Filtrado basado en contenido para artículos académicos en repositorios institucionales

José Federico Medrano[1]

Facultad de Ingeniería - Universidad Nacional de Jujuy, Jujuy 4600, Argentina
jfedericomedrano@gmail.com

Resumen La mayor parte del tiempo, los investigadores deben filtrar varios documentos académicos para encontrar aquellos relevantes para su investigación. Este filtrado muchas veces es engorroso y requiere emplear una considerable cantidad de tiempo. En la búsqueda de este tipo de material resulta útil contar con un listado de objetos relacionados no sólo con la temática buscada, sino también material que pueda estar relacionado semánticamente con el objeto de la búsqueda. Sería deseable contar con este tipo de funcionalidad en los repositorios institucionales, por esta razón en este trabajo se realiza una comparación de técnicas de filtrado basadas en el contenido semántico (*Term Frequency – Inverse Document Frequency* (TF-IDF), *Latent Semantic Indexing/Latent Semantic Analysis* (LSI/LSA) y *Word Mover's Distance* (WMD)) de una búsqueda realizada, como una propuesta para un sistema de recomendación de material académico empleando como conjunto de datos los registros almacenados en un repositorio institucional. Los resultados propuestos por dichas técnicas fueron evaluados de forma manual y comparados contra los resultados arrojados por el propio motor de búsqueda del repositorio y por los resultados entregados por *Google Scholar*. El esquema planteado mejora en gran medida los resultados actuales.

Palabras Clave: filtrado basado en contenido, PLN, sistema de recomendación, similitud semántica

1. Introducción

Un sistema de recomendación se define como una estrategia de toma de decisiones para los usuarios en entornos de información complejos [27]. Este tipo de sistemas se presentan como técnicas y herramientas de software que proporcionan sugerencias de "artículos" que estiman serán de utilidad para un usuario [4,29]. Las sugerencias están relacionadas con varios procesos de toma de decisiones, tales como qué artículos comprar, qué música escuchar o qué noticias leer. Un "artículo" es el término general utilizado para indicar lo que el sistema recomienda a los usuarios. Los sistemas de recomendación se dirigen principalmente a personas que carecen de suficiente experiencia personal o competencia para evaluar el número potencialmente abrumador de artículos alternativos que por ejemplo un sitio web puede ofrecer [30].

Un tema que resulta de mucho interés son las recomendaciones que ofrecen los motores y bases de datos académicas de acceso libre como *Google Scholar*[1], *Microsoft Academic*[2] y *ResearchGate*[3]. En estas bases de datos los elementos sugeridos son artículos o material académico, basados mayormente en las relaciones que posee el contenido buscado con el material almacenado, aquí entran en juego cuestiones como la popularidad de los elementos, tal es el caso de la aplicación del algoritmo *PageRank* de *Google* [21] para los resultados de las búsquedas y posteriores recomendaciones, o la utilización de mecanismos como *word embedding* [18], el nuevo esquema utilizado por *Microsoft Academic* para ofrecer una lista de publicaciones relacionadas [17].

Existen otros elementos de difusión y preservación donde se aloja material académico, estos son los Repositorios Institucionales (RI), este tipo de bases de datos han sido estudiadas [16,20,15] por la relevancia que poseen como elementos para organizar, gestionar, difundir, preservar y ofrecer acceso libre a la producción científica de una comunidad. Los RI poseen un motor de búsqueda interno, el cual ofrece un conjunto de resultados a partir de una búsqueda por palabra o frase, autor o fecha; buscando dentro de todas las colecciones o una colección en particular. Para los RI que utilizan como software de base a *DSpace*, el motor de búsqueda, dependiendo de la versión, es *Lucene* o *SOLR* o combinaciones de ambos. Estos motores permiten un gran número de opciones al momento de realizar la búsqueda [25], pero las búsquedas sencillas, por palabra o frase, serán las analizadas en este trabajo.

Una vez ingresada una búsqueda en un RI, los resultados pueden estar organizados por relevancia, título, fecha de publicación entre otros criterios. Las palabras ingresadas en la opción de búsqueda son buscadas en los campos título, autores, resumen, series, sponsor e identificador de cada registro. Estas búsquedas se transforman en consultas a una base de datos donde los términos introducidos son cotejados con los campos de los metadatos almacenados, el motor devuelve un conjunto de resultados donde la palabra o frase buscada figura en algún o algunos de los campos mencionados. Más allá de la existencia de estos términos en los registros, los resultados no presentan algún tipo de relación más estrecha, por ejemplo relaciones semánticas de contenido, es más al ingresar a alguno de los resultados de esta búsqueda, el RI no ofrece un listado de registros relacionados.

En este sentido, el presente trabajo aborda la problemática de evaluar y proponer un mecanismo que permita a un RI ofrecer publicaciones relacionadas semánticamente en base a una búsqueda realizada, de este modo, el usuario podría seguir navegando entre los resultados asumiendo que las sugerencias entregadas por la aplicación están realmente relacionadas con el objeto de su búsqueda.

El trabajo presentado se estructura de la siguiente manera, la Sección **2** realiza una breve introducción a los tipos de sistemas de recomendación más comunes

[1] https://scholar.google.com
[2] https://academic.microsoft.com/
[3] https://www.researchgate.net/

y sus aplicaciones, la Sección **3** presenta la definición formal del problema que se intenta resolver, la Sección **4** expone una explicación detallada del método de recomendación basado en contenido que se propone, por su parte en la Sección **5** se presentan los distintos experimentos realizados con el modelo desarrollado, la Sección **6** ofrece una breve validación de los resultados ofrecidos y por último la Sección **7** expone las conclusiones alcanzadas.

2. Trabajos relacionados

Existen dos modelos muy populares de sistemas de recomendación, los *sistemas de filtrado basados en el contenido* y los *sistemas basados en filtrado colaborativo*. Los primeros son los más populares [30], donde el contenido desempeña un papel principal en el proceso de recomendación, en el que las calificaciones de los usuarios y las descripciones de los atributos de los elementos se aprovechan para hacer predicciones. La idea básica es que los intereses del usuario se puedan modelar sobre la base de las propiedades (o atributos) de los elementos que han calificado o accedido en el pasado. Los "elementos" suelen ser textuales, por ejemplo, correos electrónicos [22] o páginas web [32]. La "interacción" generalmente se establece mediante acciones, como descargar, comprar, crear o etiquetar un artículo. Los elementos están representados por un modelo de contenido que contiene las características de los elementos. Las características suelen estar basadas en palabras, es decir, palabras sueltas, frases o n-*grams* [3].

Los sistemas basados en *filtrado colaborativo* [30,31] se refieren al uso de calificaciones de múltiples usuarios de forma colaborativa para predecir las calificaciones faltantes. El funcionamiento de este tipo de sistemas radica en la teoría que a los usuarios les gusta lo que les gusta a los que piensan de forma similar, donde dos usuarios se consideraron con ideas afines cuando calificaron elementos por igual. Cuando se identificaron usuarios de ideas afines, los elementos que un usuario calificó positivamente se recomendaron al otro usuario, y viceversa. Comparado con los sistemas de filtrado basados en contenido, los sistemas de filtrado colaborativo ofrecen tres ventajas [3]. Primero, el filtrado colaborativo es independiente del contenido, es decir, no se requiere procesamiento de elementos que puede estar propenso a error [23,31]. En segundo lugar, debido a que los humanos hacen las clasificaciones, el filtrado colaborativo toma en cuenta las evaluaciones de calidad reales [7]. Finalmente, se supone que el filtrado colaborativo debe proporcionar recomendaciones fortuitas porque las recomendaciones no se basan en la similitud de los ítems, sino en la similitud del usuario [23,14]. Esto ilustra uno de los principales problemas del filtrado colaborativo: se requiere la participación del usuario, pero a menudo la motivación para participar es baja, esta situación se conoce como el problema de "arranque en frío" (*coldstart*).

Algunas de las aproximaciones relacionadas con sistemas de recomendación de contenido académico o relacionado a la educación están bien clasificadas y resumidas en [3], los autores mencionan algunos enfoques como sistemas de recomendación de libros, sistemas de recomendación educativa, servicios de alerta

académica, búsqueda de expertos, resumen automático de artículos académicos, recomendadores de conjuntos de datos académicos y detección de plagio. En el mismo sentido, otras de las aproximaciones es el enfoque de [33] el cual propone un sistema de recomendaciones para artículos académicos que combina el análisis de citas y el análisis de redes. Por otro lado, [9] plantea un sistema de recomendación basado en un esquema híbrido que emplea redes de coautoría y técnicas basadas en contenido. Por su parte [34] presenta un sistema de recomendación entre dominios con transferencia de información coherente (CIT). En [26] se presenta una técnica llamada AKR (*Author-specified Keywords based Retrieval* - recuperación basada en palabras claves especificadas por el autor), la misma proporciona un conjunto diverso de publicaciones como parte de la lista de lectura inicial en base a las palabras claves que el autor especifica.

3. Definición del problema

Se define el problema de la siguiente manera:

Dado como entrada un artículo p de un área específica del conocimiento A que posee un conjunto C de artículos, encontrar un conjunto P de artículos que sean los más relacionados a p. Aquí, cada artículo p_i perteneciente a cualquier C_j está representado por un conjunto de metadatos M.

Como se trabajará sobre un Repositorio Institucional, los metadatos de los registros almacenados serán los datos de entrada. Para este estudio, la relación que existe entre el artículo p y el conjunto P tiene que ver con el contenido semántico, es decir, estos registros estarán relacionados semánticamente con la búsqueda realizada.

4. Enfoque propuesto

Los RI de acceso abierto que implementan el protocolo OAI-PMH[11], respetan el formato de metadatos *Dublin Core*[4] de manera obligatoria, por esta razón cada registro posee 15 campos de metadatos y 3 de encabezado. Los campos utilizados para este estudio son los *títulos* (campo *title*) y las *descripciones* (campo *description*, empleado para almacenar el resumen del elemento digital) por ser los más representativos de cada registro. Dentro de los metadatos del registro no se encuentra el cuerpo del artículo sino un enlace al mismo, debido a esto queda fuera del alcance del esquema propuesto evaluar el contenido del cuerpo del artículo. Los datos sobre los cuales se trabajó fueron recolectados en un trabajo previo [16], por este motivo los datos almacenados respetan el formato antes mencionado.

Según el estudio de [13], las palabras que ocupan diferentes posiciones en un documento tienen diferentes poderes discriminatorios, por ejemplo, una palabra

[4] http://dublincore.org/

que aparece en el título suele ser más significativa que una palabra que aparece en el texto del cuerpo. Una aplicación de esta idea fue conducida por [19] aplicando una ponderación arbitraria tres veces más fuerte a los términos del título que a los términos del cuerpo del texto, y dos veces más fuerte para los términos del resumen. Mientras que [8] experimentó con diferentes pesos para el contenido de artículos y el contexto de citas. Descubrieron que un peso igual para ambos campos alcanzaba la mayor precisión (en este caso la medida de precisión estaba relacionada con la medida de coincidencia). Atendiendo a estas consideraciones, se decidió llevar a cabo esta investigación aplicando el mismo peso para ambos campos de la tupla {*titulo, descripción*}.

Para este trabajo se decidió utilizar la librería *gensim* [28] del lenguaje *Python*[5]. Los esquemas de evaluación seleccionados para el desarrollo de los experimentos fueron: *TF-IDF* [1], *Latent Semantic Indexing/Latent Semantic Analysis* (LSI/LSA) [12,6] y *Word Mover's Distance* (WMD) [10,18].

La solución planteada funciona de la siguiente manera: Inicialmente se descartan los registros que no poseen título y descripción, todos los registros poseen un campo/metadato *subject* el cual, por lo común alberga detalles del área de conocimiento del registro, razón por la cual este campo es utilizado para dividir el conjunto total en subconjuntos más manejables. Cada subconjunto posee un conjunto de tuplas {título, descripción}, estas tuplas son procesadas para eliminar signos de puntuación, pasar las palabras a minúsculas y eliminar las *stop words*, estas *stop words* fueron obtenidas a partir de la librería *nltk*[6]. Seguido a esto, los registros fueron *tokenizados* para iniciar la labor de construcción del diccionario de términos, el cual será utilizado por las técnicas de representación de documentos seleccionadas (TF-IDF, LSI/LSA y WMD) que se encargarán de evaluar la similitud del artículo seleccionado contra el corpus procesado. Cada modelo de evaluación entregará los 10 registros del corpus que posean las mejores evaluaciones de similitud, es decir, los registros más parecidos evaluando los términos del título y descripción del artículo seleccionado. En este punto el modelo de evaluación ya ha sido entrenado con los subconjuntos de datos (división por área del conocimiento) y está listo para realizar predicciones.

En el modo normal de funcionamiento, el usuario realizaría una búsqueda inicial con algún término o frase, el motor del repositorio entregaría los resultados de la búsqueda, y una vez que el usuario elige un registro del listado resultante (este registro es preprocesado y tokenizado al igual que el conjunto de registros), el esquema aquí propuesto, realizaría la búsqueda del material relacionado con el *titulo + descripción* del artículo elegido. Se espera obtener una mejora en las búsquedas ya que es más probable, que un registro esté relacionado semánticamente con otros registros por un número mayor de palabras que si el esquema estuviera limitado solo a los títulos de las publicaciones.

[5] https://www.python.org/
[6] https://www.nltk.org/

5. Experimentos realizados

Los datos sobre los cuales se basaron los experimentos son los registros del Servicio de Difusión de la Creación Intelectual (SEDICI) [5] que es el repositorio institucional de la Universidad Nacional de La Plata[7], el año pasado, este repositorio ocupaba la posición 77°en el ranking mundial de repositorios web, 5°en Latinoamérica y 1°en Argentina [24]. El repositorio en ese entonces superaba los 54.000 registros.

Se dividió el conjunto total de registros del repositorio en áreas temáticas para que el procesamiento y tratamiento de la información sea manejable debido al gran número de registros del repositorio. Para realizar los experimentos se recuperaron los registros del área "Ciencias Informáticas", es decir, los registros que poseen dicha frase dentro del campo *subject*, esto dio un total de 9482 registros cuyos campos *título* y *descripción* son no vacíos. Se eligió esta área de conocimiento debido a la facilidad de contar con expertos para la evaluación de los resultados. Este conjunto de registros fue procesado y tokenizado para crear el diccionario de términos que luego fue utilizado para generar cada uno de los tres modelos de recomendación según las técnicas empleadas (TF-IDF, LSI/LSA y WMD).

El experimento se inició seleccionando al azar 8 artículos del área *Ciencias Informáticas*, estos artículos no formaron parte del cuerpo total de 9482. Cada título completo de estos registros fue introducido en la opción de búsqueda del SEDICI para obtener los primeros 50 resultados ordenados por relevancia. Ocho profesionales informáticos graduados participaron del experimento del siguiente modo: (1) a cada sujeto se le asignó uno de los artículos elegidos al azar y el listado correspondiente de 50 resultados arrojados por el motor de búsqueda del repositorio, cada resultado, incluido el artículo de la búsqueda, fue provisto con el título y la descripción, (2) se le solicitó que evaluarán cada uno de los 50 registros como *relacionado* o *no relacionado* y conformaran una lista ordenada de elementos recomendados de mayor a menor relevancia, (3) de cada lista se obtuvieron los 10 primeros registros para conformar la lista dorada (*gold list*) de cada uno de los 8 artículos iniciales.

De igual manera, por cada uno de los 8 artículos, se realizó una búsqueda en *Google Scholar (GS)* primero con el título del artículo completo e indicando que los recuperara del sitio web del SEDICI ingresando el comodín: **site:sedici.unlp.edu.ar**, luego se realizaron consultas empleando la técnica n-*grams*. Con esta estrategia se generan consultas extrayendo subsecuencias de n palabras del título del artículo, para los experimentos se tomó $n = 2$. Antes de aplicar dicha estrategia se preprocesó el texto del título, removiendo caracteres especiales, removiendo *stop words* y pasando los términos a minúsculas. Para todos los casos, del listado de resultados y ordenados por relevancia, se tomaron los primeros 10 resultados para evaluarlos contra la lista dorada generada por los expertos.

[7] http://sedici.unlp.edu.ar/

Por último, se emplearon cada uno de los modelos generados con las técnicas TF-IDF, LSI/LSA y WMD, ingresando por un lado los títulos de los registros, en otros casos solo la descripción y por último la tupla {título, descripción} de cada uno de los 8 registros para computar la similitud semántica entre estos y el corpus de evaluación generado previamente. Las consultas generadas con la estrategia n-*grams* sobre los términos del título de los artículos que fueron aplicadas para recuperar los registros de GS, fueron igualmente aplicadas al motor de búsqueda del SEDICI y en los modelos de evaluación TF-IDF, LSI/LSA y WMD. La idea de utilizar este tipo de técnicas es útil para comprobar si puede o no existir algún sesgo al momento de consultar con el título completo o solo con algunas palabras.

6. Validación

Para evaluar los resultados de la solución propuesta contra los resultados entregados por el motor de búsqueda del SEDICI y los entregados por *Google Scholar*, se adoptó la métrica de evaluación *recall* [2]. La sensibilidad o exhaustividad (*recall*) es la fracción de instancias o elementos relevantes recuperados por una consulta:

$$Recall = \frac{Elementos\ Relevantes\ Recuperados}{Total\ de\ Elementos\ Recuperados} \tag{1}$$

A continuación se presentan y analizan los resultados obtenidos con los experimentos llevados a cabo. Para evaluar comparativamente las alternativas mencionadas, se decidió de forma arbitraria tomar los primeros 10 resultados como el total de elementos recuperados en cada consulta realizada tanto al motor de búsqueda del repositorio, a GS y con los modelos de evaluación generados. Dentro de los primeros 10 resultados se computarán los elementos relevantes recuperados por cada alternativa. Se toma este criterio principalmente porque el usuario común tiende a buscar contenido relacionado entre los primeros resultados de una búsqueda, por ello se considera que 10 es un número razonable, asumiendo además que el usuario solo leerá los títulos y resúmenes de dichos resultados y no el cuerpo completo del documento. La Tabla 1 reúne los resultados promedios de

Tabla 1. Valores relativos de *recall* para las distintas alternativas planteadas

	SEDICI	GS	TF-IDF	LSI/LSA	WMD
Título	0.35 (0.8)	0.41 (0.5)	0.46 (1.15)	0.31 (0.73)	0.48 (0.81)
Título (2-*grams*)	0.32 (0.8)	0.39 (0.5)	0.41 (1.10)	0.28 (0.83)	0.48 (0.93)
Descripción	–	–	0.61 (1.28)	0.38 (0.85)	0.55 (0.85)
Título + Descripción	–	–	0.65 (1.31)	0.41 (0.89)	0.57 (0.85)

exhaustividad (*recall*) calculados para cada una de las alternativas empleadas,

entre paréntesis se incluye el tiempo promedio empleado en segundos para cada experimento. Estos experimentos fueron llevados a cabo realizando consultas a los distinto motores (SEDICI y GS) y modelos generados (TF-IDF, LSI/LSA y WMD) consultando por el título completo del artículo (primera fila), empleando 2-*grams* con los términos del título (segunda fila), tomando los términos de la descripción del artículo (tercera fila) y tomando tantos los términos del título como los de la descripción (cuarta fila). A simple vista, el esquema que mejor resultados ofrece es el TF-IDF empleando en conjunto el título y la descripción con un valor de 0.65 de *recall*. El segundo esquema que ha entregado buenos resultados es el WMD con un valor de 0.57 de *recall*. En las filas tercera y cuarta, para el SEDICI y GS no existe un valor ya que para estos motores sólo se ha consultado por el título del artículo sin emplear los términos de la descripción.

Al emplear la estrategia n-*grams* no se obtuvieron diferencias sustanciales en relación a los otros esquemas, esto quiere decir que para realizar consultas empleando solamente palabras del título, basta ingresar dos palabras del mismo para obtener resultados aceptables en lugar de emplear el título completo. Además hay que destacar que el empleo de las *descripciones* de los registros, mejoró en todos los resultados en todos los casos, esto es un claro indicio de las relaciones semánticas entre los documentos recuperados ya que las descripciones representan un resumen del cuerpo del artículo.

7. Conclusiones

En este trabajo se presentó una solución preliminar a un problema específico y real, poder recomendar contenido académicos semánticamente relacionado en un repositorio institucional. Si bien, hasta el momento no se intentó resolver este problema en este tipo de herramientas, el modelo aquí presentado otorga una alternativa viable y una mejora sustancial a la funcionalidad ofrecida por el repositorio. La aplicación de los algoritmos seleccionados han mostrado resultados muy alentadores, pues las relaciones entre documentos aquí encontradas e inferidas, sería una tarea realmente compleja de resolver con simples consultas a una base de datos.

La aplicación de técnicas de Inteligencia Artificial, y más específicamente, técnicas de Procesamiento del Lenguaje Natural al problema presentado, demuestran el gran potencial de estas para resolver problemas de cualquier índole y la flexibilidad de adaptarse a nuevos entornos, tal es el caso del esquema TF-IDF que en este caso entregó los mejores resultados.

La búsqueda de contenidos académicos relacionados es un tema muy importante y un instrumento valioso para la comunidad científica, además, teniendo en cuenta la importancia y la relevancia que tiene el SEDICI como un repositorio referente a nivel nacional y latinoamericano, contar con herramientas y/o servicios que mejoren sus prestaciones le aportan un gran valor institucional.

El estudio aquí presentado deja abierto el panorama para un sin fin de aplicaciones y futuras líneas de investigación. La combinación de técnicas en conjunto con las aquí presentadas puede resultar en algo más fructífero, sin ir más lejos, el

empleo de palabras claves del mismo documento o la combinación de esquemas basados en la recuperación de documentos referenciados, o el empleo de redes de coautoría, o la evaluación del cuerpo del documento o empleando un esquema de ponderación para las palabras del título y descripción pueden llegar a otorgar una mejora considerable.

Referencias

1. Aizawa, A.: An information-theoretic perspective of tf–idf measures. Information Processing & Management **39**(1), 45–65 (2003)
2. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
3. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems : a literature survey. International Journal on Digital Libraries **17**(4), 305–338 (2016). https://doi.org/10.1007/s00799-015-0156-0
4. Burke, R.: The adaptive web. chap. Hybrid Web Recommender Systems, pp. 377–408. Springer-Verlag, Berlin, Heidelberg (2007)
5. De Giusti, M.R., Oviedo, N., Lira, A.J., Sobrado, A., Martínez, J.P., Pinto, A.V.: Sedici-desafíos y experiencias en la vida de un repositorio digital. In: Conferencia sobre Bibliotecas y Repositorios Digitales (Colombia, 2011). vol. 1 (2011)
6. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391 (1990)
7. Dong, R., Tokarchuk, L., Ma, A.: Digging friendship: paper recommendation in social network. In: Proceedings of Networking & Electronic Commerce Research Conference (NAEC 2009). pp. 21–28 (2009)
8. Huang, S., Xue, G.R., Zhang, B.Y., Chen, Z., Yu, Y., Ma, W.Y.: Tssp: A reinforcement algorithm to find related papers. In: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence. pp. 117–123. IEEE Computer Society (2004)
9. Hwang, S.Y., Wei, C.P., Liao, Y.F.: Coauthorship networks and academic literature recommendation. Electronic Commerce Research and Applications **9**(4), 323 – 334 (2010), special Section: Social Networks and Web 2.0
10. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: International Conference on Machine Learning. pp. 957–966 (2015)
11. Lagoze, C., Van de Sompel, H.: The open archives initiative: Building a low-barrier interoperability framework. In: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries. pp. 54–62. ACM (2001)
12. Landauer, T.K., Dumais, S.T.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review **104**(2), 211 (1997)
13. Manning, C.D., Raghavan, P.: Introduction to Information Retrieval. Cambridge University Press. (2008)
14. McNee, S.M., Kapoor, N., Konstan, J.A.: Don't look stupid: avoiding pitfalls when recommending research papers. In: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work. pp. 171–180. ACM (2006)
15. Medrano, J.F., Figuerola, C.G., Berrocal, J.L.A.: Repositorios digitales en españa y calidad de metadatos. Scire: representación y organización del conocimiento **18**(2), 109–121 (2012)

16. Medrano, J.F.: Calidad en repositorios digitales en argentina, estudio comparativo y cualitativo. In: VII Conferencia Internacional BIREDIAL-ISTEC'17 y XII SIBD (La Plata, 2017) (2017)
17. MicrosoftResearch: Feature improvement: Related publications (2018), https://tinyurl.com/y9x8ptoc, consultado el: 2018-03-25
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Neural Information Processing Systems pp. 3111–3119 (2013)
19. Nascimento, C., Laender, A.H., da Silva, A.S., Gonçalves, M.A.: A source independent framework for research paper recommendation. In: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. pp. 297–306. ACM (2011)
20. Orduña-Malea, E.: Impacto de los repositorios a través de técnicas cibermétricas: el caso general de latinoamérica y especial de costa rica. In: III Conferencia Bibliotecas y repositorios digitales de América Latina (BIREDIAL) (2013)
21. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
22. Paik, W., Yilmazel, S., Brown, E., Poulin, M., Dubon, S., Amice, C.: Applying natural language processing (nlp) based metadata extraction to automatically acquire user preferences. In: Proceedings of the 1st international conference on Knowledge capture. pp. 116–122. ACM (2001)
23. Palopoli, L., Rosaci, D., Sarné, G.M.: A multi-tiered recommender system architecture for supporting e-commerce. In: Intelligent Distributed Computing VI, pp. 71–81. Springer (2013)
24. Pinto, A.: Sedici en el ranking webometrics (2017), https://tinyurl.com/yb4uyqel, consultado el: 2018-03-30
25. Prasad, A., Patel, D.: Lucene search engine: An overview. DRTC-HP International **10** (2005)
26. Raamkumar, A.S., Foo, S., Pang, N.: Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems. Information Processing & Management **53**(3), 577–594 (2017)
27. Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S.M., Konstan, J.A., Riedl, J.: Getting to know you: learning new user preferences in recommender systems. In: Proceedings of the 7th international conference on Intelligent user interfaces. pp. 127–134. ACM (2002)
28. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010)
29. Resnick, P., Varian, H.R.: Recommender systems. Communications of the ACM **40**(3), 56–58 (1997)
30. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. In: Recommender systems handbook, pp. 1–35. Springer (2011)
31. Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S.: Collaborative filtering recommender systems. In: The adaptive web, pp. 291–324. Springer (2007)
32. Seroussi, Y.: Utilising user texts to improve recommendations. In: International Conference on User Modeling, Adaptation, and Personalization. pp. 403–406. Springer (2010)
33. Son, J., Kim, S.B.: Academic paper recommender system using multilevel simultaneous citation networks. Decision Support Systems **105**, 24 – 33 (2018)
34. Zhang, Q., Wu, D., Lu, J., Liu, F., Zhang, G.: A cross-domain recommender system with consistent information transfer. Decision Support Systems **104**, 49 – 63 (2017)