Actas de las XX JENUI. Oviedo, 9-11 de julio 2014
ISBN: 978-84-697-0774-6
Páginas: 141-148

MiniTREC: un modelo de aprendizaje basado en proyectos para la asignatura de Recuperación de Información.

Javier Lacasta
Depto. de Informática. e Ing. de Sistemas.
Universidad de Zaragoza
Zaragoza
jlacasta@unizar.es

Javier Nogueras-Iso
Depto. de Informática. e Ing. de Sistemas.
Universidad de Zaragoza
Zaragoza
jnog@unizar.es

Resumen

La asignatura de Recuperación de información es una asignatura de nueva impartición dentro de la mención de Computación del Grado en Ingeniería Informática en la Universidad de Zaragoza. La complejidad e interrelación de los contenidos de esta asignatura dificulta la realización de prácticas en las que los alumnos prueben los modelos y algoritmos descritos en teoría. Para facilitar la asimilación de los conceptos teóricos se ha decidido utilizar una metodología de aprendizaje basado en proyectos. Este artículo describe la experiencia de aplicación de dicha metodología, los resultados obtenidos, problemas encontrados y áreas de mejora.

Abstract

Information Retrieval is a new course of the Computer Science Degree Program in Computing Engineering Studies at the University of Zaragoza. The complexity and interrelation of this course's contents makes difficult the definition of practices that allow students to test the models and algorithms described in theory classes. To facilitate the understanding of concepts, it was decided to use a project based learning methodology. This paper describes the application experience of this methodology, the results obtained, problems found and areas of improvement.

Palabras clave

Recuperación de Información, Crawlers, Búsqueda en la Web, Web Semántica, Ontologías

1. Introducción

La recuperación de información es un campo de la ciencia de la computación centrado en proponer soluciones para encontrar recursos (documentos, imágenes, etc.) de naturaleza heterogénea que satisfagan una cierta necesidad de información dentro de grandes colecciones de datos [6, 5]. Este campo del conocimiento ha ido ganando relevancia con el paso de los años con el aumento de repositorios digitales de información y la necesidad de mejores sistemas de clasificación y búsqueda de información.

Con el rediseño de los planes de estudios para su adecuación al Espacio Europeo de Educación Superior, se vio la necesidad de incluir una nueva asignatura dentro del Grado de Ingeniería en Informática de la Universidad de Zaragoza que cubriera esta área del conocimiento (no impartida en el plan de estudios anterior). La nueva asignatura se añadió como cierre de la materia obligatoria de Aprendizaje y Recuperación de Información dentro de la mención de computación.[1] En Aprendizaje Automático, el alumno ha recibido formación sobre las técnicas de clasificación de información más comunes. El planteamiento de Recuperación de Información es completar la formación del alumno con las técnicas para localizar recursos de interés sobre grandes fuentes de datos locales o en la web [6]. En la asignatura también se incluyen los conceptos y técnicas básicas de la Web Semántica para diseñar sistemas de recuperación semánticos [1].

Esta asignatura describe los métodos, medidas y arquitecturas más comúnmente usados en la construcción de sistemas de recuperación de información fiables y eficientes. Sin embargo, la construcción de un sistema de recuperación funcional, por muy sencillo que sea, requiere la adecuada integración de muchos de estos elementos. Esto dificulta la propuesta de trabajos prácticos simples en los que los alumnos puedan experimentar con los conceptos descritos en las clases teóricas. En este contexto, se vio que el modelo de aprendizaje basado en proyectos [7] era ideal para la realización de los trabajos prácticos de la asignatura.

Esta metodología se empezó a usar hace más de 40 años en la enseñanza de la medicina para fomentar

[1] http://titulaciones.unizar.es/asignaturas/30233/index13.html

la motivación de los estudiantes [12]. Actualmente se considera especialmente adecuada en el contexto de la educación superior [13] y lleva años aplicándose en la enseñanza de la informática en múltiples asignaturas [2, 9]

En la enseñanza de recuperación de información hay numerosos trabajos acerca de contenidos, estructura de la asignatura, y descripción de los trabajos que tienen que realizar los alumnos [4, 8, 3]. Sin embargo, no se han encontrado trabajos que se centren en la aplicación de una metodología de aprendizaje basado en proyectos a esta asignatura.

Este artículo describe la experiencia de aplicación de dicha metodología el primer año de impartición de la asignatura. Se describe la planificación del proyecto, tareas, resultados obtenidos, problemas encontrados y áreas de mejora. Este análisis puede ser de utilidad a otros docentes que quieran implementar una metodología basada en proyectos en una asignatura con una estructura similar a la aquí descrita.

Se decidió realizar un único proyecto a lo largo del curso, llamado MiniTrec, en el que cada alumno construyera y evaluase diferentes sistemas de recuperación de información. Para que los sistemas desarrollados fueran comparables se estableció que todos los alumnos usasen la misma colección de documentos (trabajos académicos del repositorio digital de la Universidad de Zaragoza) y respondieran a las mismas necesidades de información. La comparación de los sistemas realizados facilita la discusión sobre la idoneidad de las técnicas utilizadas para la resolución del problema propuesto. Dada la complejidad y amplitud de las tareas, se decidió que el trabajo se realizase en grupos de dos personas.

La sección 2 describe las principales características del proyecto propuesto. Después, la sección 3 detalla cada una de las tareas que componen el proyecto. El artículo termina con una descripción de los resultados obtenidos a través de esta experiencia docente durante el curso 2013-2014, y unas conclusiones sobre dicha experiencia.

2. Descripción del proyecto

Tal y como se ha comentado en la introducción, el objetivo del proyecto es que los alumnos desarrollen y evalúen distintos sistemas de recuperación de información. Para permitir la comparación de resultados todos ellos procesan la misma colección de datos y resuelven las mismas necesidades de información, pero usan aproximaciones técnicas distintas. Concretamente, se estableció el desarrollo de los siguientes sistemas:

- Sistema de recuperación de información tradicional sobre texto, basado en la librería Lucene de Apache.
- Sistema de recuperación web, basado en las herramientas Nutch y Solr de Apache.
- Sistema de recuperación semántico, que transforme los documentos de la colección en descripciones semánticas de recursos (RDF [10]), y permita el uso de un lenguaje de consulta semántico (SPARQL [11]).

Para la evaluación de la calidad de los sistemas creados, al igual que en las conferencias TREC[2] se decidió construir un banco de pruebas que fijasen un conjunto de necesidades de información y se indicase las respuestas adecuadas a cada una de ellas. Esto permite que los alumnos puedan conocer directamente lo bien o mal que se comportan los sistemas que han desarrollado. Además, esto permite que se puedan comparar entre si usando las medidas numéricas comúnmente establecidas.

Dicho banco de pruebas no se ha proporcionado como parte del material de la asignatura, sino que ha sido construido por los alumnos de forma cooperativa para que adquieran experiencia en la construcción de dichos bancos de pruebas. A través de varias tareas, los alumnos han recopilado (entre todos) un conjunto de necesidades de información, y las valoraciones (juicios de relevancia) de los documentos de la colección según esas necesidades.

Dada la amplitud del trabajo propuesto, se ha establecido su peso en la evaluación de la asignatura en un 50 %. El otro 50 % corresponde a un examen teórico. Además se ha añadido la restricción de tener que obtener una calificación superior a 5 en el trabajo para aprobar la asignatura. Dentro del trabajo se ha establecido un baremo del peso de cada una de las tareas:

- Propuesta de necesidades de información para el banco de pruebas: 5 %.
- Establecimiento de juicios de relevancia sobre un subconjunto de la colección: 5 %.
- Construcción de un sistema de recuperación tradicional sobre la colección: 30 %.
- Construcción de un sistema de recuperación web sobre la colección: 20 %.
- Construcción de un sistema de recuperación semántico sobre la colección: 30 %.
- Evaluación de los sistemas desarrollados: 10 %.

La planificación del trabajo está condicionada por la organización del contenido teórico de la asignatura. Por ejemplo, nos pareció más coherente ver en los primeros temas las técnicas de recuperación tradicionales y explicar después la evaluación de los sistemas, ya que así se podía poner en contexto. Esto ha implicado que se haya tenido que construir el sistema de recuperación tradicional antes de tener construido el banco de prue-

[2] Text Retrieval Conference. `http://trec.nist.gov/`

Figura 1: Portal *Zaguán* de la Universidad de Zaragoza.

bas. La Tabla 1 muestra el conjunto de hitos de entrega de información/datos por parte de los profesores y de resultados por parte de los alumnos. Puede observarse que esta planificación llega hasta la semana 20. Esto es debido a que se cuentan las semanas de vacaciones del primer cuatrimestre, y a que se ha decidido que la última tarea se entregue al finalizar el periodo de exámenes, en vez de al principio.

La colección de recursos que hemos elegido para que los alumnos procesen ha sido el Repositorio Digital de la universidad de Zaragoza. Esta colección está disponibles a través del Portal Web *Zaguán*.[3] Este portal es el resultado de un proyecto de digitalización cuyo objetivo es manejar el contenido digital en formato textual (PDF, en su mayoría) de la Universidad de Zaragoza en una única plataforma, de forma que puedan realizarse búsquedas por colecciones. El proyecto comenzó en 2008 e incluye una colección de copias digitales del Fondo Antiguo. Las líneas de trabajo actuales pasan por incluir, además, Trabajos Académicos, libros, tesis, informes, pre-prints y artículos. En la figura 1 se puede ver la captura del portal mostrando la colección de Trabajos Académicos.

De todo el contenido de *Zaguán* se decidió utilizar las siguientes colecciones:

- **Trabajos Académicos**:[4] Aquí se incluyen Proyectos Final de Carrera (trabajos de fin de estudios anteriores al plan Bolonia), Trabajos de Fin de Grado (trabajos de estudios de los grados actuales del plan Bolonia) y Trabajos de Fin de Master (trabajos de fin de estudios de Master).
- **Tesis Doctorales**:[5] Contiene las tesis doctorales de la Universidad de Zaragoza desde el curso 2008/2009.

[3] `http://zaguan.unizar.es/`
[4] `http://zaguan.unizar.es/collection/Trabajos\%20academicos`
[5] `http://zaguan.unizar.es/collection/Trabajos\%20academicos`

Estas colecciones agrupan un número razonable de documentos (3449) que están descritos de forma bastante detallada a través de fichas bibliográficas (metadatos). Dichas fichas contienen descripciones semiestructuradas en formato MARCXML[6]. Esto permite la construcción de sistemas de recuperación completamente funcionales que apliquen las diferentes técnicas descritas a lo largo de la asignatura. Como alternativa al uso de las fichas bibliográficas, se pensó en usar directamente los documentos de las memorias que están en formato PDF. Finalmente se descartó esta opción por la complejidad añadida de tener que procesar los PDF. Hay librerías que lo realizan, pero aun así supone una carga de trabajo adicional nada despreciable para el alumno.

Inicialmente se planteó que el acceso a las fichas de la colección lo realizasen los alumnos a través de un proceso de crawling del repositorio, o a través de los protocolos establecidos a tal efecto. Por ejemplo, *Zaguán* usa el protocolo OAI / PMH (Open Archives Initiative - Protocol for Metadata Harvesting).[7] Sin embargo, se vio que eso presentaba serios problemas en la organización de la asignatura ya que impedía que los alumnos tuvieran acceso a la colección de datos hasta haber visto dichos sistemas (finales de Noviembre). Teniendo todo esto en cuenta, la descarga de las fichas la realizamos los profesores, y a los alumnos se les proporcionó directamente la colección en ficheros de texto. Para mantener la relación entre los ficheros descargados y su url original, se nombró los ficheros descargados de forma que cada fichero incluyese el identificador del registro en la colección Zaguán. Esto permite deducir fácilmente la URL de consulta del documento en el portal *Zaguán* y que los alumnos puedan acceder a los datos completos originales. Por ejemplo, si el documento se denomina oai_zaguan.unizar.es_1877.xml, la URL de consulta en el portal *Zaguán* es: `http://zaguan.unizar.es/record/1877`.

3. Tareas del proyecto

3.1. Selección de las necesidades de información

En vez de proporcionarles directamente a los alumnos el conjunto de necesidades de información, se planteó que fueran ellos mismos los que las definieran. Esto se ha hecho teniendo como objetivo que analizaran en detalle la colección de datos, se plantearan como sería mejor publicarla, y pensaran en qué tipo de consultas serían más útiles para los posibles usuarios de su sistema.

[6] `ttp://www.loc.gov/standards/marcxml/`
[7] `http://www.openarchives.org/pmh/`

Hitos de los alumnos (equipos)	Hitos de los profesores
Primera convocatoria	
	Semana 3: Asignación del trabajo.
	Semana 6: Entrega de la colección de documentos sobre la que ejecutar los sistemas de recuperación de información.
Semana 7: Entrega de las necesidades de información.	
	Semana 7: Entrega de las necesidades de información con las que se evaluarán los sistemas.
Semana 11: Entrega del sistema de recuperación tradicional (trabajando sobre una colección dada) y el listado ordenado de documentos obtenido por cada necesidad de información.	
	Semana 11: Asignación del subconjunto de documentos que hay que juzgar por cada equipo y necesidad de información.
Semana 13: Entrega de los juicios de relevancia.	
Semana 14: Entrega del sistema de recuperación web.	
	Semana 14: Entrega de los juicios de relevancia consensuados.
Semana 20: Entrega final (sistema de recuperación semántico, programa de evaluación y breve memoria sobre los resultados de la evaluación).	
Segunda convocatoria	
Día del examen: Fecha límite para la entrega de trabajos en segunda convocatoria.	

Cuadro 1: Hitos del trabajo

```
<?xml version="1.0" encoding="UTF-8"?>
<informationNeeds>
  <informationNeed>
    <identifier>1</identifier>
    <text>¿Qué articulos existen sobre TSS
    (Time Sharing System), un sistema
    operativo para procesadores IBM?
    </text>
  </informationNeed>
</informationNeeds>
```

Figura 2: Ejemplo de necesidad de información.

Se le pidió a cada grupo que propusiera 5 necesidades de información para la colección de datos usada. Dichas propuestas tenían que realizarse mediante una descripción textual en lenguaje natural que sería la entrada de los sistemas de recuperación posteriormente implementados. Como ejemplo de necesidades de información se les proporcionó la lista de temas de las conferencias TREC[8], y un ejemplo del formato en el que tenían que describirse las necesidades (ver Figura 2). De entre todas las propuestas recogidas los profesores seleccionamos las 7 que nos parecieron más interesantes. Aquí se valoró tanto el tema de la consulta como los elementos involucrados, complejidad, errores que incluían, e idioma.

[8] http://trec.nist.gov/data/topics_eng/index.html

3.2. Juicios de relevancia

Para poder medir la bondad de un sistema de recuperación es necesario contar con unos juicios de relevancia de cada elemento de la colección respecto a cada una de las necesidades de información definidas. Dichos juicios permiten usar diferentes métricas que miden la calidad de los resultados obtenidos por cada sistema de recuperación.

La generación de dichos juicios es una labor costosa, ya que implica revisar toda la colección de documentos para todas las consultas. Por tanto su creación se planteó como una tarea colaborativa para realizar por los alumnos. Para limitar aún más el trabajo que tenían que realizar, se redujo la revisión a la unión de los 30 primeros documentos devueltos por los alumnos en sus sistemas de recuperación tradicional (dada la planificación se construyeron primero). Aquí se tomó como hipótesis que entre esos primeros documentos estarían todos los relevantes. El resto de documentos se consideraron no relevantes. Naturalmente, esta hipótesis no es correcta, pero ha permitido limitar mucho el número de documentos que hay que evaluar haciendo la tarea abordable. La única consecuencia negativa de esta aproximación es la distorsión a la baja en algunas de las medidas de calidad.

Dichos documentos se proporcionaron a los alumnos para que determinaran su relevancia de forma binaria (relevante o no relevante). Cada equipo emitió juicios de relevancia sobre pequeños subconjuntos del

conjunto seleccionado. Los documentos se distribuyeron entre los alumnos de forma que cada documento tuviera al menos dos juicios para cada consulta. Esto supuso alrededor de 200 evaluaciones por alumno. Al combinar los juicios, si al menos uno de ellos era positivo, se consideró que el documento analizado era relevante para la necesidad de información indicada.

Para estandarizar la tarea de evaluación se le proporcionó a los alumnos un documento Excel indicando los documentos sobre los que tenía que emitir los juicios. En este fichero cada fila contenía el identificador de una necesidad de información, el nombre del fichero con la ficha MARCXML de un documento potencialmente relevante, la URL del mismo documento (para visualizarlo directamente en *Zaguán*), y una casilla vacía para introducir si el documento es relevante para la consulta (1 es relevante, 0 no lo es).

Esta aproximación para realizar los juicios de relevancia presenta el problema que depende de la colaboración de todos los alumnos para funcionar. Si faltan juicios, los datos asociados no pueden ser usados en la evaluación. Por lo tanto, se impuso la condición de que su entrega en la fecha indicada era condición indispensable para la evaluación del resto de partes del trabajo.

3.3. Sistema de recuperación tradicional

La tarea que se les planteó a los alumnos fue la de crear un sistema de indexación y búsqueda sobre la colección de *Zaguán* que permitiese responder a las 7 necesidades de información elegidas utilizando las técnicas de recuperación tradicionales vistas en teoría. Para la realización de dicho proyecto se dejó libertad a los alumnos en cuanto a la selección del esquema de indexación/búsqueda y las técnicas (tokenización, normalización, etc.). La única limitación que se impuso es que la tarea de convertir el texto de una necesidad de información en una consulta del sistema de recuperación de información (ej: un conjunto de términos, una consulta booleana, una consulta *fuzzy*, etc.) debía ser realizado por el propio sistema de recuperación de información. Adicionalmente, se les indicó que los programas de indexación y búsqueda debían ser suficientemente flexibles para aceptar como entrada otras necesidades de información con características similares.

Como preparación previa a la construcción del sistema se plantearon 2 prácticas de laboratorio en las que se introdujo a los alumnos en el uso del modelo de recuperación de Lucene y la construcción de índices sobre él. Estas prácticas introducen los conceptos básicos de uso de Lucene, pero es labor del alumno profundizar en su funcionalidad, y carencias para implementar el sistema de recuperación requerido.

Para facilitar la comparación de resultados entre los sistemas construidos, se les exigió que el formato de los resultados fuera un fichero de texto en el que en cada línea contuviera uno de los resultados de una de las consultas como <IDConsulta NombreFicheroRelevante>(ver Figura 3). Además, dichos resultados debían estar ordenados por relevancia respecto a las consultas, ya que esto permite que se puedan hacer análisis más precisos sobre los resultados.

```
02-2    oai_zaguan.unizar.es_5460.xml
02-2    oai_zaguan.unizar.es_6453.xml
...
03-5    oai_zaguan.unizar.es_10751.xml
03-5    oai_zaguan.unizar.es_7131.xml
```

Figura 3: Ejemplo de fichero de resultados.

Aparte del software y los resultados obtenidos, los alumnos tuvieron que entregar una memoria describiendo: las técnicas utilizadas en el proceso de indexación y los índices (campos) creados; las técnicas utilizadas en el proceso de procesamiento de las consultas; y el algoritmo elegido para el cálculo del ranking. El objetivo de la memoria es que razonaran las elecciones de diseño tomadas.

3.4. Sistema de recuperación web

Como complemento al sistema de recuperación tradicional se planteó otra tarea basada en la extensión de los sistemas de recuperación tradicional en la web. Dadas las limitaciones temporales de la asignatura, el objetivo en este caso no fue la construcción de un sistema completo sino la puesta en marcha de uno existente.

En esta tarea se les pidió a los alumnos que configuraran Nutch y Solr para que realizasen la descarga de la web de la colección de Zaguán y su indexación. Como restricción se les indicó que se valoraría que la configuración del crawler fuese lo más restrictiva posible. El objetivo es que solo indexase las páginas relevantes y se evitase la indexación de las copias existentes en diferentes formatos (bibtext, dc, etc.), idiomas, y otras páginas existentes en el servicio.

Como preparación para esta tarea se realizó una práctica de laboratorio en la que se presentaban los sistemas de Nutch (Crawler web) y Solr (Buscador tradicional). Esta práctica solo introduce a los alumnos en el uso de estos sistemas. Es tarea de ellos el profundizar en su funcionamiento para poder construir el sistema de recuperación solicitado.

Dado que Zaguán es un sistema vital de la biblioteca de la universidad, para evitar problemas de sobrecarga por parte de los programas de los alumnos, desde el servicio informático de la biblioteca nos proporcionaron una copia del servicio, accesible solo desde dentro de la red de la universidad.

Toda esta tarea se planteó para que no fuese necesario extender Nutch o Solr, sino que con solo la modificación de ficheros de configuración fuese posible crear el sistema. Sin embargo se dejó la puerta abierta para que si algún alumno desease integrar como plug-in algún proceso de indexación/búsqueda adicional de entre los vistos en teoría, fuese valorado positivamente.

Respecto a la indexación/búsqueda, se les dio libertad a los alumnos para que seleccionasen la configuración que considerasen más adecuada a nivel de operaciones sobre los datos, siempre y cuando Solr fuese capaz de responder a las 7 necesidades de información elegidas. Dado que en esta tarea se trabaja con herramientas finales, se les permitió que la introducción de las necesidades de información fuese de forma manual evitando así tener que integrar el procesamiento de las necesidades de información desarrollado previamente.

Aparte de los ficheros de configuración modificados y el listado de resultados obtenidos, se les pidió a los alumnos la entrega de una memoria con la descripción de la configuración seleccionada y las razones para dicha selección. Adicionalmente, dado que se eliminó la comparación de los resultados con los de los otros sistemas, se les pidió aquí que analizasen la calidad de los 10 primeros resultados obtenidos para cada consulta, y explicasen las razones de su orden.

3.5. Sistema de recuperación semántico

Siguiendo el mismo esquema usado para la construcción de un sistema de recuperación tradicional, se planteó la creación de un sistema de recuperación semántico, usando la misma colección de recursos de *Zaguán* y necesidades de información. En este contexto, la primera tarea que se les pidió a los alumnos fue la definición del modelo de datos semántico que considerasen más apropiado (ej, elementos de RDFS y OWL) para la colección de *Zaguán*. Aquí se les indicó que tenían que definir un modelo que aportase más que una simple conversión de formatos. El siguiente paso fue la creación de un programa que realizase la conversión entre los documentos XML originales en un grafo RDF de acuerdo al modelo definido. La tercera tarea consistió en la definición de las consultas SPARQL adecuadas al modelo semántico definido en base a las necesidades de información seleccionadas. Y el último paso, fue la creación de un programa que ejecutase las consultas en la colección transformada.

Como introducción a la construcción de un sistema de recuperación semántico, se plantearon dos prácticas de laboratorio en las que se presentó a los alumnos en el uso de RDF y OWL como modelos de representación, Jena como sistema de almacenamiento, y SPARQL como lenguaje de consulta. A partir de estas prácticas, los alumnos tenían que profundizar sobre las características de los lenguajes y herramientas para crear el sistema de recuperación solicitado.

Además, se estableció para los resultados el mismo formato que en el caso del sistema tradicional, facilitando de esta manera la evaluación y comparación de ambas alternativas.

Igual que en los anteriores casos, se les pidió una memoria en la que describieran su sistema semántico. En este caso se solicitó que la memoria estuviera centrada en la descripción y discusión del modelo elegido y las consultas realizadas.

3.6. Evaluación de los sistemas desarrollados

Como introducción a las tareas de evaluación de los resultados de un sistema de recuperación de información, se planteó una práctica de laboratorio para introducir las diferentes técnicas de evaluación del rendimiento.

El programa desarrollado por los alumnos en esta práctica les sirvió como base para la evaluación de los sistemas de recuperación tradicional y semántico. El objetivo planteado es que a partir del listado de los juicios de relevancia de los documentos de la colección previamente definido y el listado de resultados generado por cada uno de los dos sistemas realizasen la evaluación de los sistemas. Aquí se dejó libertad a los alumnos para que seleccionaran e implementan las métricas de evaluación que considerasen más apropiadas para cada sistema.

Como resultados, además del programa de evaluación, se pidió que los alumnos entregasen un informe donde indicasen las métricas de evaluación seleccionadas, los resultados obtenidos en cada sistema de recuperación, y un análisis y comparativa de dichos resultados, razonando sobre las ventajas y desventajas de cada sistema y sus diferencias.

4. Resultados de la asignatura

Los 38 alumnos del curso 2013/2014 se organizaron en 19 parejas para realizar el trabajo, de las cuales 16 entregaron el trabajo. Un alumno más presentó el trabajo de forma independiente a su compañero de grupo. Es decir, completaron el proyecto 33 de los 38 alumnos iniciales en 17 grupos distintos. De entre ellos, 3 grupos obtuvieron aprobado, 11 notable y 3 sobresaliente.

Los principales resultados entregados por los alumnos fueron las medidas de relevancia de los documentos de la colección según cada necesidad de información (banco de pruebas), y las medidas de calidad de cada sistema construido según dicho banco de pruebas.

Respecto a los juicios de relevancia, los alumnos realizaron 1212 juicios de relevancia entre todas las ne-

cesidades de información. Concretamente, empezando desde la primera a la última, se realizaron 210, 190, 162, 134, 185, 154 y 177 evaluaciones respectivamente. Tal y como se ha indicado anteriormente, la selección de documentos que había que valorar se realizó escogiendo la unión de los 30 primeros documentos devueltos por cada sistema de recuperación tradicional (este número se eligió por seleccionar una cantidad de documentos razonable). De esos 1212 juicios de relevancia, 958 fueron realizados por 2 jueces (2 equipos distintos). Entre los 958 documentos valorados por dos jueces hubo acuerdo en 821 ocasiones (153 documentos relevantes y 668 no relevantes) originando una proporción de acuerdo del 85,69 %[9]. En el caso de disponer de doble valoración de un documento, para la valoración final de relevancia se ha considerado que un documento es relevante si cualquiera de los jueces lo ha considerado relevante. En total, 290 documentos con doble valoración fueron considerados relevantes. Incluyendo los que solo tenían 1 valoración, se obtuvo un total de 342 documentos valorados como relevantes sobre el conjunto de las 7 necesidades de información.

Inicialmente se quería realizar una comparativa de los tres sistemas desarrollados. Sin embargo, se vio que el trabajo realizado en el sistema de recuperación web sobrepasaba la carga estimada, y que la conversión del formato de salida de Solr en el requerido por la herramienta de validación añadía todavía más carga de trabajo. Por tanto, se decidió comparar solo el sistema tradicional con el semántico.

Los resultados de cada sistema creado por los alumnos usando dicho banco de pruebas han sido heterogéneos. La figura 4 muestra la calidad de cada sistema según su MAP (Media de la precisión promedio), una medida utilizada habitualmente en este contexto. Aunque los resultados son bajos en bastantes de los casos, eso no implica una falta de trabajo por el alumnado, sino que es debido en gran parte a la heterogeneidad de criterios en el establecimiento de los juicios de relevancia. Además, el tiempo disponible para la realización de los trabajos ha limitado las opciones de implementación a las técnicas más básicas de entre las explicadas en las clases de teoría. Algún grupo ha destacado más en sus sistemas pero a costa de implementar aproximaciones mucho más complejas que le han requerido dedicar al proyecto más tiempo del establecido. Dos grupos no aparecen en la tabla al no haber entregado sus sistemas de recuperación.

[9]Asumiendo que la doble valoración proviniese de los dos mismos jueces para cada documento, la medida Kappa de acuerdo entre jueces sería 0,6. Aunque el valor es ligeramente inferior a los recomendados para realizar conclusiones firmes sobre las medidas de relevancia de un sistema, debemos tener en cuenta que en realidad tenemos 17 jueces distintos. Además, aunque el valor Kappa no sea muy alto, las comparaciones relativas de los sistemas implementados por los distintos equipos siguen siendo válidas [6].

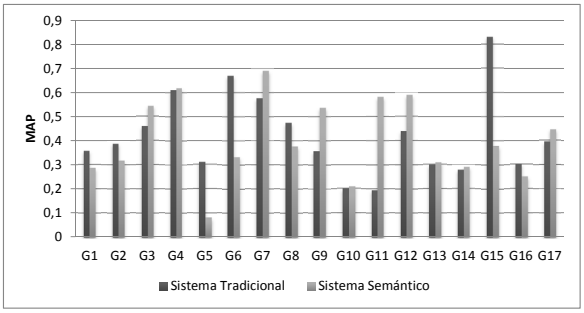

Figura 4: Medida de calidad MAP de los sistemas desarrollados por los alumnos.

En el examen teórico se evaluaron los conocimientos totales descritos a lo largo de la asignatura, los hubieran implementado o no en las diferentes tareas del trabajo. De los 38 alumnos realizaron el examen los 33 que habían presentado el trabajo, y de ellos aprobaron 32. Los resultados finales de la asignatura, ponderando al 50 % las calificaciones de trabajo y examen, han sido: 1 suspenso, 9 aprobados, 20 notables, 2 sobresalientes, y 1 matrícula de honor.

De estos resultados puede extraerse que el método aplicado ha funcionado bastante bien. No solo una gran mayoría de los alumnos han entregado el trabajo, sino que también la mayor parte de ellos han aprobado el examen teórico.

Sin embargo, la puesta en práctica de esta metodología ha supuesto numerosos problemas. Por una parte, la planificación y organización ha sido compleja. Hay una gran dependencia entre tareas y a su vez con los contenidos teóricos. Para poder realizar todo el trabajo deseado, se ha aprovechado que la asignatura se imparte en el primer semestre y disponen del periodo no lectivo de navidad cerca del final del cuatrimestre. Aun así, los plazos de entrega se han tenido que alargar hasta el final del periodo de exámenes, lo cual no es lo más conveniente y se ha notado en una menor calidad de la última entrega. Por otra parte, se han visto problemas de distribución de la carga de trabajo. Hay periodos de tiempo con una carga muy elevada de trabajo y otros en los que los alumnos tienen poco que hacer. Esto es especialmente relevante con la última entrega que se ha realizado después del periodo de exámenes.

El primer problema es de difícil solución, ya que los alumnos no pueden empezar a realizar nada del trabajo hasta haber visto las bases necesarias. Respecto al segundo, nos planteamos proporcionarles un entorno en el que las partes de programación del proyecto a desarrollar menos relevantes para la asignatura sean proporcionadas por los profesores. Con esto en vez de tener que desarrollar un sistema de cero, se podrán centrar en las partes que más aportan.

5. Conclusiones

Este artículo ha presentado el trabajo de la asignatura de Recuperación de Información del Grado en Ingeniería Informática en la Universidad de Zaragoza.

Este trabajo ha propuesto a los alumnos la construcción de diferentes sistemas de recuperación de información y su evaluación mediante un banco de pruebas común para todos los alumnos. Esto ha permitido, que los alumnos pudieran evaluar los sistemas construidos de forma objetiva e imparcial. Los resultados de la puesta en práctica de este proyecto han mostrado que ha servido para lograr que un gran número de los alumnos superasen la asignatura.

El establecimiento de esta metodología de trabajo ha supuesto un reto por el número de tareas y su complejidad. En próximos años queremos pulir los problemas encontrados y proporcionar a los alumnos un entorno más completo de trabajo para que puedan centrarse de forma más completa en el desarrollo de los sistemas, en vez de en los elementos accesorios necesarios para su construcción (herramientas, librerías, etc.).

Después de haber impartido ya el curso por primera vez, en lo referente al trabajo nos planteamos la combinación de la construcción del sistema de recuperación web con el tradicional por ser redundantes en la construcción de los índices, y plantearlo como una extensión en la que simplemente se use el Crawler Nutch para la descarga de la colección. También se ha visto que no funciona de forma del todo correcta la política de selección de las necesidades de información, ya que las proporcionadas este año no cubrían del todo todas las áreas descritas en teoría. Una alternativa que nos planteamos es reducir el número de necesidades de información seleccionadas por alumnos e incluir alguna adicional propuesta por los profesores. Esto permitiría rellenar los huecos dejados en las propuestas de los alumnos. Además, si dichas consultas se mantienen en varios años, permitiría proporcionar a los alumnos los resultados obtenidos en años anteriores como referencia respecto a los sistemas que hayan construido. Finalmente, también se estudiará la posibilidad de proporcionar a los alumnos un módulo de procesamiento de PDFs para que además del contenido de las fichas bibliográficas puedan indexar el contenido de los documentos para mejorar los resultados.

Agradecimientos

Este trabajo ha sido realizado como parte del proyecto PIIDUZ_13_171 de la Universidad de Zaragoza y financiado por el Gobierno de España a través del proyecto TIN2012-37826-C02-01.

Referencias

[1] G. Antoniou and F. van Harmelen. *A Semantic Web Primer, second edition*. MIT Press, 2008.

[2] A. Breiter, G. Fey, and R. Drechsler. Project-based learning in student teams in computer science education. *Facta universitatis - series: Electronics and Energetics*, 18(2):165–180, 2005.

[3] F. Cacheda, D. Fernández, and R. López. Experiences on a practical course of web information retrieval: Developing a search engine. In *Proceedings of the Second Int. Workshop on Teaching and Learning of Information Retrieval*, 2008.

[4] J. M. Fernandez-Luna, J. F. Huete, A. MacFarlane, and E. N Efthimiadis. Teaching and learning in information retrieval. *Information Retrieval*, 12(2):201–226, 2009.

[5] M. A. Hearst. *Search User Interfaces*. Cambridge University Press, 2009.

[6] C. D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, 2009.

[7] T. Markham. *Project based learning handbook*. Buck Institute for Education, 2003.

[8] F. McCown. Teaching web information retrieval to undergraduates. In *41st ACM technical symposium on Computer Science education*, pages 87–91, 2010.

[9] R. Puchera and M. Lehner. Project based learning in computer science - a review of more than 500 projects. In *2nd Int. Conf. on Education and Educational Psychology*, volume 29, pages 1561–1566, 2011.

[10] W3C. RDF primer. Technical report, W3C, 2004.

[11] W3C. SPARQL query language for RDF. Technical report, W3C, 2008.

[12] D. R. Woods. *Problem-based Learning: helping your students gain the most from PBL*. 1996.

[13] D. R. Woods, R. M. Felder, A. R Garcia, and J. E. Stice. The future of engineering education iii. developing critical skills. *Chemical Engineering Education*, 34:108–117, 2000.