8ª Jornada sobre la Biblioteca Digital Universitaria
JBDU2010
"Las redes sociales y las bibliotecas universitarias"

Soporte Inteligente para el Mantenimiento y Acceso Contextualizado a Repositorios Institucionales

María Alejandra Dini Mariana Anahí Varela Verónica Antúnez Ana Maguitman
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
Avenida Alem 1253 – 8000 Bahía Blanca – Argentina

Luis Herrera
Biblioteca Central de la Universidad Nacional del Sur
Avenida Alem 1253 – 8000 Bahía Blanca – Argentina

Resumen. El presente trabajo consiste en describir nuevas herramientas desarrolladas por investigadores y alumnos avanzados del Departamento de Ciencias e Ingeniería de la Computación de la Universidad Nacional del Sur en colaboración con la Biblioteca Central de dicha Institución. Las herramientas desarrolladas se sustentan en técnicas de Recuperación de Información e Inteligencia Artificial y tienen como objetivo superar ciertas limitaciones encontradas en los mecanismos actuales de mantenimiento y acceso a Repositorios Institucionales. Más específicamente, se proponen *herramientas inteligentes* para facilitar la catalogación de tesis de posgrado con el fin de facilitar su posterior acceso. Las técnicas propuestas tienen como principal componente el uso de mecanismos de búsqueda contextualizada. Estos mecanismos permiten reflejar un contexto temático en las consultas presentadas y distinguir recursos relevantes de aquellos que no lo son. Los resultados obtenidos mediante estas consultas son utilizados para generar sugerencias que facilitan la catalogación y para acceder a recursos relevantes de manera más efectiva.

1. Introducción

Los Repositorios Institucionales (RI) consisten en una colección de recursos digitales organizados para su uso a largo plazo. Para que los RI sean funcionales, las bibliotecas tienen como misión recopilar recursos en formato digital, organizarlos, preservarlos y facilitar el acceso a estos

repositorios por parte de los usuarios. Cada una de estas tareas presenta numerosos desafíos que a la vez ofrecen atractivas oportunidades de investigación y desarrollo.

Organizar información que abarca diversos tópicos es una tarea difícil y costosa para el catalogador, quien probablemente no se encuentra familiarizado con la temática heterogénea de los recursos a clasificar. Por otra parte, el acceso a dichos recursos por parte de los usuarios puede ser poco efectivo si la biblioteca digital no cuenta con los mecanismos de búsqueda apropiados para facilitar la identificación del material potencialmente relevante. En vista de estas necesidades, se han propuesto diferentes soluciones sustentadas en tecnologías desarrolladas en el área de Ciencias de la Información (Buckland 1992; Levy & Marshall 1995; Sølvberg 2001).

El presente trabajo describe herramientas de software desarrolladas para facilitar la catalogación de recursos electrónicos que son incorporados a un RI como así también para facilitar la búsqueda contextualizada de dichos recursos por parte de los usuarios. El objetivo específico de este trabajo consiste en realizar un aporte a la Biblioteca Central de la Universidad Nacional del Sur para facilitar el manejo de la información relacionada con la documentación referida a las tesis de posgrado que arriban a la biblioteca. Las herramientas propuestas en este trabajo apuntan a facilitar la tarea de mantenimiento y acceso contextualizado al repositorio de tesis.

- **Mantenimiento:** Cuando una nueva tesis arriba a la biblioteca, la misma debe ser ingresada al repositorio de tesis, asociándole un conjunto de datos que la describen. Cada tesis tiene asociado un autor, un director, un título, un subtítulo, un resumen (o abstract), códigos Dewey, una temática y una fecha de publicación. Parte de estos datos, tales como título, datos del autor, director, resumen y fecha de publicación, se encuentran explícitos en la tesis misma, mientras que otros datos, tales como los códigos Dewey, la temática y los descriptores relevantes, deben ser inferidos por el catalogador. El objetivo de esta herramienta es asistir al catalogador en la selección de los códigos Dewey, la temática y los descriptores para asignar a una tesis. El método desarrollado para alcanzar tal fin toma como punto de partida el resumen de la nueva tesis ingresada y ofrece al catalogador una serie de sugerencias respecto a los datos que pueden ser asociados a la misma.

- **Acceso contextualizado:** Típicamente los usuarios consultan los RI utilizando palabras vinculadas a campos específicos (tales como autor o título) o utilizando unas pocas palabras claves representativas del tema de interés. Sin embargo, un usuario podría estar realizando una búsqueda temática y no conocer el vocabulario específico con el que la tesis se encuentra indexada. Esta dificultad lleva a que el usuario no pueda reflejar de manera apropiada sus necesidades de información, resultando en un acceso deficiente al material relevante. La aplicación descripta en este trabajo intenta superar dicha limitación, extendiendo los mecanismos tradicionales de acceso a los RI mediante la incorporación de técnicas de búsqueda contextualizada. Dichas técnicas de búsqueda consisten en la utilización de uno o varios párrafos (extraídos de algún documento de texto representativo del tema objeto de la necesidad de información) para dar lugar a una búsqueda temática más completa y pertinente.

En este trabajo describiremos cómo las tareas de mantenimiento y acceso contextualizado son llevadas a cabo mediante la aplicación de técnicas de Recuperación de Información e Inteligencia Artificial. A continuación se introducirán los conceptos generales vinculados a la presente propuesta, tras lo cual se describirán los datos y herramientas de software utilizadas en el desarrollo de la aplicación presentada. En la sección siguiente se detallarán las técnicas desarrolladas y se describirá la funcionalidad del sistema implementado. Finalmente, se presentarán las conclusiones alcanzadas y se delinearán extensiones a desarrollar en el futuro.

2. Conceptos Generales

La presente sección tiene el propósito de introducir una serie de conceptos del área de Recuperación de la Información Clásica y Sistemas Inteligentes de Búsqueda, los cuales serán utilizados a lo largo del presente trabajo.

2.1. Modelo de Espacio Vectorial

El modelo de espacio vectorial (Salton et al. 1975) es un modelo algebraico para representar documentos y consultas como vectores en *n* dimensiones, donde cada dimensión representa un término (palabra, concepto o lema). Los términos tienen asignados pesos no binarios, los que son utilizados

para computar grados de similitud entre pares de documentos o entre documentos y consultas. Este modelo propone evaluar el grado de similitud entre dos documentos (o un documento y una consulta) como la correlación entre sus representaciones vectoriales. Dicha correlación puede se cuantificada mediante el coseno del ángulo entre los dos vectores involucrados. Para mayores detalles sobre este modelo, se remite al lector a (Salton et al. 1975; Baeza-Yates & Ribeiro-Neto, 1999).

2.2. Agrupación de documentos

La agrupación o clustering (Rousseeuw & Kaufman, 1990) consiste en agrupar un conjunto de objetos (tales como documentos) basándose en la similitud de los valores de sus atributos. Los métodos de clustering identifican regiones densamente pobladas, denominadas clusters, de acuerdo a alguna medida de distancia o similitud establecida. De esta manera se busca maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters.

El algoritmo de clustering Lingo (Osiński et al., 2004), utilizado en el presenta trabajo a través de su implementación en la herramienta Carrot2 (Osiński & Weiss, 2005), está orientado a agrupar resultados de un motor de búsqueda pero también se puede aplicar para automatizar la creación de perfiles de una entidad basada en opiniones, entre otras aplicaciones. Una vez realizado el pre-procesamiento (identificación de idioma, filtro de palabras y stemming), Lingo induce los descriptores de los clusters usando el método de Análisis de Semántica Latente (Deerwester et al., 1990) para luego asignar los documentos a los clusters usando el modelo de espacio vectorial.

2.3. Clasificación de documentos

La clasificación se utiliza para categorizar un conjunto de objetos basado en los valores de sus atributos. Por ejemplo, se podría clasificar a distintas personas para la concesión de un préstamo en riesgo bajo, medio y alto, teniendo en cuenta información histórica de las mismas.

La clasificación identifica las propiedades comunes entre un conjunto de objetos y los clasifica en diferentes clases, de acuerdo a un modelo de clasificación preestablecido. Para construir este modelo, se utiliza un conjunto de entrenamiento, en el que cada instancia consiste de un conjunto de

atributos y el valor de la clase a la cual pertenece. El objetivo de la clasificación es analizar los datos de entrenamiento y, mediante un método supervisado, desarrollar una descripción o un modelo para cada clase utilizando las características disponibles en los datos. Esta descripción o modelo permite clasificar otras instancias, cuya clase es desconocida. El método se conoce como supervisado debido a que, para el conjunto de entrenamiento, se conoce la clase de pertenencia y se le indica al modelo si la clasificación que realiza es correcta o no. La construcción del modelo se realimenta de estas indicaciones del supervisor. Para una análisis más extenso del tema remitimos al lector a (Sebastiani, 2002).

2.4. Búsqueda contextualizada

Los sistemas de búsqueda contextualizada monitorean al usuario, infieren sus necesidades de información y buscan recursos relevantes en diversos repositorios (tales como la Web o repositorios locales). Los contextos temáticos juegan un papel fundamental en los sistemas de búsqueda basados en la tarea del usuario. Desafortunadamente, aprovechar la información del contexto es una tarea difícil. Esto se debe a que los sistemas de búsqueda actuales imponen un límite a la longitud de las consultas, y aún si se permitieran consultas largas las mismas podrían volverse demasiado específicas, devolviendo muy pocos o ningún resultado. Existen varios sistemas que han adoptado diferentes enfoques para resolver el problema de la búsqueda contextualizada. Por ejemplo, Watson (Budzik et al., 2001) utiliza información de los documentos que los usuarios están accediendo para generar consultas automáticamente, recurriendo a diversas técnicas de extracción y ponderación de términos. Watson filtra los resultados recuperados, agrupa aquellos que son similares y los presenta como sugerencia al usuario. Otro sistema de búsqueda conocido como Remembrance Agent (Rhodes & Starner, 1996) puede ser integrado al editor de texto Emacs y tiene como propósito monitorear continuamente el trabajo del usuario para descubrir documentos de texto, notas y email relevantes que fueran indexados previamente. El sistema EXTENDER (Maguitman et al., 2004; Maguitman et al., 2005) aplica una técnica de búsqueda incremental para construir una descripción del contexto del usuario. Su tarea es generar descripciones breves de nuevos tópicos relevantes para un modelo de conocimiento en construcción.

Además de las propuestas mencionadas arriba existen muchos otros trabajos de investigación y sistemas destinados a resolver el problema de la búsqueda contextualizada (Armstrong et al., 1995; Lieberman, 1995; Baldonado & Winograd, 1997; Maglio et al., 2000;; Scholer & Williams, 2002; Kraft et al., 2006).

3. Datos y herramientas de software

En esta sección describimos el formato de los datos y las librerías de soporte utilizadas para el desarrollo del sistema propuesto.

3.1. Datos de utilizados

Como punto de partida, el sistema cuenta con un archivo en formato XML el cual representa a todo el conjunto de tesis existentes en la base de datos actual de la Biblioteca Central. Este archivo se encuentra definido con la estructura presentada en la Figura 1.

```
<todas_las_tesis>
<tesis>
<autor>… </autor>
<titulo>… </titulo>
<subtitulo>… </subtitulo>
<director>… </director>
<resumen>… </resumen>
<publicacion_fecha>… </publicacion_fecha>
<dewey1>… </dewey1>
<dewey2>… </dewey2>
<dewey3>… </dewey3>
<tematica1>… </tematica1>
<tematica2>… </tematica2>
<tematica3>… </tematica3>
</tesis>
<tesis>
<autor>… </autor>
<titulo>… </titulo>
<subtitulo>… </subtitulo>
<director>… </director>
<resumen>… </resumen>
<publicacion_fecha>… </publicacion_fecha>
<dewey1>… </dewey1>
<dewey2>… </dewey2>
<dewey3>… </dewey3>
<tematica1>… </tematica1>
<tematica2>… </tematica2>
<tematica3>… </tematica3>
</tesis>
…
</todas_las_tesis>
```

Figura 1: *Descripción de las tesis en formato XML*

Actualmente, la Biblioteca cuenta con más de 500 tesis de posgrado cuya descripción se encuentra disponible en formato digital. El conjunto de tesis que arriban al RI se halla en continua expansión.

3.2. Librerías de soporte

Múltiples librerías de código abierto han sido utilizadas en el desarrollo del presente trabajo. A continuación presentaremos una breve descripción de cada una de ellas.

- **Lucene[1].** La herramienta Lucene (Gospodnetic et al. 2009) es una poderosa librería que permite crear índices a partir de documentos en diversos formatos tales como PDF o XML. El índice creado por Lucene puede ser consultado por campos determinados, lo que es de gran utilidad para la aplicación en cuestión. Entre otras características, permite indexar documentos de manera incremental, eliminar stopwords (palabras muy usadas que no agregan significado, tales como "el", "la", "con", etc.) y elegir el idioma a utilizar. Además, una ventaja importante de Lucene es que soporta la concurrencia, es decir, permite que varios usuarios realicen consultas en el índice simultáneamente, así como que un usuario modifique un índice a la vez que otro lo consulta.
- **Digester[2].** La librería Digester (Janert 2002) es utilizada para procesar archivos en formato XML a través de la búsqueda de reglas. Digester ofrece una forma simple para la asignación de documentos XML a objetos Java. Se utiliza el método SAX[3] para parsear los documentos XML, lo cual es realizado través de eventos.
- **Carrot[4].** La herramienta Carrot (Osiński & Weiss, 2005) es un motor de búsqueda basado en clusters. El objetivo es organizar los resultados de las búsquedas en varios grupos de documentos relacionados. Existen varios algoritmos disponibles en el sistema, entre los cuales se encuentra un algoritmo basado en Lingo, el cual fue descripto brevemente en la sección 2.2.
- **JDOM[5].** La librería DOM es un API para leer, crear y manipular documentos XML de una manera sencilla y muy intuitiva para cualquier programador en Java. JDOM no es un analizador sintáctico (parser), sino que usa un parser para su trabajo, aportando una capa de abstracción en la manipulación de documentos XML.

[1] http://lucene.apache.org/
[2] http://commons.apache.org/digester/
[3] http://www.saxproject.org/
[4] http://search.carrot2.org/stable/search
[5] http://www.jdom.org/

4. El sistema en acción

El primer paso llevado a cabo por el sistema es indexar toda la información que contiene el archivo XML de tesis (descripto en la Figura 1) utilizando las operaciones ofrecidas por la librería Lucene y Digester. Una vez construido el índice es posible comenzar con las búsquedas de documentos basadas en contexto. El *ingreso de una nueva tesis al repositorio*, requiere de tres pasos fundamentales: (1) la construcción de consultas para la búsqueda, (2) la identificación de abstracts similares y (3) la generación de sugerencias para la catalogación del nuevo abstract.

1. **Construcción de consultas para la búsqueda contextualizada.** Se utiliza la "Técnica de la Ruleta" para abordar el problema de reflejar contextos temáticos en las consultas formuladas a un buscador. Para implementar esta técnica, deben extraerse los términos del contexto del usuario (descartando los stopwords) y cada uno de ellos se pondera de acuerdo a su importancia dentro del contexto en cuestión. Una manera simple de asignar pesos a los términos es basándose en el número de apariciones de los mismos, favoreciendo además a aquellos que aparecen en posiciones destacadas dentro del texto, tales como títulos. La "Técnica de la Ruleta" consiste en armar una ruleta con todos estos términos, para luego obtener aleatoriamente de la misma los términos que conformarán la consulta. La probabilidad de seleccionar un término para constituir una consulta es proporcional al peso de dicho término. De esta manera, un número determinado de consultas pueden ser construidas, las cuales tienden a incluir los términos más representativos del contexto temático. La aplicación de esta técnica permite explorar de manera no determinista el espacio de términos, favoreciendo a aquellos más prometedores. La "Técnica de la Ruleta" es típicamente utilizada en la programación de Algoritmos Genéticos (Holland, 1975) para seleccionar soluciones potencialmente útiles. En estos algoritmos el nivel de de aptitud de una solución es usada para asignar una probabilidad de selección. Mayor detalle sobre el método propuesto de construcción de consultas puede encontrarse en (Lorenzetti & Maguitman 2009).
2. **Búsqueda de abstracts similares.** A partir de la consulta construida anteriormente se procede con la búsqueda de documentos similares al abstract ingresado. Para lograr este objetivo se utiliza la librería Carrot2 (descripta en la sección 3.2). De esta forma los resultados obtenidos se encuentran organizados en clusters, siendo posible seleccionar aquellos más similares al abstract ingresado por el usuario. Finalmente, un documento es considerado para formar parte de la solución si la

similitud por coseno (concepto introducido en la sección 2.1) con el abstract ingresado es mayor o igual a 0.5.

3. **Generación de sugerencias.** A partir de los abstracts similares encontrados es posible generar sugerencias para la catalogación del nuevo abstract. Estas sugerencias consisten en listas de descriptores, códigos Dewey y temáticas potencialmente útiles para ser asignadas a la nueva tesis. Este mecanismo de generación de sugerencias se sustenta en la técnica de Inteligencia Artificial conocida como Razonamiento Basado en Casos (Leake 1996), donde las experiencias pasadas sirven como punto de partida para proveer soluciones en situaciones nuevas. En este caso, los abstracts ya catalogados proveen "soluciones" tentativas a través de sus descriptores, códigos Dewey y temáticas asignadas para resolver el problema de asignar estos datos a una nueva tesis.

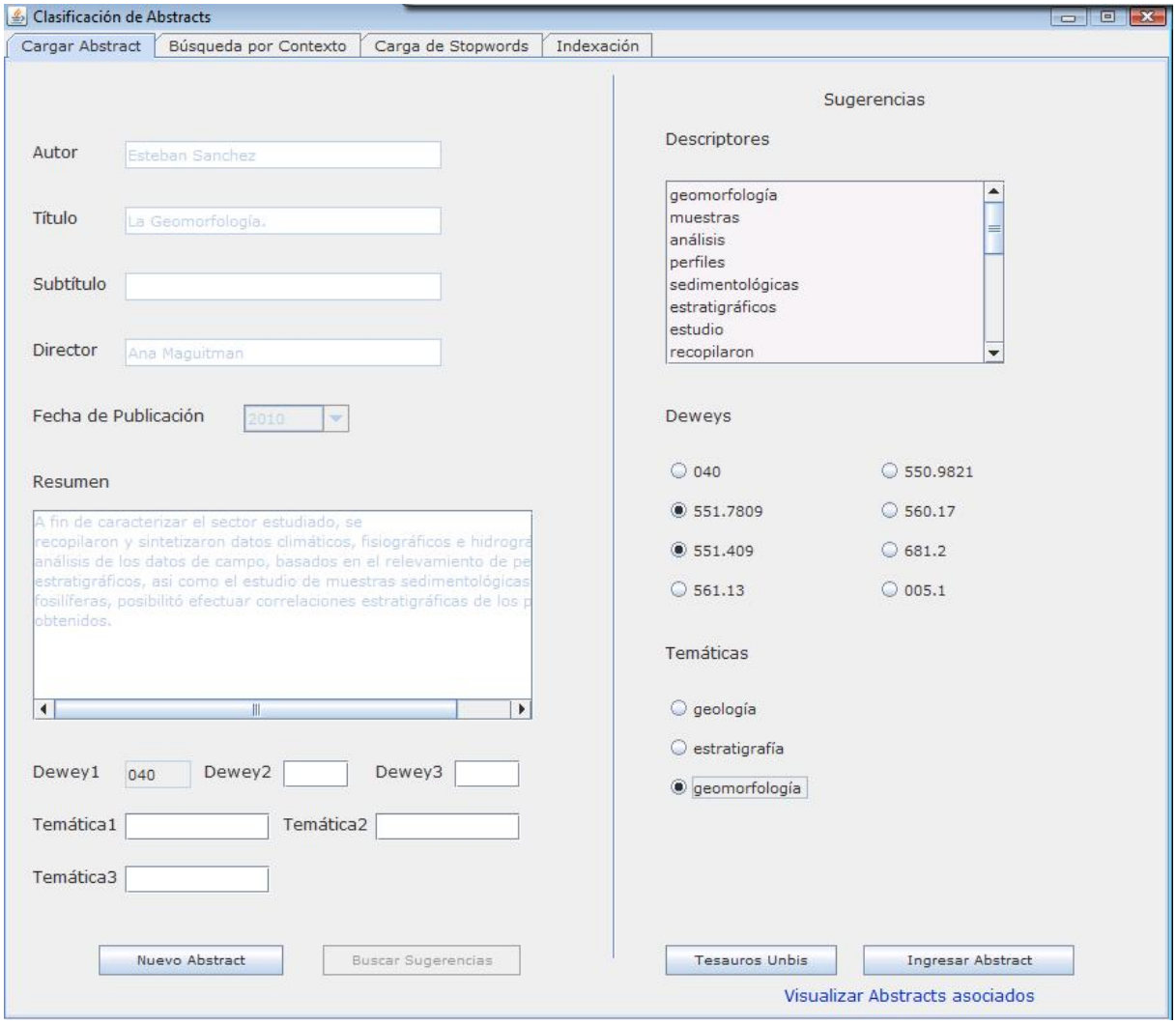

Figura 2: Generación de sugerencias para la catalogación de una nueva tesis.

Las figura 2 muestra la interfaz del sistema con las sugerencias generadas para la catalogación de una nueva tesis.

Por otra parte, la *búsqueda de tesis almacenadas en el repositorio* puede ser realizado utilizando mecanismos de búsqueda clásico o basados en contexto. En el primer caso, el usuario debe completar uno o más de los campos simples del formulario de consulta. Estos campos son referidos al autor, título, director, rango para el año de publicación, códigos Dewey y temática. La búsqueda basada en contexto, por otra parte, se realiza a partir de párrafos que representan contextos temáticos. En este caso, la construcción de consultas para la búsqueda contextualizada se realiza utilizando la "Técnica de la Ruleta", como fuera expuesto para el caso del ingreso de una nueva tesis al repositorio.

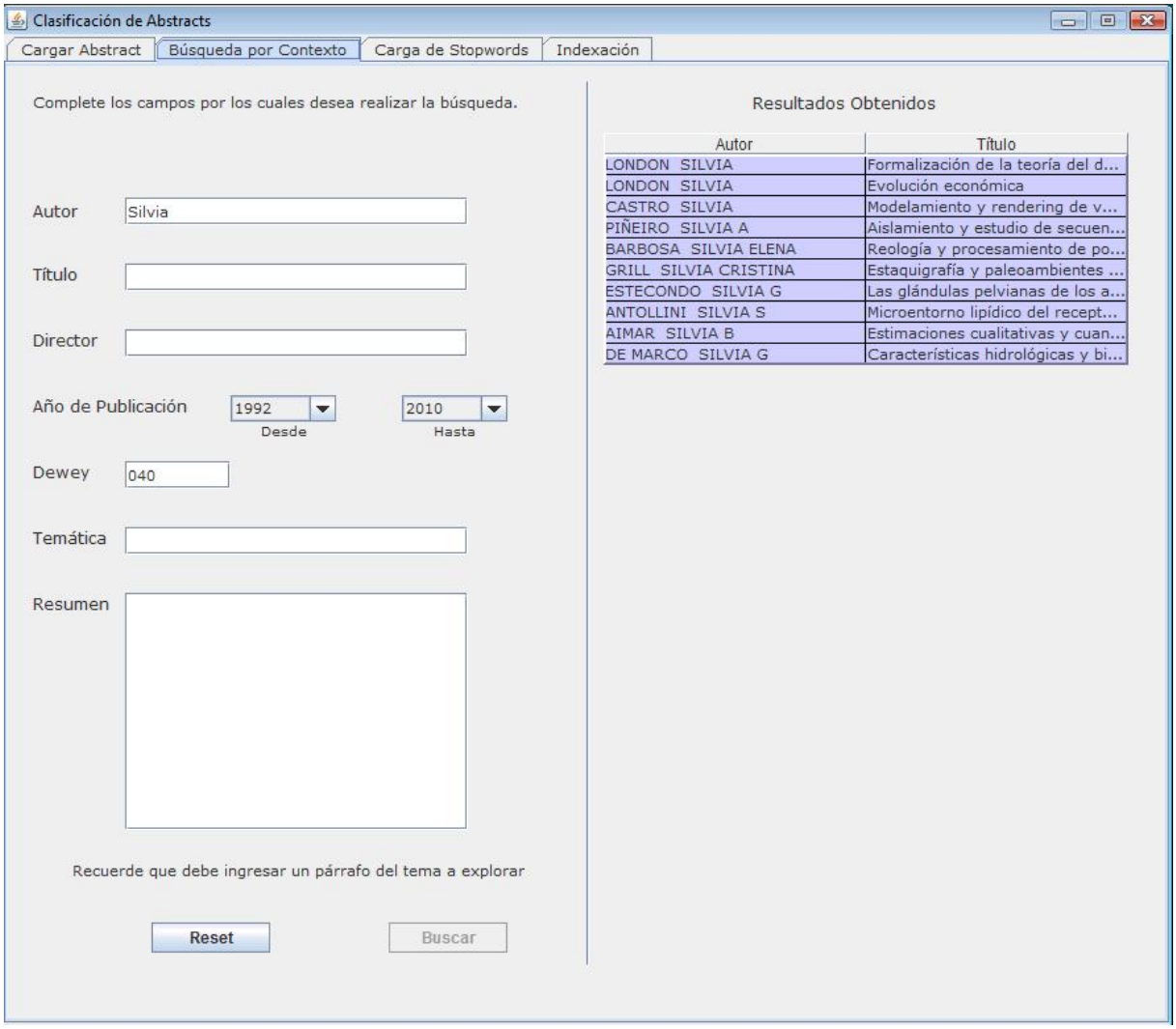

Figura 3: Resultados del proceso de búsqueda mediante consultas simples.

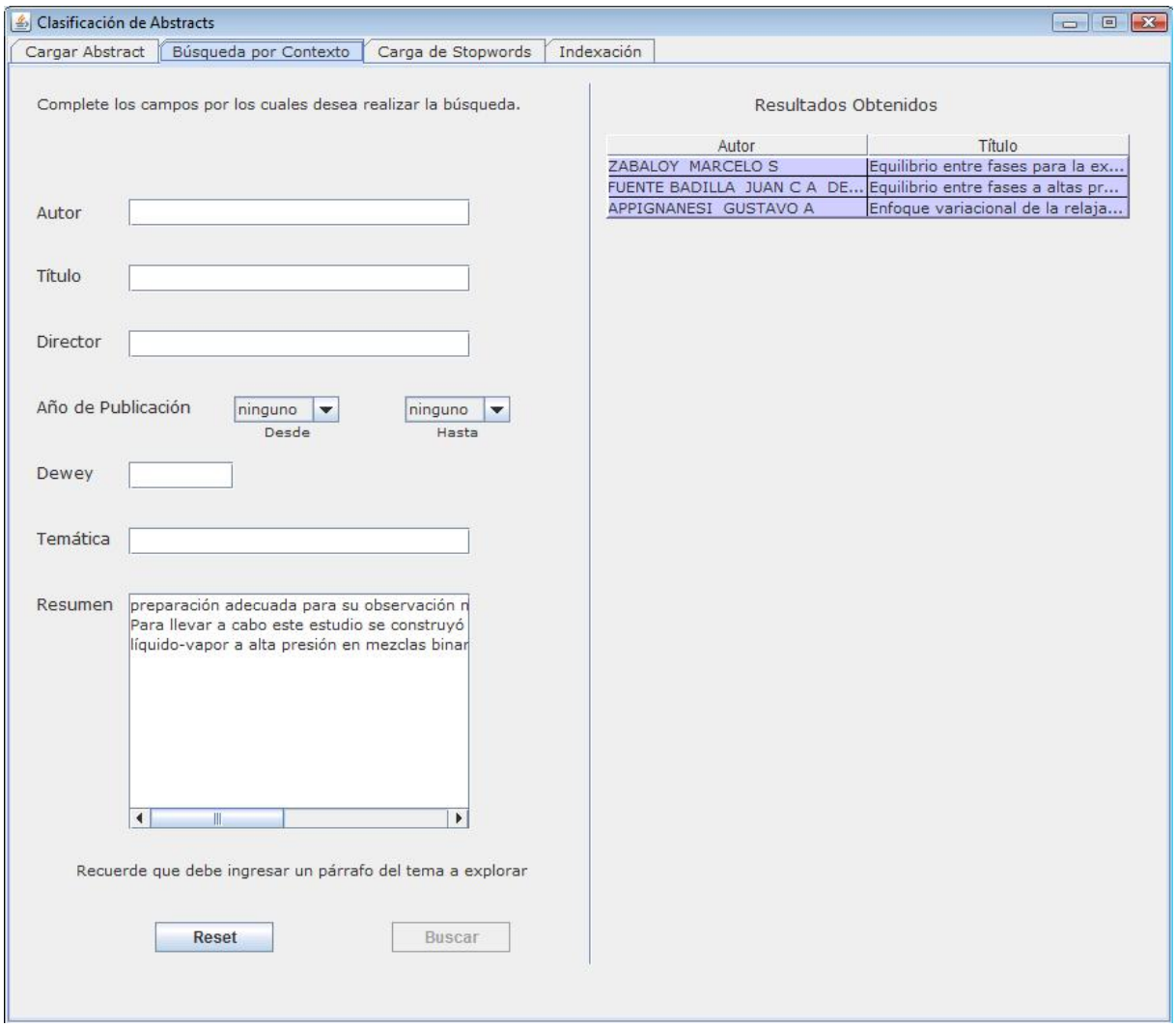

Figura 4: Resultados del proceso de búsqueda contextualizada.

Las Figuras 3 y 4 ilustran los procesos de búsqueda simple y búsqueda basada en contexto, respectivamente. Por otra parte, la Figura 5 muestra los datos que el usuario puede acceder al seleccionar una de las tesis que resultan del proceso de búsqueda.

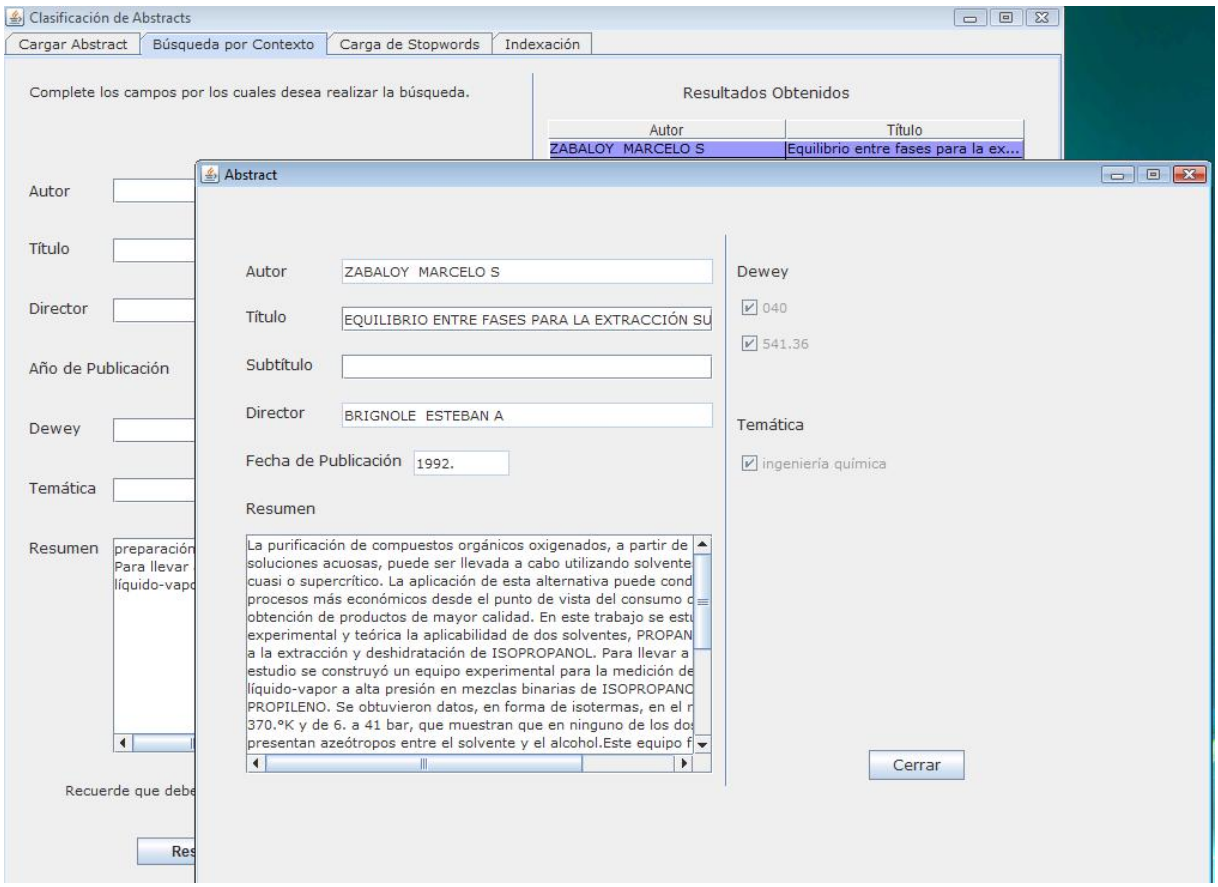

Figura 5: Acceso a los datos asociados a una tesis tras una búsqueda.

Conclusiones

En este trabajo se proponen nuevas herramientas para superar las limitaciones encontradas en los mecanismos actuales de mantenimiento y acceso a RI. Las herramientas propuestas se sustentan en métodos inspirados en técnicas de Recuperación de Información e Inteligencia Artificial. La estrategia de búsqueda contextualizada, tanto en la generación de sugerencias para la catalogación como para la búsqueda de recursos relevantes, se considera uno de los aportes más importantes de este trabajo.

Una de las posibles extensiones del trabajo realizado consistiría en el uso de recursos adicionales para mejorar las búsquedas y la generación de sugerencias. Por ejemplo, contar con un tesauro conteniendo vocabulario controlado semánticamente permitiría realizar búsquedas más representativas o generar sugerencias más complejas. Por ejemplo, los descriptores sugeridos podrían consistir en conceptos o términos compuestos tales como “evolución humana” en lugar de las palabras simples “evolución” y “humana”.

Otra posible mejora consistiría en utilizar conjuntos de datos sobre libros y artículos (es decir, no solo asociados a tesis de posgrado) con el fin de ampliar el rango de sugerencias de descriptores, códigos Dewey y temáticas que pueden ser identificadas mediante la búsqueda contextualizada.

Por último, una extensión posible es orientar la aplicación a otros idiomas distintos al español. Esto se lograría simplemente utilizando datos de las tesis en el idioma correspondiente, así como una lista de stopwords en dicho idioma.

Agradecimientos

Agradecemos a Jerónimo Spadaccioli, Fernando A. Martínez y Ricardo A. Piriz del Área de Sistemas de la Biblioteca Central de la Universidad Nacional del Sur y a Nélida Benavente, Guillermina Castellano y Marta Ibarlucea, personal no docente de dicha Biblioteca, por su tiempo, su excelente predisposición y sus ideas, las que han enriquecido enormemente este trabajo.

Bibliografía

Armstrong, R., Freitag, D., Joachims, T., & Mitchell, T. (1995). WebWatcher: A learning apprentice for the World Wide Web. In AAAI spring symposium on information gathering (pp. 6–12).

Baeza-Yates R. & Ribeiro-Neto B. (1999) Modern Information Retrieval. Addison-Wesley.

Baldonado, M. Q. W., & Winograd, T. (1997). SenseMaker: an information-exploration interface supporting the contextual evolution of a user's interests. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 11–18). ACM Press.

Buckland, M. (1992). Redesigning Library Services, A Manifesto. American Library Association, Chicago, 1992.

Budzik, J., Hammond K. J., and Birnbaum L.(2001) Information Access in Context. Knowledge Based Systems, 14(1–2):37–53.

Gospodnetic, O., Hatcher, E. McCandless M., (2009). Lucene in Action (2nd ed.). Manning Publications.

Janert, P. K. (2002). Learning and using Jakarta Digester. O'Reilly Media, Inc.

Kraft, R., Chang, C. C., Maghoul, F., & Kumar, R. (2006). Searching with Context. In WWW'06: Proceedings of the 15th International Conference Wide Web, pages 477–486, New York, NY, USA. ACM Press.

Leake, D. (1996). CBR in Context: The Present and Future", In Leake, D., editor, Case-Based Reasoning: Experiences, Lessons, and Future Directions. AAAI Press/MIT Press, 1996, 1-30.

Levy, D. M. & Marshall, C. C. (1995). Going digital: a look at assumptions underlying digital libraries Communications of the ACM. 38(4): 77-84. New York, NY, USA. ACM Press.

Lieberman, H. (1995). Letizia: An agent that assists Web browsing. In C. S. Mellish (Ed.), Proceedings of the fourteenth international joint conference on artificial intelligence. ijcai-95 (pp. 924–929). Montreal, Quebec, Canada : Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

Lorenzetti, C. M., & Maguitman, A. G. (2009) A Semi-Supervised Incremental Algorithm to Automatically Formulate Topical Queries. Information Sciences, 179(12):1881–1892. Special Issue on Web Search.

Maglio, P. P., Barrett, R., Campbell, C. S., & Selker, T. (2000). SUITOR: an attentive information system. In Proceedings of the 5th international conference on intelligent user interfaces (pp. 169–176). ACM Press.

Maguitman, A., Leake, D., Reichherzer, T., & Menczer, F. (2004). Dynamic extraction of topic descriptors and discriminators: Towards automatic context-based topic search. In Proceedings of the thirteenth conference on information and knowledge management (CIKM) (pp. 463—472). Washington, DC : ACM Press.

Maguitman, A., Leake, D., & Reichherzer, T. (2005). Suggesting novel but related topics: towards context-based support for knowledge model extension. In IUI '05: Proceedings of the 10th International Conference on Intelligent User Interfaces (pp. 207–214). New York, NY, USA : ACM Press.

Osiński S., Stefanowski, J., Weiss D. (2004). Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. Advances in Soft Computing, Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM´04 Conference (pp. 359-368). Zakopane, Poland.

Osiński S., Weiss D. (2005). Carrot2: Design of a Flexible and Efficient Web Information Retrieval Framework. Springer Lecture Notes in Computer Science, vol. 3528, (pp. 439-444), Proceedings of the third International Atlantic Web Intelligence Conference (AWIC 2005), Łodź, Poland.

Rhodes, B. & Starner, T. (1996). The remembrance agent: A continuously running automated information retrieval system. In The proceedings of the first international conference on the practical application of intelligent agents and multi agent technology (PAAM '96) (pp. 487–495). London, UK.

Rousseeuw, P.J.; Kaufman, L. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley.

Salton, G., Wong, A & Yang C. S. (1975), A Vector Space Model for Automatic Indexing,"Communications of the ACM, 18 (11): 613–620.

Scholer, F., & Williams, H. E. (2002). Query Association for Effective Retrieval. In Proceedings of the Eleventh International Conference on Information and Knowledge Management (pp. 324–331). New York, NY, USA: ACM Press.

Sebastiani, F. (2002), Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47.

Sølvberg, I. T. (2001) Digital libraries and information retrieval, Lectures on information retrieval, Springer-Verlag New York, Inc., New York, NY.