Prototipo de búsqueda y comparación que aplica técnicas de recuperación de información en bases de datos relacionales

Claudio Camacho, Walter Singer y Rosanna Costaguta

Departamento de Informática, Facultad de Ciencias Exactas y Tecnologías
Universidad Nacional de Santiago del Estero
Avda. Belgrano (S) 1912, Santiago del Estero, 4200, Argentina
claudiocamacho@yahoo.com; singerwalter@gmail.com; rosanna@unse.edu.ar

Resumen. Una de las actividades más importante de una organización comercial es la adquisición de productos, ya que el encargado de compras debe tener en cuenta diversos aspectos como precios, ofertas, disponibilidad, tiempo de entrega, etc., a fin de tomar una decisión. Esto se vuelve una tarea complicada cuando la organización posee información de numerosos artículos que provienen de distintos proveedores, donde cada uno de ellos administra sus propias listas de precios y catálogos de productos, con nomenclaturas diferentes (código de artículo, descripción, etc.). Es así, por ejemplo, que al momento de comparar precios entre los distintos proveedores no es fácil identificar productos equivalentes. Debido a lo expuesto resulta conveniente que el Sistema de Información de la organización posea ciertas características de búsqueda que permitan responder eficientemente a estas cuestiones. En el presente trabajo se propone el desarrollo de un prototipo de herramienta de búsqueda aplicable sobre bases de datos relacionales que, mediante técnicas de recuperación inteligente de información, permita al usuario realizar consultas de productos de manera fácil y eficiente a fin de aprovechar mejor las oportunidades del mercado.

Palabras clave: Técnicas de Recuperación Inteligente de Información, Bases de Datos Relacionales, Sistemas de Información, Organizaciones Comerciales

1 Introducción

En la actualidad el éxito de una organización comercial se debe en gran medida a su Sistema de Información (SI), el cual es el esqueleto de la misma. Muchas de estas organizaciones, poseen una gran cantidad de datos que no interviene en las operaciones diarias y que es imprescindible procesar para optimizar la toma de decisiones. En particular, el recurso más importante que posee la organización es su base de datos, en la cual se encuentran almacenados numerosos registros que se deben administrar de manera eficiente. Frecuentemente, con el paso del tiempo, el volumen de las bases de datos crece llegando a tener miles y hasta millones de registros. Este incremento se debe a que se han ido incorporando nuevos clientes, proveedores, listas de precios, etc. En este contexto el SI debe evolucionar para poder hacer frente a las

nuevas oportunidades de negocio que se presenten, de tal modo que ayude al crecimiento de la organización.
Desde la perspectiva del desarrollador de software, el problema radica en la forma en que se van a procesar los datos equivalentes y/o redundantes, considerando que generalmente las organizaciones cuentan con un Sistema Gestor de Base de Datos (SGBD) convencional que responde a consultas puntuales y estructuradas, brindando respuestas totalmente deterministas. Este problema implica obtener conocimiento implícito, y es aquí donde las técnicas de Recuperación de Información (RI) son útiles (Hernández Orallo *et al.*, 2004). Las técnicas de RI, son ampliamente utilizadas en el ámbito de la Web para búsqueda de información en un conjunto de documentos, publicaciones, páginas web, diarios digitales, etc. El objetivo principal de estas técnicas es el de responder a consultas de usuario semiestructuradas o no estructuradas (cercanas a lenguaje natural) donde el usuario ingresa palabras clave y el sistema devuelve un conjunto de resultados acordes con un grado de relevancia, tal como lo hacen los motores de búsqueda en Internet. En esta tarea las técnicas de clasificación son importantes ya que permiten agrupar los datos de acuerdo con un criterio determinado. Estos clasificadores tienen la propiedad de aprender del dominio y a medida que aprenden devuelven resultados más exactos (Manning *et al.*, 2009).
En el presente proyecto se desarrollará una herramienta de RI para el SI de la organización comercial local objeto de estudio en este trabajo. Esta propuesta está siendo desarrollada como tesis final de graduación por dos de los autores de este trabajo para obtener sus títulos de Licenciado en Sistemas de Información. Este artículo se organiza como sigue. En la próxima sección se plantea brevemente el problema a resolver, en la sección 3 se sintetizan los pasos metodológicos propuestos, la sección 4 contiene antecedentes relevantes y la sección 5 algunas conclusiones.

2 Planteamiento del problema

Son muchas las tareas que se realizan a diario en una organización comercial, una de las más importantes es la que involucra al área de compras. En particular, en dicha área el problema surge cuando se debe abastecer el stock de productos, decidiendo cómo, cuándo cuánto y a quién comprar un determinado producto. Tomar esta decisión implica comparar listas de precios y de ofertas de diferentes proveedores que comercializan un mismo producto, los cuales generalmente poseen una nomenclatura propia en la identificación de los artículos, que pueden no coincidir con los almacenados en la Base de Datos (BD). Esta situación provoca redundancia de datos dentro de la BD y hace más compleja la tarea de identificar productos equivalentes o iguales. Incluso puede ocurrir que los datos sobre un determinado artículo (código, precio, proveedor que lo comercializa, etc.) estén distribuidos en diferentes tablas y/o campos de un tabla dentro de la BD, ya sea con un mismo código o no, con otro nombre, o identificado por una concatenación de diferentes campos.

3 Metodología

Metodológicamente el proyecto se estructura en tres etapas principales. En la primera etapa se efectuará la búsqueda, recolección, clasificación y lectura comprensiva de bibliografía, antecedentes y material de referencia vinculado con las técnicas de recuperación de información (modelos probabilístico, vectorial, booleano, basados en algebra de relaciones, etc.) y técnicas de clasificación de datos y documentación (clustering, mapas autoorganizativos, árboles de decisión, método de clasificación de Naive-Bayes, método de clasificación k-nearest neighbor, redes neuronales, etc.). En la segunda etapa se diseñará el prototipo de herramienta de recuperación de información teniendo en cuenta las técnicas analizadas anteriormente, codificándose el mismo. En la tercera etapa se pondrá en funcionamiento el prototipo creado y se contrastarán sus respuestas con las del sistema anterior. La evaluación se llevará a cabo teniendo en cuenta aspectos tales como tiempo de respuesta, grado de precisión y complejidad de las consultas efectuadas. A la fecha se ha finalizando la primera etapa metodológica del proyecto, contándose ya con un primer diseño de la herramienta.

4 Antecedentes relacionados

A continuación se describen brevemente dos trabajos relevantes para esta propuesta ya que integran sistemas de recuperación de información con sistemas de bases de datos. Sin embargo, cabe destacar que no se encontraron antecedentes de trabajos en el ámbito comercial como el que se propone realizar.

MEDLINE cuenta con una BD donde cada registro almacena la referencia bibliográfica de un artículo científico publicado en una revista médica, conteniendo además datos básicos como título, autores, etc., posibilitando su recuperación a través de Internet (García *et al.*, 2006). Los autores desarrollaron dos sistemas de indexación y búsqueda que mejoran la capacidad búsqueda y recuperación de la información en MEDLNE.

Hiemstra *et al.* (2009) extienden la estructura de un Sistema de Gestión de Base de Datos (SGBD) semiestructurada (XML) para agregar funciones de recuperación de información a las consultas estándares. La investigación se dividió en dos etapas: Diseño de la Arquitectura de la base de datos y Optimización de las consultas. Se formuló un modelo dividido en tres capas: Física, Lógica y Conceptual. La primera etapa de desarrollo abarca las dos primeras capas, definiéndose un pequeño número de primitivas de recuperación en la capa lógica del SGBD, como una extensión de la arquitectura actual, para proveer consultas que combinen ambos enfoques (estructurado y no estructurado). Para evitar la redundancia en el procesamiento de las consultas se propone crear reglas de optimización asociadas a las primitivas de estas consultas combinadas. Esta tarea es la que se realizará en la segunda etapa. La propuesta está aún en fase de desarrollo pero los autores esperan obtener un prototipo de SGBD donde se incrementen las capacidades estándar de las consultas XML y se devuelvan resultados adecuadamente ranqueados.

5 Conclusiones

En la actualidad, la globalización está provocando que países como la Argentina, con economías inestables y carentes de políticas a largo plazo, se vuelvan lugares muy poco confiables para invertir en nuevos emprendimientos, y donde es muy difícil mantener en funcionamiento el comercio interno debido a las constantes variaciones que se producen. En este entorno, particularmente las pequeñas empresas se ven en la obligación de ser altamente competitivas en lo que respecta a los precios de sus productos. Esto lleva, entre otras cosas, a que se deban analizar las ofertas de precios de los diversos proveedores teniendo en cuenta muchos factores, como por ejemplo precio de costo, fletes, tiempo de entrega de la mercadería, disponibilidad de artículos, etc. Esta tarea consume mucho tiempo sobre todo a la hora de comparar condiciones entre diferentes proveedores debido a que estas organizaciones tienen en su stock miles de productos diferentes cuya nomenclatura (códigos, descripciones, etc.) puede variar de un proveedor a otro. Cabe destacar que la mayoría de las empresas prácticamente pasan por alto el proceso de comparación de condiciones de venta de los artículos que desean comprar. En algunos casos se reduce este trabajo solo a unas pocas comparaciones teniendo en cuenta las listas de los proveedores más reconocidos. En otros casos, se pasa por alto totalmente el análisis de condiciones y se realizan las compras de un determinado artículo siempre al mismo proveedor, dejando de lado a otros proveedores que podrían tener una mejor oferta, llegando de esta manera a perder las ventajas que se obtendrían si se realizara un análisis adecuado.
Debido a lo expuesto las PyMES encuentran cada vez más necesaria la implementación de SI que les permitan realizar estas tareas en forma eficiente, lo cual les brindaría ventajas respecto a sus competidores. Así, el desarrollo de una herramienta que permita a los encargados de la toma de decisiones comparar condiciones (por ejemplo, precios) en forma rápida e identificar posibles oportunidades de negocios es muy importante. En particular la herramienta a construir se integrará al SI de una organización comercial del medio, que tendrá como objetivo mejorar la búsqueda e identificación de productos dentro de la BD de la empresa a fin de permitir al encargado de compras de la organización, tomar decisiones acertadas de manera fácil, rápida y oportuna.

Referencias

1. Hernandez Orallo, J., Ramirez Quintana, Ma J., Ferri Ramirez, C. (2004). Introducción a la Minería de Datos. Madrid, España: Prentice Hall.
2. Manning, C., Raghavan, P., Schütze, Hinrich. (2009). An Introduction to Information Retrieval. Cambridge, England: Cambridge University Press.
3. García, F., Fernández, Ch. y Azancot, M. (2006) “Desarrollo de un sistema de indexación y búsqueda sobre la base de datos de biomedicina MEDLINE”. http://biblioteca.universia.net/html_bura/ficha/params/id/45165231.html
4. Hiemstra, D., Vries, A., Blok, H., Keulen, M., Jonker, W. y Kersten, M. (2003) “CIRQUID: Complex Information Retrieval queries in a Database”. Reporte Interno. Netherlands: Centre for Telematics and Information Technology,