



TRABAJO FIN DE MÁSTER  
MÁSTER UNIVERSITARIO EN INGENIERÍA INFORMÁTICA

# Desarrollo de un prototipo de motor de búsqueda que incorpore técnicas bibliométricas para mejorar la recuperación

**Autor**

Aythami Estévez Olivas

**Director**

Juan Manuel Fernández Luna



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE  
TELECOMUNICACIÓN

---

Granada, 4 de septiembre de 2018



# Índice general

<b>1. Introducción</b>	<b>1</b>
1.1. Recuperación de información . . . . .	1
1.1.1. Relevancia y similitud . . . . .	1
1.1.2. Las tres dimensiones de la RI . . . . .	2
1.1.3. Componentes de un sistema de RI . . . . .	3
1.1.4. Modelos . . . . .	4
1.1.5. Contexto histórico . . . . .	5
1.2. Bibliometría . . . . .	6
1.2.1. Definición . . . . .	6
1.2.2. Medidas . . . . .	6
1.2.3. Limitaciones . . . . .	8
1.3. Combinando ambas disciplinas . . . . .	9
1.3.1. Sistemas de RI actuales con medidas bibliométricas . . . . .	9
1.3.2. Trabajos relacionados . . . . .	10
1.4. Enfoque planteado . . . . .	11
<b>2. Objetivos</b>	<b>13</b>
<b>3. Planificación</b>	<b>15</b>
3.1. Gestión de recursos . . . . .	15
3.1.1. Personal . . . . .	15
3.1.2. Hardware . . . . .	15
3.1.3. Software . . . . .	16
3.2. Metodología . . . . .	16
3.3. Planificación temporal . . . . .	18
<b>4. Diseño</b>	<b>19</b>
4.1. Modelo de datos . . . . .	20
4.1.1. <i>Author</i> . . . . .	20
4.1.2. <i>Abstract</i> . . . . .	21
4.2. Arquitectura del sistema . . . . .	22
<b>5. Técnicas y herramientas [Aún en progreso]</b>	<b>23</b>

<b>6. Desarrollo</b>	<b>25</b>
6.1. <i>Sprint</i> 1 . . . . .	25
6.2. <i>Sprint</i> 2 . . . . .	26
6.3. <i>Sprint</i> 3 . . . . .	27
6.4. <i>Sprint</i> 4 . . . . .	28
6.5. <i>Sprint</i> 5 . . . . .	29
6.5.1. Obtención de datos . . . . .	30
6.6. <i>Sprint</i> 6 . . . . .	34
6.7. <i>Sprint</i> 7 . . . . .	34
6.8. <i>Sprint</i> 8 . . . . .	34
<b>7. Conclusiones y Trabajos Futuros</b>	<b>35</b>
<b>Bibliografía</b>	<b>40</b>
<b>Anexos</b>	<b>41</b>
<b>A. Glosario</b>	<b>43</b>
<b>B. Lista de Acrónimos</b>	<b>45</b>
<b>C. Manual de uso</b>	<b>47</b>

# Índice de figuras

1.1. Componentes de un sistema RI [1] . . . . .	3
1.2. Donut de Altmetric . . . . .	8
3.1. Ejemplo de tablero de un <i>sprint</i> . . . . .	17
4.1. Clase <i>Author</i> . . . . .	20
4.2. Clase <i>Abstract</i> . . . . .	21
4.3. Diagrama de arquitectura final con las principales tecnologías usadas en cada nodo . . . . .	22
6.1. Fragmento de la lista priorizada creada . . . . .	26
6.2. Ejemplo de resumen de uno de los papers . . . . .	27
6.3. Aspecto tabla ranking UGRinvestiga . . . . .	30
6.4. Distribución de nacionalidades entre los autores recuperados .	32
6.5. Fragmento de la distribución de ciudades entre los autores españoles recuperados . . . . .	33
6.6. Representación del clustering de autores llevado a cabo en la limpieza de datos . . . . .	33



# Índice de cuadros

3.1. Especificaciones del equipo utilizado . . . . .	15
3.2. Planificación inicial de tareas . . . . .	18





# Capítulo 1

## Introducción

### 1.1. Recuperación de información

La **Recuperación de Información (RI)** es una disciplina que trata de modelar, diseñar e implementar sistemas capaces de promocionar acceso basado en contenidos. [2]

En general un sistema de RI recibe una petición o consulta del usuario y debe devolver de entre su conjunto de información las unidades relacionadas con la consulta.

Estas unidades pueden representar cualquier tipo de elemento: ficheros de texto, imágenes, archivos de audio, etc. De forma genérica se denominan **documentos** así como al conjunto de información se le denomina **colección**.

#### 1.1.1. Relevancia y similitud

La **relevancia** hace referencia a la relación entre la consulta de un usuario y los documentos recuperados. Por ello intuitivamente se entiende que un documento es relevante para una consulta concreta si contribuye a satisfacer la necesidad de información expresada por la misma. Aunque el concepto pueda parecer claro, la relevancia no es para nada absoluta, es una media subjetiva que depende de varios factores como quien la valore o como se haya planteado la consulta inicial.

Respecto a la **similitud** esta es una medida de semejanza entre documentos o entre documentos y consultas. Otra vez más nos encontramos ante una medida relativa y que se puede medir de diversas maneras: comparación de cadenas de texto, uso de un mismo vocabulario, que dos documentos pertenezcan al mismo autor, que dos documentos tengan múltiples referencias

comunes...[1]

### 1.1.2. Las tres dimensiones de la RI

Se puede decir que las tres dimensiones principales de la RI son:[2]

#### Acceso a la información

Como puede acceder un usuario a los datos. Existen diversos paradigmas de búsqueda:

- **Clasificación:** cada documento pertenece a clases y estas se pueden usar como jerarquías.
- **Agrupamiento:** documentos se agrupan en conjuntos.
- **Filtrado:** se selecciona un subconjunto de documentos.
- **Recomendación:** los documentos se presentan al usuario basados en su interacción previa con el sistema.
- **Resumen:** fragmentos de documentos utilizados para reducir la información presentada al usuario, muy típico de motores de búsqueda web.

#### Tipos de información

Hoy en día vivimos en una marea de información muy heterogénea y creciente lo que hace complicado definir los distintos tipos, por citar los principales actualmente:

- **Documentos textuales** como paginas web o PDF.
- **Partes de un documento**, capítulos o secciones de este.
- **Búsqueda de información multimedia:** como canciones o vídeos a partir de propiedades perceptibles o incluso de otros elementos multimedia, la búsqueda de imágenes en Google imágenes por ejemplo.
- **Búsqueda de e-mails:** implementado en cualquier cliente de correo web o nativo.
- **Búsqueda geográfica:** por nombre del lugar, sitios cercanos...

## Colección

Guarda relación con los documentos que pueden ser buscados y se pueden clasificar tres tipos en función del tamaño:

- **Personal:** ficheros del dispositivo del usuario.
- **Corporativa:** documentos de una empresa, algo más compleja, supone búsqueda en múltiples ubicaciones conectadas en red.
- **Web:** cualquier documento web, el volumen de datos y la infraestructura es descomunal.

### 1.1.3. Componentes de un sistema de RI

De manera general un sistema de RI se compone de los elementos que se observan en la siguiente figura. Ha de contar una interfaz de usuario encargada de recoger las peticiones y mostrar los resultados. Será necesario interpretar estas consultas para convertirlas en términos que el sistema pueda entender.

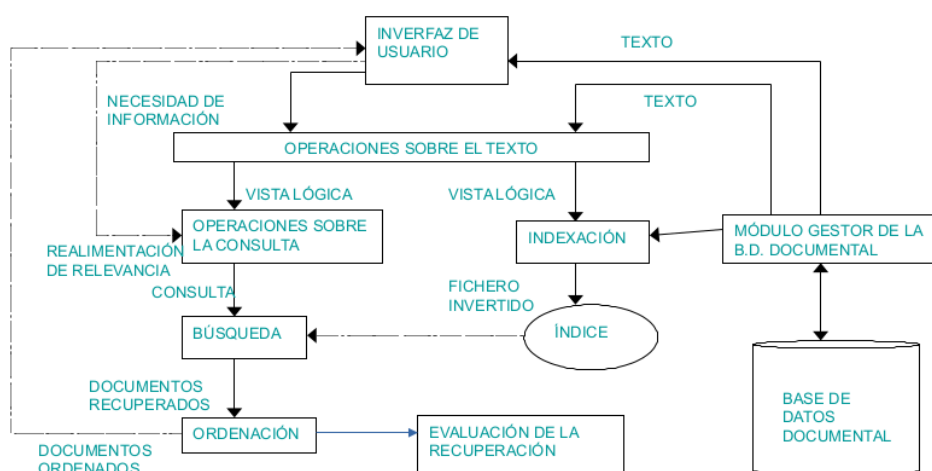


Figura 1.1: Componentes de un sistema RI [1]

Una vez hecho esto el algoritmo de búsqueda encontrará los documentos que sean relevantes para la consulta, tras esto será necesario puntuar estos documentos recuperados y devolver esta lista puntuada al usuario.

Para realizar este proceso de búsqueda y priorización se servirá del **índice**, una o varias estructuras de datos que permiten optimizar la búsqueda sobre una colección. En su forma más básica se trata de una estructura que relaciona términos con los documentos en los que aparecen, de modo que

si una consulta contiene esos términos se devuelven los documentos que los contienen.

#### 1.1.4. Modelos

Se puede definir un modelo de RI como una especificación de la forma de representar documentos, consultas y realizar comparaciones entre ambos [1]. El objetivo final es calcular una puntuación para cada documento dada una consulta específica que determine el grado de relevancia de este, utilizando esa medida se puede llevar a cabo una ordenación o ranking de los documentos.

Los modelos clásicos a pesar de ser los más básicos sirven como base para crear otros más complejos como alguno de los que se hablará posteriormente. Estos modelos clásicos son:[2]

El **Modelo booleano** basado en teoría de conjuntos y lógica booleana. Se define el conjunto  $V$  como todas las palabras clave de la colección,  $V = \{t_1, t_2, \dots, t_M\}$ , así mismo se define el conjunto  $D$  como el conjunto de todos los documentos de la colección  $D = \{d_1, d_2, \dots, d_n\}$ .

Cada documento  $d_i$  se representa por tanto como un conjunto de términos que aparecen en él, este es un subconjunto de  $V$ . Las consultas en este modelo se representan mediante las operaciones booleanas típicas AND, OR y NOT.

El **Modelo vectorial** modela los documentos y consultas como vectores de términos en un espacio vectorial de dimensión definida por el número de términos de la colección.

Cada documento es por tanto un vector en dicho espacio vectorial. Usando los conjuntos descritos en el modelo previo se puede definir un documento  $d_i$  como el vector de términos  $\vec{d}_i = (w_{1,i}, w_{2,i}, \dots, w_{M,i})$  donde  $w_{i,j}$  representa el peso del término  $i$  en el documento  $j$ .

Queda por determinar el esquema de pesos y la función de similitud entre vectores.

El **Modelo probabilístico** pretende expresar la relevancia de los documentos utilizando la teoría de probabilidades, por tanto se define  $P(Rel|d, q)$  como la probabilidad de que dado un documento  $d$  y una consulta  $q$  el documento sea relevante con cierta probabilidad.

Simplificando la notación a  $P(Rel|d)$  y usando el Teorema de Bayes<sup>1</sup> podemos transformar el cálculo de esa probabilidad en:

$$P(Rel|d) = \frac{P(d|Rel)P(Rel)}{P(d)} \quad (1.1)$$

La función de similitud en el modelo probabilístico se expresa como:

$$\text{sim}(q, d) = \frac{P(\text{Rel}|d)}{P(\overline{\text{Rel}}|d)} \quad (1.2)$$

Es decir el ratio entre la probabilidad de que un documento sea relevante y no para  $q$ .

Aplicando la transformación anterior esta función queda como:

$$\text{sim}(q, d) = \frac{P(d|\text{Rel})}{P(d|\overline{\text{Rel}})} \frac{P(\text{Rel})}{P(\overline{\text{Rel}})} \approx \frac{P(d|\text{Rel})}{P(d|\overline{\text{Rel}})} \quad (1.3)$$

Donde  $\frac{P(\text{Rel})}{P(\overline{\text{Rel}})}$  se conoce como característica independiente de consulta al suponer la relación entre las probabilidades de escoger un documento al azar y que este sea relevante o irrelevante, se suele eliminar como simplificación.

Por otro lado  $\frac{P(d|\text{Rel})}{P(d|\overline{\text{Rel}})}$  supone la relación entre la probabilidad sabiendo que un documento es relevante este sea  $d$  y su inversa. Estas probabilidades resultan más fáciles de calcular y son las utilizadas en estos modelos.

### 1.1.5. Contexto histórico

La historia de la RI comienza antes de la era digital. Como ejemplo las tablas de contenidos e índices de un libro forman un sistema de RI a pequeña escala, los índices relacionan términos de indexación con su ubicación en el documento de forma similar a un glosario de términos, ver 2 .

Esto es lo que se conoce como búsqueda **pre-coordinada** donde los términos de búsqueda o consultas están definidas de antemano, esto hace que se pueda organizar muy bien la información (como se hace en una biblioteca por ejemplo), pero requiere que el usuario conozca estos términos y no supone un método muy escalable teniendo en cuenta el volumen de información manejado en sistemas actuales. Por ello surgió la **post-coordinación** en la década de 1950 que se basa en definir las consultas en el momento de la búsqueda, dándole libertad al usuario.

A este último enfoque pertenecen la mayoría de los sistemas actuales, los cuales recibieron un enorme impulso con el desarrollo de la web iniciado en 1989 por Tim Berners-Lee en el CERN. Los primeros buscadores web de la forma que los conocemos hoy surgieron entorno a 1994, sistemas como Lycos o Altavista.

La búsqueda de documentos en la web siempre ha resultado un reto debido a su naturaleza heterogénea, la propuesta de unos jóvenes estudiantes

de la universidad de Stanford en 1998 supuso una revolución y asentó lo que hoy conocemos como buscador. Esa propuesta fue el algoritmo *PageRank*[3] y esos chicos eran Larry Page, Sergey Brin, Rajeev Motwani y Terry Winograd; los dos primeros fundaron poco después una empresa llamada Google, que a día de hoy es el buscador más utilizado. [4]

Hoy en día existe una dependencia casi absoluta por los motores de búsqueda para navegar por internet, lo que hace de este un tema tan crucial en el que aún se sigue trabajando, con retos aún por delante como el enorme tamaño que ha alcanzado la web, la heterogeneidad de contenido con cada vez más contenidos multimedia o el acceso desde cualquier tipo de dispositivo y desde cualquier lugar.

## 1.2. Bibliometría

### 1.2.1. Definición

La **bibliometría** se puede definir como el análisis estadístico de publicaciones escritas. Sus métodos se suelen utilizar para ofrecer un análisis cuantitativo de la literatura académica [5]. Se relaciona mucho con la **cienciometría** que se puede entender como el estudio cuantitativo de la ciencia de forma general [6].

Específicamente para este trabajo resulta destacable el enfoque de poder medir la importancia de trabajos científicos de forma cuantitativa y como esto se puede utilizar para mejorar los resultados de un sistema RI. Yendo al uso más particular que se llevara a cabo durante el desarrollo de este trabajo ambas disciplinas proponen una serie de medidas particulares, algunas de las cuales comentaré en el siguiente apartado.

### 1.2.2. Medidas

#### Número de citas

Esta es una de las métricas más directas y sencillas, se basa la premisa de que un documento científico es más relevante si cuenta con mayor número de citas. Actualmente se encuentra disponible en casi cualquier plataforma bibliográfica como Google Scholar, Scopus o Web of Science.

Sobre esta métrica básica se han construido muchas posteriormente como el impacto de citas (media de citas por documento), el impacto de citas ponderado por el campo (la anterior pero ponderado por materia, tipo de publicación y año), *Highly cited papers* (documentos en el top 1% de citas ponderado por campo y año)[7], *CiteScore* (media de citas anuales para

todos los documentos de una revista concreta durante los últimos 3 años), *SCImago Journal Rank* (medida que pondera el número de citas de una revista científica con el prestigio de las que la citan) o *SNIP* (normalización del número de citas de un paper por su impacto en la materia) [8].

### Análisis de co-citación

Medida de similaridad que establece que dos documentos serán semejantes si aparecen citados conjuntamente con frecuencia por otros documentos. Sobre esta premisa se puede considerar un índice de co-citación que es simplemente el número de citas co-citaciones de 2 documentos.

También existen métricas más complejas como el análisis de co-citación por proximidad que incorpora a la idea previa el hecho de si dos citas aparecen en la misma sección del texto estas estarán más relacionadas que si una aparece en la introducción y otra en las conclusiones por ejemplo. [9]

### Índice h

Este índice se puede aplicar a un autor científico, el índice h de un autor es x si este tiene x artículos con al menos x citas [7]. Se utiliza ampliamente para medir la productividad de un investigador.

Tiene diversas variantes como el índice g (índice h para el número de citas medio), índice m (corrección del índice h con el tiempo) o el índice Py (número medio de publicaciones por año).

### Acoplamiento bibliográfico

Muy relacionado con el análisis de co-citación, el acoplamiento bibliográfico o *bibliographic coupling* basa su concepto de similaridad en que 2 documentos serán semejantes si comparen citas.

En base a esta idea se puede crear una métrica simple que sea el número de referencias comunes entre dos documentos, o medidas más complejas como el acoplamiento bibliográfico entre autores que supone el acoplamiento entre el conjunto de citas de todos los trabajos de un autor con otro [10].

### Altmetrics

Conjunto de medidas relativas al mundo *online*, veces que un documento ha sido descargado, compartido, mencionado en las redes sociales, blogs, wikipedia... [11]

Existen diversas compañías que ofrecen productos sobre estas métricas siendo una de las principales Altmetric. Entre sus productos destaca el "donut" que se puede apreciar en la siguiente imagen. Este corresponde con una medalla que puede colocar en la página de un artículo y permite de un vistazo la atención que está generando este trabajo. También incluye información en detalle de donde se está hablando del mismo incluyendo comentarios en redes sociales.

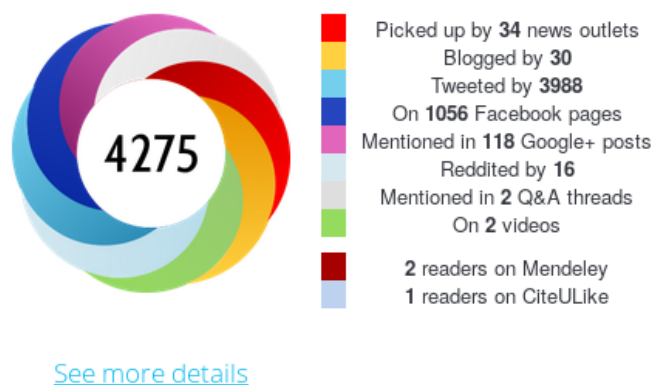


Figura 1.2: Donut de Altmetric

### 1.2.3. Limitaciones

Estas medidas ayudan a dar una idea de la importancia de un documento o de la productividad de un autor, pero no deben de ser tomadas como únicos criterios ya que presentan sus limitaciones.

Respecto a las medidas basadas en número de citas son muy dependientes de las fuentes de las que tomen datos así como su cobertura, por otro lado que un artículo sea muy citado no quiere decir que este sea muy influyente, se puede citar un trabajo para criticarlo. También hay que tener en cuenta las «auto citas» citas de un autor a otros trabajos suyos que pueden distorsionar estas medidas [7].

No se puede comparar categóricamente dos trabajos o autores de cualquier disciplina usándolas, ya que en distintos campos se cita de manera distinta.

En las medidas que dependan de citaciones (prácticamente todas) hay que considerar un factor temporal, para que un trabajo adquiera relevancia es necesario cierto tiempo, por ello se recomienda dejar fuera las publicaciones recientes (de los últimos tres años)[7].



### 1.3. Combinando ambas disciplinas

En los últimos años han surgido diversos trabajos que proponen combinar ambas disciplinas, utilizar la bibliometría para mejorar la recuperación de información. Esto no se puede aplicar de manera genérica y directa a cualquier sistema de RI ya que la bibliometría está muy centrada en el mundo de la académico y de la investigación, sin embargo con algunas adaptaciones las ideas de la bibliometría se encuentran muy presentes en los sistemas de RI, por ejemplo el algoritmo PageRank hace una analogía entre páginas web con artículos científicos así como hiperenlaces con citas en los mismos [3].

#### 1.3.1. Sistemas de RI actuales con medidas bibliométricas

Donde resulta obvio que esta combinación puede ser beneficiosa es los sistemas RI especializados en literatura académica siendo los más importantes:

- Google Scholar: buscador gratuito especializado en literatura científica, contiene algunas medidas bibliométricas con el número de citas o el índice-h de un autor. Sin embargo ha sido muy criticado por intentar incluir el mayor número de artículos posibles sin importar la calidad de los mismos [12] o dar demasiada importancia al número de citas lo que hace que sea complicado descubrir nuevos trabajos que no son muy conocidos [13].
- Web of Science (WoS): Sistema de RI por suscripción más utilizado, habitualmente las instituciones académicas lo tienen contratado, como es el caso de la Universidad de Granada a través de la Fundación Española para la Ciencia y la Tecnología (FECYT). Agrupa múltiples bases de datos de diversas índoles llegando a tener más de 100 millones de documentos [14]. Fue uno de los primeros sistemas en aparecer y actualmente pertenece a Clarivate Analytics. A diferencia de Google Scholar los contenidos son revisados por expertos antes de ser incluidos [14].
- Scopus: Similar a WoS, en este caso pertenece a la editorial Elsevier. También se puede usar desde la UGR gracias a la FECYT. Este sistema es más reciente, cuenta con unos 70 millones de documentos [15] y también tiene un proceso de revisión de contenido. Dispone de algunas otras medidas de bibliometría además del número de citas o índice-h, como los otros sistemas descritos, cuenta con las medidas *CiteScore*, *SCImago Journal Rank* y *SNIP* (ver 1.2.2 para más información).

### 1.3.2. Trabajos relacionados

Desde un punto de vista académico resultan especialmente interesantes los trabajos de los talleres *Bibliometric-enhanced Information Retrieval* (BIR). Estos congresos se celebran anualmente desde 2014 y tienen como objetivo "hacer consciente a la comunidad de RI de sus posibles vínculos con la bibliometría" [16]. Voy a destacar algunos de ellos a continuación.

Ante el creciente volumen de trabajos científicos se hace complicado para investigador mantenerse al día en su propio campo porque resulta imposible leer todos los trabajos, por ello se suele utilizar algún sistema de recomendación 3. Pero los sistemas existentes no contemplan bien el uso de medidas bibliométricas para distinguir los trabajos relevantes, por ello en [17] se propone reordenar la salida del sistema de recomendación Mr.DLib utilizando como criterio varias medidas derivadas del número de lectores de un artículo, un tipo de *altmetric* (ver 1.2.2). Las conclusiones de este trabajo demuestran que se encuentra una mejora utilizando las medidas de cienciometría solas o en combinación con el ordenamiento normal basado en texto que usando solo el ordenamiento del sistema RI, siendo la métrica que mejor resultado obtuvo el número absoluto de lecturas sin normalización. Esta mejora no es muy significativa como cabría esperar, los autores achacan esto a la falta de cobertura de datos bibliométricos en la colección.

Otro enfoque interesante es el de [18], donde se propone utilizar medidas bibliométricas como variables independientes de consulta o *static features*, ver el modelo probabilístico 1.1.4. Estas características se conocen como priores en el modelo probabilístico de lenguaje utilizado y modifican la probabilidad de que un documento sea relevante para una consulta dada, multiplicando dicha probabilidad por un factor. Para realizar su estudio han utilizado la colección bibliográfica de prueba iSearch, que desgraciadamente parece no estar disponible ya. De esta colección seleccionaron el subconjunto de artículos con al menos alguna cita dentro de la colección (para poder llevar a cabo un análisis de co-citación).

Como priores han seleccionado la proporción de citas dentro del subconjunto de documentos entre el número total de estas, el *PageRank* calculado en el subconjunto y el clustering de co-citación <sup>1</sup>, siendo todas las varian-tes también suavizadas con una versión logarítmica. Aunque el concepto es interesante y el método exhaustivo los resultados obtenidos no son significativos, se achaca esto al bajo tamaño de la colección de prueba, con 863 documentos.

Siguiendo un modelo vectorial pero aplicando una variación del conocido

---

<sup>1</sup>creando un grafo no dirigido con peso de citas donde los nodos son documentos y las aristas citas, sobre el que se aplica un algoritmo de clustering y se calculan los priores del cluster mediante validación cruzada de 5 iteraciones

esquema de pesos  $tf*idf$  se encuentra [19]. En el esquema de ponderación para los términos original  $tf$  representa la frecuencia de un término en el documento e  $idf$  el número de documentos donde aparece el término a la inversa. Con ello se consigue dar mayor importancia a los términos que aparezcan menor número de veces ya que serán más discriminantes e importantes para una búsqueda.

El modelo propuesto en este paper apuesta por buscar documentos en función de otros, es decir los términos de consulta son otros documentos de la colección, los cuales se conocen como semillas. Para la ponderación de peso en cada documento ahora  $tf$  es el número de documentos co-citados con el documento semilla e  $idf$  la inversa del número total de citas entre documentos de la colección. Por último cita algunos ejemplos sobre colecciones de prueba y discute el modelo planteado sin llegar a evaluarlo realmente. De este, se dice que sería especialmente útil para usuarios con conocimiento previo del dominio, por ejemplo para llevar acabo una investigación bibliográfica sobre algún autor.

Basado parcialmente en el trabajo previo, el siguiente artículo [20] propone la creación de un *framework* 4 para llevar a cabo una investigación bibliográfica siguiendo un enfoque híbrido combinando información textual y de citación. A partir de la selección de algunos documentos semilla, el sistema crea el espacio de citas <sup>2</sup>, sobre el cual se filtra poniendo condiciones, el resultado de este filtrado se refina buscando términos y frases comunes con las semillas en el resumen de los documentos.

Utilizan su sistema para comparar con diversos trabajos de revisión bibliográfica realizados manualmente con el objetivo de lograr obtener el mismo conjunto final de documentos relevantes revisando menos trabajos con lo que se ahorraría un tiempo significativo. Sus resultados resultan prometedores, usando todas las combinaciones de 3 documentos semilla entre los seleccionados finalmente por cada revisión logran recuperar todos los documentos finales disponibles en Scopus (plataforma en la que realizan el estudio) reduciendo el número de documentos totales recuperados para esa revisión en hasta el 80 % dependiendo de la revisión concreta.

## 1.4. Enfoque planteado

Como se ha podido comprobar en el apartado previo existen numerosas propuestas para combinar estas la bibliometría y la RI, pero estos trabajos se realizan en un ámbito más académico y muchas veces no llegan a materializarse en sistemas reales. Los principales sistemas de RI para la recuperación

---

<sup>2</sup>Documentos citados por las semillas, aquellos que las citan a ellas y documentos con relación de co-citación con estas

de literatura académica integran algunas medidas básicas pero no llegan a implementar los modelos más complejos planteados desde la investigación a pesar de que sus resultados sean esperanzadores.

Esto hace que este campo siga estando en desarrollo y resulte interesante plantear nuevos modelos a la par que testarlos.

Por ello este TFM servirá para desarrollar un prototipo de modelo que combine ambas ramas de la teoría de la información y sirva para evaluar el rendimiento de estos planteamientos. El objetivo es comparar un sistema de RI clásico con uno que incorpore técnicas bibliométricas intentando medir su viabilidad y potencial mejora en los resultados recuperados.

En los últimos tiempos el acceso a artículos se ha incrementado exponencialmente gracias a algunas de las plataformas descritas previamente, a su vez la información bibliográfica que incorporan también ha aumentado sustancialmente lo que hace cada vez más plausible la implementación de este tipo de sistemas. La irrupción de las *altmetrics* resulta destacable, a pesar de ser unas medidas bibliométricas bastante recientes su capacidad para determinar la popularidad de los trabajos científicos es muy significativa.

Es especialmente interesante el enfoque de las técnicas híbridas como las planteadas en el último trabajo que permiten por un lado no divergir del modelo mental de sistema de RI del usuario, un buscador textual, pero incluir las potenciales ventajas de usar otro tipo de medidas.

El sistema propuesto utilizará como base un modelo clásico realizando una reordenación *a priori* de los resultados en función de algunas medidas directas como el número de citas y el número de lectores en una plataforma (incluyendo con ello alguna *altmetric*). Se plantea utilizar una realimentación inconsciente del usuario basada en considerar como relevantes los documentos del listado inicial que el usuario descargue tras leer el resumen y utilizándolos como semillas para refinar la búsqueda de manera transparente para él. Esta reordenación *a posteriori* se servirá del grafo de citación de estos documentos semilla para seleccionar los que documentos que tengan una relación fuerte.

## Capítulo 2

# Objetivos

En este capítulo recogeré de manera sintetizada los objetivos del proyecto, lo cual ayudará a comprender la funcionalidad del sistema a desarrollar así como definir su alcance.

- **Obtención de artículos o *papers* de autores de la escuela:** como colección de datos del sistema RI a desarrollar se ha decidido utilizar el conjunto de todos los trabajos de autores de la Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación, ya que la familiaridad de los datos permite analizar si los resultados obtenidos tienen sentido. Esto puede servir para comprender relaciones como que un par de autores del mismo grupo de investigación o departamento aparezcan juntos en los resultados, muy relacionado con esto se encuentra el siguiente objetivo.
- **Evaluar si la mejora producida al aplicar medidas bibliométricas al sistema:** El objetivo anterior favorece que esta evaluación se pueda llevar a cabo, aunque esta no es una tarea sencilla. La evaluación de sistemas RI se suele llevar a cabo mediante colecciones de prueba donde existe un conjunto de documentos, un conjunto de consultas y un conjunto de resultados que deberían de retornar las mismas [21]. Al crear un sistema nuevo con una colección de documentos no utilizada no podemos comparar los resultados obtenidos con los que se esperarían, este problema se debe a la relatividad del concepto de relevancia. Por ello las colecciones de pruebas suelen incluir valoraciones de relevancia de expertos en la materia.
- **Desarrollar un sistema que sea usable:** Muchos de los enfoques que he visto durante mi proceso de documentación no pasan de modelos teóricos, prototipos o sistemas en los que la usabilidad y la orientación al usuario brillan por su ausencia. Aunque este no sea el punto

objetivo principal del sistema me parece muy importante que se tenga en cuenta al usuario durante todo el proceso de desarrollo, ya que un sistema puede ser increíble pero si los usuarios no lo entienden o no le saben sacar partido no sirve de nada. Por ello pretendo diseñar una interfaz de usuario que sea simple de usar y emplee metáforas y componentes ampliamente conocidos por los usuarios.

- **Conocer y emplear tecnologías punteras en el ámbito de la RI:** durante la asignatura del máster Gestión de Información en la Web (GIW) se vieron algunas pinceladas de las herramientas empleadas para llevar a cabo estos sistemas así como su base teórica. Esto me llamo realmente la atención ya que todos utilizamos sistemas de búsqueda, pero no tenía ni idea de como se podía implementar uno. A pesar de haber desarrollado un pequeño sistema como parte de sus prácticas me quedé con la ganas de ver herramientas más potentes. Este es el principal motivo por el que me decanté a la realización de este proyecto, a nivel personal me gustaría cumplir con creces este objetivo.

## Capítulo 3

# Planificación

### 3.1. Gestión de recursos

En este apartado describiré brevemente los recursos utilizados para llevar a cabo este proyecto teniendo en cuenta personal, hardware y software.

#### 3.1.1. Personal

El único recurso humano que ha contribuido a la realización del proyecto soy yo mismo actuando como los diversos roles que llevan a cabo el proceso de desarrollo de un proyecto de estas características.

#### 3.1.2. Hardware

Respecto al hardware utilizado para este TFM solo he requerido mi ordenador portátil personal, cuyas características se destacan en la siguiente tabla:

CPU	Intel ®Core™i7-4700MQ CPU @ 2.40GHz x 8
RAM	8 GB RAM DDR3
Almacenamiento	HDD 750 GB (5400 RPM)

Cuadro 3.1: Especificaciones del equipo utilizado

Esto ha bastado para el desarrollo, pero sería necesario conseguir un servidor en condiciones para llegar a poner en producción el sistema. Tampoco sería necesario nada muy potente ya que incluso en mi propio ordenador los tiempos empleados en la búsqueda son bastante aceptables.

### 3.1.3. Software

Se han empleado utilidades de software libre en la totalidad del proyecto. Aquí enumerare el principal software empleado y su función primordial, para una descripción más detalladas de las herramientas empleadas ver el capítulo Técnicas y herramientas [Aún en progreso]:

- **Debian:** Sistema Operativo (SO).
- **Python:** Lenguaje de programación usado en las primeras fases del proyecto y en servidor *backend* 5.
- **JavaScript:** Lenguaje de programación interpretado en el que se ha escrito el *frontend* 6.
- **Elasticsearch:** Servidor de búsqueda.
- **MongoDB:** Base de datos NoSQL.
- **React:** *Framework* para el desarrollo de interfaces de usuario.
- **Searchkit:** *Framework* que incluye un conjunto de componentes React para la comunicación con Elasticsearch.
- **TeXstudio:** Entorno integrado de escritura en  $\text{\LaTeX}$  utilizado para generación de la documentación.
- **Docker:** Software de virtualización para basado en contenedores. Permite gestionar de forma simple la gestión y despliegue de una infraestructura software.
- **Visual Studio Code:** Editor de código creado por Microsoft utilizado para toda la programación del proyecto.

## 3.2. Metodología

Para desarrollar este proyecto se ha empleado una metodología ágil similar a *SCRUM* 7 algo más relajada. Es una modelo particular ya que yo mismo soy el desarrollador, el coordinador (rol del *Scrum master*) y la persona encargada de definir las tareas y evaluar el cumplimiento de los mismas (el *Product owner*).

Me he decantado por este modelo ya que permite más flexibilidad al enfrentarse a problemas en entornos desconocidos, como es este proyecto, y he tenido buena experiencia con él tanto en mi TFG como durante mi vida laboral. Se basa en la descomposición del proyecto completo en pequeños



subproyectos o *sprints*, en los que define de manera acotada las tareas y objetivo del mismo. Permitiendo el refinamiento iterativo del producto final así como el aprovechamiento del conocimiento obtenido a lo largo de los sprint para mejorar los venideros, al contrario que otros modelos de desarrollo más clásicos que resultan más estáticos y rígidos.

Con el objetivo de almacenar y versionar todo el material producido en este proyecto he utilizado la plataforma *GitHub* y el sistema de control de versiones *git*.

Para seguir el progreso de cada uno de los *sprints* del proyecto he utilizado los tableros que ofrece el propio *GitHub projects* donde cada una de sus tareas o tarjetas corresponden con *issues* como se puede ver en la siguiente imagen.

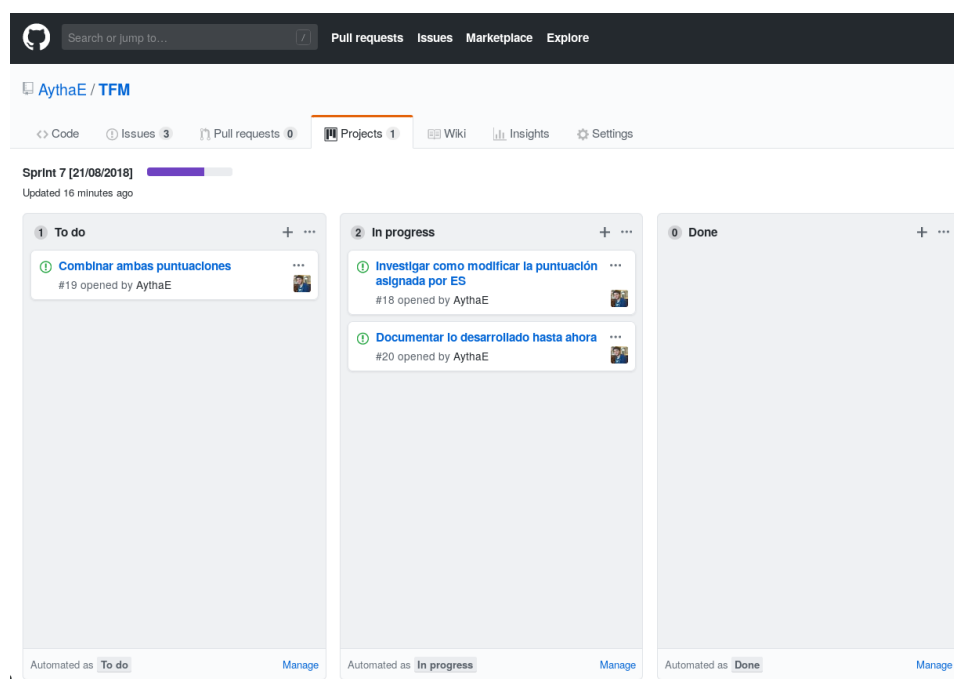


Figura 3.1: Ejemplo de tablero de un *sprint*

Además de esto he utilizado un archivo *Markdown* a modo de diario donde ir apuntando cosas interesantes según iban surgiendo, el estado actual de desarrollo o algunas tareas para completar próximamente, dicho fichero se llama *Diario.md*.

### 3.3. Planificación temporal

Desde la asignación del TFM, en diciembre de 2017, me percaté que iba a ser realmente complicado alcanzar la primera convocatoria con la carga de trabajo que suponía el máster. Por lo que decidí tomarlo con calma y llegar a septiembre, pero tras todo finalizar el curso comencé la realización de mis prácticas, lo cual unido a lo extenuado que había acabado el año me impidió alcanzar otra vez el objetivo.

En Octubre de 2017 comienzo a trabajar, lo cual suponen muchas horas menos al día y me llevo un tiempo adaptarme por lo que no fue hasta Febrero de 2018 cuando me lo empecé a tomar en serio. Teniendo en cuenta mi limitada disponibilidad horaria que apenas me permitía dedicarle 1-2 horas entre diario esboqué una planificación con el objeto de entregar el proyecto en Julio de 2018, dicha planificación inicial se recoge en la siguiente tabla.

<b>Tarea</b>	<b>Inicio</b>	<b>Fin</b>	<b>Duración</b>
Investigación	19/02/2018	23/04/2018	8 semanas
Obtención de datos	23/04/2018	07/05/2018	2 semanas
Procesado de datos	07/05/2018	21/05/2018	2 semanas
Búsqueda básica	21/05/2018	04/06/2018	2 semanas
Búsqueda con bibliometría	04/06/2018	02/07/2018	4 semanas
Refinamiento	02/07/2018	09/07/2018	1 semana

Cuadro 3.2: Planificación inicial de tareas

Desgraciadamente las vacaciones que he tomado entre medias junto con algunos imprevistos no me permitieron llegar a tiempo aunque ya tenía el proyecto encaminado.

## Capítulo 4

# Diseño

Al ser este un sistema RI el diseño ha de girar en torno a los datos, por ello comencé explorando diversas alternativas para utilizar como fuente de datos.

Como punto inicial para la extracción de los datos de los autores de la E.T.S.I.I.T. partí del Ranking UGRinvestiga [22]. En este se lista a todos los autores de la facultad junto con sus citas e índice h total y los últimos 5 años.

Sin embargo no hay modo de obtener datos de los artículos concretos por lo que barajé múltiples bases de datos bibliográficas, además de las ya comentadas en la sección Sistemas de RI actuales con medidas bibliométricas también exploré Research Gate.

Los resultados de mi análisis fueron los siguientes:

- Google Scholar y Research Gate parecen tener bastantes datos en comparación a otras opciones, sin embargo no ofrecen una API por lo que la extracción de datos solo se podría llevar a cabo mediante *web scraping* 8. Lo cual no es un proceso sencillo ya que muchas de estas web contienen mecanismos de detección de bots como los CAPTCHAs [23] y además es un método muy sensible al cambio, en caso de que produzca alguna modificación en la web en cuestión es probable que toque reescribir el método de extracción de datos.
- WoS y Scopus sí disponen de una API, tras realizar algunas pruebas y leer opiniones sobre ambas me decanté por utilizar Scopus como fuente de datos ya que su API me resultó más cómoda de utilizar y encontré datos fácilmente de algunos autores de la E.T.S.I.I.T. (cosa que me costó bastante más en WoS).

## 4.1. Modelo de datos

Siguiendo el modelo de Scopus decidí separar la búsqueda de autores y la de artículos ya que son dos tipos de datos muy diferenciados e intuitivamente se comprende que un sistema de RI funcionará mejor si los datos son homogéneos.

Por tanto mi modelo de datos está constituido por 2 entidades distintas pero relacionadas: Autores (*Author*) y Artículos (*Abstract*).

En las siguientes secciones se describen brevemente ambas entidades.

### 4.1.1. *Author*

Esta entidad es resultado de la combinación de la información disponible en el Ranking UGRinvestiga [22] y en Scopus, por eso podemos ver en el siguiente diagrama con el prefijo **ugr** las medidas de dicho ranking y con el de **scopus** los de la plataforma homónima.

Respecto a los atributos que finalizan en 5 estos hacen referencia a los valores de dichas medidas de los últimos cinco años mientras que este sufijo no aparece las medidas son totales.

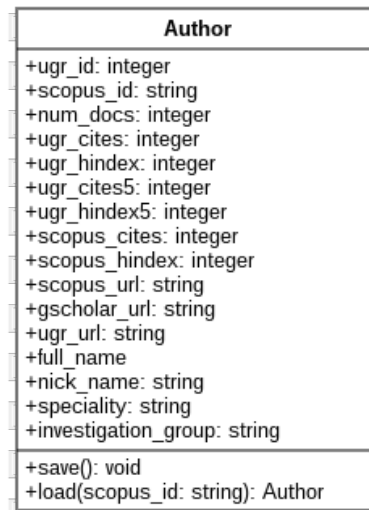


Figura 4.1: Clase *Author*

#### 4.1.2. *Abstract*

La entidad *Abstract* corresponde casi directamente con los datos obtenidos a través de la API de Scopus, a excepción de:

- Las referencias han sido limitadas a citas internas, es decir entre los artículos que están disponibles en la colección, con la idea de usar esta información en el proceso de búsqueda.
- Los autores, que únicamente aparecían listados con su `scopus_id`, han sido enriquecidos con el nombre de los mismos así como el `ugr_id` en caso de estar disponible este último (este identificador estará disponible si el autor es alguno de los que forman parte de la colección de autores).

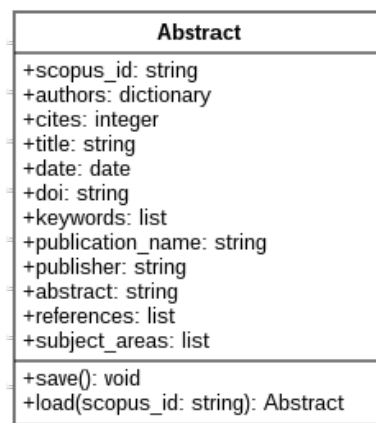


Figura 4.2: Clase *Abstract*

## 4.2. Arquitectura del sistema

La arquitectura final del sistema es la de una aplicación web distribuida:

- Un **cliente web** que se ejecuta en el navegador y constituye la interfaz de usuario, este se encarga de recoger las consultas, enviarlas al servidor de *Elasticsearch* (ES) y mostrar los resultados de las mismas sirviéndose de tecnologías como *Searchkit*, *ReactJS* o *MaterialUI*.
- Un **servidor Elastic** donde se ejecuta ES y el cual contiene los índices de autores y *abstracts* así como los datos indexados para la búsqueda (que no son todos los listados en el modelo de datos previamente).
- Un **servidor de monitorización** del servidor Elastic utilizando el proyecto de software libre Cerebro. Este permite monitorizar en tiempo real diversas medidas del servidor como la carga, espacio utilizado o estado además de múltiples funciones como un cliente REST que permite la consulta a bajo nivel sobre ES muy utilizado para depuración o durante el desarrollo para hacer pequeñas pruebas.
- Un **servidor de aplicación** que utiliza *Flask* para montar algunos microservicios auxiliares que permiten recuperar el resto del modelo de datos que no se encuentra indexado así como aportar algunas estructuras de datos precalculadas para facilitar la búsqueda.
- Un **servidor mongo** que lleva una BD *MongoDB* la cual contiene la totalidad de los datos de ambas colecciones y sirve como fuente de datos para el servidor de aplicación.

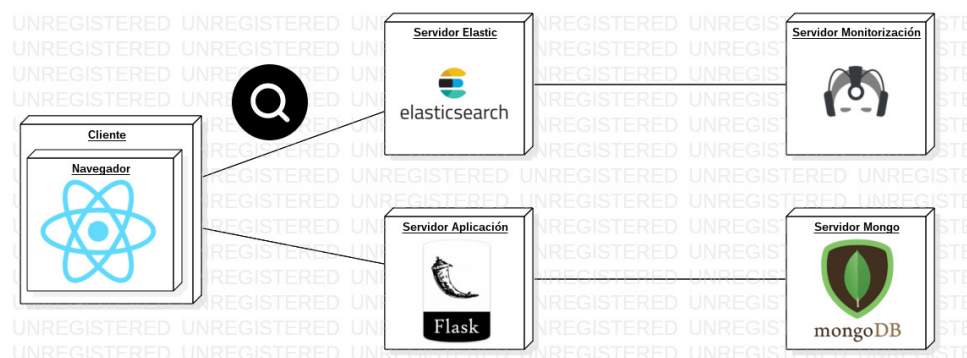


Figura 4.3: Diagrama de arquitectura final con las principales tecnologías usadas en cada nodo

## Capítulo 5

# Técnicas y herramientas [Aún en progreso]

Comentar en más detalle todas las herramientas utilizadas con enlaces de cada una.

- **Debian:** Sistema Operativo (SO).
- **Python:** Lenguaje de programación usado en las primeras fases del proyecto y en servidor *backend* 5.
- **JavaScript:** Lenguaje de programación interpretado en el que se ha escrito el *frontend* 6.
- **Elasticsearch:** Servidor de búsqueda.
- **MongoDB:** Base de datos NoSQL.
- **React:** *Framework* para el desarrollo de interfaces de usuario.
- **Searchkit:** *Framework* que incluye un conjunto de componentes React para la comunicación con Elasticsearch.
- **TeXstudio:** Entorno integrado de escritura en  $\text{\LaTeX}$  utilizado para generación de la documentación.
- **Docker:** Software de virtualización para basado en contenedores. Permite gestionar de forma simple la gestión y despliegue de una infraestructura software.
- **Visual Studio Code:** Editor de código creado por Microsoft utilizado para toda la programación del proyecto.
- Git

- GitHub
- Pandas
- PyMongo
- scopus-api
- Beautiful Soup
- Matplotlib
- MaterialUI
- elasticsearch-py
- Star UML
- Cerebro
- Flask
- Gimp



## Capítulo 6

# Desarrollo

En este capítulo se detallara el proceso de desarrollo seguido dividido en *sprints* como ya se ha comentado en la sección Metodología.

### 6.1. *Sprint* 1

Este primer *sprint* tiene como objetivo general la investigación básica sobre la RI. Sus objetivos concretos son:

- Leer el libro Recuperación de Información: un enfoque práctico y multidisciplinar [2].
- Buscar los talleres BIR centrándose en sus editoriales para realizar un listado priorizado por interés de los diversos artículos.

Siguiendo la recomendación de mi tutor (co-autor del libro) me centré en los capítulos "*1 Introducción a la recuperación de información*", "*2 Indexación de documentos y procesamiento de consultas*", "*3 Modelos de recuperación de información clásicos*" y "*10 Técnicas de modificación de la consulta*".

Esta lectura me hizo adquirir unas bases más teóricas a lo que ya había estudiado en la asignatura del máster Gestión de Información en la Web donde obtuvimos una nociones básicas de lo que supone la RI y sus vertientes realizando alguna práctica.

Los artículos de los talleres *Bibliometric-enhanced Information Retrieval* que encontré más destacados están detallados en el apartado Trabajos relacionados de la Introducción. Dichos trabajos se encuentran disponibles gratuitamente en <http://ceur-ws.org/> bajo el amparo de la Universidad Técnica de Aquisgrán (*RWTH Aachen University*) en Alemania.

Cada una de las ediciones de estos talleres se encuentra estructurada dividida en diversos trabajos. Por un lado un editorial que resume la edición y todos los trabajos aceptados en la conferencia así como los trabajos individuales, alguna *keynote* o presentación y algunas demostraciones.

En este *sprint* me dediqué a leer dichos editoriales clasificando por aparente interés los artículos de cada BIR generando una lista priorizada utilizada como orden en el que estudiar los trabajos. En la siguiente imagen se puede apreciar el aspecto de dicha lista generada en *Markdown*.

### Lista priorizada de papers

- ☑ **BIR2014\_1**: Meaning of citations, use of words in standard parts. Find interesting distribution of words
- ☑ **BIR2014\_5**: iSearch analysis. Extract connection between citation based and topic relevance rankings checking if this collection is suitable to perform these kind of analysis
- ☑ **BIR2015\_1**: Potentiality of interdisciplinary research between IR and bibliometrics. Ask several research questions summarizing other investigations in between of this disciplines.
- ☑ **BIR2015\_4**: bibliometric-enhanced search features based on Marcia Bates model with 2 specific examples.
- ☑ **BIR2014\_2**: use citations a mechanism for systematic IR of scientific literature. Helpful in situations in which one needs a comprehensive overview of a research topic.
- ☑ **BIR2014\_4**: Use citation and co-citation as previous probabilities for relevancy. Measures: citation count, PageRank & co-citation clustering. Very query-dependent, suggestion to improve using this citation info & content features
- ☑ **BIR2017\_11**: how to improve scientific recommender systems using bibliometry Measures.
- ☑ **BIR2016\_7**: Alternative to the bag of words model (TF\*IDF) used to find similar documents to a giving one based on citations. The tf and idf measures re-defined using citation information.
- ☑ **BIR2015\_5**: Combining bibliographics and IR studying how polyrepresentation could be used to improve interactive search. Alternatives to ranked lists using bibliographics and contextual information. Based in clustering.

Figura 6.1: Fragmento de la lista priorizada creada

## 6.2. *Sprint 2*

Para el segundo *sprint* planteo el indagar en la RI e ir introduciendo la bibliometría. Sus objetivos concretos son:

- Leer los primeros 17 papers de los BIR resumiendo y extrayendo ideas interesantes para el proyecto.
- Buscar información sobre medidas bibliométricas (Citas, índice h combinado...)

A partir de la lista priorizada fui leyendo los primeros artículos, el orden de esta lista lo fui alterando ya que con frecuencia al indagar en el trabajo este perdía o incrementaba su interés. Con el objetivo de poder aprovechar más estas lecturas fui creando una especie de resúmenes en los que anotaba los puntos más importantes que se trataban en el artículo, otros artículos

relacionados de los que se hablaba o los resultados obtenidos con su trabajo. En la siguiente imagen se puede apreciar uno de estos resúmenes”:

## BIR2014\_5

### "On the Connection Between Citation-based and Topical Relevance Ranking: Results of a Pretest using iSearch"

Zeljko Carevic and Philipp Schaer

- **IR limitations.** Buckley, C.: Why current IR engines fail. *Inf. Retr.* 12, 6, 652–665 (2009).
- **Science Citation Index increase performance in scholarly IR systems.** Pao, M.L.: Term and citation retrieval: A field study. *Inf. Process. Manag.* 29, 1, 95–112 (1993).
- **iSearch test collection.** Lykke, M. et al.: Developing a Test Collection for the Evaluation of Integrated Search. In: Gurrin, C. et al. (eds.) *Advances in Information Retrieval*. pp. 627–630 Springer, Berlin, Heidelberg (2010).
- **Novel IR approach using Bibliometric & IR with a tf\*idf schema.** White, H.: Some new tests of relevance theory in information science. *Scientometrics.* 83,3, 653–667 (2010).
- **TF\*IDF schema using co-citation as measure of similarity:**
  - TF: #citations of the seed document together with the target document to analyze
  - DF: #overall citations for the target document in all iSearch collection
  - Method:
    - a. Given seed document A retrieve all documents that cite A
    - b. List of candidates created taken all citations from the documents in step 1
- **Result:** Using only internal references and relevant documents retrieve not enough documents to perform an analysis.
- **Further works:** Expand co-citation analysis using title, authors, journal and publication year along with the cites itself

Figura 6.2: Ejemplo de resumen de uno de los papers

En este momento me empecé a introducir en el mundo de las medidas bibliométricas, ya que no tenía conocimiento previo alguno de que existiera esta disciplina siquiera, mediante la lectura de artículos iba descubriendo distintas medidas así como sus posibles aplicaciones a la RI cuyo resultado final se encuentra sintetizado en la sección Bibliometría de la Introducción.

## 6.3. *Sprint 3*

Ya en este punto decidí que estaba listo para ir documentando todo lo que había aprendido por ello en este *sprint* me puse como objetivo escribir la introducción de este TFM. Los objetivos concretos son:

- Escribir la introducción del TFM.
- Pensar un enfoque para el proyecto a desarrollar decidiendo que clase de sistema se desarrollaría.
- Continuar leyendo algunos artículos más de los BIR.

Para asimilar y reflexionar sobre todo lo que había leído me puse a escribir la introducción de este trabajo ya que ello me ayudaría a pensar

un enfoque correcto. También quería poder mostrar algo a mi tutor para obtener algo de *feedback* por su parte.

Continué leyendo algunos trabajos más con lo que dí por concluida mi proceso de investigación. Uno de esos últimos trabajos fue [20] el cual me gustó especialmente ya que pasaban de un modelo teórico a algo más práctico, accediendo a la API de Scopus e incluyendo ejemplos de su implementación en un repositorio de *GitHub* lo que me llevo a plantear mi modelo híbrido combinando el sistema habitual de un sistema RI con reordenamiento de resultados *a priori* usando medidas bibliométricas y un ordenamiento *a posteriori* utilizando un grafo de citación entre los documentos.

Desgraciadamente estos enfoques dependen ampliamente de la cobertura de medidas bibliométricas disponibles, por ejemplo me hubiera encantado poder probar una reordenación previa utilizando alguna *altmetric* como el número de lecturas o descargas de un artículo, pero la plataforma que he utilizado para extraer los artículos no dispone de dichas medidas.

## 6.4. *Sprint 4*

Una vez sintetizado lo investigado y teniendo una idea aproximada de lo que pensaba desarrollar, en este *sprint* me dispuse a realizar diversas pruebas que fueran definiendo las tecnologías a emplear, en concreto:

- Buscar soluciones para montar sistemas RI.
- Investigar como conectarse a la API de Scopus o WoS.

Estuve haciendo alguna pruebas con algunos *frameworks* de búsqueda. Como he dicho previamente ya había utilizado *Lucene* como parte de las prácticas de la asignatura GIW, pero me pareció demasiado bajo nivel así que me centré en investigar otras alternativas. Encontré que las principales, que casualmente utilizaban por debajo *Lucene*, eran ***Solr*** y ***Elasticsearch***.

Buscando alguna comparativa [24] y comentarios de usuarios en plataformas tan reputadas como *StackOverflow* [25] parecía que se recomendaba ES por ser más sencillo de usar por lo que realicé un breve tutorial [26] que me gustó bastante ya que resulta realmente simple de usar y basta únicamente con realizar consultas sobre una API REST por lo que basta con hacer peticiones HTTP con algún cliente simple como *curl*. Todo esto hizo que me decidiera por este servidor de búsqueda.

Como ya comenté en el apartado 4 realicé un análisis de diversas fuentes de datos y me decanté por seleccionar alguna que dispusiera de API, en concreto Scopus ya que me resultó más rápido encontrar artículos que me servirían para mi colección documental. Por lo que comencé buscando la

implementación que habían realizado en el artículo [20], ya que en el mismo comentan que se encuentra disponible en *GitHub* [27]. Vi que dicha implementación era de muy bajo nivel, extraía los artículos mediante peticiones HTTP y parseaban el XML obtenido directamente. Pensé que tenía que haber un mejor método para esto, alguna librería que ya implementara los métodos de la API y abstrayera de esas tareas.

Así encontré la implementación oficial ofrecida por Elsevier, `elsapy` [28], un modulo en *python* que cuenta con diversas clases para modelar los tipos de documentos y permite obtener y parsear los datos de la API de manera transparente. También encontré el módulo `scopus` [29] una implementación alternativa que había comenzado a desarrollarse algo antes que la oficial, parecía más activa y contaba con mejor documentación incluyendo ejemplos bastante útiles.

Funcionalmente `scopus` era muy similar a la versión oficial, pero estaba mejor modelada, se centraba solo en Scopus, que era lo que yo necesitaba, y tenía algunos añadidos muy útiles como uso de una caché por defecto con el objetivo de no repetir peticiones ya hechas, si no servirlos desde un documento almacenado en disco, lo que hacía increíblemente veloz la recuperación de entidades ya consultadas previamente así como gestión transparente de API *key*, basta con introducirla en la primera ejecución y la propia librería la guarda en un fichero interno el cual usa para su consulta el resto de ocasiones. Todo esto hizo que me decantara por esta última sobre la implementación oficial.

Para poder realizar mis pruebas tuve que solicitar una API *key* en el portal de Elsevier developers <https://dev.elsevier.com/myapikey.html> creando una cuenta gratuita. Para poder descargar los datos de Scopus desde fuera de la red de la universidad me ha sido necesario utilizar el servicio de VPN de la UGR siguiendo el siguiente tutorial [30].

## 6.5. *Sprint 5*

Tras toda la documentación e investigación previas en este punto realicé el proceso de obtención de datos. Dicho proceso fue dividido en

- Extraer información de los autores del ranking UGRinvestiga.
- Usar esos autores obtenidos para buscarlos en Scopus.
- Descargar todos los *abstracts* de los autores obtenidos en el punto previo.
- Preprocesar y limpiar los datos obtenidos.

### 6.5.1. Obtención de datos

Teniendo en cuenta que no parecía haber opción de descarga de los datos del Ranking UGRinvestiga y que el formato parecía bastante estático con todos los datos en una tabla mi primer enfoque fue realizar un *web scrapping* utilizando la librería BeautifulSoup ya comentada previamente. El aspecto de la tabla se puede ver en la siguiente imagen.

<b>Tipo de Ranking:</b> Disciplina <b>Periodo:</b> Histórico					
 Ingenierías Informática y de Telecomunicación					
Rank	Nombre	Citas	h-index	Perfil	Grupo
1	FRANCISCO HERRERA TRIGUERO	61092	123	Perfil	TIC186
2	ENRIQUE HERRERA VIEDMA	25156	76	Perfil	TIC186
3	JOSE LUIS VERDEGAY GALDEANO	12607	48	Perfil	TIC169
4	OSCAR CORDON GARCIA	12027	52	Perfil	TIC186
5	SALVADOR GARCIA LOPEZ	11163	37	Perfil	TIC186
6	MIGUEL DELGADO CALVO-FLORES	9288	47	Perfil	TIC111
7	MARIA AMPARO VILA MIRANDA	8639	45	Perfil	TIC174
8	MANUEL LOZANO MARQUEZ	6806	31	Perfil	TIC186
9	ALBERTO LUIS FERNÁNDEZ HILARIO	6725	32	Perfil	TIC186
10	JUAN JULIAN MERELO GUERVOS	6464	33	Perfil	TIC024

Figura 6.3: Aspecto tabla ranking UGRinvestiga

Realizando este proceso extraje datos de históricos y de los últimos 5 años de los autores de la facultad, además de aplicar este proceso al catálogo de grupos de investigación de Tecnologías de la Información y la Comunicación de la UGR para realizar un *macheo* por código de grupo con los datos del ranking.

Posteriormente dí con el proyecto OpenData de la Universidad de Granada el cual entre otros *datasets* ofrecía el mencionado ranking actualizado a fecha 28 de Mayo de 2018 [31] en formato Excel, fácilmente legible y con algunos datos adicionales como el departamento de los autores, la url de Google Scholar de cada autor o su *nick* el cual podía ser muy útil para buscarlos en Scopus.

Tras cargar estos últimos datos obtuve **214 autores** de la facultad. A continuación pasé al problema de buscar estos autores en Scopus, este problema no es nada trivial, ya que con frecuencia los autores no firman sus trabajos con su nombre completo, si no con parte de él o incluso de manera distinta entre unos trabajos y otros. Por ejemplo mi tutor, "JUAN MANUEL FERNÁNDEZ LUNA" suele firmar como Fernández Luna, Juan M. o como Fernández Luna, J.M.

El formato necesario para la búsqueda de autores en Scopus es nombre, apellidos y afiliación (la UGR en este caso) por lo tanto primero me dispuse a extraer el nombre y apellidos a partir de la cadena completa que disponía, esto puede parecer sencillo, pero la riqueza del lenguaje español hace que existan apellidos compuestos o personas con segundos nombres lo que dificulta la tarea.

Estos problemas con los nombres me hicieron montar el siguiente algoritmo para intentar buscar un autor en Scopus:

---

**Algorithm 1** Obtiene los autores de Scopus a partir del ranking UGR

---

```

for all author in author_list do
   $f\_name, l\_name \leftarrow get\_name(author.nick)$ 
   $success, auth\_result \leftarrow author\_query(f\_name, l\_name)$ 
  if not success then
     $f\_name, l\_name \leftarrow get\_name(author.full\_name)$ 
     $success, auth\_result \leftarrow author\_query(f\_name, l\_name)$ 
    if not success then
       $f\_name, l\_name \leftarrow get\_name(author.name\_wo\_1\_lname)$ 
       $success, auth\_result \leftarrow author\_query(f\_name, l\_name)$ 
      if not success then
         $f\_name, l\_name \leftarrow get\_name(author.name\_wo\_2\_lname)$ 
         $success, auth\_result \leftarrow author\_query(f\_name, l\_name)$ 
        if not success then
          return No results found
  return  $auth\_result$ 

```

---

Donde  $f\_name$  y  $l\_name$  hacen referencia a nombre y apellidos respectivamente y  $name\_wo\_2\_lname$  hace referencia al nombre completo del autor eliminando el segundo apellido.

Con este algoritmo he logrado encontrar algún resultado en Scopus para **202 de los 214** autores buscados ( $\sim 94.39\%$  de los autores), a pesar de esto el número total de autores recuperados son 397, por lo que habrá que realizar una limpieza de los mismos.

### Limpieza de datos de autores

Si observamos la distribución de **nacionalidades de los autores** recuperados, la cual se ve en la siguiente imagen se aprecia que hay una mayoría abrumadora de Españoles (como cabría esperar) pero también algunas nacionalidades inesperadas como Palestina o Australiana. Por ello consideraré que todo autor cuya nacionalidad no sea española debe haber sido recuperado incorrectamente y será eliminado al tratarse de ruido, haciendo esto pasamos de 397 autores de Scopus representando a 202 autores de la UGR

a un total de 381 autores de Scopus respresentando a **198 autores de la UGR** ( $\sim 98.02\%$  de los 202 previos).

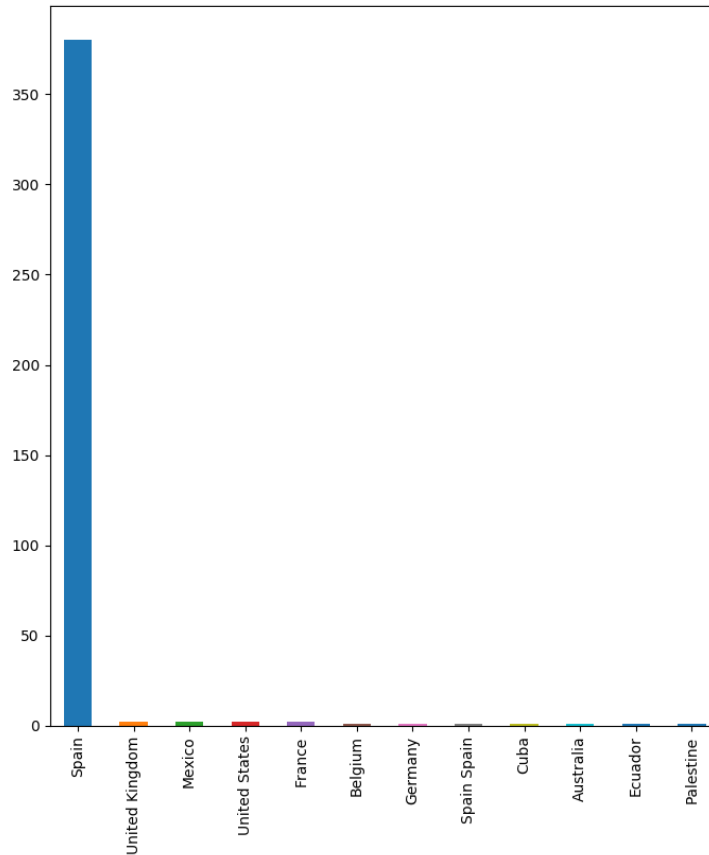


Figura 6.4: Distribución de nacionalidades entre los autores recuperados

Continuando el filtrado si mostramos la distribución de **ciudades de los autores** restantes (ver la siguiente imagen) vemos que Granada o sus variantes aglutinan a la gran mayoría de los autores, por ello filtraré todos aquellos que no sean Granada lo cual me deja con un total de 334 autores de Scopus (**188 autores de la UGR**  $\sim 94.95\%$  de los 198 previos).

Por último he realizado un filtrado por **área del conocimiento** descartando aquellos autores que no tengan al menos un trabajo englobado en el área de Ciencias de la Computación (*computer science*) con lo que finalmente me quedo con 199 autores de Scopus (**164 de la UGR**,  $\sim 87.23\%$  de los 188 previos)



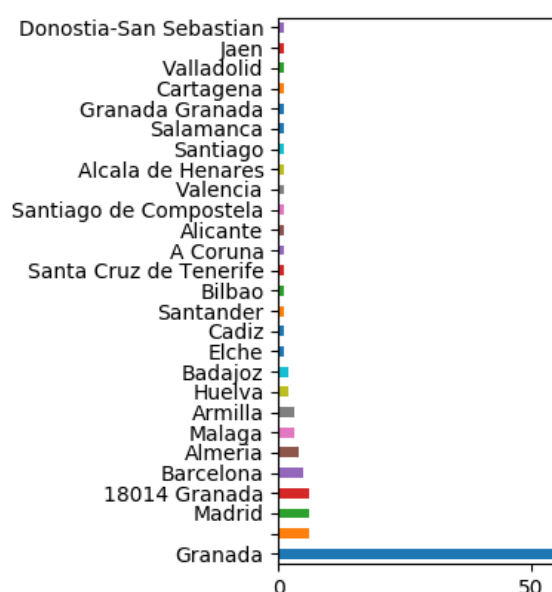


Figura 6.5: Fragmento de la distribución de ciudades entre los autores españoles recuperados

En la siguiente gráfica represento un *clustering de autores* UGR donde aparecen etiquetadas las siguientes clases: la clase *comp\_spa\_gr* representa a los autores que tienen algún autor en Scopus el cual es Español, de Granada y tiene algún artículo de Ciencias de la computación; la clase *spa\_gr* igual con excepción del área del conocimiento, la clase *spa* que tiene algún autor solo español y la clase *other* que no cumple ningún criterio. Por tanto cada clase se ha ido eliminando en cada filtrado hasta quedarnos solo con la primera, la cual representa al  $\sim 81.19\%$  de los autores totales de la UGR (164 de 202 autores posibles).

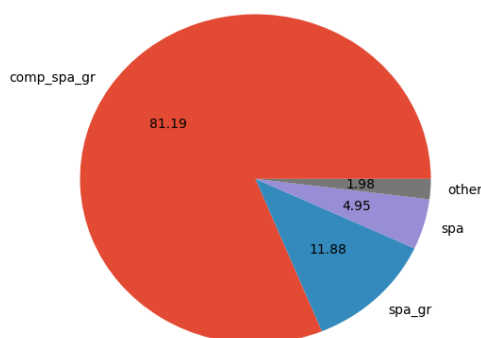


Figura 6.6: Representación del clustering de autores llevado a cabo en la limpieza de datos

### Obtención de abstracts

Una vez estuve contento con la limpieza de autores me dispuse a llevar a cabo el grueso de la obtención de datos, la descarga de todos sus *abstracts*

Como curiosidad de este *sprint* me gustaría destacar que durante mi investigación de los datos devueltos por Scopus sobre los *abstracts* me percaté que la librería `scopus` no parseaba las palabras clave de los mismos aunque estas aparecían en los XML obtenidos de la API, estas palabras clave podían funcionar bien en los sistema de búsqueda, como había leído durante mi proceso de documentación, por ello añadí esa funcionalidad a la librería y dicho cambio fue incorporado a la misma en su versión 0.9

### 6.6. *Sprint 6*

### 6.7. *Sprint 7*

### 6.8. *Sprint 8*

## Capítulo 7

# Conclusiones y Trabajos Futuros



# Bibliografía

- [1] J. M. F. Luna, “Apuntes de la asignatura Gestión de Información en la Web del Máster en Ingeniería Informática.”
- [2] J. F. H. G. Fidel Cacheda Seijo, Juan Manuel Fernández Luna, *Recuperación de Información: Un enfoque práctico y multidisciplinar*. 1ª ed.
- [3] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web.,” Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.
- [4] statcounter, “Search engine market share worldwide.” <http://gs.statcounter.com/search-engine-market-share#monthly-201705-201805-bar>. [Online; accedido 23-Junio-2018].
- [5] N. De Bellis, *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Scarecrow Press, 2009.
- [6] L. Leydesdorff and S. Milojevic, “Scientometrics,” *CoRR*, vol. abs/1208.4566, 2012.
- [7] L. Byl, J. Carson, A. Feltracco, S. Gooch, S. Gordon, T. Kenyon, B. Muirhead, D. Seskar-Hencic, K. MacDonald, M. T. Özsu, and P. Stirling, “White paper on bibliometrics, measuring research outputs through bibliometrics,” tech. rep., University of Waterloo, 01 2016.
- [8] University of Leeds, “What are bibliometrics?.” [https://library.leeds.ac.uk/info/1406/researcher\\_support/17/measuring\\_research\\_impact/1](https://library.leeds.ac.uk/info/1406/researcher_support/17/measuring_research_impact/1). [Online; accedido 23-Junio-2018].
- [9] B. Gipp and J. Beel, “Citation Proximity Analysis (CPA) - A New Approach for Identifying Related Work Based on Co - Citation Analysis,” in *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)* (B. Larsen and J. Leta, eds.), vol. 2, (Rio de Janeiro, Brazil), International Society for Scientometrics and Informetrics, Jul. 2009. ISSN 2175-1935.

- [10] Z. Dangzhi and S. Andreas, “Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis,” *Journal of the American Society for Information Science and Technology*, vol. 59, no. 13, pp. 2070–2086.
- [11] University of Leeds, “What are altmetrics?.” [https://library.leeds.ac.uk/info/1406/researcher\\_support/17/measuring\\_research\\_impact/3](https://library.leeds.ac.uk/info/1406/researcher_support/17/measuring_research_impact/3). [Online; accedido 23-Junio-2018].
- [12] J. Beall, “Google scholar is filled with junk science.” <https://web.archive.org/web/20141107175135/http://scholarlyoa.com/2014/11/04/google-scholar-is-filled-with-junk-science/>. [Online; accedido 23-Junio-2018].
- [13] J. Beel and B. Gipp, “Google Scholar’s Ranking Algorithm: An Introductory Overview,” in *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI’09)* (B. Larsen and J. Leta, eds.), vol. 1, (Rio de Janeiro, Brazil), International Society for Scientometrics and Informetrics, Jul. 2009. ISSN 2175-1935.
- [14] C. Analytics, “Web of science: It’s time to get the facts.” [https://cdn.clarivate.com/wp-content/uploads/2017/05/d6b7faae-3cc2-4186-8985-a6ecc8cce1ee\\_Crv\\_WoS\\_Upsell\\_Factbook\\_A4\\_FA\\_LR\\_edits.pdf](https://cdn.clarivate.com/wp-content/uploads/2017/05/d6b7faae-3cc2-4186-8985-a6ecc8cce1ee_Crv_WoS_Upsell_Factbook_A4_FA_LR_edits.pdf). [Online; accedido 23-Junio-2018].
- [15] ELSEVIER, “How scopus works - content.” <https://www.elsevier.com/solutions/scopus/how-scopus-works/content>. [Online; accedido 23-Junio-2018].
- [16] “Editorial,” in *Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR 2014)*, Amsterdam, The Netherlands, April 13, 2014., pp. 1–4, 2014.
- [17] S. Siebert, S. Dinesh, and S. Feyer, “Extending a research-paper recommendation system with bibliometric measures,” in *Proceedings of the Fifth Workshop on Bibliometric-enhanced Information Retrieval (BIR) co-located with the 39th European Conference on Information Retrieval (ECIR 2017)*, Aberdeen, UK, April 9th, 2017., pp. 112–121, 2017.
- [18] H. Zhao and X. Hu, “Language model document priors based on citation and co-citation analysis,” in *Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR 2014)*, Amsterdam, The Netherlands, April 13, 2014., pp. 29–36, 2014.

- [19] H. D. White, “Bag of works retrieval: Tf\*idf weighting of co-cited works,” in *Proceedings of the Third Workshop on Bibliometric-enhanced Information Retrieval co-located with the 38th European Conference on Information Retrieval (ECIR 2016)*, Padova, Italy, March 20, 2016., pp. 63–72, 2016.
- [20] M. J. Sarol, L. Liu, and J. Schneider, “Testing a citation and text-based framework for retrieving publications for literature reviews,” in *Proceedings of the 7th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2018) co-located with the 40th European Conference on Information Retrieval (ECIR 2018)*, Grenoble, France, March 26, 2018., pp. 22–33, 2018.
- [21] P. R. Christopher D. Manning and H. Schütze, “Information retrieval system evaluation.” <https://nlp.stanford.edu/IR-book/html/htmledition/information-retrieval-system-evaluation-1.html>.
- [22] UGR, “Rankings de investigadores ugrinvestiga según citación.” [http://investigacion.ugr.es/ugrinvestiga/static/BuscadorRanking/\\*/buscar?tipo=&rama\\_c=&disciplina\\_c=TELE\\_D&especialidad\\_c=&indicador=&periodo=](http://investigacion.ugr.es/ugrinvestiga/static/BuscadorRanking/*/buscar?tipo=&rama_c=&disciplina_c=TELE_D&especialidad_c=&indicador=&periodo=). [Online; accedido 3-Septiembre-2018].
- [23] H. Else, “How i scraped data from google scholar.”
- [24] Slant, “Solr vs elasticsearch.” [https://www.slant.co/versus/355/358/~solr\\_vs\\_elasticsearch](https://www.slant.co/versus/355/358/~solr_vs_elasticsearch). [Online; accedido 3-Septiembre-2018].
- [25] S. Opel, “Solr vs elasticsearch - stack overflow.” <https://stackoverflow.com/a/10213568>. [Online; accedido 3-Septiembre-2018].
- [26] J. Abrahamsson, “Elasticsearch 101 – a getting started tutorial.” <http://joelabrahamsson.com/elasticsearch-101/>. [Online; accedido 3-Septiembre-2018].
- [27] M. J. S. (@janinaj), “lit-review-search.” <https://github.com/janinaj/lit-review-search>. [Online; accedido 4-Septiembre-2018].
- [28] E. Developers, “elsapy: A python module for use with elsevier’s apis: Scopus, sciencedirect, others.” <https://github.com/ElsevierDev/elsapy>. [Online; accedido 4-Septiembre-2018].
- [29] scopus api, “scopus: Python-based api-wrapper to access scopus.” <https://scopus.readthedocs.io/en/latest/>. [Online; accedido 4-Septiembre-2018].

- [30] CSiRC, “Configuración del servicio vpn ssl.” <https://csirc.ugr.es/informatica/RedUGR/VPN/ConfVPNSSL/>. [Online; accedido 4-Septiembre-2018].
- [31] O. UGR, “7ª edición de los rankings de investigadores ugrinvestiga según citación del vicerrectorado de investigación y transferencia tras la revisión realizada por la comunidad investigadora de la ugr..” <http://opendata.ugr.es/dataset/rankings-de-investigadores-ugrinvestiga-segun-citacion/resource/4937db08-d757-469d-9bee-96fee9f798f8>. [Online; accedido 4-Septiembre-2018].
- [32] Wikipedia, “Software framework.” [https://en.wikipedia.org/wiki/Software\\_framework](https://en.wikipedia.org/wiki/Software_framework). [Online; accedido 31-Agosto-2018].
- [33] Wikipedia, “Front-end y back-end.” [https://es.wikipedia.org/wiki/Front-end\\_y\\_back-end](https://es.wikipedia.org/wiki/Front-end_y_back-end). [Online; accedido 31-Agosto-2018].
- [34] M. G. Software, “Scrum.” <http://www.mountaingoatsoftware.com/agile/scrum>. [Online; accedido 1-Septiembre-2018].
- [35] Wikipedia, “Web scraping.” [https://en.wikipedia.org/wiki/Web\\_scraping](https://en.wikipedia.org/wiki/Web_scraping). [Online; accedido 3-Septiembre-2018].



# Anexos



## Apéndice A

# Glosario

**Teorema de Bayes** Teorema perteneciente a la teoria de probabilidades que permite calcular la probabilidad condicional de dos eventos  $A$  y  $B$  en base su probabilidad condicional inversa y su probabilidad marginal

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Donde  $P(A|B)$  es la probabilidad condicional de que dado un evento  $B$  se produzca  $A$ ,  $P(A)$  la probabilidad márginal o incondicional de que suceda  $A$ ,  $P(B|A)$  la probabilidad condicional de que dado  $A$  suceda  $B$  y  $P(B)$  la probabilidad marginal de  $B$ . 4

**términos de indexación** concepto asociado a una serie de palabras donde el concepto es definido o discutido. 5

**sistema de recomendación** Tipo de sistema de RI en el que el usuario no expresa directamente su necesidad de información, si no que se le muestran items "similares" a los que ya ha consultado. 10

**framework** Abstracción que provee un entorno reutilizable y genérico que puede ser utilizado para facilitar el desarrollo software [32]. 11, 16, 23, 28

**backend** Parte de un sistema informático encargada del procesamiento o almacenamiento de información [33]. 16, 23

**frontend** Parte de un sistema informático encargada de la interacción con el usuario [33]. 16, 23

**Scrum** Metodología ágil de desarrollo de software basa en la descomposición del proyecto en sprints en los que definen los objetivos en lugar de

como hacer cada paso del proyecto. Existen tres roles en un proyecto Scrum:

- Equipo de desarrollo: formado por varios integrantes habitualmente
- ScrumMaster: parte del equipo de desarrollo pero con un rol especial que puede ser entendido como el "entrenador" del equipo
- Product owner: representa el cliente o los usuarios finales que son los que finalmente usaran el software desarrollado.

Cada uno de los sprint cuentan con un conjunto de tareas definidas a cumplir conocidas como sprint backlog y se suele hacer uso de tableros para seguir el progreso de cada una de las mismas. [34]. 16

***web scrapping*** Extracción de datos de una página web de forma automática manejando directamente los ficheros HTML de las páginas web [35]. 19, 29

## Apéndice B

# Lista de Acrónimos

**API** *Application Programming Interface*. 19, 21, 28, 29, 34

**BD** Base de Datos. 22

**BIR** *Bibliometric-enhanced Information Retrieval*. 9, 25, 26, 27

**E.T.S.I.I.T.** Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación. 19

**ES** *Elasticsearch*. 22, 28

**FECYT** Fundación Española para la Ciencia y la Tecnología. 9

**GIW** Gestión de Información en la Web. 14, 25, 28

**HTTP** *HyperText Transfer Protocol*. 28

**Mr.DLib** *Machine-readable Digital Library*. 10

**PDF** *Portable Document Format*. 2

**REST** *Representational State Transfer*. 28

**RI** Recuperación de Información. 1, 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 19, 25, 26, 27, 28, 43

**SO** Sistema Operativo. 16, 23

**TFG** Trabajo de Fin de Grado. 16

**TFM** Trabajo de Fin de Máster. 12, 15, 17, 27

**UGR** Universidad de Granada. 9, 29, 30, 31, 32, 33

**VPN** *Virtual Private Network*. 29

**WoS** Web of Science. 9, 19, 28

**XML** *Extensible Markup Language*. 28, 34

Apéndice C

Manual de uso





