Hacia una integración de un sistema de búsqueda de respuestas sobre la inteligencia empresarial mediante el uso de ontologías

Sandra Roger[1,2], **Antonio Ferrández**[2], **Jesús Peral**[2*]

[1] University of Comahue, Argentina
Buenos Aires 1400 - 8300 Neuquén - Argentina
Tel/Fax (54) (299) 4490312 ext. 435 /(54) (299) 4490313

[2] Natural Language Processing and Information Systems Group
Department of Software and Computing Systems University of Alicante, Spain
Carretera San Vicente S/N - 03080 Alicante - España
Tel/Fax (34) 965903400 ext. 2385/ (34) 965909326
sroger@dlsi.ua.es, antonio@dlsi.ua.es, jperal@dlsi.ua.es

Resumen

El objetivo general de las aplicaciones de inteligencia empresarial (*Business Intelligence*, a partir de ahora BI) es permitir a sus usuarios entender y analizar los datos existentes en sus organizaciones para adquirir conocimiento útil y lograr así una mejor toma de decisiones. El corazón de las aplicaciones de BI son los almacenes de datos (*Data Warehouse*, a partir de ahora DW), los cuales integran diferentes recursos de datos, principalmente bases de datos estructuradas. Sin embargo, una nueva tendencia a utilizar la Web como fuente de informaciónsobre el entorno de las organizaciones ha emergido.

Como parte de esta línea de investigación, estamos trabajando en la aplicación de un sistema de búsqueda de respuesta (*Question Answering*) como herramienta vinculante a los DW para la obtención de información que ayude en la toma de decisiones, continuando, de esta manera, con los avances obtenidos en [2].

1. INTRODUCCIÓN

Un buen DW contiene toda la información generada por una organización. Estos datos pueden ser enriquecidos con información proveniente del exterior, como por ejemplo páginas Web de otras organizaciones.

En la jerarquía de una organización, la persona encargada de la toma de decisiones necesita tener presente diferentes indicadores. Estos indicadores pueden ser de múltiples y diversas fuentes que van, por ejemplo, desde un listado de venta hasta información obtenida de la Web. En este sentido, a lo que BI apunta es a encontrar los indicadores necesarios para la toma de decisiones en cada organización.

En la actualidad, la mayor parte de las aplicaciones que precisan gestionar grandes colecciones de documentos, aplican técnicas de recuperación de información (IR). En este sentido se propone combinar dos áreas distintas de estudio: los almacenes de datos y los sistemas de búsqueda de respuestas (QA). De esta manera se pretende obtener información sobre la Web mediante el uso de un sistema de

*This paper has been partially supported by the Spanish Government project number TIN2009-13391-C04-01, by the Generalitat Valenciana project number PROMETEO/2009/119 and by the University of Comahue under the project 04/E084.

búsqueda de respuesta. La forma de integrar los DW con el sistema de QA es de tal manera que uno retroalimente al otro para lograr el enriquecimiento de los datos, y en consecuencia, de la información obtenida.

2. MARCO DE TRABAJO

Los sistemas de IR permiten el acceso a las fuentes de datos, pero carecen de comprensión semántica de los mismos, es decir, no analiza el significado de la información recuperada de los documentos. Los sistemas de IR sólo devuelven documentos (información no estructurada) que no puede alimentar directamente a las aplicaciones de BI. A diferencia de los sistemas de IR, los sistemas de QA, aumentan la precisión de los resultados debido a que éstos realizan un análisis más profundo del texto y sus resultados sí pueden ser estructurados en una base de datos con el fin de ser procesado por un sistema de BI. El enfoque propuesto permite un mejor conocimiento de los datos y el primero en integrar un sistema de QA con los DW. Por otra parte, también se realiza un tratamiento sobre cualquier tipo de datos no estructurados (por ejemplo XML, HTML, o PDF) y facilita el acceso independientemente de su origen (la intranet de la empresa o la Web)

El enfoque propuesto se basa en los siguientes pasos que se realiza en una forma semi-automática:

- Una ontología de dominio es obtenida partir del modelo UML de los sistemas de DW, el cual se usa para el modelado multidimensional del DW [5].
- Esta ontología es alimentada por los contenidos del sistema DW.
- Esta ontología es fusionada y mapeada con la ontología superior utilizada por el sistema de QA.
- El sistema de QA se adapta a los nuevos tipos de consultas que son requeridos por los usuarios.
- Este sistema de QA alimentará al DW con la nueva información extraída de las consultas planteadas en la Web.

Con estos pasos, la integración de los sistemas de DW y QA es realizada por medio de una ontología. Esta ontología es utilizada para compartir conceptualizaciones, la cual es alimentada por los DW (paso 2). Esto servirá para que la ontología creada enriquezca a los sistemas de QA para incrementar su precisión. EL DW es beneficiado por la integración de ésta ontología con una de nivel superior (paso 3) y la información provista por los sistemas de QA (paso 5). De esta manera, esta propuesta difiere de otros enfoques [7, 6] los cuales utilizan a las ontologías sólo para comunicar los resultados entre los sistemas de DW y QA.

Sistema de QA

El sistema de QA utilizado en esta investigación es el sistema AliQAn. Este sistema ha participado en distintas instancias de la competencia internacional CLEF[1], tanto en tareas monoligüe [9] como *cross-lingual* [3]. AliQAn [8] es un sistema QA de dominio abierto basado en el uso intensivo de herramientas de NLP[2].

AliQAn consiste de dos fases principales (Figura 1). La fase de indexacción y la de búsqueda. En ambas fases, tanto en los documentos del corpus como de las preguntas, el mismo proceso de NLP es

[1] http://www.clef-campaign.org/ visitada el 4 de abril de 2010

[2] *Natural Language Processing*

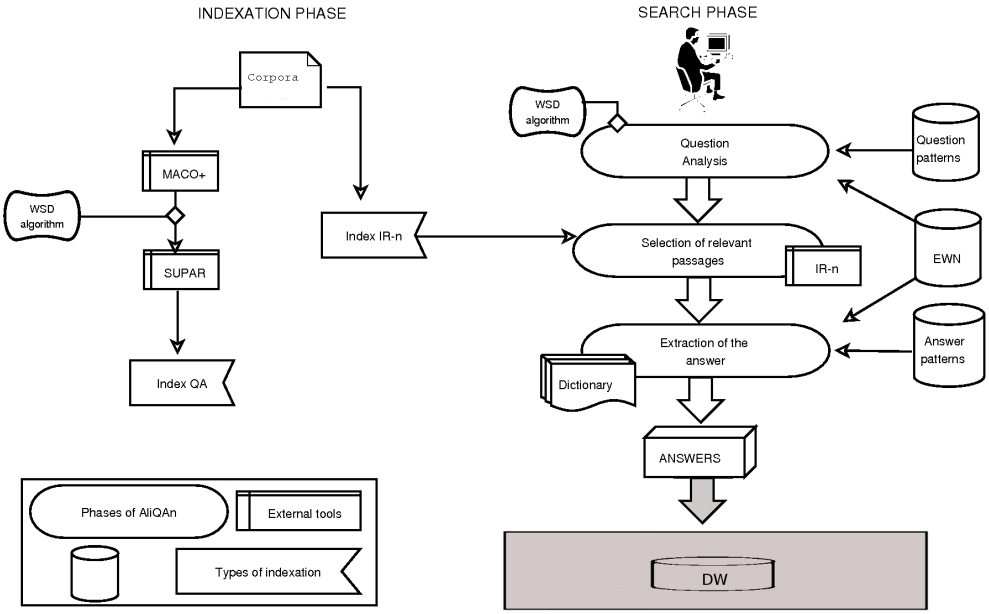

Figura 1: Arquitectura de AliQAn

aplicado: el analizador morfosintáctico Maco+[3] o TreeTagger[4], un analizador sintáctico SUPAR [1] y un algoritmo de desambiguación [4].

3. CONCLUSIONES Y TRABAJOS FUTUROS

Hoy en día, los DWs desempeñan un papel decisivo en las aplicaciones de BI, debido al hecho de que pueden proporcionar muchos años de información histórica en una forma precisa para la toma de decisiones. Esta información histórica puede ser estructurada (por ejemplo, bases de datos transaccionales) o no estructurada (por ejemplo, informes internos o correos electrónicos).

Tradicionalmente, las soluciones de BI se han centrado en datos estructurados, pero no se ha prestado suficiente atención a los datos no estructurados. Sin embargo, fuentes no estructuradas de datos se están volviendo más y más importantes para potenciar el proceso de toma de decisiones. En concreto, los datos (no estructurados) provenientes tanto de dentro de la empresa (por ejemplo, la informes o mensajes de correo electrónico del personal de la empresa almacena en el intranet de la compañía) como del exterior (por ejemplo, de las Webs de la empresa competidores).

Desafortunadamente, la investigación en esta dirección sólo se refirió al uso de la utilización de IR para el manejo datos no estructurados. El principal inconveniente de estos sistemas es que no analizan el significado de la información en los documentos, por lo que sólo devuelven los documentos los cuales no sirven para alimentar directamente las aplicaciones de BI.

Para superar esta situación y obtener un mejor conocimiento, se propone obtener el primer modelo para la integración de DW y los sistemas de QA. Este modelo supera los enfoques anteriores, ya que sistemas de QA incrementan la precisión de los resultados a través de una comprensión más profunda del texto y los resultados obtenidos enriquecen las DWs mejorando considerablemente el proceso de la toma de decisiones.

Nuestro modelo también supera los enfoques basados en la extracción de información porque esta

[3]http://garraf.epsevg.upc.es/freeling/demo.php visitada el 4 de abril de 2010

[4]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ visitada el 4 de abril de 2010

tecnología no facilita la procesamiento de enormes cantidades de documentos (por ejemplo, la Web o la intranet de la compañía). Por otra parte, se limita a un conjunto predefinido de plantillas, mientras que la tecnología de QA funciona como una herramienta típica de IR, pero mejorando sus resultados como se ha mencionado anteriormente.

Por otra parte, esta propuesta de integración se logra por medio de una ontología que representa los beneficios que esta integración produce tanto en tecnologías de QA como DW, a diferencia de otros anteriores enfoques, que utilizan ontologías sólo para la comunicación e intercambio de datos.

El modelo presentado se basa en cinco pasos que se llevan a cabo de una manera semi- automática. Para hacer uso de las evaluaciones e implementaciones se ha utilizando el sistema de QA denominado AliQAn, con el que se ha participado en varias competiciones CLEF, tanto en tareas monolingües como cross-lingual.

Como proyectos futuros, se estudiará el pre-procesamiento de las páginas web con el fin de manejar adecuadamente tablas. Por otra parte, vamos a estudiar cómo las diferentes etapas de este enfoque puede ser automatizado, por ejemplo, cómo una consulta inicial en la DW sistema puede generar diferentes consultas y combinando adecuadamente las respuestas del sistema, ayudar positivamente al proceso de toma de decisiones.

REFERENCIAS

[1] A. Ferrández, M. Palomar, and L. Moreno. An Empirical Approach to Spanish Anaphora Resolution. *Machine Translation. Special Issue on Anaphora Resolution In Machine Translation*, 14(3/4):191–216, December 1999.

[2] A. Ferrández and J. Peral. The benefits of the interaction between data warehouses and question answering. In *International Workshop on Business IntelligencE and the WEB (BEWEB'10)*, Lausanne, Switzerland, 2010.

[3] S. Ferrández, P. López-Moreno, S. Roger, A. Ferrández, J. Peral, X. Alvarado, E.Ñoguera, and F. Llopis. Monolingual and cross-lingual qa using aliqan and brili systems for clef-2006. *Workshop of Cross-Language Evaluation Forum (CLEF)*, 2006.

[4] S. Ferrández, S. Roger, A. Ferrández, A. Aguilar, and P. López-Moreno. A new proposal of word sense disambiguation for nouns on a question answering system. *Advances in Natural Language Processing. Research in Computing Science. ISSN: 1665-9899. 7th International Conference, CICling*, 18:83–92, February 2006.

[5] S. Luján-Mora, J-Trujillo, and I. Song. A UML profile for multidimensional modeling in data warehouses. *Data Knowl. Eng.*, 59(3):725–769, 2006.

[6] T. Priebe and G. Pernul. Ontology-based Integration of OLAP and Information Retriebal. In *Proceedings of the 14th International Workshop on Database and Expert Systems Applications (DEXA'03)*, page 610, 2003.

[7] T. Priebe and G. Pernul. Towards integrative enterprise knowledge portals. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM'03)*, pages 216–223, 2003.

[8] S. Roger, S. Ferrández, A. Ferrández, J. Peral, F. Llopis, A. Aguilar, and D. Tomás. AliQAn, Spanish QA System at CLEF-2005. *In Workshop of Cross-Language Evaluation Forum (CLEF). ISSN: 0302-9743 - Lecture Notes in Computer Science - Accessing Multilingual Information Repositories*, 4022(1):457–466, 2005.

[9] S. Roger, K. Vila, A. Ferrández, M. Padiño, J.M. Gómez, M. Pucho-Blasco, and J. Peral. Using AliQAn in Monoligual QA@CLEF-2008. *Evaluating Systems for Multilingual and Multimodal Information Access*, 5706(1):333–336, 2009.