	**Hacia la construcción de un sistema de Extracción de Información chino como soporte de innovación en los servicios bibliotecarios.** Zhang Zhixiong, Li Sa, Wu Zhengxin, Lin Ying Biblioteca de la Academia de las Ciencias China Pekín 100080 China Traducido por: Marina Jiménez Piano
Jornadas:	**97 Information Technology with Audiovisual and Multimedia and National Libraries (part 2)**
Traducción simultánea:	No

WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL 20-24 August 2006, Seoul, Korea http://www.ifla.org/IV/ifla72/index.htm

Resumen

Siendo conscientes de la importancia de la Extracción de Información (EI) para favorecer la innovación en muchas áreas de los servicios bibliotecarios, los autores construyen un sistema de extracción de información chino para procesar con eficacia los enormes recursos de información chinos. Basados en experiencias y comparaciones de algunos populares sistemas de EI, los autores proponen una solución de EI china, que haga pleno uso del sistema GATE (General Architecture for Text Engineering) de la Universidad de Sheffield, intentando desarrollar un *plug-in* chino para procesar los recursos de información chinos basados en una estructura GATE. Después de más de un año de trabajo, los autores implementaron este sistema.
Este artículo analiza la estructura del sistema GATE, describe la solución de EI (Extracción de información) china basada en ésta, centrándose en tres problemas claves en el proceso de implementación del sistema de extracción de información chino, que son el problema de la segmentación o tokenización[i], los *gazetteers* profesionales y la identificación de las entidades con

[i] Nota del traductor. El *tokenizing* o segmentación en palabras de un texto, no tiene traducción en español, se ha traducido a veces como tokenización

nombres. (1) La segmentación es un problema debido a que la estructura de la lengua china es muy flexible y realizar la segmentación de palabras en chino es muy difícil. Para resolver este problema el software de código libre llamado ICTCLAS (Institute of Computing Technology, Chinese Lexical Análisis systems) de CAS (Academia de las Ciencias china) se integró en el sistema. (2) En el sistema GATE para ayudar a la identificación de las entidades con nombres se proporciona una lista de diccionarios geográficos. Pero la lista proporcionada por GATE está pensada para el reconocimiento de las entidades con nombre ingleses, no chinos. Como base para una buena identificación las entidades chinas con nombres, los autores han recolectado y desarrollado alrededor de 1000 mega bites de *gazetteers* profesionales para ser usados en el sistema GATE. (3) El sistema GATE usa gramáticas JAPE (Java Annotation Patterns Engine) para escribir reglas de reconocimiento de NE (Nombres de entidad). Puesto que la gramática del chino es muy diferente de la del inglés las reglas JAPE que proporciona GATE no son idóneas para los textos chinos. Los autores mantuvieron en uso las gramáticas JAPE y rescribieron unas 100 normas para adaptarlas a la identificación de las entidades con nombres chinas.
También llevaron a cabo un experimento en el cual el sistema de extracción chino recuperó con éxito miles de piezas de novedades de ciencia y tecnología. Los autores creen que este sistema es una prueba significativa y que sienta las bases para futuros trabajos de investigación.

Palabras claves: Extracción de Información china; Biblioteca Digital, Procesamiento del lenguaje natural; GATE; Innovación en los servicios bibliotecarios

1 Introducción

En 2001 la Academia de las Ciencias china (CAS) inició el programa Biblioteca Digital de la Ciencia Nacional china (CSDL)[1] y la biblioteca de la CAS fue una de lasque implementó el CSDL. La misión del CSDL es desarrollar y mantener un entorno digital integrado para los investigadores y los graduados que trabajen en los institutos de investigación CAS del país, proporcionar servicios de información de integrados fiables para ayudar al lector a usar eficientemente recursos de alta calidad.
Después de casi cinco años de funcionamiento, el CSDL llegó a ser uno de los proyectos más notables de biblioteca digital de toda China, con abundantes recursos de información y una amplia gama de servicios de información:

- CSDL proporciona a sus usuarios abundantes recursos de información digital. Incluye recursos como texto completo de revistas STM (scientific, technical, medical), actas de congresos, tesis y disertaciones (ETDs), patentes, obras de referencia y libros electrónicos. En lo que se refiere solo revistas electrónicas, el CSDL cubre en la actualidad más de 6000 revistas STM occidentales, y 10.000 chinas. CSDL establece también un sistema de suministro que permite a sus usuarios obtener artículos de 15000 revistas en el plazo de un día.
- CSDL desarrolla una amplia gama de servicios de información con servicios en red, incluyendo catálogos colectivos, búsquedas federadas en bases de datos, suministro de documentos, referencia digital, personalización de “Mi biblioteca” y autentificación remota.
- Lleva a cabo muchos programas de formación y difusión para ayudar a los profesores y estudiantes a comprender y usar los servicios CSDL

Hoy día el CSDL ha llegado a ser una de las instituciones claves de investigación para los investigadores y graduados de la CAS . Están tan acostumbrados a usar el CSDL que un fallo de sus servicios en red ha llegado a convertirse en el peor desastre para nuestra biblioteca.

Además, con el rápido desarrollo de la ciencia y la tecnología china los requerimientos de los investigadores y graduados han cambiado muy rápidamente. Frente a la enorme cantidad de información académica y de otra información de investigación, los usuarios del CSDL consideran que el uso de los métodos de recuperación de información tradicionales no son suficientes, debido a que el número de documentos recibidos como respuesta a cualquier consulta es enorme. Por tanto necesitan:

- Deshacerse del ruido informativo de modo que puedan identificar correctamente los resultados potencialmente interesantes y localizar certeramente, extraer, recopilar y usar el conocimiento contenido en la bibliografía electrónica disponible.
- Realizar una visual de conjunto efectiva de los desarrollos recientes de sus áreas de interés, incluyendo la realización de resúmenes precisos y personalizados para los investigadores
- Revelar las relaciones significativas de la información, extraer vetas más ricas del material de investigación electrónico y descubrir nuevo conocimiento de la información digital

Desde otro punto de vista, los bibliotecarios del CSDL necesitan también mejorar su servicio normalizado. Además de la recuperación y el suministro de información, los bibliotecarios del CSDL reflexionan acerca de cómo convertir la biblioteca digital en un repositorio de conocimiento, intentando encontrar soluciones adecuadas para hacer buen uso de gran cantidad de bibliografía académica y datos guardados en el CSDL y desarrollar herramientas automáticas para analizar las grandes colecciones textuales.
La Extracción de Información es la tecnología emergente que cubre nuestras necesidades.

2 La EI y su función potencial en la innovación en los servicios en la biblioteca

Desde 2004, el CSDL inició varios proyectos relacionados con el uso de la tecnología de Extracción de Información (E.I.) en el entorno de biblioteca digital, intentando aplicar EI para traer innovación a los servicios bibliotecarios. Desde 2005, tenemos también la colaboración del National Social Sciences Foundation of China (NSSF), centrada en la implementación la extracción de conocimiento de los recursos digitales.

2.1 Extracción de información (EI)

Extracción de Información es un término que ha comenzado a utilizarse para la actividad de extraer automáticamente tipos de información previas especificadas de los textos de lenguaje natural[2]. Sus objetivos son extraer conocimiento estructurado, dependiente del contexto, de la información existente, generalmente texto no estructurado, con el fin de mejorar el uso y la reutilización de esta información. Hamish define la extracción de información como un proceso que toma los textos (y a veces el habla) como entrada y que produce formatos fijos, datos no ambiguos como salida[3]. EI también puede verse como la actividad de poblar una fuente estructurada de información (o base de datos) desde una fuente de información no estructurada o de texto libre. Esta fuente estructurada de información (o base de datos) se usa entonces para otros propósitos: para la búsqueda o el análisis usando consultas convencionales de bases de datos o técnicas de minería de datos; para generar un resumen; para construir índices en los textos fuente.
Iniciativas del gobierno de Estados Unidos como el Message Understanding Conference (MUC)[4], TIPSTER[5] y ACE (Automatic Content Extraction)[6] promueven el desarrollo de tecnología de Extracción de Información y prepara el terreno para la creación de algunos sistemas en curso de Extracción de Información. El programa MUC divide la Extracción de
Información en cinco tareas:

- Identificación de entidades con nombre (NE). Encontrar y clasificar nombres, lugares, etc.