II Jornadas de Tratamiento y Recuperación de Información, Madrid, Septiembre 2003, p. 228-232.

Sistema para la Indización Semiautomática (SISA) de Artículos de Revista de Biblioteconomía y Documentación

Gil-Leiva, Isidoro
Universidad Politécnica de Valencia
Facultad de Informática, Camino de Vera s/n, 46022 Valencia

isgil@har.upv.es

Resumen. Se presenta un sistema de indización semiautomática para artículos de revista de Biblioteconomía y Documentación. El sistema utiliza un conjunto de fuentes y heurísticas para proponer los términos de indización.

Palabras clave: Indización automática; vocabulario controlado; artículos de revista; Biblioteconomía y Documentación

1. INTRODUCCIÓN

Las investigaciones sobre la automatización de la indización comenzaron a finales de los años cincuenta con H.P. Luhn. Desde entonces se han realizado numerosas y variadas propuestas para acometer el proceso intelectual que supone la indización. Aunque existen diversas definiciones sobre indización podemos establecer que es un proceso guiado por el documentalista que permite recorrer, tanto a los documentos como a las preguntas, un trayecto iniciado desde puntos enfrentados. Este proceso consiste en el análisis y la selección de los conceptos esenciales, así como la asignación de los implícitos -si fuera necesario-, y el almacenamiento de los mismos en lenguaje natural o su conversión en términos normalizados y controlados que permitan recuperar los documentos en el momento deseado.

La terminología utilizada en la literatura para referirse al proceso de la automatización de la indización es variada, pudiendo encontrar estas denominaciones, entre otras: «Automated assisted indexing», «Automated indexing», «Automated support to indexing», «Automatic support to indexing», «Computer aided indexing», «Computer assistance in indexing», si bien la más utilizada es «Automatic indexing». La definición de la automatización de la indización se debe acometer desde una triple perspectiva: a) Programas informáticos que asisten en el proceso de almacenamiento de los términos de indización, una vez obtenidos de modo intelectual. (Indización Asistida por Ordenador Durante el Almacenamiento); b) Sistemas que analizan los documentos de modo automático, pero los términos de indización propuestos los valida y edita -si es necesario- un profesional (Indización Semiautomática); y c) Programas sin ningún tipo de validación, es decir, los términos propuestos se almacenan directamente como descriptores de dicho documento. (Indización Automática).

Las metodologías empleadas en la automatización de la indización desde finales de los años cincuenta hasta la actualidad han ido variando. En los primeros momentos, se utilizaba casi exclusivamente la estadística para obtener los términos de indización representativos de los documentos; pero a partir de los años ochenta se fueron incorporando en las propuestas para la automatización de la indización Técnicas de Procesamiento del Lenguaje Natural como herramientas para conseguir las raíces de las palabras, etiquetadores morfológicos, así como analizadores sintácticos, entre otras. Pero lo habitual es que las propuestas o prototipos presentados por los investigadores incluyan una combinación de ambas aproximaciones, es decir, cálculo de la frecuencia y herramientas, más o menos complejas, para el procesamiento del lenguaje natural.

Los avances en la indización automática se han ido utilizando en determinadas unidades documentales que manejan gran cantidad de información, y por tanto, es necesario automatizar, en la medida de lo posible, los procesos de análisis y tratamiento para agilizar los procesos. De este modo, han surgido prototipos como "Shapire" desarrollado por la Biblioteca Nacional de Medicina de los Estados Unidos (Hersh y Greenes, 1990); en el centro de documentación de la NASA (Silvestre, Genuardi y Klingbiel, 1994); o más recientemente, en el Laboratorio Europeo de Física de Partículas (CERN) de Ginebra (Montejo Ráez, 2001), entre otros.

Se presenta una demostración de un Sistema de Indización Semiautomática (SISA) implementado en Java para el análisis de artículos científicos de Biblioteconomía y Documentación.

2. FUENTES UTILIZADAS EN SISA

Vocabulario controlado: Actualmente está compuesto por casi 3000 términos, de los cuales 2200 son descriptores y 800 no descriptores.

Palabras vacías: Está conformado por 273 palabras que no tienen capacidad para transmitir tema o asunto alguno.

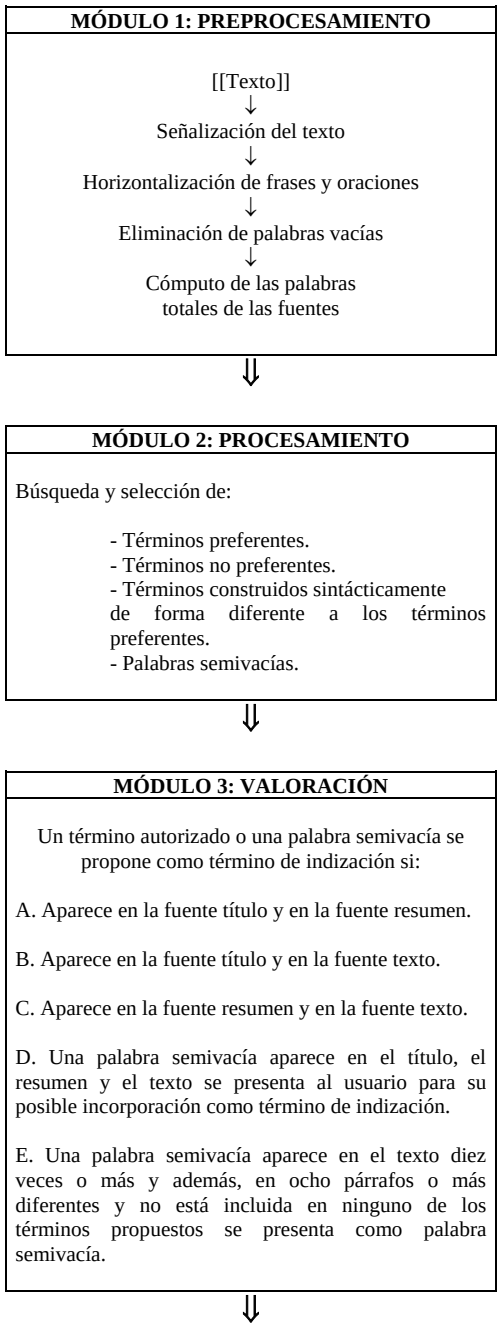
Documentos para indizar: Los documentos para ser indizados con SISA deben estar en formato txt y con un conjunto de etiquetas que delimitan el comienzo y fin del título del artículo, del resumen y del texto completo.

Ejemplo:

#CTI#
Rendimiento de los sistemas de recuperación de información en la web: evaluación de servicios de búsqueda (search engines)
#FTI#

#CR#
Se han evaluado diez servicios de búsqueda: Altavista, Excite, Hotbot, Infoseek, Lycos, Magellan, OpenText, WebCrawlcr, WWWWorm, Yahoo. Se formularon 20 preguntas a cada uno de los 10 sistemas evaluados por lo que se realizaron 200 consultas...
#FR#
#CTE#
1 Introducción
El trabajo que se presenta...
#FTE#

3. MÓDULOS DE SISA



PRESENTACIÓN DE LOS RESULTADOS PARA SU VALIDACIÓN Y EDICIÓN		
Título: ###################### **Resumen:** ##############################		
TÉRMINOS PROPUESTOS		**PALABRAS SEMIVACÍAS**
	AÑADIR **SUPRIMIR** ⇐	

El algoritmo de SISA busca cada uno de los descriptores incluidos en el vocabulario controlado en el artículo que se pretende indizar y retiene el lugar en donde ha sido localizado (título, resumen o texto). Posteriormente, una vez analizado el artículo se proponen unos términos de indización y unas palabras semivacías (palabras que no están en el vocabulario controlado ni en el listado de palabras vacías) de acuerdo a una serie de heurísticas.

En la figura 1 y figura 2 se ofrece la interfaz de SISA.

BIBLIOGRAFÍA

[1] GIL-LEIVA, Isidoro. (1999): La automatización de la indización de documentos. Gijón, Trea.

[2] HERSH, W.R. y GREENES, R.A. (1990): "SAPHIRE, an information retrieval system featuring concept matching automatic indexing, probabilistic retrieval, and hierarchical relationships", en *Computers and Biomedical Research*, vol. 23, pp. 410-425.

[3] MONTEJO RÁEZ, Arturo. (2001): "Proyecto de indexado automático para documentos en el campo de la física de altas energías", en *Boletín de Sociedad Española para el Procesamiento del Lenguaje Natural*, nº 27, septiembre, pp. 295-296.

[4] SILVESTER, J.P., GENUARDI, M.T. y KLINGBIEL, P.H. (1994): "Machine-aided indexing at NASA", en *Information Processing & Managemen"*, vol. 30, nº 5, pp. 631-645.

Figura 1: Explicación de la interfaz de SISA (v.1)

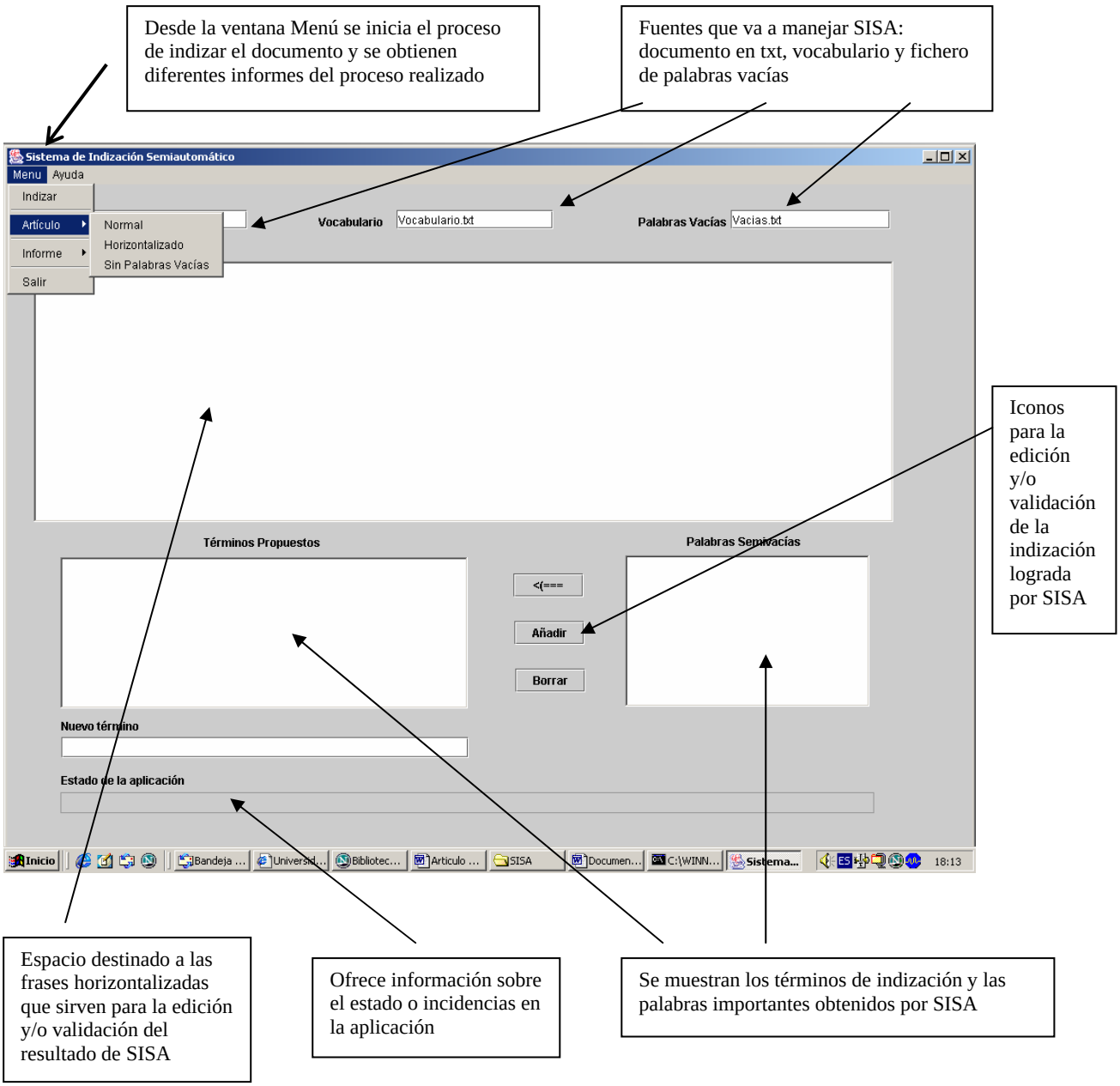


Figura 2: Pantalla de SISA una vez que ha indizado un documento: