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 (FCEyT)
Universidad Nacional de Santiago del Estero (UNSE)
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, y tiempo de entrega, entre otras cuestiones, 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 obviamente administra sus propias listas de precios y catálogos de productos, con nomenclaturas diferentes (código de artículo, descripción, etc.). Es así que al momento de, por ejemplo, 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 este artículo se presenta un prototipo de herramienta de búsqueda aplicable sobre bases de datos relacionales, que utiliza técnicas de recuperación inteligente de información. Las pruebas realizadas arrojaron resultados muy satisfactorios.

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

1 Introducción

En la actualidad es cada vez mayor la utilización de Sistemas de Información (SI) en las distintas áreas de las organizaciones comerciales, así como también el incremento en los volúmenes de datos almacenados en sus Bases de Datos (BD). Debido a esto, la obtención de información a través de un determinado proceso se hace cada vez más ineficiente ya que el tiempo de respuesta de las consultas sobre los datos, es cada vez mayor y con resultados menos precisos.

En particular en las organizaciones comerciales, en el área de compras es donde se concentran las decisiones que se deben tomar con respecto a la mejor opción de compra, y es por ello que generalmente se deben comparar los precios de los artículos en las listas de precios de los diferentes proveedores que comercializan un mismo

producto. Realizar este proceso en forma manual requiere mucho tiempo, ya que cada artículo posee una nomenclatura propia (código, descripción, marca, etc.) para cada uno de los proveedores, y es aquí donde surge la necesidad de utilizar métodos y técnicas de búsqueda y clasificación más eficientes que ayuden a los usuarios a identificar artículos equivalentes o iguales.

Dado lo expuesto, en este artículo se presenta una herramienta de RI y clasificación de los resultados para el SI de una organización comercial del medio, cuyo uso permitió optimizar los procesos de búsqueda y comparación por parte del usuario. El prototipo desarrollado utiliza librerías de RI (Sphinx 0.9.9-rc2) y técnicas de clasificación. La presente investigación fue desarrollada como trabajo final de graduación de dos de los autores, a fin de obtener el título de Licenciado en Sistemas de Información.

Este artículo se organiza como sigue. En la sección 2 se describe brevemente la problemática que dio origen a este trabajo. La sección 3 contiene antecedentes relevantes. La sección 4 describe el prototipo desarrollado. La sección 5 documenta las pruebas efectuadas y el análisis de los resultados obtenidos. La sección 6 enuncia algunas conclusiones sobre el trabajo realizado.

2 Planteamiento del problema

Son muchas las tareas que se realizan en una organización comercial, una de las más importantes es la que involucra al área de compras. En esta área el encargado debe analizar y comparar artículos en extensas listas de artículos de distintos proveedores teniendo en cuenta ciertos factores como precio, marcas, tiempo de entrega de la mercadería, disponibilidad en stock en los proveedores, entre otros. Esta es una tarea que suele consumir un tiempo considerable sobre todo en el momento de identificar los artículos. Esta labor se vuelve aun más compleja si se tiene en cuenta que la nomenclatura (códigos, descripciones, marca, etc.) que se utiliza varía de un proveedor a otro. Existen casos en los que, por ejemplo, se suprime parte del código del artículo, en otros casos se agrega un código complementario al código del artículo, y hasta en otros casos el Código del artículo se sustituye completamente y el mismo es agregado como parte de la descripción del artículo. Así, puede ocurrir que un mismo artículo se presente dentro de la BD de maneras diferentes, con lo que la tarea de filtrar estos datos a través de una consulta estructurada es casi imposible dada la cantidad de variantes con la cuales se puede identificar un artículo. En la Tabla 1 puede observarse un ejemplo extraído del catálogo real de Cables Marca *"NGK"*. El artículo *"CABLE DE BUJIA VW GOL*", cuyo código es *"ST-V02",* aparece cargado tres veces en la tabla artículos de la organización ya que es suministrado por tres proveedores diferentes. Sin embargo, como puede observarse, el código de artículo ST-V02 aparece cargado como código de artículo en el caso del primer proveedor, como parte del código de artículo en el caso del segundo proveedor y como parte de la descripción del artículo para el tercer proveedor. Así, como puede observarse en el ejemplo de la Tabla 1, un mismo artículo se presenta dentro de la BD de la organización de maneras diferentes, y por tanto, la tarea de filtrar estos datos a través

de una consulta estructurada es casi imposible por la cantidad de variantes con la cuales se puede identificar un artículo dado.

Tabla 1. Fragmento de la BD de la organización

Código	Descripción de Artículo	Cód. Proveedor
STV02	CABLE DE BUJIA	1
…	…	…
0605STV02	CABLE BUJIA SEN/ GOL	5
…	…	…
NGK-02	CABLE DE BUJIA NGK ST V02	12
…	…	…

3 Antecedentes relacionados

A continuación se describen brevemente dos trabajos relevantes que integran sistemas de recuperación de información con sistemas de bases de datos estructuradas o semiestructuradas. Sin embargo, cabe destacar que no se encontraron antecedentes en el ámbito comercial como el que se desarrolla aquí.

En el ámbito de la salud se encontró un proyecto aplicado sobre la BD biomedicinal MEDLINE [1]. Esta 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. Los autores desarrollaron dos sistemas de indexación y búsqueda utilizando dos tecnologías aplicadas al tratamiento de los datos (LUCENE[1] y PostgreSQL[2]) que mejoró la capacidad de búsqueda y recuperación de información en MEDLINE. Una vez construidos ambos sistemas de búsqueda, éstos fueron evaluados en cuanto a rendimiento para luego decidir cuál era el más apropiado para manejar la base de datos de MEDLINE. Los autores eligieron a LUCENE porque está optimizada para bases de datos textuales y por tanto ofrece mejores posibilidades para tratar datos no estructurados.

En [2] se extiende 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 abarcó 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 los autores proponen crear reglas de optimización asociadas a las primitivas

[1] Lucene es un grupo de librerías escritas en JAVA que brinda las primitivas necesarias para el tratamiento de datos textuales en la recuperación de información.

[2] PostgreSQL es un SGBD similar a MySQL u Oracle, pero más robusto que estos dos últimos cuando se necesita operar con BD grandes

de estas consultas combinadas. Esta tarea es la que se realizará en la segunda etapa. Como resultado [3] se obtuvo un prototipo que realiza búsquedas en la colección de documentos de la base de datos y en la estructura de la misma (etiquetas XML) para obtener así información más certera.

4 El prototipo desarrollado

El prototipo se desarrolló teniendo en cuenta dos procesos principales: Proceso de RI y Proceso de Clasificación, los mismos a su vez están divididos en subprocesos. El proceso de RI está compuesto por los subprocesos: *Depuración de datos, Indexación de datos y Búsqueda*. El proceso de clasificación está compuesto por los subprocesos: *Entrenamiento del clasificador y Clasificación de los resultados*. Cada uno de estos subprocesos contempla la utilización de diferentes algoritmos y librerías de RI, así como también lenguajes de programación como SQL y Delphi. El motor de BD que se utilizó fue MySQL Server 5.0.

4.1. Los subprocesos de RI

La *depuración de datos* implicó realizar tareas de limpieza sobre la BD, como ser la eliminación de caracteres extraños (por ej. &, %, #, ã, Ø, ©, Œ) y espacios innecesarios. Este subproceso de depuración es necesario para optimizar los resultados de la indexación. La depuración se realiza a nivel de capa de datos, es decir, como procedimientos almacenados en la BD y convocados por el prototipo, pero ejecutados por el motor de la BD. Cada vez que se ingresan nuevos registros o se modifican registros existentes, el prototipo ejecuta este subproceso para depurar solamente estos registros.

La *indexación* genera un conjunto de archivos externos a la BD, cuya estructura es propia de la herramienta de RI utilizada (Sphinx). Este subproceso se realiza siempre que se agreguen nuevos datos a la BD. El prototipo, accede a los datos a través de una consulta sobre la tabla de interés, en la cual se seleccionan los campos que se desean indexar.

El *subproceso de búsqueda* es el encargado de procesar las consultas de los usuarios utilizando el índice generado por el subproceso de Indexación. El prototipo convoca al subproceso de búsqueda para realizar una búsqueda, así, a partir de una consulta de usuario utiliza los archivos índice y devuelve un listado de artículos relevantes. Luego genera una tabla índice dentro de la BD con el campo identificador de la tabla de datos, seguidamente realiza una unión de la tabla índice con la tabla de datos, y de esta manera puede mostrar información más detallada de los registros encontrados como resultado final de la consulta.

4.2. Los subprocesos de Clasificación

El subproceso de clasificación utiliza una modificación de los algoritmos propuestos por [4]. Aquí se clasifica cada registro de la lista obtenida por el subproceso de búsqueda teniendo en cuenta una categoría en particular. En primera instancia, se deduce o clasifica la consulta ingresada por el usuario para obtener cual es la

categoría más probable a la que pertenece. En segunda instancia, el clasificador utiliza dicha categoría para clasificar la lista de registros. El criterio para ambos pasos es similar. Así el subproceso de clasificación se separa en dos subprocesos, en el primero se clasifica la consulta del usuario y se obtiene la categoría más probable, y en el segundo subproceso se ordenan los registros encontrados mediante la búsqueda y utilizando la categoría encontrada en el proceso anterior. Durante el entrenamiento del clasificador se utilizó el algoritmo de Naives Bayes [4]. A continuación se presenta el modelo bayesiano utilizado en el subproceso.

Sean C un conjunto de clases tal que C={P,NP} donde P="Pertenece" y NP="No Pertenece", C'={c_1',c_2',…,c_m'} un conjunto de categorías definidas por el usuario en base a algún criterio subjetivo en donde cada una de éstas categorías toma los estados P y NP, X un conjunto de todos los artículos de la BD de la organización tal que $X=\{x_1,x_2,...,x_n\}$ y $V=\{t_1,t_2,...,t_k\}$ un conjunto de términos seleccionados en el proceso de indexación de la BD, se define la red bayesiana mostrada en la Figura 1 para realizar la clasificación de una lista de artículos en donde C_i representa las categorías definidas por el usuario y t_1, …, t_k son los términos claves que fueron indexados en el proceso de indexación.

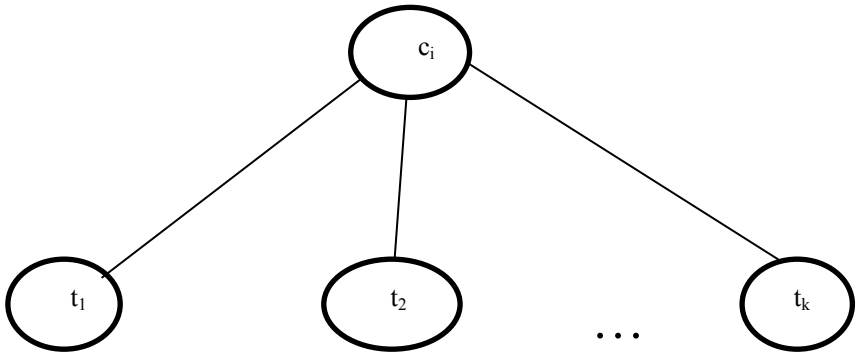

Fig. 1. Red bayesiana para la clasificación de listado de registros

Teniendo en cuenta la ecuación (1), la probabilidad a priori de c_i ($P(c_i)$) y las probabilidades a posteriori de t_j ($P(t_j/c_i)$) son estimadas en el proceso de entrenamiento sobre un conjunto inicial de artículos (registros). Para cada nuevo registro el prototipo convoca al subproceso de clasificación para la clase *P* y *NP*, luego obtiene el máximo valor de probabilidad entre P(*P*) y P(*NP*) para determinar si pertenece o no a una categoría $c_k^{'}$ determinada.

$$c_{MAP} = \arg\max_{c\in\square} P(c \mid d) = \arg\max_{c\in\square} P(c) \prod_{1\le k\le n_d} P(t_k \mid c) \qquad (1)$$

En el subproceso de entrenamiento, los parámetros que se deben tener en cuenta son: las categorías elegidas por el usuario, el conjunto de registros de entrenamiento y las dos clases, Pertenece (P) y No Pertenece (NP). Cada categoría posee estas dos clases. El algoritmo irá determinando las probabilidades P y NP para cada palabra de un conjunto de entrenamiento seleccionado a priori, considerando cada categoría

definida. Cabe aclarar que estas palabras se obtuvieron mediante un algoritmo de selección de términos relevantes.

Cuando se preparan los datos para el entrenamiento debe tenerse en cuenta que existe la posibilidad de que el usuario pueda ingresar nuevas categorías en cualquier momento. Además, luego de una clasificación de los resultados, el usuario podría clasificar manualmente cada registro para mejorar la exactitud del clasificador.

El subproceso de clasificación identifica cual es la categoría más probable a la que pertenece la consulta, para esto se utiliza un algoritmo que permite clasificar e identificar entre todas las categorías cual es la más probable. El algoritmo devuelve un conjunto de registros *L*. Para clasificar la lista de resultados *L* devueltos por el buscador, se convoca a rutinas que permiten obtener un coeficiente de ordenamiento para cada registro de la lista. Este coeficiente indica la relevancia que tiene un registro con respecto a la consulta del usuario. Una vez obtenido el conjunto de coeficientes asociados a cada registro del listado *L*, éste es ordenado en forma ascendente utilizando dicho coeficiente. Se obtiene así un listado de registros ordenados por relevancia. La Tabla 2 muestra los procesos y subprocesos que componen el prototipo desarrollado.

Tabla 2. Componentes del prototipo desarrollado

Procesos	Subprocesos	Descripción
Proceso de RI	Depuración	Elimina caracteres extraños.
	Indexación	Busca en la BD palabras claves y los indexa guardando en un archivo externo.
	Búsqueda	Utiliza el archivo índice generado para realizar búsqueda además de operaciones relacionales para obtener información más detallada sobre cada registro (artículos).
Proceso de Clasificación	Entrenamiento	Se realiza con un conjunto inicial de registros previamente clasificados por el usuario, de manera encontrar una función que permita clasificar registros cuya categoría se desconozca.
	Clasificación de los resultados	Determina a que categoría pertenecen nuevos registros de la lista de resultados devueltos por el proceso de búsqueda.

5 Experimentación y análisis de resultados

Las pruebas se realizaron considerando tres escenarios diferentes (*Sistema Actual, Prototipo de búsqueda sin clasificador y Prototipo de búsqueda con clasificador*). En los mismos se utilizaron como casos de prueba diez consultas especialmente formuladas por el encargado de compras de la organización, por ser ejemplos de búsquedas típicas en situaciones reales. Para evaluar los resultados obtenidos para los casos de prueba en cada uno de los tres escenarios propuestos se definieron cinco indicadores de rendimiento. La Tabla 3 muestra la unidad de medida y una pequeña descripción de cada uno de los indicadores utilizados.

Tabla 3. Indicadores de rendimiento utilizados

Indicadores	U. de medida	Descripción
Tiempo de respuesta (TR)	Milisegundos	Tiempo que el usuario debe esperar antes de obtener los resultados de su consulta
Cantidad de resultados (CR)	Registros	Cantidad de resultados devueltos por el sistema
Velocidad de respuesta (VR)	Registros por milisegundo	Cantidad de registros por milisegundo que devuelve el sistema
Precisión (P)	Valor real entre 0 y 1	Proporción de registros relevantes dentro del conjunto de registros recuperados
Exhaustividad (E)	Valor real entre 0 y 1	Proporción de registros relevantes de la BD que fueron recuperados

A fin de facilitar la comparación de los resultados de las pruebas realizadas con el SI actual (escenario 1), con el prototipo de búsqueda desarrollado pero sin el clasificador (escenario 2) y con el prototipo de búsqueda incluyendo el clasificador (escenario 3), los mismos se resumen en la Tabla 4.

Como puede observarse, el indicador *Tiempo de respuesta (TR*) es el que más evidencia aporta para demostrar que el uso del prototipo es más eficiente. Utilizando el prototipo sin clasificador los tiempos de búsqueda se reducen hasta alcanzar una diferencia de velocidad media (Vm) 120 veces mayor a la del sistema actual, y cuando se utiliza el prototipo con clasificador la diferencia de velocidad media es aproximadamente 11 veces mayor a la del sistema actual de la organización. Si bien en los dos escenarios que se utilizó el prototipo de búsqueda TR disminuyó considerablemente, al utilizar el clasificador el TR es mayor por cuanto el prototipo debe realizar operaciones adicionales para cumplir con el objetivo de clasificar la lista de resultados antes de ser devuelta al usuario. Con respecto al indicador Cantidad de resultados (CR) puede afirmarse que no se observan cambios significativos entre los valores arrojados en cada escenario, existiendo sólo una variación de 30 registros en promedio.

Comparando el indicador *Velocidad de respuesta (VR)* se observa que el prototipo de búsqueda sin clasificador es en promedio 13 veces más rápido que el prototipo con clasificador. Esto resulta lógico ya que como se explicó el prototipo de búsqueda que incluye al clasificador debe realizar tareas adicionales para clasificar los resultados antes de mostrárselos al usuario. La diferencia de velocidad media se calcula utilizando la ecuación (2):

$$Vm = \left(\sum_{1}^{n} \left({tr1_i}/{tr2_i} \right) \right)/n \qquad (2)$$

Siendo *tr1* y *tr2* los tiempos de respuesta del sistema actual y del prototipo respectivamente, y *n* la cantidad de pruebas realizadas. El resultado obtenido de la división entre tr1 y tr2 indica cuantas veces más rápido es el prototipo de búsqueda que el sistema actual. Si $tr1/tr2 > 1$ entonces el prototipo es más rápido que el SI, si $tr1/tr2 = 1$ indica que el prototipo y el SI tienen el mismo TR, si $tr1/tr2 < 1$, el prototipo es más lento que el SI actual. En la ecuación (2) se busca obtener un valor

Vm que indicará cuanto más rápido es el prototipo de búsqueda que el SI actual y debido a esto se calcula el promedio de los resultados de las divisiones de los TR.

Tabla 4. Comparación de los indicadores obtenidos en los tres escenarios de prueba

Casos de prueba	Escenarios	Indicadores de rendimiento				
		TR	**CR**	**VR**	**P**	**E**
amort* bora	SI actual	15702	57	0,004	0,614	1
	Prototipo de RI	140	45	0,321	0,511	0,657
	Prototipo de RI con Clasif.	2780	45	0,016	0,511	0,657
rotula golf	SI actual	19030	52	0,003	0,731	1
	Prototipo de RI	172	32	0,186	0,813	0,684
	Prototipo de RI con Clasif.	2125	32	0,015	0,813	0,684
bujia corsa	SI actual	18186	52	0,003	0,500	1
	Prototipo de RI	62	21	0,339	0,714	0,577
	Prototipo de RI con Clasif.	1281	21	0,016	0,714	0,577
cable bujia gol*	SI actual	19905	99	0,005	0,909	1
	Prototipo de RI	203	36	0,177	1,000	0,400
	Prototipo de RI con Clasif.	1906	36	0,019	1,000	0,400
correa 6pk* gol	SI actual	19530	86	0,004	1,000	1
	Prototipo de RI	235	39	0,166	1,000	0,453
	Prototipo de RI con Clasif.	2390	39	0,016	1,000	0,453
rot* gol thompson	SI actual	19811	52	0,003	1,000	1
	Prototipo de RI	79	28	0,354	1,000	0,538
	Prototipo de RI con Clasif.	1468	28	0,019	1,000	0,538
correa dist* gol*	SI actual	19201	110	0,006	0,536	1
	Prototipo de RI	250	78	0,312	0,372	0,492
	Prototipo de RI con Clasif.	3999	78	0,020	0,372	0,492
bujia golf	SI actual	15218	62	0,004	0,468	1
	Prototipo de RI	188	24	0,128	0,708	0,586
	Prototipo de RI con Clasif.	1718	24	0,014	0,708	0,586
filtro aire gol*	SI actual	1998	79	0,040	0,709	1
	Prototipo de RI	297	65	0,219	0,677	0,786
	Prototipo de RI con Clasif.	3499	65	0,019	0,677	0,786
optica corsa vic	SI actual	20076	32	0,002	0,688	1
	Prototipo de RI	234	13	0,056	0,538	0,318
	Prototipo de RI con Clasif.	781	13	0,017	0,538	0,318

Respecto al indicador *Precisión (P)* se observa que en dos casos de prueba no existen diferencias en los resultados obtenidos usando el SI actual y el prototipo de

búsqueda (con y sin clasificador), en cuatro casos el uso del prototipo mejora los resultados pero en otros cuatro los empeora. Dada esta situación se decidió calcular un valor promedio para el indicador, obteniéndose 0,715 para las pruebas con el SI actual y 0,733 con el prototipo. Estos resultados son un indicio de que el prototipo de búsqueda desarrollado es un poco más eficiente que el SI actual en la obtención de resultados relevantes.

En cuanto al indicador *Exhaustividad (E)*, se observó que el SI actual tiene mejor desempeño. Esto se debe en gran medida a la forma en la que se realizan las búsquedas en dicho sistema. Cabe destacar que este indicador por sí solo no demuestra nada, ya que si bien los valores llegan al máximo, se tiene que analizar la cantidad total de registros devueltos por la consulta y el tiempo de respuesta de la misma. En particular, los resultados arrojados por el prototipo de búsqueda, con y sin clasificador, presentan un valor medio es de 0,61 que se ubica por encima de la media general del indicador. Este valor medio podría mejorar ya que dependen en gran medida del conocimiento y practica que adquiera el usuario en la formulación de las consultas, es decir, cuanto más precisa sea la consulta del usuario, mejores resultados devolverá el prototipo, ya sea por la experiencia del mismo o por el uso de atajos que brinda el prototipo como son los comodines.

6 Conclusiones

El desarrollo del prototipo permitió al SI de la organización comercial incrementar su capacidad de búsqueda, ya que el mismo fue integrado satisfactoriamente a dicho SI. Luego del análisis de las pruebas realizadas al SI y al prototipo de búsqueda se pudo observar que se han mejorado notablemente los tiempos de respuesta, así como también la calidad de los resultados de las consultas de usuario. Esto se debe a que con la herramienta de RI utilizada (Sphinx) y la aplicación de técnicas de clasificación sobre los resultados, se logró obtener una lista refinada y ordenada de resultados considerando su relevancia respecto a la consulta de usuario. Además, es de resaltar que la precisión en los primeros puestos del ranking mejora a medida que el usuario realiza las consultas y el entrenamiento continuo del prototipo.

Con el prototipo de RI desarrollado la organización comercial objeto de estudio logró obtener familias de artículos semánticamente similares en los cuales se encontraron registros sintácticamente diferentes, es decir, el prototipo devolvió artículos relevantes de diferentes proveedores cuyos códigos poseen diferente nomenclatura. Además, se pudo comprobar que incorporando tecnologías de la información como las técnicas de Minería de Datos y la Inteligencia Artificial a las técnicas convencionales, se resuelven estas situaciones de manera más eficiente que utilizando solamente procedimientos, técnicas y métodos convencionales.

Actualmente se está considerando aplicar otras técnicas de clasificación, como por ejemplo, el vecino más cercano [4], a fin de comprobar si es posible mejorar aún más el desempeño del prototipo desarrollado.

Referencias

1. García, F., Fernández, Ch., Azancot, M. (2007). Desarrollo de un sistema de indexación y búsqueda sobre la base de datos de biomedicina MEDLINE. Recuperado el 13 de marzo de 2013 en http://biblioteca. universia.net/html_bura/ ficha/params/id/45165231.html
2. Hiemstra, D., Vries, A., Blok, H., Keulen, M., Jonker, W., Kersten, M. CIRQUID: Complex Information Retrieval queries in a Database. Recuperado el 13 de marzo de 2013 en http://doc.utwente.nl/47223/ 1/hiemstra03cirquid.pdf
3. Johan, L., Vojkan, M., Ramirez, G., de Vries, A., Hiemstra, D., Blok, H. (2005). "*TIJAH: Embracing IR Methods in XML Databases*", Information Retrieval Journal 8, Kluwer Academic Publishers, ISSN 1386-4564, pp. 547-570. Recuperado el 13 de marzo de 2013 en http://www.cs.utwente.nl/~hiemstra/papers/irj05.pdf
4. Manning, C., Raghavan, P., Schütze, H. (2009). An Introduction to Information Retrieval. Cambridge, England: Cambridge University Press.
5. Tolosa, G, Bordignon, F. (2008). Introducción a la Recuperación de Información: Conceptos, modelos y algoritmos básicos. (Universidad Nacional de Luján, Buenos Aires). Recuperado el 13 de marzo de 2013 en http://www.tyr.unlu.edu.ar/tallerIR/2008/docs/Introduccion-RI-v9f.pdf
6. Brisaboa, N., Farina, A., Pedreira, O., Reyes, N. (2007). Indexación dinámica para la recuperación de información basada en búsqueda por similitud. Recuperado el 13 de marzo de 2013 en http://www.sistedes.es/sistedes/ pdf/2007/JISBD-07-brisaboa-indexacion.pdf
7. Hernandez Orallo, J., Ramirez Quintana, M., Ferri Ramirez, C. (2004). Introducción a la Minería de Datos. Madrid, España: Prince Hall.
8. Artayer, L. (2006). Construcción de un prototipo de un Sistema de Información Basado en Ontología Trabajo final para optar por el título de Licenciado en Sistemas de Información. Universidad Nacional de Santiago del Estero