SulaIR: Una Plataforma Interactiva para la Enseñanza y Aprendizaje de la Recuperación de Información

Juan M. Fernández-Luna[1], Juan F. Huete[1], Patricia Olvera[1], Adrián Peña[1], M. del Carmen Rodríguez-Hernández[2], Julio C. Rodriguez-Cano[3]

[1]Departamento de Ciencias de la Computación e Inteligencia Artificial
E.T.S.I. Informática y de Telecomunicación, CITIC-UGR
Universidad de Granada, 18071 – Granada, Spain
[2] Departamento de Informática y Matemática. Universidad de Holguín, Cuba.
[3] Centro de Desarrollo Territorial de Holguín. Universidad de Ciencias Informáticas, Cuba
{jmfluna,jhg}@decsai.ugr.es, olverar@correo.ugr.es, aderiand@gmail.com, mrodriguezh@facinf.uho.edu.cu, jcrcano@uci.cu

Resumen En este trabajo presentamos la herramienta *SulaIR*[1], una aplicación diseñada con el objetivo general de ayudar a los dos actores participantes en el proceso de formación (profesor y alumno), en la enseñanza y el aprendizaje de la Recuperación de Información. *SulaIR* es una aplicación fácil de utilizar, donde se visualiza el proceso completo de la recuperación y permite además al usuario interactuar con cada una de sus etapas.

1. Introducción y Motivación

La Recuperación de Información (RI), disciplina que queda enmarcada dentro del campo de la Informática y la Documentación [1], está tomando mucho auge en nuestros días debido a la necesidad de organizar y acceder de una manera eficiente a la ingente cantidad de información que se pone a disposición de los usuarios en Internet. Además, en cualquier institución de tamaño mediano el volumen de documentos que se maneja suele ser muy alto y se necesita gestionarlos con técnicas específicamente diseñadas para tal fin, y por tanto, los conocimientos en esta área cada vez son más requeridos en el mundo empresarial y académico.

En los estudios que ya se están extinguiendo de Ingeniería Informática y Licenciatura de Documentación, apenas si incluían asiganturas que describieran los fundamentos de esta disciplina. Esta situación está cambiando lentamente con la implantación de los nuevos grados de Informática y Documentación, incluso en los másteres, tanto de investigación como profesionalizantes, podemos

[1] Adaptación del topónimo Sulayr ("montaña del sol"), nombre con el que los hispano–árabes denominaban a Sierra Nevada, incorporándole las siglas IR, referidas a la traducción en inglés del término Recuperación de Información (Information Retrieval).

ir encontrando materias que tratan de los fundementos como de las tendencias actuales de investigación.

Por otro lado, la enseñanza teórica de la RI se ha centrado en una pura transmisión de las técnicas principales para resolver los problemas más relevantes relacionados con ella, utilizando como base clásica el libro de texto. Desde un punto de vista práctico, lo normal es emplear bibliotecas software, como *Luke/Lucene*, *Terrier*, *Minion*, *Indri/Lemur*, etc., para que el alumno pueda desarrollar los ejemplos que se han visto en clase. Ya E. Fox en 1998 [6] pone de manifiesto la necesidad de herramientas que apoyen el proceso de enseñanza y aprendizaje.

Este enfoque tiene dos problemas importantes: el primero es que se "obliga" al alumno a entender la RI sólo desde una perspectiva teórica, antes de poder aplicarla en sus prácticas; y el segundo, es que la parte práctica supone, además del hecho de entender los problemas, soluciones y técnicas del objeto de estudio, emplear bibliotecas software complejas en lenguajes de programación para implementar las soluciones correspondientes, con la dificultad añadida que ello conlleva (estudio de nuevos lenguajes, de la forma de utilizar las bibliotecas de clases, etc.). Todo esto incrementa la complejidad del aprendizaje y hace que el alumno no se centre en la problemática relacionada con la RI, sino que su atención quede diluida por la necesidad de adquirir otros conocimientos previos y externos necesarios para lograr el objetivo final, el cual, teniendo en cuenta estos condicionantes, en pocos casos se alcanza.

Por tanto, se hace necesario contar con un software que permita asentar los principales conceptos estudiados en teoría sobre RI, utilizable tanto por el profesor como por el estudiante, sin necesidad de que este último tenga conocimientos de programación. Con ella el alumno podrá revisar y fijar dichos conceptos de una manera interactiva, experimentando con cada uno de los módulos tantas veces como desee, adaptando su aprendizaje a un ritmo apropiado. Una vez adquiridos dichos conocimientos esenciales, entonces el alumno estará capacitado para realizar otras tareas prácticas de mayor complejidad. Entre éstas, y para aquellos que tengan un nivel de programación aceptable, podrán extender a *SulaIR* con nuevos módulos y funcionalidades.

SulaIR está disponible en http://utai-citic91.ugr.es/sulaIR/index.html para su descarga y uso público. Es una herramienta de software libre, desarrollada con *Java* como aplicación de escritorio. Los objetivos específicos que se persiguen son:

- Apoyar al alumno en la tarea del aprendizaje de los conceptos fundamentales de la RI mediante la interacción sencilla y amigable con un sistema de RI didáctico.
- Facilitar al docente su labor de enseñanza de la disciplina de la RI mediante el soporte de las nuevas tecnologías.
- Desarrollar una plataforma fácil de manejar e intuitiva, disponible permanentemente, accesible desde Internet, de cualquier tipo de ordenador o sistema operativo, y extensible con nuevos módulos.

Este trabajo queda organizado como sigue: tras esta introducción y motivación, la sección 2 ofrece una visión general de *SulaIR*; la sección 3 revisa algunas experiencias que se han llevado a cabo para apoyar la docencia de la RI, y finalmente concluye con las conclusiones y trabajos futuros.

2. Breve Descripción de SulaIR

SulaIR es una herramienta que permite visualizar el proceso completo de RI e interactuar con el sistema en cada una de las etapas en que podemos dividir dicho proceso (preprocesado de documentos, indexación, consulta, recuperación y realimentación por relevancia). El usuario podrá ver en cada etapa qué acciones realiza internamente el sistema de RI, llevándolas a cabo paso a paso y pudiendo parar el proceso y retomarlo en cualquier momento. También permitirá inspeccionar en detalle el fichero invertido creado. La interfaz de usuario y la colección pueden gestionarse en español o inglés.

Más concretamente, podemos hablar de tres módulos principales:

1. *Preprocesamiento de la colección.* Inicialmente está preparado para indexar documentos XML de la colección Wikipedia, por tanto, el primer paso será eliminar marcas y obtener un texto plano. Seguidamente, se procederá a realizar un análisis léxico, eliminación de palabras vacías (existen listas por defecto, pero el usuario puede introducir la suya propia) y segmentación (stemming). En estas tareas el usuario verá cómo palabra a palabra se va procesando todo el texto hasta obtener un texto final formado por las raíces de las palabras originales que no se han eliminado en procesos previos (en caso de que haya elegido aplicar stemming).
2. *Indexación.* Este módulo es el encargado de la construcción y visualización del índice invertido de la colección. Desde el punto de vista de la implementación de esta parte, no se ha utilizó ninguna de las bibliotecas actualmente existentes, con el objetivo de ganar flexibilidad a la hora de integrar la visualización del proceso con las estructuras de datos del índice. Cuando el usuario está indexando una colección, se observa cómo se van creando las listas de documentos donde aparecen los diferentes términos de la colección (Figura 1). Además se muestra cómo se calculan los pesos. Una vez concluido este proceso, el usuario puede navegar por el índice y obtener gráficos con estadísticas básicas de la colección, términos y documentos.
3. *Recuperación.* El modelo de recuperación que implementa *SulaIR* es el Espacio Vectorial. En este caso, el usuario formula una consulta y, paso a paso, puede observar cómo se va procesando cada término de la misma, cómo se van realizando las acumulaciones en los documentos donde aparecen dichos términos, y cómo se calcula finalmente el grado de relevancia. Por último, los documentos se muestran ordenados de forma decreciente según dicho valor y el usuario puede inspeccionar su contenido (Figura 2). Alternativamente, el módulo de consulta se ha replicado para que funcione como un buscador normal, sin que el usuario tenga posibilidad de ver el funcionamiento interno

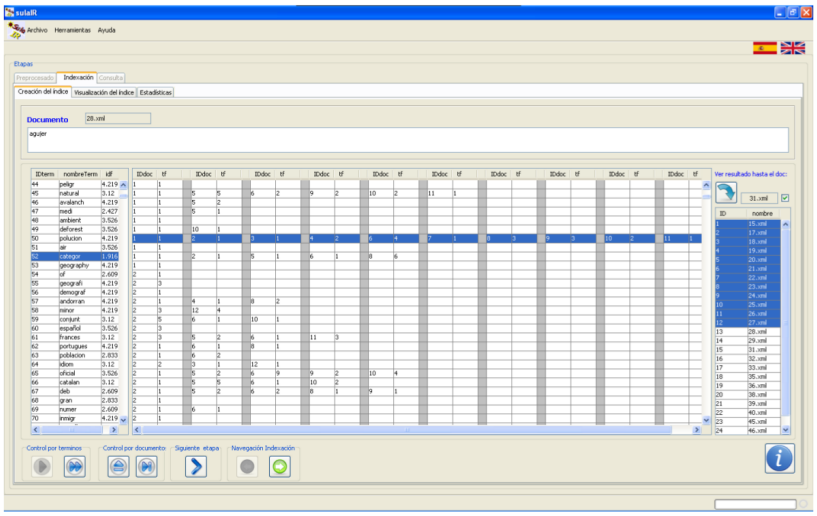

Figura 1. Ventana de visualización del proceso de creación del índice

de la recuperación. Por último, el módulo de recuperación ejemplifica la realimentación por relevancia, basada en el algoritmo de Rocchio, modificando los valores de los parámetros que lo controlan, observando cómo se hace la expasión de consultas y el repesado de los términos.

Los usuarios pueden avanzar en el proceso, desde el preprocesado hasta la realimentación, parar en cualquier momento, almacenar el trabajo realizado y retomarlo desde donde se dejó, o alternativamente, empezar el proceso de nuevo.

3. Otros Sistemas de Apoyo a la Enseñanza y al Aprendizaje de la RI

En la literatura relacionada con la enseñanza y el aprendizaje de la RI (véase [2] para una revisión detallada del estado del arte en este campo) aparecen claros ejemplos que intentan afianzar los conceptos estudiados en teoría antes de pasar a las prácticas por medio de sistemas de apoyo a la docencia. El primer ejemplo es *IR Toolbox* [3], una herramienta experimental de enseñanza para aprender sobre sistemas de RI. El estudiante puede simular todo el proceso de recuperación sin tener que programar, aunque tiene una interfaz de usuario muy básico (basado en mandatos) y es poco flexible al ser difícil añadir nuevos módulos.

El segundo ejemplo [4] se centra sólo en la formulación de consultas, exactamente en el entrenamiento del alumno en la creación de consultas más precisas a la necesidad de información del usuario. Para este fin, requieren al alumno que formule consultas sobre un tópico dado. En la colección de prueba sobre la que trabaja el sistema de RI, se conocen los juicios de relevancia (los documentos

Figura 2. Ventana de visualización del proceso de recuperación

relevantes a dicho tópico). Tras evaluar los resultados de la consulta con respecto a dichos juicios, el alumno conoce el rendimiento de su consulta, y si es malo, formula otra más cercana a la temática entre manos, refinando así su consulta y adquiriendo buenos hábitos a la hora de buscar información.

Finalmente, destacar el trabajo realizado por Brusilovsky y col. [5], los cuales han desarrollado una aplicación web basada en la interacción y visualización de los modelos Booleano y Vectorial. Este sería el trabajo más cercano al nuestro. Las principales diferencias con *SulaIR* es que la configuran como demos aislados en páginas Web y que la forma de visualizar la recuperación es muy diferente a la nuestra. *SulaIR* es más interactivo y permite tener una visión global del proceso.

4. Conclusiones y Trabajos Futuros

Este trabajo ha presentado la plataforma para la enseñanza–aprendizaje de la RI, *SulaIR*, la cual permite al docente ejemplificar el proceso completo de recuperación y al alumno poner en práctica los conceptos principales de esta disciplina. La aplicación se basa en una visualización sencilla e intuitiva de los procesos principales que intervienen en la recuperación documental.

Actualmente se está trabajando en añadir a *SulaIR* un módulo de búsqueda booleana y otro de modelos del lenguaje para enseñar cómo funcionan dichos modelos de recuperación; un módulo que ejemplifique el funcionamiento de una araña de un motor de búsqueda para la Web, otro que integre la técnica presentada en [4] para el entrenamiento de la formulación de consultas y dos que implementen las tareas de agrupamiento y clasificación documental. También

se modificará el módulo de preprocesado e indexación para que pueda indexar varios tipos de documentos, en especial aquellos que sigan el formato TREC. Por último, también se tiene previsto realizar un estudio con estudiantes sobre la usabilidad de *SulaIR*.

Agradecimientos. Este trabajo ha sido cofinanciado por la Universidad de Granada mediante los proyectos de innovación docente "SulaIR: Una plataforma de apoyo al aprendizaje y la enseñanza de la recuperación de información" (2010-20), de 2010, y "SulaIR II: Una plataforma de apoyo al aprendizaje y la enseñanza de la recuperación de información" (2011-206), de 2011, y el Ministerio de Ciencia de Innovación mediante el proyecto TIN2008-06566-C04-01.

Referencias

1. F. Cacheda, J.M. Fernández-Luna and J.F. Huete (2011). Recuperación de Información. Un enfoque práctico y multidisciplinar. Ra-Ma.
2. J.M. Fernández-Luna, J.F. Huete, A. Macfarlane, E. Efthimiadis (2009). Teaching and learning in information retrieval. Information Retrieval, 12(2): 201–226.
3. E. Efthimiadis and Nathan G. Freier (2007). IR-Toolbox: an experiential learning tool for teaching IR. SIGIR 2007, 914.
4. E. Halttunen and E. Sormunen (2000). Learning information retrieval through an educational game. Is gaming sufficient for learning? Edu. for Inf., 18, 289-311.
5. P. Brusilovsky, J.W. Ahn, and E. Rasmussen (2010). Teaching Information Retrieval with Web-based Interactive Visualization. Journal of Education for Library and Information Science 51 (3), 187–200.
6. E. A. Fox. Effects on Education, and a Proposal for Collection of Tools. IR Tools Workshop (http://fox.cs.vt.edu/talks/1998/IRtools.htm).