Criterios de búsqueda y extractores de datos aplicados en los portales de Bibliotecas Digitales BTC y BDBComp

Luis Alejandro Vargas[1], Alberto Laender[2], German Montejano[3]
[1]Departamento de Sistemas – Facultad de Ingeniería – Universidad Nacional de Jujuy
[2]Departamento de Ciencia da Computacão - Universidade Federal de Minas Gerais - Brasil
[3]Facultad de Ciencias Físico-Matemáticas y Naturales – Universidad Nacional de San Luis
[1]avargas@fi.unju.edu.ar, [2]laender@dcc.ufmg.br, [3]gmonte@unsl.edu,ar

Resumen

En la Web encontramos que la información crece constantemente y parte de ella está disponible a través de servicios especializados de Bibliotecas Digitales. Efectuar búsquedas en cada uno de los portales de Bibliotecas Digitales consumiría bastante tiempo, por la cual nos encontramos con el inconveniente de disponer de toda esa información en el instante que la necesitamos o la precisamos. Concentrar la información proveniente de diferentes fuentes de datos, obviamente, relativas a una misma área de interés, beneficiaria en la búsqueda de información en la que el usuario está interesado.

Proponemos desarrollar extractores de datos que hagan uso de los criterios de búsqueda que el usuario introduce en el portal BTC (BTC Biblioteca de Trabajos Científicos, portal desarrollado por la Facultad de Ingeniería, UNJu, Argentina), aplicarlos en el portal BDBComp (DBDComp Biblioteca Digital Brasileira de Computacão desarrollado por el Departamento de Ciencia da Computacão, UFMG, Brasil), y así lograr obtener los datos-resultados de diferentes páginas. Los criterios de búsqueda son ingresados en cualquier de los siguientes idiomas: español, portugués e inglés, en el portal BTC, y la traducción se efectúa via ONLINE a otros idiomas mediante Google Translator, donde también aplicaremos el concepto de extracción de datos. Dichos procesos son llevados a cabo en forma transparente para el usuario que efectúa la consulta. Los resultados son formateados, clasificados según el idioma de escritura y visualizados mediante archivos XML dentro del portal BTC, sitio donde se va a concentrar la información.

Palabras Clave: Biblioteca Digital, Extracción de Datos. Wrapper.

Contexto

El portal BDBComp (Biblioteca Digital Brasileira de Computacão), portal que ofrece servicios de búsqueda de documentos científicos en sus bases de datos. Comprende 6060 trabajos publicados en diferentes eventos y periódicos de computación realizado en Brasil. También incluye trabajos publicados referentes a los siguientes periódicos: JBCS, RITA, IP e INFOCOMP.

El portal BTC (Biblioteca de Trabajos Científicos), portal que almacena artículos publicados en la VI Jornada de Ciencia y Tecnologia de las Ingenierías del Noa, llevadas a cabo en la provincia de Jujuy – Argentina, el año 2010, con un total de 160 trabajos en sus correspondientes bases de Datos. El portal tiene muy pocos artículos incorporados a su Base de Datos, y por tal motivo proponemos realizar la búsqueda

de información no solo en dicho portal, sino efectuar la búsqueda de información en otras bibliotecas digitales, como ser la BDBComp, en su servicio de búsqueda de artículos por el concepto Títulos.

1.- Introducción

En [1] nos determina tres clases de tareas relacionadas al manejo de información en la Web, siendo i) Modelado y consultas en la web, ii) Extracción de Información e integración y iii) construcción y reconstrucción de sitios en la web. En base a esto, determinamos extraer una representación estructurada de los datos de páginas HTML, que nos brinda la BDBComp, después de realizar las respectivas consultas, a través de su motor de búsqueda, y aplicamos el concepto de crear nuevos sitios.

Los criterios de búsqueda que el usuario introduce en el portal BTC (A),

pueden ser aplicados en el portal BDB Comp (B), para efectuar la búsqueda de artículos científicos. Los resultados devueltos por portal B son extraídos, mediantes extractores de datos, analizados y presentados dentro del portal A. Figura 1.

Para recuperar la información de un portal a otro, se crean diferentes archivos con formato XML. En internet hay un gran incremento de la cantidad de datos en formato XML [2], siendo considerando el formato de elección para el intercambio de datos, ya que es flexible para representar diferentes tipos de información, sobre todo datos semi-estructurados [3].

Abordaje de la Propuesta: Extracción de Datos.

Extraer la información correcta, es de suma importancia, por la cual en su mayoría las aplicaciones utilizan wrappers.

Los wrappers son programas capaces de reconocer y extraer objetos de interés dentro de páginas fuentes de la web. En [4] se efectúa una caracterización de los diferentes tipos de extractores de datos, según la técnica principal utilizada para la generación de wrappers. En [5] establece tres enfoques: la dificultad de un extractor de información, la técnica utilizada en la extracción, y por último el esfuerzo del usuario en el proceso de llevar dicho extractor a otro dominio.

El conjunto de páginas generadas del sitio BDB Comp son dinámicas, donde la diferencia de las páginas devueltas sólo se puede apreciar en el contenido.

Pasos a efectuar, para la extracción de información.

1.- Identificar los objetos-información a extraer de la BDBComp.

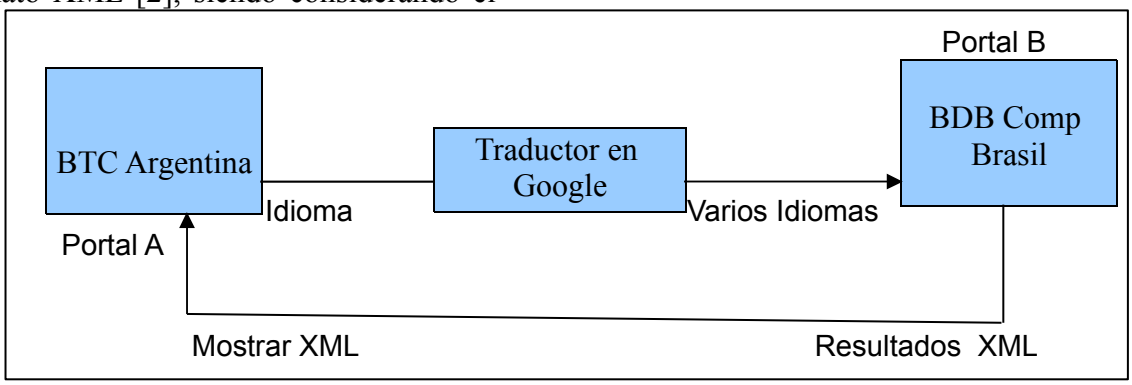

Figura 1. Criterios de Búsqueda en los dos portales

Objetos a Extraer son:

Total de Registros (R),Autores (A),título (T), Evento (E), Año (Ñ) y Link del trabajo (L)

El Conjunto de Objetos:
$O = \{R, \{A_1 \dots A_n\}, T, E, Ñ, L\} \in S_o$,
donde L, tiene como valores {Null, enlace del trabajo}
$A_1 \dots A_n \in A$ siendo A un conjunto de Autores.
$S_0 \in S$ siendo S conjunto de páginas resultados de una consulta a la BDBComp
$S \subseteq B$, siendo B Base de datos de BDBComp.

2.- Identificar en S entre qué patrones se almacena la información S_o, y así poder efectuar la extracción de los objetos O.
Determinar un lenguaje de programación para la extracción de datos o la utilización de wrappers.

3.- Almacenar la información. Recordemos que estamos en presencia de datos semi-estructurados, en la cual los objetos $O_1..O_n$ pueden tener o no valores. Los datos son almacenados en estructuras XML[8]. Tales archivos son utilizados para el traspaso de información de un portal a otro.

4.- Aplicar algún criterio de validación de la información extraída del portal BDBComp.

2.- Lineas de Investigación y Desarrollo

Las lineas de investigación son:

– Aplicación de Técnicas de Recuperación de Información (RI).

– Estudio de problemas relacionados a la búsqueda, extracción, consulta, modelado, almacenamiento, transformación, e integración de datos disponibles en la web.

– Desarrollo de un prototipo de una Biblioteca Digital que almacena documentos científicos.

3.- Conclusiones y Trabajos Futuros

Desarrollamos el prototipo de una Biblioteca Digital utilizado para almacenar los artículos científicos que se presentaron en la VI Jornadas de Ciencia y Tecnología de las Ingenierías del NOA, y así poder aplicar los diferentes resultados de la extracción de datos de la presente investigación.

Desarrollamos un prototipo de extractores de datos en tiempo real, que extrae el 100% de los datos de las páginas del portal B, efectúa la clasificación de la información y la visualización de los resultados en el portal A, generando diferentes archivos con formato XML.

Logramos que el proceso de la extracción sea RESISTENTE, ya que con nuevas páginas de consultas So, tomadas de la misma fuente Web B, donde el formato HTML ha cambiado, pero el contenido de las páginas siguen siendo el mismo [6].

No podemos decir que el proceso sea **ADAPTABLE**, ya que en el presente trabajo solamente nos limitamos a extraer los datos de una sola fuente B.

Como futuro trabajo sería interesante aplicar verificación de la calidad de los datos extraídos, a través de cálculo de medidas de similitud probabilísticas, utilizando posicionamiento y una estructura de los datos en las páginas de origen [7]. Para llevar a cabo dicho punto se realizaron adaptaciones en los formatos de salida de los archivos XML que se han generado, almacenando los valores de las posiciones de cada uno de los datos que se han extraído, y sobre ellos aplicar los

criterios de evaluación.

4.- Formación de Recursos Humanos

El contexto del presente trabajo es llevado a cabo por Ing. Luis Alejandro Vargas que desarrolla su tesis de Postgraduación de la Maestría en Ingeniería de Software. En el desarrollo del prototipo de la Biblioteca Digital BTC se trabajó en coordinación con alumnos de la Facultad de Ingeniería (UNJu). La orientación del trabajo se realiza en conjunto con el Laboratorio de Banco de Dados, dirigida por el Prof. Dr. Alberto Laender, de la Universidade Federal de Minas Gerais, Brasil y el Mg. German Montejano de la Universidad Nacional de San Luis, Argentina. Se prevee una mayor interacción y colaboración con el Laboratorio de Base de Dados de la Universidad Federal de Minas Gerais. Se espera que a partir de los logros obtenidos de la presente linea de trabajo, se incorporen otros trabajos ya sea a nivel de Maestrias o tesinas de grado de la Carrera de Ingenieria Informática o Licenciatura en Informática de la Facultad de Ingeniería de la Universidad Nacional de Jujuy.

Bibliografía

[1] Daniela Florescu, Alon Levy, Alberto Mendelzon. Databases Techniques for the World -Wide Web: A Survey. SIGMOD Record 27 (3): 59-74, 1998.

[2] Dongwon Lee, Wesley W. Chu, Comparative Analysis of Six XML Schema Languages. ACM SIGMOD Record Volume 29 (3): 76-87, 2000

[3] D. Chamberlin, Xquery: An XML query language. IBM Systems Journal, vol 41, N° 4, 2002

[4] A.H.Laender, B. Ribeiro-Neto, A.S. Da Silva, and J.S. Teixeira. A Brief Survey of Web Data Extraction Tools. SIGMOD Record, 31(2): 84-93, 2002.

[5] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled Shaalan. A Survey of Web Information Extraction Systems. Journal IEEE Transactions on Knowledge and Data Engineering. Volume 18 (10):1411-1428, 2006

[6] P. B. Golgher, A. S. da Silva, A. H. F. Laender, and B. A. Ribeiro-Neto. Bootstrapping for Example-Based Data Extraction. In Proceedings of the Tenth ACM International Conference on Information and Knowledge Management, pages 371-378, Atlanta, Georgia, 2001.

[7] Olga Regina Fradico de Oliveira. Uma Abordagem para Verificação Automática da Qualidade de Dados Extraídos da Web. Tesis de PosGraducación- 2003 UFMG, Belo Horizonte. Brasil

[8] Alberto H.F. Laender, Mirella M. Moro, Cristiano Nascimiento, Patricia Martins. An X-Ray on Web-Available XML Schemas. Sigmod Record, March 2009. Vol 38 N° 1.

[9] Alberto H.F. Laender, Altigran S. da Silva. Cooperative Research on Web Data Management at UFMG and UFAM – A Brief Report. Web Congress, Latin American, pag: 144-150. IEEE Computer Society.