Evaluación de la calidad de la Información extraída por wrappers, de un sitio web

VARGAS, Alejandro, SÁNCHEZ RIVERO David,
VALDEZ Angel, BERNECHEA Miguel, CASTILLO Natalia & COLQUI Reinaldo
Evaluación y Calidad Web (EvalCalWeb) / Facultad de Ingeniería /
Universidad Nacional de Jujuy
Av. Italia y Av. Martiarena / S. S. de Jujuy / Provincia de Jujuy
Tel. 388-4221591
lavargas@fi.unju.edu.ar, vdsanchezrivero@fi.unju.edu.ar, a.roberto.valdez@gmail.com,
pi77co@hotmail.com, natyc48@hotmail.com, mycorreo20@yahoo.com.ar

Resumen

La complejidad creciente de la estructura y la cantidad de datos presentes en un sitio web determinado, torna necesaria la existencia de herramientas para la recuperación de información (RI), la cual se considera pertinente y adecuada, para su posterior análisis. En tal sentido los wrappers, programas para extracción de datos de la web, cumplen tal función, y pueden ser generados, mediante herramientas, en forma automática o desarrollados en forma artesanal (utilizando los lenguajes de programación python o perl, por ejemplo). Los wrappers son los encargados de transformar la información semi-estructurada (presente en un sitio web) en información estructurada, a través del lenguaje XML (eXtensible Markup Language).

El carácter dinámico de los sitios web posiblemente degrade la calidad de la información extraída por los wrappers, programas que trabajan en base a ciertos criterios, como ser color, posición en la página, fuente, tags, entre otros; los cuales pueden cambiar por el dinamismo propio del sitio.

Los resultados del presente trabajo, van a permitir establecer un criterio de evaluación y comparación de las calidad de los datos extraídos de un sitio web, a medida que este presenta cambio y/o modificaciones.

Palabras clave: Extracción de datos. Datos semi-estructurados. Calidad de datos. Medidas de calidad. Wrappers.

Contexto

El proyecto parte de un trabajo anterior denominado “Criterios de búsqueda y extractores de datos aplicados en los portales de Bibliotecas Digitales BTC y BDBComp”, llevado a cabo por integrantes del grupo de trabajo, en conjunto con el Departamento de Ciencia da Computacão - Universidade Federal de Minas Gerais – Brasil, en convenio con la Universidad Nacional de San Luis – Argentina, en el año 2010.

El proyecto de investigación se desarrolla en la Facultad de Ingeniería, de la Universidad Nacional de Jujuy, aprobado por Resolución del Concejo Superior 0167/12. El Proyecto posee Categoría “B”, otorgado por la Secretaría de

Ciencia y Técnica y Estudios Regionales, dependiente de la Universidad Nacional de Jujuy.

Introducción

Internet es un gran repositorio de datos, con altas tasas de crecimiento día tras día, donde la información no mantiene un lineamiento de estructura de datos estándar, y no hay forma de manipularla y sobre todo se hace difícil llegar a tener un control sobre ella, en forma eficiente. Estos datos no pueden ser restringidos a través de un esquema tradicional y es conveniente contar con un formato de estructura muy flexible para el intercambio de datos entre bases de datos dispares [Buneman, 1997]. Los tipos de datos expuestos en la web, son llamados Semi-Estructurados, ya que tienen, como característica principal, el poseer una estructura implícita, donde la extracción de los mismos se hace en forma manual o en forma automática, a través de programas denominados wrappers. Existen patrones para la extracción de datos, que determinan los elementos relevantes a extraer, como ser el tipo de fuente, el color, la localización dentro de las páginas web, interpretación de tags y otras. Abiteboul [Abiteboul, 1997] determina un nuevo abordaje para la especificación de patrones de extracción de información de la web.

A pesar de la existencia de nuevas herramientas de búsqueda en la web, el acceso a datos semi-estructurados sigue siendo inadecuado, ya que no se cuenta con el uso de estructuras de datos estándar, y por lo tanto diversas áreas, relacionadas con la temática de Base de Datos, enfrentan nuevos desafíos en la revisión de temas tradicionales como la integración de la información, modelos de datos y lenguajes de consulta aplicados en el contexto de la Web [Lima et al., 1999]. De acuerdo con los casos de estudio planteados por Florescu [Florescu et al., 1998], orientados al gerenciamiento de la información en la Web, en los cuales describe tres clases de tareas: Extracción e Integración de la Información; Modelos de datos y lenguajes de consulta para la Web; y Construcción y reestructuración de Sitios Web.

Los wrappers son programas capaces de reconocer y extraer objetos de interés dentro de páginas o sitios web. Laender [Laender et al., 2002] detalla una caracterización de los diferentes tipos de extractores de datos, según la técnica principal utilizada para la generación de wrappers. El trabajo de Chang [Chang et al., 2006] establece tres enfoques: la dificultad de un extractor de información, la técnica utilizada, y por último, el esfuerzo del usuario en el proceso de llevar dicho extractor a otro dominio.

Los wrappers son responsables de retornar los resultados extraídos en forma estructurada, y para realizarlo pueden utilizar el formato XML [Bray et al., 2000], pueden

ser programados en lenguaje python, perl, u otros lenguajes similares.

No solo se debe tener en cuenta la extracción de los datos de las páginas web y completar, de esa manera, el nuevo repositorio inicial de datos; sino también, y no es un tema menor, verificar la calidad de la información extraída, ya que la misma puede estar incorrecta, o bien que los datos no sean pertinentes de acuerdo a los temas de interés buscados [Golpher et al., 2001]. Las comparaciones para determinar la calidad de la extracción de datos pueden estar incorporadas dentro del wrapper, o pueden ser aplicadas en forma independiente sobre los resultados. Podemos mencionar la que utilizan heurísticas, o determinar la calidad a través de funciones estadísticas, de inducción, de posicionamiento de los objetos en las páginas origen y su estructura u otras.

Líneas de Investigación y Desarrollo

Las líneas de investigación son:

- Aplicación de Técnicas de Recuperación de Información.
- Estudio de problemas relacionados a la búsqueda, extracción, consulta, modelado, almacenamiento, transformación e integración de datos disponibles en la web.

Se investigarán, en principio, diferentes criterios de evaluación de la información extraída por los wrappers, y se dispondrá de la aplicación de funciones y/o algoritmos que sirvan como patrones de evaluación para los resultados obtenidos.

Resultados y Objetivos

En el trabajo anterior "Criterios de búsqueda y extractores de datos aplicados en los portales de Bibliotecas Digitales BTC (Biblioteca de Trabajos Científicos) y BDBComp (Biblioteca Digital Brasileira de Computacão)", el grupo de trabajo desarrolló un prototipo de extractores de datos en tiempo real, el que extrae el 100% de los datos de las páginas de un repositorio de datos, además se efectuó la clasificación de la información y la visualización de los resultados, generando diferentes archivos con formato XML.

El equipo de investigación se encuentra desarrollando actividades de investigación y desarrollo en el diagnóstico, análisis y descripción de las herramientas que extraen datos de la web. Como así también se está trabajando en la búsqueda, descripción e instalación de herramientas, basadas en software libre, para la extracción de datos y así poder generar los wrappers, sobre un sitio determinado.

Finalmente se procura establecer un estado de la cuestión y analizar en detalle los procesos de ejecución de los wrappers y así poder comparar criterios de evaluación de datos extraídos de un sitio web.

Formación de Recursos Humanos

El proyecto de investigación se encuentra conformado por dos docentes investigadores categorizados, tres alumnos y un egresado de la carrera de Ingeniería Informática, en proceso de formación, abocados en tareas de investigación y programación.

En el presente año, se realizaron diferentes reuniones con alumnos y egresados interesados en incorporarse al proyecto y se establecieron líneas de investigación para generar Proyectos Finales, en la carrera de Ingeniería en Informática de la Facultad de Ingeniería de la Universidad Nacional de Jujuy.

Por otra parte el desarrollo de las tareas de investigación ha generado, en el año 2012, un anteproyecto de tesis en la Maestría en Ingeniería de Software, de la Universidad Nacional de San Luis.

Se prevee una mayor interacción y colaboración con el Laboratorio de “Base de Dados” de la Universidad Federal de Minas Gerais, dirigida por el Prof. Dr. Alberto Laender, de la Universidade Federal de Minas Gerais, Brasil

Referencias

[Buneman, 1997] Buneman, Peter. Semistructured Data. In Proc. Of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), pages 117-121, Tucson, Arizona, 1997.

[Abiteboul, 1997] Abiteboul, Serge, Querying semi-structured data. In Proc. Of the Int. Conf. On Database Thery (ICDT), Delphi, Greece, 1997.

[Lima et al., 1999] Lima, F; Casanova, M.A. & Melo, R. N., Revisitando Técnicas de Bancos de Dados no contexto da Web, 1999.

[Florescu et al., 1998] Florescu, D., Levy, A. & Mendelzon, A., "Database Techniques for the World-Wide Web: A Survey", SIGMOD Record, vol. 27, n. 3, Setembro 1998.

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

[Chang et al., 2006] Chang Chia-Hui; Kayed, M.; Girgis, M.R. & Shaalan, K., A survey of Web Information Extraction Systems. Journal IEEE Transactions on Knowledge and Data Engineering. Volume 18(10):1411-1428, 2006.

[Bray et al., 2000] Bray T., Paoli J, Sperberger C.M., and Maler E.; Extensible Markup Language (XML) version 1.0 (second edition). Technical Report REC-xml-20001006, World Wide Web Consortion. W3C, October 2000.

[Golpher et al., 2001] Golpher, P.B.; da Silva, A.S.; Laender, A.H.F.; Ribeiro-Neto, B.A., Bootstrapping for Example-Based Data Extraction. In Proceeding of the Tenth ACM International Conference on Information and Knowledge Managemente, pages 371-378, Atlanta, Georgia, 2001.