Incorporación de módulos lingüísticos a la indización y la consulta de Bases de Datos de Texto

Gustavo Crispino
crispino@fing.edu.uy
Instituto de Computación - Facultad de Ingeniería
Universidad de la República (Uruguay)
Julio Herrera y Reissig 565 - CP 11300
Montevideo - Uruguay

Ricardo Morán
Informática
Administración Nacional de Telecomunicaciones
Colonia 1920 4to. piso
Montevideo - Uruguay

Resumen

En este artículo se presentan los beneficios de la inclusión de módulos lingüísticos en los procesos de indización y consulta de textos, con el objetivo de mejorar la calidad del resultado de las consultas.
Se analizan las ventajas a partir de las pruebas realizadas con un prototipo que se desarrolló a tales efectos. Éste utiliza un conjunto de recursos para la construcción de los módulos de indización y consulta : un diccionario de palabras del castellano para vincular distintas variantes de una palabra con un mismo representante canónico, una gramática para la resolución de ambigüedades léxicas y un analizador.

Palabras clave : Inteligencia Artificial, Ingeniería Lingüística, Lenguaje Natural, Bases de Datos de Texto

1. Introducción

La explotación de Bases de Datos de Texto es un área en expansión. La "explosión" informativa implica la posibilidad de acceso a grandes cantidades de documentos no estructurados (texto) en distintos tipos de actividades : investigación, actividades empresariales, legales, etc. El acceso a redes internacionales y la disponibilidad masiva de computadoras personales multiplican la información disponible en texto. La respuesta informática a los problemas planteados por consultas a cantidades masivas de texto se ha centrado en distintos aspectos : modelos de datos, algoritmia y estructuras de datos para búsquedas eficientes, integración con bases de datos relacionales, interfaces ergonómicas con el usuario, etc.

Sin embargo, es habitual que los productos desarrollados para indizar y consultar estas bases, al buscar por palabra, grupo de palabras o "frases" no recuperen todas las variantes morfológicas de los términos propuestos para la búsqueda.
Para resolver este problema se propone, en general, una búsqueda por prefijos. Por ejemplo, para buscar los textos que contengan las palabras "auto", "automóvil", "automóviles", etc., tomar el prefijo "auto" para hacer la consulta. Como resultado de una búsqueda por este prefijo podrían obtenerse, además de las esperadas, palabras tales como "autor", "automático", "autoridad", "autorización", "automatización", etc., por lo que podrían recuperarse documentos no relevantes a los efectos de la búsqueda, fenómeno que se conoce con el nombre de "ruido".
También se propone la posibilidad de formular reglas para la formación de plurales, pero esta propuesta no constituye una solución dada la abundancia de excepciones, en particular para el castellano, y por otra parte, no contempla la variación de género, las desinencias verbales, los aumentativos y diminutivos, etc.

A un nivel mayor de dificultad, existen diversos modos de realización lingüística para la expresión de un concepto, por lo que puede ocurrir que la forma que se utilizó en el texto no coincida con los términos de la consulta. Por ejemplo, una consulta que incluya "... cancelación automática de deudas ... " podría no recuperar un documento que contenga la expresión "... la deuda quedará automáticamente cancelada ...". Esto provoca el fenómeno conocido como silencio, es decir, la omisión de documentos relevantes en la recuperación.
Estos fenómenos han sido estudiados en numerosos trabajos. Algunas definiciones extraídas de autores que plantearon esta problemática permitirá aclarar más estos conceptos.

El **ruido** es la tasa de documentos no pertinentes que sin responder a la pregunta formulada son seleccionados siguiendo esencialmente combinaciones accidentales de características de los términos propuestos para caracterizar la búsqueda [CHAUMIER84].
El **silencio** es la tasa de documentos pertinentes de acuerdo a una consulta formulada, que están en la base documental y no son seleccionados como resultado de la búsqueda [CHAUMIER84].
La **recuperación** es una métrica utilizada para evaluar la capacidad que tiene el método de búsqueda para obtener documentos pertinentes, es decir, considerados útiles por quien hizo la consulta. Es el cociente entre el total de documentos pertinentes obtenidos y el de documentos pertinentes de la base. Por lo tanto, puede interpretarse como la probabilidad de que un documento pertinente sea obtenido [SALTON83].
La **precisión** es una métrica utilizada para evaluar la capacidad que tiene el método de búsqueda para no obtener documentos no pertinentes. Es el cociente entre el total de documentos pertinentes obtenidos y el de documentos obtenidos. Por lo tanto, puede interpretarse como la probabilidad de que un documento obtenido sea pertinente [SALTON83].

A nivel comercial se han desarrollado algunos productos tendientes a enriquecer con técnicas lingüísticas los procesos de indización y consulta, como por ejemplo *Topic-Aleth*, surgido de la conexión establecida entre *Aleth* de la empresa *GSI-Erli* y *Topic* de la empresa

Verity [NORMIER95 et al.], pero hasta ahora no existe ningún producto de este tipo para el castellano.

En este trabajo se presentan las principales características de dos módulos lingüísticos, uno para incluir en un proceso de indización y otro para incluir en un proceso de consultas, y algunos de los resultados obtenidos a partir de la experimentación llevada a cabo.

2. La indización

La indización consiste en la creación y mantenimiento de índices a partir de las palabras presentes en un texto.
Según Chaumier, la **indización** es la expresión más o menos condensada de las características de un documento en los términos de un lenguaje propio y fuertemente restringido con respecto al lenguaje natural [CHAUMIER84].
La idea básica de la indización automática consiste en utilizar aspectos formales del texto para determinar cuál es el tema que se trata en el mismo; algunos de estos aspectos tienen que ver con la ubicación y frecuencia de un término o conjunto de términos en el texto. Puede ser hecha sobre todo el documento o sobre un resumen del mismo, y se efectúa según distintos criterios; por ejemplo, tomando la frecuencia con que un término aparece en un documento para determinar su importancia.
Los sistemas de Bases de Datos de Texto permiten habitualmente indizar "a palabra", excluyendo aquéllas que pueden ser consideradas "vacías" (artículos, preposiciones, etc.) por carecer de interés en la consulta de documentos.
Por lo tanto, en estos sistemas, se resuelve la indización sin determinar qué palabras o grupo de palabras del texto son más relevantes para "representar" el tema tratado en el documento, y sin establecer vínculos entre diferentes variantes morfológicas de las palabras.
A los efectos de obtener mejores resultados en las consultas (desde el punto de vista de las métricas definidas por Salton en [SALTON83]) se aborda aquí este último aspecto.

En el prototipo desarrollado se destacan los siguientes componentes que apuntan en la dirección anteriormente señalada :

- sustituir, por su correspondiente representante canónico, las palabras formadas por flexión o derivación; por ejemplo, un verbo conjugado por su correspondiente infinitivo, los sustantivos en femenino y/o plural por su correspondiente en masculino singular, etc.

- obtener todas las categorías gramaticales posibles de una palabra y determinar cuál es la correcta en base al contexto establecido por su ubicación en la frase; por ejemplo,

bajo	(adjetivo, adverbio, preposición, sustantivo, verbo)
enfrente	(adverbio, verbo)
harto	(adjetivo, adverbio, verbo).

Es importante destacar la conveniencia de incluir un módulo en la indización. Si se incluyera solamente el módulo de la consulta, al producir por cada término de búsqueda todas sus variantes morfológicas se estaría multiplicando la cantidad de ecuaciones de búsqueda pudiendo hacer sumamente ineficiente el proceso. Habiendo almacenado durante la indización las palabras en su forma canónica, es suficiente con que el módulo de la consulta dé el mismo tratamiento a las palabras usadas para formular la interrogación.

Para incorporar estas técnicas se utilizó un **procesador lingüístico**, es decir, un conjunto de recursos que permiten efectuar el análisis. Este conjunto se compone de un **diccionario** [DICTILLE], una **gramática** y un **analizador**, los que serán presentados en la sección 4.

3. La consulta

Como se expresó anteriormente, los sistemas de Bases de Datos de Texto indizan habitualmente "a palabra". Una consulta por una palabra suele no ser suficientemente específica; es así que, en general, se realizan consultas a través de expresiones constituidas por varios términos vinculados por operadores.
Los operadores más frecuentemente implementados en estos productos son los *operadores lógicos* ("and", "or", "not"), los *operadores relacionales* (">", "<", "=", etc.) y los *operadores de proximidad* (dentro del mismo párrafo, dentro de la misma frase, a determinada distancia medida en número de palabras, etc.)

En este prototipo se implementó un módulo de consultas que actúa de manera simétrica con el módulo de indización y permite interrogar

- por prefijo
- por una secuencia de palabras especificando una distancia máxima entre ellas
- por frase.

Para el caso de las consultas por frase, cada uno de los términos de la consulta (excepto las palabras vacías) es transformado en su correspondiente representante canónico antes de realizar la búsqueda en las estructuras de índices.

4. Implementación de los módulos lingüísticos

Se presentarán en esta sección los aspectos más destacables de estos módulos y de las estructuras por ellos manejadas.

4.1. Procesamiento de la cadena de entrada

Durante el procesamiento del texto se hace un tratamiento especial de ciertos símbolos, que si bien en los sistemas de Bases de Datos de Textos no es relevante a los efectos de la incorporación de términos a la estructura de índices, deben ser considerados en el módulo de

indización de este prototipo para la aplicación de reglas de desambiguación. Tal es el caso de los caracteres ".", ",", ":", "-", "!", "¿", y "?" que marcan límites para la aplicación de las reglas de la gramática de desambiguación.

Para la indización de los textos se siguieron los siguientes procesos :

- transformación de mayúsculas en minúsculas y tratamiento de caracteres especiales
- identificación del "canónico" de cada palabra para lo cual se usaron las tablas del diccionario, se incorporó un tratamiento de pronombres enclíticos y un tratamiento para los verbos irregulares
- tratamiento de desambiguación
- eliminación de palabras "vacías"

4.2. Estructura de índices

Puesto que no se consideró central para los objetivos de este prototipo la eficiencia en lo relativo a tiempos de acceso a los índices y espacio físico necesario para su almacenamiento, se optó por una implementación tradicional de archivos invertidos que permitiera llevar a cabo la experimentación a la vez que no demandara esfuerzos adicionales de implementación.
Cada entrada del archivo índice de palabras no vacías contiene la siguiente información : la palabra tal como aparece en el texto, el representante canónico, la categoría gramatical, el número de documento y el ordinal dentro de ese documento.

4.3. Estructura del diccionario

El diccionario Dictille 1.0 [DICTILLE] es un paquete compuesto por

- Un archivo *pdes.dat* que contiene formas invariables - preposiciones, conjunciones, adverbios-, pronombres, adjetivos determinativos, etc., y sus correspondientes homógrafos; así como las de verbos con flexión anormal (*ir*) o muy frecuentes (*haber*). Todas ellas con sus respectivas descripciones morfológicas y categoriales.

- Un archivo *dic.dat*, que contiene lemas, categoría morfológica -sustantivo, adjetivo, verbo-, sus respectivas raíces, y su modelo de flexión morfológica.

- Dos tablas de modelos de flexión : verbal y nominal.

- Dos tablas de todas las desinencias verbales y nominales.

4.4. Obtención del representante canónico

En el proceso de indización se realiza una transformación de cada palabra no vacía en su representante canónico. De esta manera se tienen vinculados dentro de la estructura de índices las diferentes variantes morfológicas de una palabra.
En el momento de la consulta se aplica la misma transformación, por lo que al hacer la interrogación usando un término en particular se podrá acceder a través del representante

canónico almacenado en la estructura de índices a todos los variantes morfológicas de ese término que aparecen en los textos.
Para realizar estas transformaciones se implementó un analizador morfológico que utiliza la información del diccionario Dictille. Este analizador tiene además una función para separar los pronombres enclíticos unidos a los verbos. Por ejemplo, analiza la palabra "firmándoselo" como firmando-se-lo, reconociendo los pronombres enclíticos "se" y "lo" y almacenando en el índice, junto con la palabra "firmándoselo" su representante canónico "firmar", para lo cual utiliza la información del diccionario Dictille que permite establecer la vinculación entre "firmando" y "firmar".

4.5. Gramática para desambiguación

En el proceso de obtención del representante canónico puede ocurrir que una palabra, tal como aparece en el texto, pertenezca a más de una categoría gramatical. Por ejemplo, la palabra "sobre" puede ser una preposición, una conjugación del verbo "sobrar", o un sustantivo (cubierta en la que se incluye una carta). Si se indizara tal como aparece en el texto o si se descartara como palabra vacía (que es lo que hacen habitualmente los programas que indizan texto) se estaría realizando un análisis incompleto.
En este prototipo se analiza, utilizando una gramática, el contexto de la frase o expresión en que aparece la palabra para determinar cuál es su categoría gramatical. Para este ejemplo puede ocurrir :

a) que sea una preposición, en cuyo caso se descarta por ser palabra vacía

b) que sea una conjugación del verbo "sobrar", en cuyo caso se le asocia como representante canónico el infinitivo "sobrar".

c) que sea un sustantivo, en cuyo caso se le asocia como representante canónico el sustantivo masculino singular "sobre".

El módulo de consulta procede de la misma manera.

La gramática utilizada es una gramática libre de contexto, cuyas reglas indican -en base a heurísticas, análisis estadísticos y reglas de sintaxis- las posibles combinaciones de categorías gramaticales.
Existen reglas que permiten desambiguar

determinante/verbo
preposición/verbo
preposición/verbo/sustantivo
sustantivo/verbo
etc.

Se definieron e implementaron las reglas de desambiguación en base a la palabra anterior y la palabra posterior a la ambigua, utilizando como contexto de aplicación de las reglas los caracteres " ", ",", ".", "-", "!", "?", "¿" y la ocurrencia de una o más líneas en blanco.
Se presentan a continuación algunos ejemplos de reglas :

- determinante/verbo (por ejemplo : una, unas)

si la palabra que precede a la ambigua es pronombre personal o si la palabra que le sigue es un determinante se trata de verbo; en caso contrario se toma como determinante

- preposición/verbo/nombre (por ejemplo : sobre)

si la palabra precedente es un pronombre personal, se trata de un verbo; si la palabra precedente es un pronombre posesivo, un determinante o alguna de las palabras de la lista "éste, éstos, ésta, aquél, aquéllos" es un sustantivo; en cualquier otro caso se toma como preposición

- nombre/verbo (por ejemplo : cargo)

si la palabra precedente es un determinante se trata de un sustantivo; si la palabra precedente es un pronombre personal se trata de un verbo.

5. Experimentación

Se desarrolló el prototipo en lenguaje *C* en una estación de trabajo *Sun* bajo sistema operativo *Unix*.
Se realizó una experimentación en base a un "corpus" de artículos periodísticos.
A partir de esta experimentación se pueden verificar algunas ventajas de este tipo de procesamiento respecto a una indización que no tenga en cuenta los elementos aquí tratados.
Por ejemplo,

- ante la consulta " ...generación de obligaciones ..." se obtuvo, entre otros, un texto que contiene " ...generan la obligación..." como consecuencia de las siguientes transformaciones :

- del sustantivo generación se pasa al verbo generar y de allí a su forma conjugada "generan";
- se eliminan respectivamente las palabras vacías "de" (en la consulta) y "la" (en la indización);
- del sustantivo "obligaciones" se obtiene su forma singular "obligación".

- análogamente, ante la consulta "...elaboración de arreglos..." se obtuvo, entre otros, un texto que contiene "... elaboraron y ejecutaron un arreglo ..."

Un caso particularmente interesante de esto se verifica comparando consultas por prefijo (disponible en general en los productos para acceder Bases de Datos de Texto) cuando en las mismas se involucran verbos irregulares.
Por ejemplo,

- ante una consulta que contiene "...conferir ..." intentando recuperar distintas conjugaciones del verbo, como por ejemplo "confieren", se realizó una consulta, sin utilizar

los módulos lingüísticos por el prefijo "conf". Se obtuvieron documentos que contienen las palabras "conferencia", "confesando", "confianza", "conflicto" y "confusión", mientras que con la indización realizada utilizando los módulos lingüísticos se consulta por "conferir" obteniendo todas las ocurrencias en el texto de conjugaciones de dicho verbo, como por ejemplo "confieren", eliminando el "ruido" producido con la búsqueda por prefijos.

- análogamente, ante una consulta por prefijos para obtener documentos que contenga "...desplegar..." incluyendo sus variantes conjugadas, se obtuvieron documentos que contienen las palabras "desplazamientos", "desplazando" y "desplazarse". Haciendo la consulta en base a la indización realizada utilizando los módulos lingüísticos se consulta por "desplegar" obteniendo todas las ocurrencias en el texto de conjugaciones de dicho verbo, como por ejemplo "despliegue", eliminando el "ruido" producido con la búsqueda por prefijos.

6. Conclusiones

Si bien no ha finalizado aún la etapa de experimentación, en base a las pruebas realizadas, podemos afirmar que la incorporación de módulos lingüísticos mejora la calidad del resultado de las consultas a Bases de Datos de Texto.

Por un lado, la gramática para desambiguar la categoría gramatical de una palabra con ambigüedad léxica aumenta la *precisión* (disminuye el "ruido") ya que impide la indización incorrecta basada en una interpretación errónea de la categoría gramatical a la que pertenece la palabra analizada.
Por otro lado, la indización por representante canónico aumenta la *recuperación* (disminuye el "silencio") ya que permite obtener en una misma consulta más variantes de una palabra que si se indiza "a palabra" o "a expresión".
Analizando pues ambas métricas, concluimos que se mejora cada una de ellas sin degradar la otra.

El trabajo futuro se sitúa a dos niveles.

En primer lugar, finalizar la experimentación y analizar las dificultades que deben superarse para incorporar estos módulos a programas que manejan Bases de Datos de Texto.

En segundo lugar, y teniendo en cuenta los resultados promisorios obtenidos en la fase de experimentación en que actualmente nos encontramos, estudiar la incorporación de recursos lingüísticos adicionales. A modo de ejemplo, la posibilidad de establecer vínculos semánticos entre los términos de un texto, con herramientas del tipo WordNet ([WORDNET]), y/o la posibilidad de disponer de bases de conocimiento lexicales como las que se están desarrollando en los proyectos Acquilex ([ACQUILEX]), Eagles ([EAGLES]), y Sparkle ([SPARKLE]), entre otros, permitiría mejorar los resultados.

Referencias bibliográficas

[ACQUILEX] Acquilex : The Acquisition of Lexical Knowledge
www.cl.cam.ac.uk/Research/NL/acquilex/projdesc.html

[CHAUMIER84] Chaumier, Jacques : Les techniques documentaires.
Entreprise Moderne d'Edition - Paris, 1984

[DICTILLE] Diccionario DICTILLE 1.0. ACTA. Comunicación y Publicación.
España

[EAGLES] Expert Advisory Group on Language Engineering Standards
http://www.ilc.pi.cnr.it/EAGLES/home.html

[WORDNET] A Lexical Database for English
Princeton University. DARPA/ITO Research Areas.
http://www.ito.darpa.mil/Summaries95/B370--Princeton.html

[NORMIER95 et al.] Normier, Etienne; Le Loarer, Pierre
Intégration des techniques documentaires et linguistiques :
Topic-Aleth
Génie linguistique 95, Montpellier (France)

[SALTON83] Salton, Gerard; McGill, Michael
Introduction to modern information retrieval
McGraw-Hill Book Company, 1983

[SPARKLE] Proyecto Sparkle
http://www.ilc.pi.cnr.it/sparkle.html