Proyecto de indexado automático para documentos en el campo de la Física de Altas Energías

Arturo Montejo Ráez
Laboratorio Europeo de Física de Partículas (CERN)
Ginebra (Suíza)

Resumen Se describe aquí el sistema *HEPindexer*, un indexador automático para documentos sobre Física de Altas Energías. En su primera fase se ha conseguido la proposición de palabras clave primarias usando el tesauro del laboratorio alemán DESY. Los resultados, utilizando un enfoque estadístico, esperanzan la consecución de una herramienta eficaz de ayuda en el proceso de indexado.

Palabras clave: *indexación automática, modelo de espacio vectorial, modelo estadístico, tesauro.*

1 Introducción

Este proyecto consiste en el desarrollo de un sistema automático de indexado por asignación. El indexado por asignación consiste en la selección de palabras clave dentro de un léxico controlado (en nuestro caso un tesauro) que describan y resuman los conceptos más importantes tratados en un texto dado. El sistema propone palabras clave según el tesauro del laboratorio alemán DESY (Deutsche Elektronen-Synchrotron) a partir de artículos completos en inglés relacionados con Física de Altas Energías.

El sistema imlementado toma documentos en los formatos PDF, PostScript y texto plano (ASCII) y genera una lista de palabras clave *primarias*, pues en su primera fase no se consideran las palabras clave complementarias (o *secundarias*).

2 Trabajos anteriores

El indexado automático basado en la frecuencia de palabra se remonta a los años 50 y los trabajos de Luhn [4] y Baxendale [1]. Posteriormente han aparecido otros muchos trabajos y algunos sistemas de indexado por asignación integrados como sistemas de ayuda. Entre estos sistemas podemos citar BIOSIS de Vleduts-Stokolov [8], el sistema MeSH de ayuda al indexado de la Biblioteca Nacional de Medicina Estadounidense [5] y, sobre todo, el NASA MAI System [3]. Éste último es uno de los sistemas que mayor éxito han tenido y cuyo uso viene a demostrar la efectividad de la ayuda automática en el proceso de indexado.

3 El problema del indexado

El servidor *weblib.cern.ch* cubre más de 430,000 referencias bibliográficas y 170,000 documentos en formato electrónico, relacionados con el CERN y con la Física de Altas Energías. El uso de palabras clave para su clasificación y búsqueda es muy importante, siendo el laboratorio de DESY el encargado de etiquetar los documentos gracias a su equipo de indexadores (ver ejemplo en figura 1). Pero el creciente volumen de artículos hace que el esfuerzo empleado se vea desbordado. Para ello un sistema de ayuda a la indexación manual es requerido en este entorno.

```
electron positron, colliding beam
K*(892), hadronic decay
mass spectrum, (K0(S) pi-)
antimatter
experimental results
electron positron --> tau+ tau-
K*(892) --> K- pi0
10.6 GeV-cms
```

Figura 1: *Ejemplo de palabras clave propuestas por DESY para un artículo.*

4 Descripción del sistema

El modelo utilizado sigue el modelo de espacio vectorial propuesto por Salton [7]. Los textos se procesan aplicando eliminación de palabras vacías [2], *stemming* [6] y limitando

el vector a los términos con mayor peso según su frecuencia y su frecuencia inversa de documento (*idf*). A partir de la colección de entrenamiento, en la cual se dispone del texto íntegro asociado a las claves DESY, se calculan, usando métodos estadísticos, las siguientes relaciones:

Relación entre un término y un documento. Corresponde con el vector del texto íntegro.

Relación entre una palabra clave y un documento. Se obtiene a partir de la colección de entrenamiento. Es, básicamente, la frecuencia de la clave primaria en las claves DESY del documento.

Relación entre una palabra clave y un término. Usando los documentos como nexo de unión, se calcula el valor del peso entre cada clave y cada termino. Este valor se normaliza haciendo uso de lo que hemos denominado *frecuencia inversa de palabra clave* (IKF) que penaliza aquellos términos que se relacionan con muchas palabras clave.

Esta última relación es la matriz que se usa para calcular un *ranking* de claves a partir de un nuevo documento de texto. El sistema propone las claves con mayor peso como posibles claves primarias para el texto en cuestión.

5 Resultados

El sistema se entrenó a partir de una colección de 2,400 documentos y los primeros tests se realizaron sobre una colección de 1,200 documentos. Los resultados muy satisfactorios (actualmente 53,8% en precisión y 59,7% cobertura, ver figura 2).

6 Conclusión

Su interfaz web permite el amplio uso de la herramienta la cual, aún en su primera fase, promete convertirse en una valiosa ayuda a la indexación realizada manualmente. Actualmente es sistema se encuentra en producción en el servidor de documentos del CERN.

Este sistema demuestra que un algoritmo estadístico sencillo puede ser usado en un entorno con vocabulario técnico, donde la ambigüedad es baja dada la especialización del área.

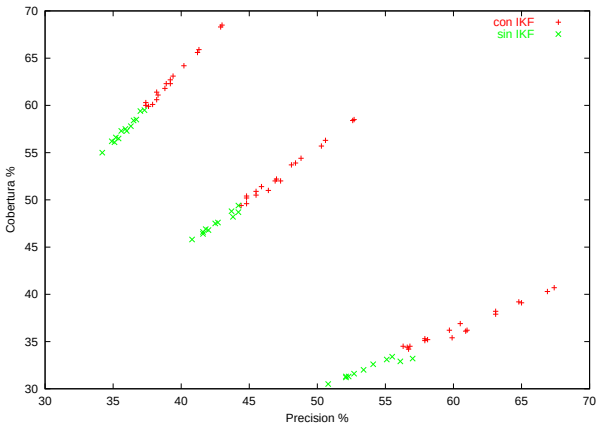

Figura 2: *Influencia de IKF sobre la precisión y la cobertura*

7 Trabajo futuro

Aún quedan problemas por resolver, como la generación de claves sobre desintegración de partículas y otros problemas que sólo aparecen en el ámbito de la Física de Altas Energías. Se pretende en futuras fases el uso de recursos lingüísticos (como WordNet), así como el estudio del algoritmo para proponer claves secundarias. Este sistema ya ha despertado el interés de expertos de la NASA trabajando con el actual NASA MAI System.

Referencias

[1] P. Baxendale. Machine-made index for technical literature — an experiment, 1958.

[2] William B. Frakes and Ricardo Baeza-Yates. Information retrieval: Data structures and algorithms, 1992.

[3] Paul H. Klingbie June P. Silvester, Michael T. Genuardi. Machine-aided indexing at nasa, 1994.

[4] H. Luhn. A statistical approach to mechanized encoding and searching of literary information, 1957.

[5] National Library of Medicine, Bethesda, US. *Medical Subject Headings (MeSH)*, 1993.

[6] M.F. Porter. An algorithm for suffix stripping, 1980.

[7] G. Salton. A vector space model for automatic indexing, 1975.

[8] Natasha Vieduts-Stokolo. Concept recognition in an automatic text-processing system for the life sciences, 1987.