores importantes como el número de autores por artículo o la reputación de la revista donde se han publicado los artículos, entre otros. Por lo tanto, no penaliza el posible fraude en estos 10 autores. En cambio, la reputación propuesta tiene en cuenta estos factores y otros, lo que penaliza la reputación general de estos autores.

4.2.3. Clustering de autores basado en la reputación propuesta y en el h-index

El tercer experimento estudia el clustering del conjunto de 45 autores presentado previamente en la Sección 4.3.2. Este agrupamiento se basa en sus trayectorias de reputación y los resultados se comparan con el basado en el índice h. Nótese que el clustering no se basa en las puntuaciones de reputación, sino en las trayectorias de reputación. De esta manera los resultados se enriquecen con la información de toda la trayectoria científica. Para ello se realiza un grid-based análisis y el conocido algoritmo DTW para calcular distancias a partir de trayectorias. Estas distancias se utilizan para alimentar un algoritmo de hierarchical clustering. El dendrograma resultante con información de clusters se muestra en la Figura 4.6.

Se puede observar que se han generado 3 clusters diferentes. Las tablas 4.13 y 4.14 muestran la comparación de los clusters generados con la métrica propuesta y los clusters

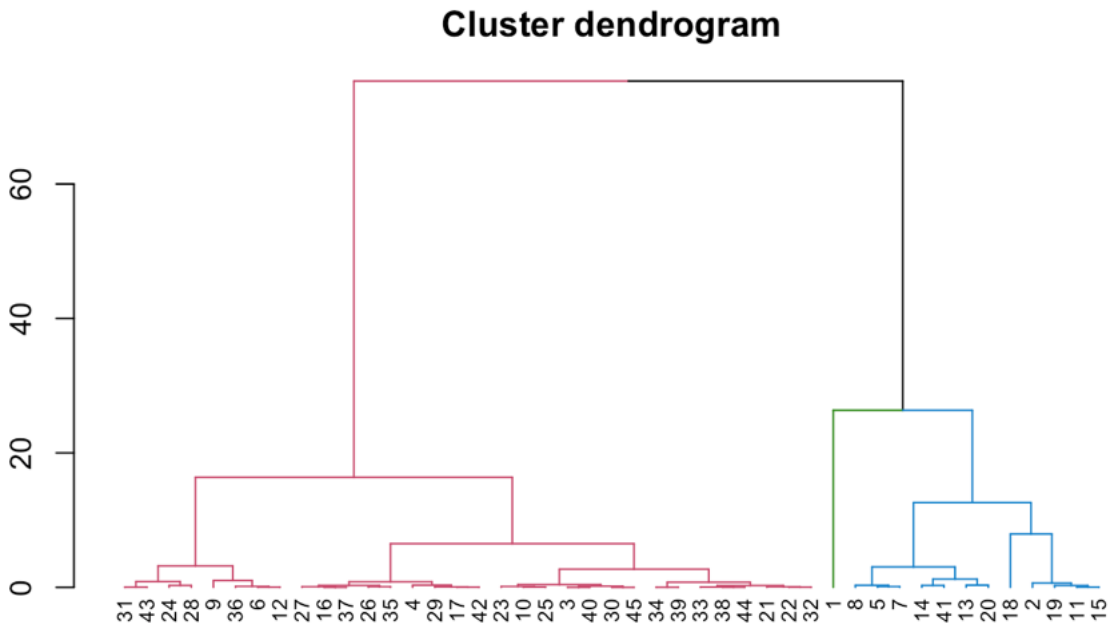

Figura 4.6: *Clustering jerárquico basado en la reputación de 45 autores.*

basados en el índice h. Al perfilar estos clusters, se puede indicar que el primer autor de la tabla es un investigador top, los 12 autores etiquetados como B tienen una reputación alta y, finalmente, el resto de los autores tienen una reputación media. El clustering basado en las trayectorias de reputación es bastante similar al clustering basado en los índices h con algunas diferencias. Respecto a estas diferencias, solo un autor etiquetado como C en los clusters del índice h ha sido etiquetado como B en los clusters de las trayectorias de reputación. Este autor es Syantan Ghosal, y la razón es que la novedad de sus obras es alta. Con respecto a los autores etiquetados como B en los clusters de índice h, 8 de 18 han sido etiquetados como C en los clusters de las trayectorias de reputación. La mayoría de ellos han sido degradados debido a la calidad y novedad de sus obras. Sin embargo, Xavier Marie y Prat N. han sido degradados porque Semantic Scholar no ha indexado sus artículos más influyentes. En cuanto a los autores etiquetados como A en los clusters del índice h, D.E. Walling se ha etiquetado como B en los clusters de las trayectorias de reputación. Esto se debe a dos razones principales, la novedad de sus obras no es alta y la calidad de sus últimos trabajos ha disminuido abruptamente.

En conclusión, este clustering ha demostrado que proporciona resultados aceptables para clasificar a los autores según su reputación. Las métricas propuestas actúan como filtro descartando casos extremos y fraude. Por lo tanto, se puede decir que este framework es un prototipo funcional para analizar y agrupar a los autores por sus trayectorias científicas.

Author name	Trajectories cluster	H-index cluster
Andrew N. Meltzoff	A	A
D.E. Walling	B	A
Xavier Marie	C	B
Prat N.	C	B
Terrance E. Boult	B	B
Douglas Irwin	C	B
Sefaattin Tongay	B	B
Sebastien Marcel	B	B
Timothy J. Hatton	C	B
Salvador García	C	B
Jan Luiten Van Zanden	B	B
Cindy Morris	C	B
Joao Carlos Setubal	B	B
Tom Heskes	B	B
David Tax	B	B
Francesc Gallart	C	B
Víctor Izquierdo-Roca	C	B
Jean Greenberg	B	B
Giorgio Giacinto	B	B
David Studholme	B	B
Polemio M	C	C
Marco Petitta	C	C
David S. Jacks	C	C
Antonio Alonso Ayuso	C	C
Zanoni Dias	C	C

Tabla 4.13: *Clusters de las trayectorias y del índice h. (Parte 1)*

4.3. Domains Classifier based on Risky Websites (DOCRIW)

Esta sección aborda un conjunto de experimentos que explican el diseño del framework DOCRIW y evalúan el rendimiento general del sistema.

El primer experimento, presentado en la Sección 4.3.1, muestra una batería de pruebas realizada para justificar la selección de los elementos incluidos en el *Machine Learning Model* (ver Fig. 3.16) . Estos elementos son: la similaridad entre dominios, el algoritmo ML, el umbral de probabilidad proporcionado por el algoritmo y un conjunto de dominios de referencia. En este caso, se utilizan 1500 dominios previamente etiquetados para entrenar y probar el modelo (750 *non-risky* y 750 *risky*).

El propósito del segundo experimento es validar el rendimiento del *Machine Learning Model*. La sección 4.3.2 describe un experimento que aborda dos problemas diferentes. En

Author name	Trajectories cluster	H-index cluster
Katja Fiehler	C	C
Rainer Fremdling	C	C
Tine de Moor	C	C
Teodoro Estrela	C	C
Nerantzis Kazakis	C	C
Laura Ding	C	C
Matthieu Boisgontier	C	C
Eduardo M. García Roger	C	C
Giovanni Pardini	C	C
Olga Petrucci	C	C
Bert de Vries	C	C
Bas Van Leeuwen	C	C
Ángela Aurora Pasqua	C	C
Ioannis Kougias	C	C
Thorsten Rissom	C	C
Tjeerd Dijkstra	C	C
Syantan Ghosal	B	C
Kelvin S. Oie	C	C
Bishnupriya Gupta	C	C
Isaac M. de Diego	C	C

Tabla 4.14: *Clusters de las trayectorias y del índice h. (Parte 2)*

el primero se evalúa el rendimiento para la clasificación de dominios *risky*. Para ello se han utilizado 200 dominios extraídos del índice *Risky Domains* del módulo *Knowledge Base*. En el segundo, se evalúa el rendimiento para clasificar dominios *non-risky*. En este caso, se han probado 200 dominios de prestigio. Este segundo experimento no utiliza el *Domains Extraction and Validation Module*, por lo que solo se evalúa el clasificador.

El tercer experimento, descrito en la Sección 4.3.3, simula la funcionalidad de etiquetado completa del sistema. Utiliza 100 dominios *risky*, 100 dominios *non-risky* y 20 dominios inactivos para proporcionar las etiquetas predichas correspondientes (*risky* y *non-risky*) y sus probabilidades.

4.3.1. Entrenamiento y evaluación del modelo de aprendizaje máquina

Este experimento se utiliza para seleccionar y evaluar los elementos adecuados del *Machine Learning Model*. Se ha realizado con 1,500 dominios ya etiquetados como *risky* o *non-risky* por expertos del dominio. El conjunto de datos se ha dividido en train (70 %) y test (30 %). El conjunto de train se ha utilizado para entrenar un conjunto de algoritmos de ML que predicen las etiquetas de los dominios de entrada. Los dominios del conjunto de test se han utilizado como entrada del modelo para evaluar el rendimiento. Así, se

han seleccionado las medidas de similaridad, los pesos para combinar estas medidas, el algoritmo ML y el umbral de probabilidad de corte que maximizan el rendimiento global. Los conjuntos de datos de train y test se dividieron aleatoriamente 10 veces, y se realizó una ejecución del experimento en cada secuencia. Por lo tanto, se presentan la media y la desviación estándar de las medidas de rendimiento.

La distancia *Levenshtein* se usa como medida de similaridad para el *domain name*. Las otras similaridades se calcularon evaluando si dos dominios tienen el mismo valor para la variable correspondiente (similaridad = 1) o no (similaridad = 0). Así, se obtuvieron seis medidas de similaridad diferentes. La similaridad global se calculó como una suma ponderada de las seis similaridades individuales. Los mejores pesos seleccionados para calcular la similaridad global fueron 0,5, 0,15, 0,25, 0,05, 0,05 y 0, correspondientes a nombre de dominio, ciudad, país, fecha de creación, fecha de vencimiento y correo electrónico, respectivamente. Por tanto, en este caso la similaridad de correo electrónico no se incluyó en el cálculo de similaridad global. Estos valores fueron seleccionados durante la fase de train y evaluados en la fase de test.

Se han evaluado varios algoritmos de ML diseñados para ofrecer una buena respuesta como clasificadores binarios. Estos son los más típicos en la literatura del dominio. Así, un algoritmo basado en LR (Menard, 2002), dos algoritmos de bagging que usan árboles de decisión (*Random Forest* (RF) (Friedman *et al.*, 2001) y *Extremely Randomized Trees* (ERT) (Friedman *et al.*, 2001), dos algoritmos de boosting (*Adaboost* (AB), (Freund y Schapire, 1999) y *Gradient Boosting* (GB)) (Natekin y Knoll, 2013) y un algoritmo basado en vectores de soporte (*Support Vector Machine* (SVM)) (Moguerza y Muñoz, 2006). Adicionalmente, se han incluido otros algoritmos para medir el rendimiento de los anteriores: *k-Nearest Neighbour* (kNN) (James *et al.*, 2013), *Naïve Bayes* (NB) (Friedman *et al.*, 2001) y *Linear Discriminant Analysis* (LDA) (Friedman *et al.*, 2001). Una breve descripción de estos métodos:

- AdaBoost es un método de ensemble que entrena y despliega árboles en serie. AdaBoost implementa boosting, donde un conjunto de clasificadores débiles se conecta en serie de modo que cada clasificador débil intente mejorar la clasificación de las muestras que fueron clasificadas incorrectamente por el clasificador débil anterior. Al hacerlo, el impulso combina clasificadores débiles en serie para crear un clasificador fuerte. (Freund y Schapire, 1999)

- Extremely Random Tree es lo mismo que Random Forest con la excepción de que los umbrales de decisión utilizados para dividir los nodos también se eligen aleatoriamente en lugar de seleccionar los más discriminatorios. (Friedman *et al.*, 2001)

- Gradient boosting produce un modelo de predicción mediante un conjunto de modelos de predicción débiles, normalmente árboles de decisión. Construye el modelo por etapas como lo hacen otros métodos boosting y los generaliza al permitir la

optimización de una función de pérdida diferenciable arbitraria. (Natekin y Knoll, 2013)

- K-Nearest Neighbors es un método no paramétrico que almacena todos los casos disponibles y clasifica los nuevos casos en función de una medida de similaridad (por ejemplo, funciones de distancia). (James *et al.*, 2013)

- Linear Discriminant Analysis es un método de aprendizaje que permite encontrar una combinación lineal de características que separan dos clases. Se utiliza como clasificador lineal o para tareas de reducción de dimensiones antes de la clasificación. (Friedman *et al.*, 2001)

- Logistic Regression es un método estadístico que utiliza una función logística para modelar una variable dependiente binaria, Y, a partir de una o más variables de respuesta, X. (Menard, 2002)

- Naïve Bayes es un algoritmo de aprendizaje simple que utiliza la regla de Bayes junto con una fuerte suposición de que los atributos son condicionalmente independientes, dada la clase. Si bien esta suposición de independencia a menudo se viola en la práctica, Naïve Bayes ofrece una precisión de clasificación competitiva. (Friedman *et al.*, 2001)

- Random Forest es un conjunto de clasificadores de árboles de decisión aleatorios, que hace predicciones combinando las predicciones de los árboles individuales. (Friedman *et al.*, 2001)

- Support Vector Machines son clasificadores lineales particulares que se basan en el principio de maximización del margen. Realizan la minimización del riesgo estructural, lo que mejora la complejidad del clasificador con el objetivo de lograr un excelente rendimiento de generalización. El SVM realiza la tarea de clasificación al construir, en un espacio dimensional superior, el hiperplano que separa de manera óptima los datos en dos categorías. (Moguerza y Muñoz, 2006)

Nótese que las complejidades de almacenamiento y computacionales de estos algoritmos son diferentes. Sin embargo, el objetivo de este experimento es seleccionar un modelo de ML único. Por lo tanto, este problema no afecta el rendimiento relativo del framework DOCRIW.

Se ha aplicado el método Grid Search (Lameski *et al.*, 2015) para seleccionar los valores óptimos de los parámetros para los modelos ML. Este método prueba cada algoritmo con diferentes valores de sus parámetros y compara los resultados obtenidos. Además, se han utilizado otras técnicas para encontrar los valores óptimos, como el diagrama Out-Of-Bag (OOB) Error Rate para Random Forest (Friedman *et al.*, 2001). Los parámetros que

optimizan los algoritmos de ML son los siguientes. LR hace uso de Ridge Regression (L2) (Gruber, 2017) como función de regularización y un parámetro de penalización igual a 10. RF incluye 400 estimadores (árboles), log2 (logaritmo en base 2) como la función que calcula el número de variables por árbol, y el coeficiente de Gini (X.-X. Zhang *et al.*, 2018) como criterio de selección. ERT utiliza los mismos parámetros que RF excepto por el criterio de selección, que se ha configurado en Entropía en lugar del coeficiente de Gini. AB incluye 300 estimadores y una tasa de aprendizaje igual a $0,1$. GB usa los mismos parámetros que AB, agregando una longitud de profundidad igual a 3. SVM usa un kernel lineal y un parámetro de penalización igual a 1. Finalmente, kNN se ha configurado para 5 vecinos, una importancia ponderada para los vecinos más cercanos y la distancia euclidia. NB y LDA no tienen parámetros.

Las métricas de rendimiento consideradas para probar el *Machine Learning Model* son (Zhu *et al.*, 2010): *accuracy*, *sensitivity* y *specificity*. El *accuracy* es la proporción de dominios (*risky* y *non-risky*) que están correctamente identificados por el método ML:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}, \tag{4.1}$$

donde:

- *True Negative* (TN) = websites *non-risky* correctamente clasificados como *non-risky*.
- *True Positive* (TP) = websites *risky* correctamente clasificados como *risky*.
- *False Negative* (FN) = websites *risky* erróneamente etiquetadas como *non-risky*.
- *False Positive* (FP) = websites *non-risky* erróneamente etiquetadas como *risky*.

La *sensitivity* es la proporción de dominios *risky* que se identifican correctamente como tales:

$$Sensitivity = \frac{TP}{TP + FN}. \tag{4.2}$$

Finalmente, la *specificity* es la proporción de dominios *non-risky* que están correctamente identificados como tales:

$$Specificity = \frac{TN}{TN + FP}. \tag{4.3}$$

La tabla 4.15 muestra las métricas de rendimiento de todos los algoritmos de ML evaluados. El algoritmo LR fue el mejor para cada medida de rendimiento. El umbral de probabilidad de corte que maximiza el rendimiento fue de 0,6. Es decir, los dominios con una probabilidad pronosticada de ser *risky* inferior a $0,6$ se clasifican como *non-risky*. Por el contrario, los dominios con una probabilidad pronosticada de ser *risky* mayor o igual a $0,6$ se clasifican como *risky*.

Por tanto, el *Machine Learning Model* ha sido construido y probado. Sus elementos son: el algoritmo LR que utiliza una combinación ponderada predefinida de medidas de

ML Model	Accuracy	Sensitivity	Specificity
AB	0,80 (0,02)	0,77 (0,07)	0,83 (0,05)
ERT	0,81 (0,01)	0,75 (0,03)	0,87 (0,04)
GB	0,84 (0,01)	0,79 (0,01)	0,89 (0,03)
KNN	0,82 (0,01)	0,79 (0,04)	0,85 (0,03)
LDA	0,72 (0,03)	0,72 (0,05)	0,72 (0,02)
LR	0,89 (0,01)	0,85 (0,01)	0,92 (0,02)
NB	0,68 (0,02)	0,67 (0,06)	0,69 (0,04)
RF	0,82 (0,01)	0,80 (0,01)	0,85 (0,02)
SVM	0,86 (0,01)	0,85 (0,01)	0,87 (0,01)

Tabla 4.15: *Resultados (Media y Deviación Estándar) para los algoritmos de ML del Experimento 1.*

similaridad individuales y el umbral de probabilidad de corte y el conjunto de dominios utilizados para aprender el modelo.

4.3.2. Validación del modelo de aprendizaje máquina

El propósito del segundo experimento es validar el rendimiento del *Machine Learning Model* construido en el primer experimento. Para ello se han llevado a cabo dos validaciones diferentes. La primera evalúa la clasificación de 200 dominios predefinidos como dominios *risky*. La segunda evalúa el rendimiento de clasificar 200 dominios predefinidos como dominios *non-risky*.

Dominios risky

Esta primera validación se ha realizado con 200 dominios recogidos del índice *Risky Domains* del módulo *Knowledge Base*. Por lo tanto, todos los dominios son *risky*. El objetivo principal de esta validación es verificar si el sistema etiqueta todas las entradas como dominios *risky*, como debería ser. Para ello se han utilizado dos módulos del framework DOCRIW: el módulo *Host-based Variables Extraction* y el módulo *Classification*. De esta forma, se valida el *Machine Learning Model*. La tabla 4.16 muestra la etiqueta de predicción y la probabilidad asignada por el sistema para un extracto de estos dominios.

El sistema ha logrado un *accuracy* de 0,86 para la clasificación de estos 200 dominios *risky*. Por lo tanto, se puede concluir que el clasificador tiene un buen rendimiento. Además, el *accuracy* al clasificar dominios *risky* debería ser similar a la *sensitivity* alcanzada en el primer experimento, que es de $0,85$. Nótese que si el módulo *Domains Extraction and Validation* se hubiera incluido en este experimento, alcanzaría un valor de *accuracy* de 1. Esto se debe a que todos estos dominios se habrían encontrado en el índice *Risky Domains* y se etiquetarían automáticamente como *risky*.

Domain name	Predict. Label	Prob. (risky)
1080p-torrents.kickass-torrent.biz	risky	0,99
5movies.to	risky	0,97
ddlvalley.me	risky	0,97
filetram.com	risky	0,84
filmstreaminghd.biz	risky	0,90
foumovies.com	risky	0,90
heroturko.net	risky	0,94
limetorrents.co	risky	0,99
madefittoday.com	non-risky	0,49
putlockers.ws	non-risky	0,46
sipeliculas.com	risky	0,99
sockshare.io	risky	0,98
torrentdownloads.unblocked.live	risky	0,99
uwatchfree.tv	risky	0,98
zooqle.com	risky	0,73

Tabla 4.16: *Extracto de la clasificación de dominios* risky *del índice* Risky Domains.

Además, se detectaron algunos problemas con dominios inactivos para generar las variables basadas en host ya que las fuentes de información web no brindan información sobre ellos. Este problema se soluciona usando solo dominios activos para realizar este experimento. En todo el framework DOCRIW, está controlado por el módulo *Domains Extraction and Validation*. Comprueba si los dominios están activos o no. En el segundo caso, los etiqueta como inactivos (es decir, no se proporciona ninguna etiqueta *risky* o *non-risky* para los dominios inactivos).

Dominios non-risky

Esta segunda validación se ha realizado con 200 dominios de prestigio, todos ellos previamente etiquetados por expertos como *non-risky*. Esta vez, el objetivo es verificar si el clasificador etiqueta todas estas entradas como dominios legales. La tabla 4.17 muestra la etiqueta de predicción y la probabilidad asignada por el sistema para un extracto de estos dominios.

El sistema ha logrado un *accuracy* de 0,88 para la clasificación de estos 200 dominios *non-risky*. Aunque el *accuracy* es menor que el *specificity* del primer experimento $(0,9)$, los resultados son suficientemente buenos. En total, 24 dominios han sido clasificados como *risky*. 13 de estos 24 dominios han logrado una probabilidad entre 0,6 y 0,61, por lo que el clasificador tampoco está seguro de que estos dominios sean realmente *risky*.

Nótese que cuando un nombre de dominio contiene subcadenas utilizadas en los nombres de dominio *risky*, podrían clasificarse como *risky*. Especialmente cuando las variables

Domain name	Predict. Label	Prob. (risky)
adidas.es	non-risky	0,02
amazon.com	non-risky	0,04
atleticodemadrid.com	non-risky	0,23
audi.es	non-risky	0,02
bancosantander.es	non-risky	0,01
carrefour.es	non-risky	0,17
disney.es	non-risky	0,22
elmundo.es	non-risky	0,04
facebook.com	non-risky	0,55
google.es	non-risky	0,07
hboespana.com	non-risky	0,40
linkedin.com	risky	0,63
telecinco.es	non-risky	0,01
uber.com	risky	0,62
urjc.es	non-risky	0,01

Tabla 4.17: *Clasificación de los dominios non-risky.*

basadas en host no proporcionan información adicional. Ejemplos de este problema son *linkedin.com* y *uber.com*, que aparecen en la tabla 4.17. *link* y *ube* aparecen en varios nombres de dominio *risky* usados para entrenar el algoritmo LR. Además, ambos están alojados en California, EE. UU. (variables de ciudad y país). Estos valores basados en host tampoco ayudan a caracterizar los dominios como *non-risky* (solo el 35 % de los dominios *non-risky* utilizados para entrenar el algoritmo LR están alojados en EE. UU.). Sin embargo, tampoco está del todo claro que estos dominios sean *risky*. Por lo tanto, sus probabilidades de ser *risky* han sido 0,63 y 0,62, respectivamente. Obsérvese que la probabilidad de corte es 0,6.

Respecto al tema de los dominios inactivos, no es común encontrar dominios *non-risky* inactivos. Los dominios *non-risky* suelen tener una larga vida. En cambio, los cierres de dominios *risky* son más frecuentes, debido a las actividades ilegales realizadas por ellos. Como se mencionó anteriormente, el framework DOCRIW controla esta situación.

4.3.3. Simulación del sistema en producción

Este experimento simula la funcionalidad completa de clasificación de dominios. Esta funcionalidad abarca el *Domain Labeling Process*. Así, el objetivo es ejecutar el proceso completo para simular el funcionamiento del framework DOCRIW en la etapa de producción. Por lo tanto, los módulos involucrados en el experimento son: el módulo *Domains Extraction and Validation*, el módulo *Host-based Variables Extraction* y el módulo *Classification*.

Profundizando en el experimento, se evaluaron un total de 220 dominios que no habían sido considerados previamente por el framework (es decir, 100 dominios *risky*, 100 dominios *non-risky* y 20 dominios inactivos). Se ha logrado un *accuracy* de 0,86.

Respecto a los dominios *risky*, 81 de los 100 dominios se han clasificado correctamente. Cinco de ellos han sido etiquetados directamente por el módulo *Domains Extraction and Validation* (por ejemplo, *ugtorrent.com*), ya que fueron almacenados en el índice *Risky Domains*. Por lo tanto, el resto de los módulos no se consideran y estos dominios se clasificaron correctamente con un valor de probabilidad de 1. En relación a los dominios *non-risky*, 89 de 100 han sido debidamente clasificados. Esto prueba que el sistema está diseñado para minimizar el error al clasificar dominios *non-risky*. Finalmente, todos los dominios inactivos han sido bien clasificados. También han sido etiquetados por el módulo *Domains Extraction and Validation*.

La tabla 4.18 muestra los resultados de la clasificación de un extracto de estos dominios. Presenta el nombre de dominio, la etiqueta real, la etiqueta predicha y la probabilidad de ser clasificado como dominio *risky*.

En conclusión, el framework DOCRIW ha demostrado que proporciona resultados aceptables para clasificar dominios según su riesgo en base a la reputación del dominio web. El módulo de *Domains Extraction and Validation* actúa como un filtro descartando los dominios inactivos y aquellos que ya han sido almacenados en el índice de *Risky Domains*. Esto permite minimizar las características que no aportan valor y que tienen que ser evaluadas por los otros dos módulos que forman parte del *Domain Labeling Process*. Así, se puede decir que DOCRIW es un prototipo funcional para detectar y clasificar dominios potencialmente riesgosos en base a su reputación.

Domain name	Actual Label	Predict. Label	Prob. (risky)
3hdmovies.com	inactive	inactive	1,00
acmefilm.ee	risky	non-risky	0,43
ariamovie7.site	risky	risky	0,92
bigcinema.tv	inactive	inactive	1,00
canon.es	non-risky	non-risky	0,05
cisco.com	non-risky	non-risky	0,34
edreams.es	non-risky	non-risky	0,08
fnac.es	non-risky	non-risky	0,01
freemovieswatchonline.co	inactive	inactive	1,00
harley-davidson.com	non-risky	non-risky	0,31
hdfilme.tv	risky	risky	0,78
hornyblog.eu	inactive	inactive	1,00
iberia.es	non-risky	non-risky	0,08
marca.com	non-risky	risky	0,61
mobilemoviescorner.com	risky	non-risky	0,58
nvidia.es	non-risky	non-risky	0,04
seedpeer.me	risky	risky	0,73
serviwin.com	inactive	inactive	1,00
sony.es	non-risky	non-risky	0,24
templestowepub.com	risky	risky	0,76
ugtorrent.com	risky*	risky	1,00
usabit.com	risky	risky	0,87
vodafone.es	non-risky	non-risky	0,06
watchonline.red	risky	risky	0,83
xdownload.pl	risky	risky	0,97

Tabla 4.18: *Extracto de la clasificación usando los módulos implicados en el* Domain Labeling Process*. La etiqueta* risky *con asterisco significa que viene del índice* Risky Domains*.*

Capítulo 5

Conclusiones

En esta tesis hemos destacado la importancia del cálculo de la reputación por diversas razones. Por un lado, el cálculo de la reputación ayuda a evaluar la calidad de productos o servicios, así como el trabajo de autores, lo cual es fundamental para tomar decisiones informadas. Por otro lado, el cálculo de la reputación es una herramienta valiosa para tomar decisiones en entornos inciertos, lo que resulta especialmente relevante en los mercados financieros. Además, la reputación también puede aumentar la confianza en las relaciones comerciales y ser un factor clave en la competencia empresarial.

Asimismo, en esta tesis hemos destacado la importancia del contexto en el cálculo de la reputación y la necesidad de elegir el sistema de cálculo adecuado en función del ámbito de aplicación. Para ello, se ha realizado una revisión exhaustiva de la literatura y se han propuesto nuevos métodos para dos ámbitos de aplicación concretos: el científico e Internet. De esta forma, se contribuye al avance del conocimiento en estas áreas y se abre la puerta a nuevas investigaciones en el futuro.

5.1. Principales contribuciones

Las principales contribuciones de esta investigación se han logrado mediante el desarrollo de tres frameworks que proponen nuevas métricas para superar las limitaciones de las métricas revisadas en el estado del arte. Estos frameworks representan una mejora significativa en el campo de las métricas de evaluación de sistemas, y se espera que tengan un impacto positivo en la construcción de sistemas basados en el conocimiento. Con la propuesta de estas nuevas métricas, se ofrece una perspectiva innovadora y útil para mejorar la evaluación de sistemas, lo que se traduce en beneficios tanto para la industria como para la investigación. Además, se ha demostrado la utilidad y efectividad de los frameworks propuestos a través de experimentos y análisis comparativos, lo que refuerza aún más su validez y relevancia en el campo.

El primero de los frameworks es UNIKO, que proporciona un conjunto de funcionalidades orientadas a satisfacer las necesidades de los investigadores relacionadas con el estudio de un dominio específico, proporcionando información sobre artículos y autores. UNIKO es un *Knowledge-Based Systems* (KBS) que almacena información e interactúa con los usuarios actuando también como un *Content-Based Recommendation System*. Por lo tanto, recopila información para artículos y autores de fuentes web utilizando rastreadores y técnicas de scrapping. Esta información se mejora mediante la obtención de puntuaciones de reputación para el propio artículo y sus autores. Luego, se realiza el análisis de sentimiento sobre el texto. Este análisis utiliza el léxico SentiWordNet (Baccianella *et al.*, 2010) como fuente principal de información. Se completa con una red neuronal convolucional (Bhavsar *et al.*, 2017). Ésta predice los valores de sentimiento para las palabras con valores neutrales en el léxico y para las palabras que no se pueden emparejar. Finalmente, se han incluido dos métodos para llevar a cabo la tarea de recomendación. El primer método se centra en los usuarios. Configuran sus etiquetas de acuerdo con sus antecedentes específicos y campos de interés. Además, para cada artículo procesado se obtienen etiquetas similares. Esto permite generar coincidencias entre ellos, mostrando al usuario los artículos de interés más similares. Dado un autor o un artículo de interés, el segundo método muestra a los usuarios autores y artículos similares, respectivamente. También es capaz de organizar un conjunto de artículos o autores de interés en función de sus palabras clave. Esto facilita la búsqueda de elementos similares y permite obtener recomendaciones para los usuarios. En cuanto a la arquitectura del sistema, comprende una base de datos ElasticSearch y dos sistemas principales que están conectados a ella: el IRC y el KMC. El IRC está compuesto por cuatro módulos que realizan las operaciones correspondientes para recopilar información de las fuentes web: *Articles Information Processor*, *Authors Information Processor*, *Reputation Calculator* and *Sentiment Calculator*. El KMC comprende dos módulos: *User Operations Manager* y *Similarity and Clustering Evaluator*. Engloban las tareas de recomendación y las principales operaciones y solicitudes realizadas por los usuarios de UNIKO. El segundo framework propuesto es FRESA, centrado en estimar la reputación de los investigadores. Se ha utilizado Semantic Scholar y otras fuentes de información web para recopilar conocimientos específicos de los autores y sus publicaciones asociadas. Esta funcionalidad se completa con la estimación de dos índices: el índice de relevancia y el índice de novedad. Se han definido según diferentes características. Para el índice de relevancia, se utilizaron siete características y se aplicaron medidas de similaridad basadas en texto para calcularlas. En el caso del índice de novedad, se utilizan técnicas simples de NLP para recopilar la información textual utilizada para estimarlo. Ambos índices se utilizan para calcular la puntuación de reputación. Mediante el cálculo de los índices propuestos a lo largo de los años para los autores, es posible analizar su trayectoria científica. Esto facilita estudiar con mayor detalle que otros índices la trayectoria científica de los investigadores. Por ejemplo, podría detectarse cuándo los autores han publicado obras muy relevantes y novedosas durante un periodo de

tiempo. De la misma manera, esta característica podría identificar cuándo esos autores reducen la cantidad de artículos publicados o su investigación se estanca en temas similares durante años. El tercer framework propuesto es DOCRIW, que clasifica los dominios web como *riesgoso* y *no riesgoso* en base a su reputación. Para ello, se utilizan fuentes de información web para recopilar información específica de dominios web potencialmente peligrosos. Esta funcionalidad se completa con un clasificador ML basado en un algoritmo LR. El clasificador ha sido entrenado a través de $\mathbf{1,500}$ dominios URL etiquetados ($\mathbf{750}$ *non-risky* y $\mathbf{750}$ *risky*). Se utilizan seis features para definir seis similaridades diferentes para medir la similaridad entre dominios. El primero usa la distancia normalizada *Levenshtein* para los nombres de los dominios web. Las otras $\mathbf{5}$ medidas de similaridad se definen a partir de la correspondencia entre las ciudades, los países, las fechas de creación, las fechas de vencimiento y los correos electrónicos. La similaridad global se produce a través de una combinación ponderada de todas estas similaridades individuales, donde los pesos representan la influencia de cada feature, definiendo la reputación del dominio web.

Se han realizado varios experimentos para evaluar el rendimiento de los tres frameworks. En el caso de UNIKO se han presentado tres experimentos relacionados con el proceso de recuperación de información y las tareas de evaluación de similaridad y agrupamiento. Por motivos de reproducibilidad, se han fijado los umbrales y los parámetros de compensación del framework. Los experimentos se han logrado una vez recopilada de las fuentes web la información de diez artículos y sus autores asociados. El primer experimento está relacionado con la reputación de los artículos y los autores. El segundo utiliza los resúmenes de los diez artículos para generar una puntuación de sentimiento. Esta puntuación se calcula usando la CNN como predictor más el léxico de SentiWordNet, y también usando solo el léxico. Esto permite mostrar los beneficios del enfoque considerado (Cambria, 2016) ilustrando las diferentes puntuaciones de reputación producidas. El tercer experimento evalúa los artículos más similares a un conjunto de artículos previamente seleccionados. Muestra las relaciones entre ellos a través de palabras clave. También se agrupan generando un clustering adecuado. En el caso de FRESA se han propuesto tres experimentos que han mostrado resultados prometedores para probar la viabilidad y para demostrar que es un enfoque más completo que los existentes. Esto se ha conseguido gracias a que se han añadido al cálculo de la reputación componentes nuevos a través del índice de relevancia y el índice de novedad. El primer experimento explica la trayectoria de reputación interpretándola de manera heurística. El segundo experimento demuestra tres cosas: la reputación mejora los índices h e i10; esta reputación evita las limitaciones de otras métricas para medir la reputación de un autor con una carrera científica corta pero de gran impacto; y esta reputación evita el fraude. Por último, también se han realizado tres experimentos para evaluar el rendimiento del sistema DOCRIW. El primero de ellos muestra una batería de pruebas para justificar la selección de elementos incluidos en el modelo de Machine Learning. El segundo valida el rendimiento del sistema evaluando la clasificación de dominios *risky* y *non-risky* en base a la reputación. Por último, se simula

la funcionalidad del sistema proporcionando las etiquetas predichas en base a los scores de reputación de cada dominio web.

5.2. Líneas de investigación futuras

Los tres frameworks propuestos son sistemas completos, pero son prototipos iniciales. Por ello, se incluirán mejoras futuras que permitirán mejorar las funcionalidades de cada uno de los frameworks y proteger las vulnerabilidades detectadas. Así como se propondrán otras métricas de reputación para otros dominios de aplicación relacionados que completen la propuesta global. Entre las mejoras que se pueden introducir en el framework UNIKO está considerar el análisis semántico latente (*Latent Semantic Analysis* (LSA)) (Hofmann, 2017) para generar nuevas relaciones entre etiquetas y también palabras clave. Esto permitirá extender el framework a nuevos niveles semánticos, pudiendo combinar información semántica de más de una fuente de información (actualmente WordNet (Miller, 1995) es el principal proveedor). Los objetivos de recopilar los textos completos de los artículos, obtener suficiente información sobre los autores y también sus posibles citas faltantes son las ramas principales para el trabajo futuro. Para mejorar el rendimiento del framework FRESA, se propone mejorar el índice de novedad ingresando información sobre los artículos de otros autores. Nótese que en este trabajo el índice de novedad se calcula considerando únicamente las publicaciones del autor. Además, el índice de relevancia podría redefinirse teniendo en cuenta el orden de los autores dentro del artículo. Este hecho actualizaría el componente que ya se ha introducido en la métrica al considerar el número de autores por artículo. Finalmente, sería interesante seguir investigando con otros algoritmos de agrupamiento para lograr seleccionar varios perfiles heterogéneos de autores con el fin de detectar futuras mejoras a incluir. Por último, para mejorar el rendimiento del sistema DOCRIW podrían probarse y compararse varias medidas de similaridad bien conocidas (W. Cohen *et al.*, 2003) (por ejemplo, edit distance, similaridad de Smith-Waterman, similaridad de Jaro-Winklers o similaridad de Monge-Elkan). Por otro lado, se podrían considerar nuevas funcionalidades que podrían brindar información relevante para los dominios web. Además, el clasificador LR podría ser reentrenado con aquellos dominios con alta probabilidad de ser *risky* o *non-risky* y también se podrían incluir técnicas de aprendizaje por refuerzo. Finalmente, sería interesante realizar más investigaciones con otros clasificadores. Aunque LR ha sido el modelo más eficiente en este experimento, otras configuraciones del resto de clasificadores, o incluso métodos de ensamble, también podrían ser efectivos.

5.3. Publicaciones

Esta tesis es el resultado de la combinación de varios trabajos que han derivado en tres artículos aceptados y/o publicados en revistas científicas y un artículo publicado en una conferencia. Los tres artículos publicados en revistas científicas son: *A Unified Knowledge Compiler to Provide Support the Scientific Community*, publicado en diciembre de 2018 en la revista *Knowledge-Based Systems*; *Knowledge-Based Approach to Detect Potentially Risky Websites*, publicado en enero del 2021 en la revista *IEEE Access*; y *Visual Framework for Scoring the Scientific Reputation of Researchers* que fue aceptado en marzo de 2023 en la revista *Knowledge and Information Systems*. También se publicó un artículo titulado *A Supervised Learning Approach to Detect Copyright Infringements* en la revista *International Conference on Information Management and Processing* (ICIMP) en enero de 2018, el cual fue presentado en la *International Conference on Information Management and Processing 2018* que tuvo lugar en el Imperial College de Londres en la misma fecha.

La investigación fue financiada por el Ministerio de Economía y Competitividad a través de varios programas de colaboración, incluyendo los programas Retos: PPI (Ref: RTC-2015-3580-7), UNIKO (Ref: RTC-2015-3521-7), SABERMED (Ref: RTC-2017-6253-1) y MODAS-IN (Ref: RTI-2018-094269-B-I00).

Referencias

Abdel-Kader, M. G., y Mentzeniot, V. (2007). The effect of corporate restructuring on the shareholders' value: The case of gec/marconi. *World Journal of Business Management*, *1*(1), 28–46.

Abraham, S., y Chengalur-Smith, I. (2010). An overview of social engineering malware: Trends, tactics, and implications. *Technology in Society*, *32*(3), 183–196.

Akerkar, R., y Sajja, P. (2010). *Knowledge-based systems.* Jones & Bartlett Publishers.

Alavi, M., y Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS quarterly*, 107–136.

Aleskerov, E., Freisleben, B., y Rao, B. (1997). Cardwatch: A neural network based database mining system for credit card fraud detection. En *Proceedings of the ieee/iafe 1997 computational intelligence for financial engineering (cifer)* (pp. 220–226).

Allen Institute for Artificial Intelligence and Semantic Scholar. (2018). *Semantic Scholar API.* `https://api.semanticscholar.org/`. ([Online: accedido 27-Feb-2018])

Baccianella, S., Esuli, A., y Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. En *Lrec* (Vol. 10, pp. 2200–2204).

Balakrishnan, H., Kaashoek, M. F., Karger, D., Morris, R., y Stoica, I. (2003). Looking up data in P2P systems. *Communications of the ACM*, *46*(2), 43–48.

Baldoni, M., Baroglio, C., Patti, V., y Rena, P. (2012). From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. *Intelligenza Artificiale*, *6*(1), 41–54.

Barnett, M. L., Jermier, J. M., y Lafferty, B. A. (2006). Corporate reputation: The definitional landscape. *Corporate reputation review*, *9*, 26–38.

Baxter, S., y Vogt, L. C. (2002). *Content management system.* Google Patents. (US Patent 6,356,903)

Bertin, M., y Atanassova, I. (2017). K-means and Hierarchical Clustering Method to Improve our Understanding of Citation Contexts. En *Proc. of the 2nd joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (birndl), tokyo, japan, ceur-ws. org* (pp. 107–112).

Bhavsar, K., Kumar, N., y Dangeti, P. (2017). *Natural Language Processing with Python Cookbook: Over 60 recipes to implement text analytics solutions using deep learning principles.* Packt Pub.

Bolton, R. J., y Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, 235–249.

Bordons, M., y cols. (1999). Evaluación de la actividad científica a través de indicadores bibliométricos. *Revista española de cardiología*, *52*(10), 790–800.

Brown, E. S., Palka, J., Helm, S. V., y Kulikova, A. (2022). The relative importance of reputation and pride as predictors of employee turnover in an academic medical center. *Health Care Management Review*, *47*(1), 66–77.

Cabral, L. (2012). Reputation on the internet. *The Oxford handbook of the digital economy*, 343–354.

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, *31*(2), 102–107.

Cappelletti-Montano, B., Columbu, S., Montaldo, S., y Musio, M. (2021). New perspectives in bibliometric indicators: Moving from citations to citing authors. *Journal of Informetrics*, *15*(3), 101164.

Carayol, N., Agenor, L., y Oscar, L. (2019). The right job and the job right: Novelty, impact and journal stratification in science. *Impact and Journal Stratification in Science (March 5, 2019).*

Carmeli, A., y Freund, A. (2002). The relationship between work and workplace attitudes and perceived external prestige. *Corporate reputation review*, *5*(1), 51–68.

Chen, K.-h., y Liao, P.-y. (2012). A comparative study on world university rankings: a bibliometric survey. *Scientometrics*, *92*(1), 89–103.

Chen, M., y Singh, J. P. (2001). Computing and using reputations for internet ratings. En *Proceedings of the 3rd acm conference on electronic commerce* (pp. 154–162).

Chiba, D., Tobe, K., Mori, T., y Goto, S. (2012). Detecting malicious websites by learning ip address features. En *2012 ieee/ipsj 12th international symposium on applications and the internet* (pp. 29–39).

Chikersal, P., Poria, S., y Cambria, E. (2015). SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning. En *Semeval@ naacl-hlt* (pp. 647–651).

Ciregan, D., Meier, U., y Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. En *Computer vision and pattern recognition (cvpr), 2012 ieee conference on* (pp. 3642–3649).

Cohen, B. A. (2017). Point of view: How should novelty be valued in science? *Elife*, *6*, e28699.

Cohen, W., Ravikumar, P., y Fienberg, S. (2003). A comparison of string metrics for matching names and records. En *Kdd workshop on data cleaning and object consolidation* (Vol. 3, pp. 73–78).

The computing research and education association of australasia. (s.f.). `https://www.core.edu.au/`. (Accedido: 2021-06-25)

CORE. (s.f.). *Core portal.* http://portal.core.edu.au/conf-ranks/.

Cortes, C., y Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

db.aa419.org. (2019). *Fake sites database.* https://db.aa419.org/fakebankslist.php. ([Online: accedido 24-Jan-2020])

Delgado-Márquez, B. L., Bondar, Y., y Delgado-Márquez, L. (2012). Higher education in a global context: Drivers of top-universities' reputation. *Problems of Education in the 21st Century*, *40*, 17.

Devaki, R., Kathiresan, V., y Gunasekaran, S. (2014). Credit card fraud detection using time series analysis. *International Journal of Computer Applications*, *3*, 8–10.

Dhamdhere, S. N. (2018). Cumulative citations index, h-index and i10-index (research metrics) of an educational institute: A case study. *International Journal of Library and Information Science*, *10*(1), 1–9.

Dolle, R. (2014). *Online reputation management* (B.S. thesis). University of Twente.

Dounis, A. I., Tiropanis, P., Argiriou, A., y Diamantis, A. (2011). Intelligent control system for reconciliation of the energy savings with comfort in buildings using soft computing techniques. *Energy and Buildings*, *43*(1), 66–74.

Egghe, L. (2006). An improvement of the h-index: The g-index. *ISSI newsletter*, *2*(1), 8–9.

Ettenson, R., y Knowles, J. (2008). Dont confuse reputation with brand. *MIT Sloan Management Review*, *49*(2), 19.

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., y Karageorgopoulos, D. E. (2008). Comparison of scimago journal rank indicator with journal impact factor. *The FASEB journal*, *22*(8), 2623–2628.

Fernández-Isabel, A., y Fuentes-Fernández, R. (2017). Extending a generic traffic model

to specific agent platform requirements. *Comput. Sci. Inf. Syst.*, *14*(1), 219–237.

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.

Fombrun, C., y Van Riel, C. (1997). The reputational landscape. *Corporate reputation review*, 1–16.

Forbes, C., Evans, M., Hastings, N., y Peacock, B. (2011). *Statistical distributions*. John Wiley & Sons.

Forrest, S., y cols. (1993). Genetic algorithms- Principles of natural selection applied to computation. *Science*, *261*(5123), 872–878.

Foster, D., McGregor, C., y El-Masri, S. (2005). A survey of agent-based intelligent decision support systems to support clinical management and research. En *Proceedings of the 2nd international workshop on multi-agent systems for medicine, computational biology, and bioinformatics* (pp. 16–34).

Freitas, A. A. (2003). A survey of evolutionary algorithms for data mining and knowledge discovery. En *Advances in evolutionary computing* (pp. 819–845). Springer.

Freund, Y., y Schapire, R. E. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, *14*(5), 771–780.

Fricke, S. (2018). Semantic scholar. *Journal of the Medical Library Association: JMLA*, *106*(1), 145.

Friedman, J., Hastie, T., y Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York.

Galois, É., y Neumann, P. M. (2011). *The mathematical writings of évariste galois* (Vol. 6). European Mathematical Society.

Gamon, M. (2006). Graph-based text representation for novelty detection. En *Proceedings of textgraphs: The first workshop on graph based methods for natural language processing* (pp. 17–24).

Garfield, E., y cols. (1994). The impact factor. *Current contents*, *25*(20), 3–7.

Garg, A., Gupta, K., y Singh, A. (2017). Survey of Web Crawler Algorithms. *International Journal*, *8*(5).

Gokulan, B. P., y Srinivasan, D. (2010). Distributed geometric fuzzy multiagent urban traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, *11*(3), 714–727.

Gómez, E. J., del Pozo, F., Ortiz, E. J., Malpica, N., y Rahms, H. (1998). A broadband multimedia collaborative system for advanced teleradiology and medical imaging diag-

nosis. *IEEE transactions on information technology in Biomedicine*, *2*(3), 146–155.

Gormley, C., y Tong, Z. (2015). *Elasticsearch: The definitive guide: A distributed real-time search and analytics engine*. O'Reilly Media, Inc.

Group, R. (2018). *Elsevier official web page*. `https://www.elsevier.com/`. ([Online: accedido 03-Feb-2018])

Gruber, M. (2017). *Improving efficiency by shrinkage: The james–stein and ridge regression estimators*. Routledge.

Guerrero-Sosas, J. D., Chicharro, F. P. R., Serrano-Guerrero, J., Menendez-Dominguez, V., y Castellanos-Bolaños, M. E. (2019). A proposal for a recommender system of scientific relevance. *Procedia Computer Science*, *162*, 199–206.

Haddaway, N. R. (2015). The use of web-scraping software in searching for grey literature. *Grey J*, *11*(3), 186–90.

Hastie, T., Tibshirani, R., y Friedman, J. (2009). Overview of supervised learning. En *The elements of statistical learning* (pp. 9–41). Springer.

He, B., Macdonald, C., He, J., y Ounis, I. (2008). An effective statistical approach to blog post opinion retrieval. En *Proceedings of the 17th acm conference on information and knowledge management* (pp. 1063–1072).

Helm, S., Liehr-Gobbers, K., y Storck, C. (2011). *Reputation management*. Springer Science & Business Media.

Herbig, P., y Milewicz, J. (1995). The relationship of reputation and credibility to brand success. *Journal of consumer marketing*, *12*(4), 5–11.

Hilas, C. S. (2009). Designing an expert system for fraud detection in private telecommunications networks. *Expert Systems with applications*, *36*(9), 11559–11569.

Hirsch, J. E., y Buela-Casal, G. (2014). The meaning of the h-index. *International Journal of Clinical and Health Psychology*, *14*(2), 161–164.

Hofmann, T. (2017). Probabilistic latent semantic indexing. En *Acm sigir forum* (pp. 211–218).

Holding, A. N. (2019). Novelty in science should not come at the cost of reproducibility. *The FEBS journal*, *286*(20), 3975–3979.

International Organization for Standardization. (2019). *Country Codes - ISO 3166*. https://www.iso.org/iso-3166-country-codes.html. ([Online: accedido 24-Jan-2020])

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

JCR. (s.f.). *Journal citations reports*. https://jcr.clarivate.com.

Jennings, N. R. (2001). An agent-based approach for building complex software systems. *Communications of the ACM*, *44*(4), 35–41.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.

Jøsang, A., y Golbeck, J. (2009). Challenges for robust trust and reputation systems. En *Proceedings of the 5th international workshop on security and trust management (smt 2009), saint malo, france* (Vol. 5).

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation*, *14*(1), 10–25.

Lameski, P., Zdravevski, E., Mingov, R., y Kulakov, A. (2015). Svm parameter tuning with grid search and its impact on reduction of model over-fitting. En *Rough sets, fuzzy sets, data mining, and granular computing* (pp. 464–474). Springer.

Langston, M., y Tyler, J. (2004). Linking to journal articles in an online teaching environment: The persistent link, DOI, and OpenURL. *The Internet and Higher Education*, *7*(1), 51–58.

Lanubile, F., Ebert, C., Prikladnicki, R., y Vizcaíno, A. (2010). Collaboration tools for global software engineering. *IEEE software*, *27*(2).

Lariviere, V., y Sugimoto, C. R. (2019). The journal impact factor: A brief history, critique, and discussion of adverse effects. En *Springer handbook of science and technology indicators* (pp. 3–24). Springer.

LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Li, W. (2004). Using genetic algorithm for network intrusion detection. *Proceedings of the United States Department of Energy Cyber Security Group*, *1*, 1–8.

López-Cózar, E. D., Orduña-Malea, E., y Martín-Martín, A. (2019). Google scholar as a data source for research assessment. En *Springer handbook of science and technology indicators* (pp. 95–127). Springer.

Lops, P., De Gemmis, M., y Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. En *Recommender systems handbook* (pp. 73–105). Springer.

Lu, J. (2009). *Video fingerprinting for copy identification: from research to industry applications.* (Vol. 7254).

Luna, F., y Stefansson, B. (2012). *Economic Simulations in Swarm: Agent-based modelling and object oriented programming* (Vol. 14). Springer Science & Business Media.

Ma, J., Saul, L. K., Savage, S., y Voelker, G. M. (2009). *Beyond blacklists: learning to detect malicious web sites from suspicious urls.* ACM.

Madden, M., y Smith, A. (2010). Reputation management and social media.

MalwareURL.com. (2019). *Malware urls database.* https://www.malwareurl.com/. ([Online: accedido 24-Jan-2020])

Mao, W. (2003). *Modern cryptography: theory and practice.* Prentice Hall Professional Technical Reference.

Marcus, M. P., Marcinkiewicz, M. A., y Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, *19*(2), 313–330.

Masood, F., Almogren, A., Abbas, A., Khattak, H. A., Din, I. U., Guizani, M., y Zuair, M. (2019). Spammer detection and fake user identification on social networks. *IEEE Access*, *7*, 68140–68152.

Meho, L. I., y Rogers, Y. (2008). Citation counting, citation ranking, and h-index of human-computer interaction researchers: a comparison of scopus and web of science. *Journal of the American Society for Information Science and Technology*, *59*(11), 1711–1726.

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Mitnick, B. M., y Mahon, J. F. (2007). The concept of reputational bliss. *Journal of Business Ethics*, *72*(4), 323–333.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American society for information science*, *48*(9), 810–832.

Moguerza, J. M., y Muñoz, A. (2006). Support vector machines with applications. *Statistical Science*, *21*(3), 322–336.

Narkhede, N., Shapira, G., y Palino, T. (2016). *Kafka: The Definitive Guide.* O'Reilly Media, Inc.

Natekin, A., y Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*.

Ng, K. W., Tsai, F. S., Chen, L., y Goh, K. C. (2007). Novelty detection for text documents using named entity recognition. En *2007 6th international conference on information, communications & signal processing* (pp. 1–5).

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., y Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, *50*(3), 559–569.

OEDI. (2018). *Spanish observatory of cyber crimes.* http://oedi.es/estadisticas/.

of Florida, U. (s.f.). *George a. smathers libraries, university of florida.* https://guides.uflib.ufl.edu/.

ORCID, I. (2018). *ORCID: Connecting Research and Researchers.* https://orcid.org/. ([Online: accedido 27-Feb-2018])

OSI. (2011). *Study by observatory of information security.* https://www.prevent.es/Documentacion/estudio_fraude_4t10.pdf.

Over, R. (1982). The durability of scientific reputation. *Journal of the History of the Behavioral Sciences*, *18*(1), 53–61.

Page, L., Brin, S., Motwani, R., y Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web.* (Inf. Téc.). Stanford InfoLab.

Pal, S. K., Talwar, V., y Mitra, P. (2002). Web mining in soft computing framework: relevance, state of the art and future directions. *IEEE Transactions on Neural Networks*, *13*(5), 1163–1177.

Pang, B., Lee, L., y cols. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135.

Pazzani, M. J., y Billsus, D. (2007). Content-based recommendation systems. En *The adaptive web* (pp. 325–341). Springer.

Poria, S., Cambria, E., y Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. En *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2539–2544).

portal, C. (s.f.). *Conference portal.* https://www.core.edu.au/conference-portal.

Priem, J., Taraborelli, D., Groth, P., y Neylon, C. (2011). *Altmetrics: A manifesto.* http://altmetrics.org/manifesto. (Accedido: 2021-06-25)

Qin, Y., Liu, J., Wu, C., y Shi, Y. (2012). uEmergency: a collaborative system for emergency management on very large tabletop. En *Proceedings of the 2012 acm international conference on interactive tabletops and surfaces* (pp. 399–402).

RAE. (s.f.-a). *Oxford english dictionary.* https://www.oed.com/.

RAE. (s.f.-b). *Real academia española.* https://dle.rae.es/.

Ravi, V., Kurniawan, H., Thai, P. N. K., y Kumar, P. R. (2008). Soft computing system for bank performance prediction. *Applied soft computing*, *8*(1), 305–315.

RELX Group. (2018). *ScienceDirect official web page.* https://www.sciencedirect.com/. ([Online: accedido 27-Feb-2018])

Resnick, P., Kuwabara, K., Zeckhauser, R., y Friedman, E. (2000). Reputation systems. *Communications of the ACM*, *43*(12), 45–48.

Reuters, T. (2018). *ResearcherID.* `http://www.researcherid.com/`. ([Online: accedido 27-Feb-2018])

Richardson, L., y Ruby, S. (2008). *RESTful web services.* O'Reilly Media, Inc.

Rosenberg, M. J. (2001). *E-learning: Strategies for delivering knowledge in the digital age* (Vol. 9). McGraw-Hill New York.

Rosenkrantz, D. J., y Stearns, R. E. (2003). NP-complete problems. *Encyclopedia of Computer Science*.

Rousseau, R. (2001). Indicadores bibliométricos y econométricos en la evaluación de instituciones científicas. *Acimed*, *9*, 50–60.

Rufener, J. (2006). *Document management systems.* Google Patents. (US Patent App. 11/536,618)

Sabater, J., y Sierra, C. (2002). Reputation and social network analysis in multi-agent systems. En *Proceedings of the first international joint conference on autonomous agents and multiagent systems: part 1* (pp. 475–482).

Sarkar, D. (2016). *Text analytics with python: A practical real-world approach to gaining actionable insights from your data.* Apress.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. *Approaches to emotion*, *2293*, 317.

Schiffman, B., y McKeown, K. (2005). Context and learning in novelty detection. En *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 716–723).

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85–117.

Sekine, S., y Ranchhod, E. (2009). *Named entities: recognition, classification and use* (Vol. 19). John Benjamins Publishing.

Sendhilkumar, S., Elakkiya, E., y Mahalakshmi, G. (2013). Citation semantic based approaches to identify article quality. En *Proceedings of international conference iccsea* (pp. 411–420).

Senin, P. (2008). Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, *855*(1-23), 40.

Shabtai, A., Moskovitch, R., Elovici, Y., y Glezer, C. (2009). Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey. *information security technical report*, *14*(1), 16–29.

Singh, M., y Markeset, T. (2009). A methodology for risk-based inspection planning of oil and gas pipes based on fuzzy logic framework. *Engineering Failure Analysis*, *16*(7),

2098–2113.

Socher, R. (2014). *Recursive deep learning for natural language processing and computer vision* (Tesis Doctoral no publicada). Citeseer.

Socher, R., Lin, C. C., Manning, C., y Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. En *Proceedings of the 28th international conference on machine learning (icml-11)* (pp. 129–136).

Sosa, J. D. T. G., Domínguez, V. H. M., Bolaños, M. E. C., y Montalvo, J. R. G. (2019). Use of an ontological model to assess the relevance of scientific production. *IEEE Latin America Transactions*, *17*(09), 1424–1431.

Springer International Publishing AG. (2018). *Springer official web page.* `http://www.springer.com`. ([Online: accedido 27-Feb-2018])

Steinbach, M., Karypis, G., Kumar, V., y cols. (2000). A comparison of document clustering techniques. En *Kdd workshop on text mining* (pp. 525–526).

Swan, J., Newell, S., Scarbrough, H., y Hislop, D. (1999). Knowledge management and innovation: networks and networking. *Journal of Knowledge management*, *3*(4), 262–275.

T. Erekhinskaya and D. Moldovan. (2013). Lexical chains on wordnet and extensions. En C. Boonthum-Denecke y G. M. Youngblood (Eds.), *Proceedings of the twenty-sixth international florida artificial intelligence research society conference* (pp. 52–57). The AAAI Press.

Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics*, *8*, 321–340.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., y Qin, B. (2014). Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. En *Acl (1)* (pp. 1555–1565).

Torrente, J., Del Blanco, Á., Marchiori, E. J., Moreno-Ger, P., y Fernández-Manjón, B. (2010). <e-Adventure>: Introducing educational games in the learning process. En *Education engineering (educon), 2010 ieee* (pp. 1121–1126).

Trewin, S. (2000). Knowledge-based recommender systems. *Encyclopedia of library and information science*, *69*(32), 180.

Veugelers, R., y Wang, J. (2019). Scientific novelty and technological impact. *Research Policy*, *48*(6), 1362–1372.

Volkwein, J. F., y Sweitzer, K. V. (2006). Institutional prestige and reputation among research universities and liberal arts colleges. *Research in Higher Education*, *47*, 129–148.

Wang, J., Veugelers, R., y Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416–1436.

Weitzel, L., Oliveira, J. P. M. d., y Quaresma, P. (2013). Exploring trust to rank reputation in microblogging. En *International conference on database and expert systems applications* (pp. 434–441).

Whois.com. (2019). *Whois domain information.* https://www.whois.com/whois/. ([Online: accedido 24-Jan-2020])

Wilensky, U., y Rand, W. (2015). *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo*. MIT Press.

Witten, I. H., Frank, E., Hall, M. A., y Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yager, R. R., y Zadeh, L. A. (2012). *An introduction to fuzzy logic applications in intelligent systems* (Vol. 165). Springer Science & Business Media.

Yu, M.-C., Wu, Y.-C. J., Alhalabi, W., Kao, H.-Y., y Wu, W.-H. (2016). Researchgate: An effective altmetric indicator for active researchers? *Computers in human behavior*, *55*, 1001–1006.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... others (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM*, *59*(11), 56–65.

Zaíane, O. R. (2002). Building a recommender agent for e-learning systems. En *Computers in education, 2002. proceedings. international conference on* (pp. 55–59).

Zhang, M., Wang, W., y Li, X. (2008). A paper recommender for scientific literatures based on semantic concept similarity. *Digital libraries: Universal and ubiquitous access to information*, 359–362.

Zhang, X., Sun, S., y Zhang, K. (2017). A Novel Comprehensive Approach for Estimating Concept Semantic Similarity in WordNet. *arXiv preprint arXiv:1703.01726*.

Zhang, X.-X., Wang, Y.-M., Chen, S.-Q., Chu, J.-F., y Chen, L. (2018). Gini coefficient-based evidential reasoning approach with unknown evidence weights. *Computers & Industrial Engineering*.

Zhou, C. V., Karunasekera, S., y Leckie, C. (2005). A peer-to-peer collaborative intrusion detection system. En *Networks, 2005. jointly held with the 2005 ieee 7th malaysia international conference on communication., 2005 13th ieee international conference on* (Vol. 1, pp. 6–pp).

Zhu, W., Zeng, N., y Wang, N. (2010). Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, *19*.