Creación de una colección de prueba de literatura científica en español para evaluar sistemas de recuperación de información

Tolosa, Gabriel H.[1]; Bordignon, Fernando R. A., Peri, Jorge A., Banchero Santiago
Universidad Nacional de Luján
Departamento de Ciencias Básicas
{tolosoft, bordi, jperi}@unlu.edu.ar, santigobanchero@gmail.com

[1] Becario de Investigación. Secretaría de Investigación y Postgrado. Universidad Nacional de Luján

Resumen

La evaluación de sistemas de recuperación requiere contar con colecciones de prueba compuestas por un corpus de documentos, un conjunto de necesidades de información (tópicos) y los juicios de relevancia. Éstas permiten evaluar diferentes estrategias y sistemas ya que permiten comprender la naturaleza de los resultados, compararlos con otros y reproducir pruebas en iguales condiciones. El proceso de armado de una colección es una tarea que requiere un importante esfuerzo humano ya que no se puede realizar – de manera completa – automáticamente.

En este trabajo se plantean los lineamientos para la construcción de una colección de prueba en español de dominio público a partir de artículos de investigación en el área de la informática y las ciencias de la computación. La creación de esta colección – destinada a la evaluación la recuperación "ad-hoc" – persigue como primer objetivo poner a disposición de la comunidad universitaria un corpus de documentos semi-estructurados que permita la evaluación de diferentes estrategias de búsqueda. Además, debido a que el tema de recuperación de información se encuentra en pleno crecimiento consideramos que en los próximos años se evaluará su incorporación como tema de grado en diferentes carreras. Es por ello es que creemos que este corpus sería un buen recurso didáctico para realizar tareas de laboratorio.

Un segundo objetivo consiste en recolectar y procesar la mayor cantidad posible de artículos científicos publicados en español y crear una colección mayor que sirva para investigación de diversos aspectos del área de recuperación de información como: extracción de información, clasificación, respuestas a preguntas, resumen automático, entre otros.

Se presenta una metodología para la selección de los documentos, la demarcación de su estructura, la creación de los tópicos y de los juicios de relevancia, junto con una primera prueba con un conjunto reducido de documentos.

Palabras clave: recuperación de información, evaluación, colección de prueba.

1 – Introducción

La evaluación de sistemas de recuperación de información (SRI) requiere contar con colecciones conocidas sobre las cuales se puedan determinar consultas y la relevancia de los documentos respecto de éstas, para luego calcular las métricas correspondientes. Estos juegos – conocidos como colecciones de prueba – se fueron desarrollando con el tiempo y evolucionaron en tamaño y calidad. Los primeros esfuerzos en su creación se deben a Cleverdon, en los denominados *Experimentos Cranfield* en el área aeronáutica entre 1957 y 1968. Si bien estas colecciones contenían unos cientos de documentos, marcaron una línea de trabajo, la cual – en la actualidad – se considera un referente en la evaluación de los SRI (*La Tradición Cranfield*). Una colección de prueba para evaluar recuperación "ad-hoc" requiere de tres componentes [3]:

1) Un conjunto de documentos que constituyen el corpus.
2) Un conjunto de necesidades de información (NI) o tópicos.

3) Juicios de relevancia que relacionan las NI con los documentos del corpus que son relevantes a éstas. Generalmente, se los conoce como *Qrels*.

Una colección de prueba es una herramienta experimental indispensable para los investigadores en recuperación de información (RI) ya que permite comprender la naturaleza de los resultados, compararlos con otros y reproducir pruebas en iguales condiciones. Para corpus grandes, el proceso de armado de la colección y la creación de los juicios de relevancia es una tarea que requiere un importante esfuerzo humano ya que no se puede realizar – de manera completa – automáticamente.

En la actualidad, existen grandes colecciones de prueba que son utilizadas por los grupos de investigación para la evaluación de sus sistemas, por ejemplo, las surgidas de las conferencias TREC [13] y de las reuniones CLEF [2]. Estas colecciones no son de dominio público sino que tienen un costo asociado y restricciones en su uso debido al esfuerzo que se requiere para su creación y a la naturaleza de sus documentos, los cuales – en algunos casos – tienen derechos de autor o resguardo por patentes.

Además, durante mucho tiempo estas colecciones se formaron con documentos en inglés y en los últimos años aparecieron esfuerzos para crearlas en otros idiomas [5] [2] [6]. No obstante, no se han encontrado proyectos de elaboración de colecciones de prueba de literatura científica en español que sean de dominio público.

En este trabajo se plantean los lineamientos para la construcción de una colección de prueba en español de dominio público a partir de los artículos de investigación en el área de la informática y las ciencias de la computación.

La creación de esta colección, destinada a la evaluación de la recuperación “ad-hoc”, persigue inicialmente objetivos relacionados con la docencia, en dos aspectos. En el primero, se pondría a disposición de los interesados un corpus de documentos semi-estructurados que permite la evaluación de diferentes estrategias de búsqueda. En el segundo, debido a que el tema de recuperación de información se encuentra en pleno crecimiento y si bien en la actualidad se aborda en seminarios o talleres, consideramos que en los próximos años se evaluará su incorporación definitiva como tema de grado en diferentes carreras. Esto último impulsa – aún más – la necesidad de contar con material de trabajo para docentes y alumnos.

Un segundo objetivo – más ambicioso aún – consiste en recolectar y procesar la mayor cantidad posible de artículos científicos publicados en español y crear una colección mayor que sirva para investigación de diversos aspectos de la RI (extracción de información, clasificación, respuestas a preguntas, resumen automático, entre otros). En este sentido, se requerirá del esfuerzo conjunto de diferentes grupos de RI a los efectos de poder realizar el proceso completo, incluyendo el trabajo humano de creación de los juicios de relevancia y la evaluación de la colección.

Una diferencia significativa respecto de las primeras colecciones de prueba es que ésta contiene documentos extensos (artículos completos) y – además – se encuentra rotulada de acuerdo a la estructura formal de un artículo de investigación.

Inicialmente, se plantea la metodología para la selección de los documentos, la demarcación de su estructura, la creación de los tópicos y de los juicios de relevancia, junto con una primera prueba con un conjunto pequeño de documentos.

2 – Metodología

La creación de la colección de prueba semi-estructurada requiere del procesamiento de los documentos a los efectos de normalizar los formatos e incrustar los rótulos o marcas de estructura propuestas. El proceso completo es el siguiente:

- Selección de los artículos.
- Normalización de formatos y conversión a texto plano.
- Parsing – mediante heurísticas – para reconocer estructura lógica: título, autor, resumen, cuerpo, conclusiones, referencias.
- Post-procesamiento manual a los efectos de corregir defectos de los procesos anteriores.

Luego de contar con el corpus debidamente procesado y marcado, se procederá a la elaboración de las necesidades de información (tópicos) y los juicios de relevancia. La creación de los tópicos está basada en la metodología ampliamente aceptada utilizada en la TREC [4] [12]. Éstos deben simular una necesidad de información real, en este caso, relacionada con la temática de corpus. Cada tópico constará con una identificación y tres componentes estructurales:

1) Un título, el cual es una breve descripción del contenido.
2) Una descripción, donde se amplía en solo una oración el contenido del tópico.
3) Una narrativa, en la cual se expresa de manera más extensa y compleja la necesidad de información, con mayor detalle y especificando el criterio a utilizar para determinar la relevancia de un documento.

Para la creación de los tópicos se definirán las reglas correspondientes y se convocará a un conjunto de colaboradores (docentes e investigadores) del área a los efectos de que generen los tópicos que luego deberán evaluar con los resultados del proceso de creación de los juicios de relevancia.

Para la creación de los juicios de relevancia también se seguirán las bases de la metodología de la TREC. Como es la tarea más dificultosa se utilizará una variante del mecanismo de combinación (*pooling*) [4] [9] en el cual solo una fracción de la colección se selecciona para evaluar manualmente. Aunque en la TREC se combinan los resultados de múltiples sistemas, de diferentes grupos de investigación, como este es un proyecto de menor escala se utilizarán sistemas de recuperación de información disponibles libremente como – por ejemplo – el Lemur Toolkit [7], SMART [10] y Terrier [8].

Si bien se empleará el método de *pooling*, se está diseñando una estrategia de combinación de resultados para reducir la cantidad de documentos que deberán ser evaluados por los asesores humanos. Además, en esta primera etapa se prevé que la evaluación sea dicotómica (relevante o no relevante) y – además – dos asesores humanos independientes juzguen los *topics* para minimizar los desacuerdos entre jueces. Esto se debe a que – si bien es un problema debido a la naturaleza subjetiva del proceso – se ha estudiado [11] y se determinó que tiene poca influencia en la efectividad relativa de los sistemas.

Finalmente, se evaluará la colección formada tanto en completitud como en consistencia. Para

ello se tendrán en cuenta los juicios emitidos por cada uno de los jueces en particular y en conjunto.

3 – Procedimientos y Resultados Iniciales

Para la colección inicial, se tomaron los documentos del CACIC de los años 2002 y 2004. El primer paso fue realizar la conversión de los mismos a formato de texto plano, lo cual requirió 2 pasos. Primero, se utilizó la herramienta de libre distribución PdfToHtml (http://pdftohtml.sourceforge.net/) para los documentos en formato PDF y Ghost (http://www.cs.wisc.edu/~ghost/) para los Postcript. Luego, se realizó la conversión definitiva con un programa propio. Un segundo programa eliminó caracteres no válidos y algunos defectos de ambas conversiones.

Luego, se filtraron manualmente aquellos documentos de la colección que no se encontraban en español (en CACIC se publica también en ingles y portugués) y aquellos en los que la conversión falló y no se pudo extraer el documento en formato textual. Los resultados de estos procesos se resumen en la tabla 1.

	CACIC2002	**CACIC2004**
Cantidad de documentos (PDF y PS)	115	158
Otros formatos (.doc)	0	14
Total documentos	115	172
Eliminados por errores de conversión	19	6
Eliminados idioma portugués	7	24
Eliminados idioma inglés	32	31
Definitivos	**57**	**111**

Tabla 1 – resultado la conversión y del filtrado de los documentos.

El paso siguiente consistió en procesar los documentos para incrustar marcas de estructura de artículo científico. Básicamente, se reconocieron las siguientes partes lógicas y se definieron etiquetas que demarcan: título, autor, resumen, cuerpo, conclusiones y referencias. Para esta tarea se escribió un *parser* que identificó tales componentes de la estructura mediante heurísticas. Éstos – a su vez – deben mantener un orden, por lo que se realiza un control de secuencia que verifica que se cumpla la estructura formal. Con cada coincidencia se coloca una marca de inicio con un identificador del componente, por ejemplo <RESUMEN>. La estructura del artículo se basa en la descripción utilizada por Bordignon y otros [1] en la implementación de un filtro que permite reconocer literatura científica (*papers*).

Para evaluar la eficiencia de esta pieza de software se codificó un script que recibe la salida del *parser* y determina cuáles fueron totalmente marcados, es decir, aquellos en los que se pudo reconocer todos los componentes de su formato. Para los documentos del CACIC 2002 los resultados fueron:

- Total de documentos procesados: 57
- Total de documentos totalmente marcados: 33
- Eficiencia del parser: 57.89%

Por otro lado, para los documentos del CACIC 2004 los resultados fueron:

- Total de documentos procesados: 111

- Total de documentos totalmente marcados: 62
- Eficiencia del parser: 55.86%

Para terminar la demarcación se codificó un módulo visual que permite editar de forma rápida cada documento y agregar, reemplazar o modificar las etiquetas incrustadas por el *parser* y reparar los posibles errores que pueden surgir del proceso automático. Si bien es una tarea manual, el software la facilita y acelera ya que – en promedio – cada documento demandó de 90 segundos de procesamiento.

El resultado actual es un corpus de 168 documentos debidamente procesados y marcados que forman la base de la colección de pruebas de literatura científica en español, semi-estructurada de acuerdo al formato propuesto.

4 – Consideraciones y trabajos en curso

La construcción de una colección de prueba en español de dominio público a partir de los artículos de investigación persigue dos objetivos que se consideran relevantes: en docencia, contar con dicha colección, como un recurso didáctico, en virtud del crecimiento del área de recuperación de información; y en investigación, crear una colección mayor que sirva para la evaluación de sistemas de recuperación de información de literatura científica semi-estructurada en español.

Además, se está trabajando en mejorar la eficiencia del *parser*, en determinar automáticamente el lenguaje en que están escritos los artículos y en mejorar las interfaces de usuario a los efectos de minimizar el tiempo de procesamiento manual de los documentos.

El paso siguiente tiene que ver con el armado de los tópicos y los juicios de relevancia, de acuerdo a la base propuesta en las conferencias TREC y ampliamente aceptada en la comunidad de recuperación de información. Se espera contar con una primera versión de esta colección de prueba durante el transcurso del año a los efectos de poder convocar a otros investigadores y grupos del área para llevar adelante el proyecto propuesto de una colección de mayor envergadura, que será de utilidad – además – en investigación.

5 – Referencias

[1] Bordignon, F.R.A., Tolosa, G.H. y Lavallén, P. "*Desarrollo de un Filtro Destinado al Reconocimiento Automático de Artículos Científicos* ". Jornadas de la Ciencia y Tecnología. Universidad Nacional de Luján. Mayo, 2004.

[2] CLEF 2004. *"Working Notes for the CLEF 2004 Workshop*", 15-17, Bath, UK. Septiembre, 2004.

[3] Cormack, G.V., Palmer, C.R., Clarke, L.A. *"Efficient Construction of Large Test Collections"*. En: Proceedings of Melbourne SIGIR 1998. Conference on Research and Development in Information Retrieval. ACM Press. Págs. 282 - 289. 1998.

[4] Harman, D. *"Overview of the Third Text REtrieval Conference (TREC-3)"*. En: Proceedings of Third Text REtrieval Conference (TREC-3), pags 1-20. NIST Special Publication 500-226. 1994.

[5] Hiemstra, D. y van Leeuwen, D. *"Creating an Information Retrieval test corpus for Dutch*", En: Selected Papers of the 12th meeting of Computational Linguistics in the Netherlands, CLIN / 2002 / ISBN 90-420-0943-8. 2002.

[6] Kando, N., editor. Proceedings of the 4th NTCIR Workshop (2003/2004) Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Summarization. National Institute of Informatics. 2004.

[7] Ogilvie, P. y Callan, J. *"Experiments Using the Lemur Toolkit*". En: Proceedings of the Tenth Text Retrieval Conference, TREC 2001. NIST Special Publication 500-250, pp. 103-108. 2001.

[8] Plachouras, V., He, B. y Ounis, L. *"University of Glasgow at TREC2004: Experiments in Web, Robust and Terabyte tracks with Terrier*". En: Proceedings of the 13th Text REtrieval Conference (TREC2004), Gaithersburg MD, USA, 2004.

[9] Rasmussen, E. *"Evaluation in Information Retrieval*". En: Proceedings of the ISMIR 2002 Conference Panel I: Music Information Retrieval Evaluation Frameworks, Paris, France. pp. 45-49. October 17, 2002.

[10] Salton, G., editor. "*The SMART Retrieval System: Experiments in Automatic Document Processing*". Prentice Hall, Englewood Cliffs, NJ, 1971.

[11] Voorhees, E. M. "*Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*". En: Information Processing and Management, 36 (5), págs. 697-716. 2000.

[12] Voorhees, E. M. *"The Philosophy of Information Retrieval Evaluation*", En: Proceedings of the 2nd Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany. 2001.

[13] Voorhees, E. M. "*Overview of TREC 2003"*. En: Proceedings of Twelth Text REtrieval Conference (TREC 2003), pags 1-13. NIST Special Publication 500-255. 2003.