

Uso del recurso lingüístico WordNet en la expansión de consultas con un modelo del usuario de recuperación de información

Francisco João Pinto

Departamento de Ciencias de Computación

Universidad de la Coruña

Campus de Elviña, S/N, 15071 A Coruña

fjoao@udc.es

Resumen

Este trabajo describe la experimentación llevada a cabo para probar la efectividad de la expansión de consultas utilizando una metodología de la evaluación basada en la simulación bajo el modelo del usuario. Realizamos diferentes experimentos de consultas utilizando la información lingüística extraída desde WordNet. Los experimentos muestran que la expansión no es capaz de mejorar la ejecución de la recuperación de información sin la selección del significado correcto de la palabra de entre el conjunto de posibles significados; a este proceso le llamaremos desambiguación del sentido de la palabra.

1. Introducción

La recuperación de datos, en el contexto de un sistema de IR, consiste básicamente en determinar que documentos de una colección contienen las palabras clave insertadas por el usuario, lo cual no es suficiente para satisfacer la necesidad de información del usuario. De hecho, un usuario de un sistema de IR está interesado en obtener información sobre un tema determinado.

El objetivo de este trabajo es probar si la expansión de consultas utilizando el modelo del usuario mejora la ejecución de la recuperación. Los usuarios de sistemas de recuperación que usan la coincidencia de palabras como base de la recuperación se enfrentan con el reto de expresar sus consultas con palabras de los

vocabularios de los documentos que desean recuperar. Esta dificultad es grave en grandes bases de datos de texto debido a que dichas bases de datos contienen muchas expresiones diferentes para referirse al mismo concepto. La habilidad para recuperar documentos de estas bases de datos es crucial en un amplio rango de aplicaciones: recuperar documentación para la ayuda en casos legales, facilitar la organización y recuperación de correspondencia y formularios en una oficina, filtrar fuentes de noticias para encontrar artículos de interés, encontrar pasajes relevantes dentro de un conjunto completo de manuales de un sistema complejo para un problema particular a mano, etc. Un método de aligerar la carga del usuario cuando selecciona las palabras de una consulta es que el sistema de recuperación expanda automáticamente la consulta añadiéndole términos que estén relacionados con las palabras proporcionadas por el usuario. Los nuevos términos pueden estar estadísticamente relacionados con las palabras originales de la consulta, es decir, los términos tienden a concurrir en documentos o pueden ser elegidos con la ayuda de los recursos lingüísticos.

El uso de información de ocurrencia para expandir vectores de consulta es atractivo debido a que las relaciones se pueden generar fácilmente a partir de los documentos, obviando la necesidad de recursos lingüísticos u otras herramientas adicionales, que son costosas de desarrollar y mantener. Desafortunadamente, tales métodos han tenido poco éxito en mejorar la eficacia de la recuperación cuando se

usan sin datos de relevancia. Sin embargo, métodos que explotan relaciones estadísticas pero no expanden la consulta, como el Latent Semantic Indexing [8], sí han sido más exitosos.

El uso de información lingüística como fuente de términos relacionados ha encontrado algún éxito en pequeños experimentos. Salton y Lesk [4] encontraron que la expansión mediante sinónimos mejora el rendimiento, pero los resultados de la expansión mediante términos más generales o más restringidos seleccionados de recursos lingüísticos jerárquicos eran demasiados inconsistentes para ser útiles en general, al realizarse sobre colecciones pequeñas. Wang, Vendendorpe y Evens [13] encontraron que una variedad de relaciones léxico-semánticas mejoran el rendimiento. Sin embargo, cada una de estas conclusiones se alcanzó en experimentos sobre colecciones muy pequeñas usando recursos lingüísticos de un único dominio. En cuanto a la utilidad de la expansión de consultas mediante relaciones léxico-semánticas en colecciones grandes que abarcan varios dominios, los trabajos más representativos son los que utilizan la colección de referencia TREC que comentaremos más adelante.

Los estudios previos con estas características en el uso de información lingüística han ofrecido distintos resultados. Voorhees [3] mostró que usar WordNet para la expansión de consultas no aumenta la efectividad de la recuperación de información. En un trabajo de Mandala [7] las relaciones almacenadas en WordNet se combinan con otras medidas de similitud basadas en dependencias sintácticas e información de concurrencia y sí mejoran considerablemente la eficacia de la recuperación de información.

En el trabajo de Ellen M. Voorhees [3] las consultas se expanden usando las relaciones codificadas en WordNet, recurso lingüístico que comentaremos en la sección 2, y son evaluadas contra los tópicos del 401 al 450 producidos por la octava conferencia TREC. Para evitar los efectos distorsionadores de partir de una consulta original con selección pobre de palabras, los términos originales de la consulta que se expanden fueron generados automáticamente a partir de los tópicos, seleccionado ade-

más los conjuntos de sinónimos de WordNet que se consideraba que resultaban los conceptos importantes del tópico. Por eso, los resultados obtenidos en este trabajo representan un límite superior en el rendimiento que se puede esperar en un procedimiento de selección totalmente automático que use la estrategia de expansión. Incluso en el mejor caso, la expansión no mejoró la efectividad de las consultas que eran completas inicialmente, es decir, de las consultas que contenían términos suficientemente representativos y que describían el tópico de interés con detalle. Consultas menos completas, consistentes en una sentencia simple describiendo el tópico de interés, sí mejoraban significativamente con la expansión.

En la sección 2 se explicarán de forma clara el modelo utilizado y cuál va a ser la medida de similitud entre las consultas y los documentos para determinar cuáles son relevantes y cuáles no. Además esta medida será útil para determinar un orden (rankear) de los documentos relevantes. En la sección 3 se describirá brevemente WordNet. En la sección 4 se presentará la aplicación desarrollada para la formulación de las consultas y diseño de los experimentos. En la sección 5 se describen los experimentos y sus resultados. Las conclusiones son presentadas en la sección 6.

2. Modelo de usuario

Uno de los problemas más importantes en recuperación de información consiste en formular la consulta para que plasme adecuadamente la necesidad informativa del usuario. Aparte de los requerimientos del sistema para formalizar la consulta, el mayor problema consiste en determinar el conjunto de palabras que expresen semánticamente esa necesidad. El problema se agrava debido al efecto de inconsistencia en la asignación subjetiva de términos a conceptos. Figuras como la sinonimia o la polisemia (u otras menos importantes, como la homonimia, antonimia, la hiperonimia, la hiponimia etc) hacen que el mismo concepto pueda expresarse con palabras diferentes y una misma palabra pueda aparecer en documentos que tratan sobre temas distintos. En

esta situación no es de extrañar que el usuario tenga que replantear su consulta para obtener mejores resultados.

Se han propuesto diversos mecanismos para construir la nueva consulta. En general, en todos ellos se realiza una ampliación de nuevos términos a la consulta inicial y un recalcado de la importancia de cada término. Esto es lo que se conoce como expansión de consulta. Se pretende ampliar el número de términos que mejor definan la necesidad informativa del usuario de acuerdo a la colección documental y al modelo de recuperación utilizado.

El interés para agilizar la expansión de las consultas se centra en consultas con muy pocos términos, pues las consultas largas suelen proporcionar buenos resultados de recuperación, al incluir más términos comunes con los documentos relevantes. De hecho, la mayor parte de las consultas que se realizan en buscadores y sistemas de información en Internet tienen de uno a tres términos [2].

2.1. Descripción del modelo

Supongamos que T , D , N son campos del tópico t de TREC [12] de la consulta q , donde T es el título, D es la descripción y N es la narrativa, que se pueden representar por $T + D + N$, eliminando los campos $N + D$, quedaríamos con sólo el título, debido a que según los experimentos realizados por Voorhees [3] demostraron que las consultas expandidas obtenidas a partir de consultas originales, más cortas, sí consigue mejoras significativas con respecto a las originales en su versión más larga. Además los experimentos de Voorhees demostraron, por lo tanto, que la expansión mediante relaciones léxico-semánticas proporcionan un beneficio pequeño cuando el usuario proporciona una consulta detallada. Aún así, los usuarios frecuentemente no proporcionan una consulta detallada. En este caso, las relaciones léxico-semánticas tienen el potencial de mejorar las consultas iniciales, aunque la consulta expandida es poco probable que sea tan efectiva como una consulta mejor formulada y proporcionada por el usuario.

Sea $\vec{q}_1 = \langle A_0, B_0, C_0 \rangle$ el vector de la consulta no expandida con sólo el título, donde A_0 ,

B_0 y C_0 son términos pertenecientes a q_1 en los cuales A_0 y C_0 son nombres. Supongamos también que a partir de la expansión eligiendo sólo los nombres que aparecen en la descripción textual del tópico utilizando el recurso lingüístico obtenemos los términos siguientes: A_1 , A_2 y A_3 como sinónimos relacionados con A_0 y C_1 , C_2 , C_3 y C_4 como sinónimos relacionados con C_0 , entonces el nuevo vector de la consulta expandida quedará del siguiente modo: $\vec{q}_2 = \langle A_0, B_0, C_0, A_1, A_2, A_3, C_1, C_2, C_3, C_4 \rangle$. Supongamos también que en la consulta: $\vec{q}_2 = \langle A_0, B_0, C_0, A_1, A_2, A_3, C_1, C_2, C_3, C_4 \rangle$ se ha seleccionado los términos A_1 , C_1 y C_2 como los correctos la consulta quedará de la siguiente manera: $\vec{q}_3 = \langle A_0, B_0, C_0, A_1, C_1, C_2 \rangle$ con esto se pretende que si en una consulta un termino original tiene diversos significados, se supone que aquellos que son distintos del correcto no influyen en eficacia de la recuperación. Supongamos además que cada documento y cada consulta son representados por un vector de la frecuencia del término $\vec{d} = \langle x_1, x_2, \dots, x_n \rangle$ y $\vec{q} = \langle y_1, y_2, \dots, y_n \rangle$ respectivamente, donde n es el número total de términos, o tamaño del vocabulario y x_i , y_i son las frecuencias del término t_i en d y q respectivamente.

Dada una colección C , la frecuencia del documento inverso (*idf*) de un término es dado por $\log(N/n_i)$, donde N es el número total de documentos en C y n_i es el número de documentos con el término i . Todos los términos en la consulta son pesados por la fórmula de pesado *TFIDF* heurística. Esto es los vectores pesados por d y q son:

$$\begin{aligned} \vec{d} &= \langle tfd(x_1)idf(t_1), \dots, tfd(x_n)idf(t_n) \rangle \\ \vec{q} &= \langle tfq(y_1)idf(t_1), \dots, tfq(y_n)idf(t_n) \rangle \end{aligned}$$

$RawTF = TF(t_i, d_j)$ indica la frecuencia del término t_i en el documento d_j e *idf* = $\log(N/n_i)$ la frecuencia del documento inverso. Si el mínimo es 0 entonces $n_i = N$, es decir, el término t_i aparece en todos los documentos; si el máximo es $\log(N)$, entonces $n_i = 1$, es decir, el término t_i aparece en un documento, siendo t_i es el número de términos índices.

La frecuencia inversa depende de la colección, necesitamos N (número de documentos

pero es un entero que depende de la colección). Finalmente el esquema de pesado *tf/idf* asigna el siguiente peso al término t_i en el documento d :

$$W_{i,d} = \text{RawTF} * \text{IDF} = \text{RawTF} * \log(N/n_i)$$

La función *TF* de la consulta está definida de forma similar como sigue:

$$W_{i,q} = \text{RawTF} * \text{IDF} = \text{RawTF} * \log(N/n_i)$$

La puntuación de un documento d con la consulta q viene dada por la fórmula

$$S(\vec{d}, \vec{q}) = \sum_{i=1}^n \text{tf}_d(x_i) \text{tf}_q(y_i) \text{idf}(t_i)$$

Una vez obtenida la consulta expandida, el campo título seleccionado es indexado utilizando las rutinas estándares del Sistema de Recuperación de Información Lemur, se computa un rank de documentos para cada consulta. Para cada consulta los documentos están ranked en orden decreciente de similitud. La puntuación de un documento d con la consulta q es dada por una expresión matemática a cima representada y posteriormente una vez obtenidos los resultados de la consulta expandida y de los distintos experimentos contemplados se procederá a la evaluación y la comparación entre ellas con las consultas sin expandir.

3. WordNet

WordNet es un diccionario MRD (Machine Readable Dictionaries), para el idioma inglés (ver [1, 10, 13]) convirtiéndose en uno de los recursos más valiosos para el procesamiento del lenguaje natural (PLN). El desarrollo de WordNet se inició en 1985 y en el Laboratorio de Ciencias Cognitivas de la Universidad de Princeton bajo la dirección del Profesor de Psicología George A. Miller. Este recurso posee una base de datos que agrupa las palabras en conjuntos de sinónimos llamados synsets y provee definiciones, comentarios y ejemplos de uso de estas palabras y sentido de las mismas. De esta manera, combina los elementos de un

diccionario (definiciones y algunos ejemplos) y los de un tesoro (sinónimos), y crea un apoyo muy importante para el análisis automático de textos y palabras.

4. Generación de consultas

El procedimiento de expansión usado en este trabajo fue el siguiente: A partir de los ejemplos de necesidades de información que la TREC incluye se realizaba una selección automática de términos. Para cada uno de esos términos, y mediante la ayuda de WordNet, se seleccionaba, también de forma automática, un conjunto de sinónimos de cada uno de estos términos originales que se consideraban del sentido más correcto para añadir al vector de la consulta. Todos los sinónimos contenidos dentro de un conjunto de sinónimos de la cadena son añadidos a la consulta. Las palabras más comunes en el idioma o stop words, como por ejemplo of, se quitan, y a las palabras restantes no se les ha aplicado un procesado morfológico consistente en quedarse con la raíz de la palabra (stemming). Para implementar la aplicación se han utilizado las tecnologías asociadas a la plataforma Visual Basic .Net. que nos permite probar las diferentes opciones de la expansión.

Para la realización de la interfaz de usuario se ha optado por una interfaz de ventanas bajo el sistema operativo Windows XP. La capa modelo está implementada en Basic .Net. A la hora de definir el ejercicio o experimento existen un gran número de posibilidades cuyo significado se detalla a continuación: En primer lugar, es necesario indicar el rango de tópicos sobre los que se realizará el ejercicio. Para ello, en el primero de los campos del formulario (Introducir tópicos (151-550)) se introduce un rango en la formulación min-max, donde min será el número de tópico de comienzo y max el número de tópico a usar en el ejercicio. Los números de los tópicos deben estar comprendidos entre 151 y 550 (ver figura 1).

Después se selecciona el algoritmo de stemmer. Seleccionar stemmer para consulta a aplicar sobre las palabras, para pasar a formar parte de las consultas. Se puede seleccionar no

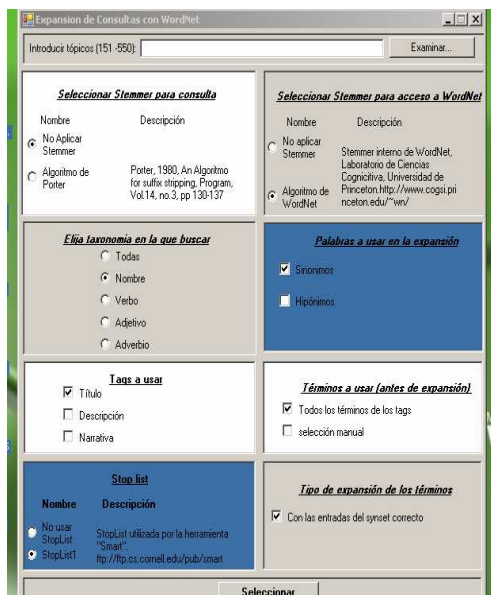


Figura 1: Interfaz de la aplicación que nos permite definir las opciones de la expansión.

aplicar ninguno, o bien aplicar el algoritmo de Porter [5]. Si se elige no aplicar ninguno, las palabras de las consultas irán completas, si se elige el algoritmo de Porter aparecerán únicamente las raíces obtenidas por este algoritmo.

La siguiente posibilidad Seleccionar Stemmer para acceso a WordNet nos permite indicar si queremos que a la hora de buscar una palabra en WordNet se busque de manera exacta y sólo se encuentre si está incluida exactamente o si se desea aplicar el procesamiento morfológico de WordNet encontrando así también las palabras cuya raíz aparece en WordNet aunque no aparezca en sí misma. Según la opción que se escoja en Elija la taxonomía en la que buscar la búsqueda de significados devolverá los resultados de buscar las palabras sólo como nombres, sólo como verbo, sólo como adjetivo, sólo adverbio o buscando todos los significados de todas las palabras en todas las taxonomías.

La siguiente sección Palabras a usar en la expansión permite indicar que palabras se añadirán a cada consulta en la expansión pudiendo-

se expandir con sinónimos, hipónimos de primer nivel, o con ambos tipos a la vez, ya que se permite la selección múltiple. Como las consultas se generan a partir de los tópicos de los títulos la siguiente opción es Tags a usar, donde se debe indicar a partir de que partes de un tópico se deben generar, pudiéndose escoger entre el Título (Headline), la Descripción (Head) o la Narrativa (Text) o cualquier combinación de ellas. Con estos tags de los tópicos (términos a usar antes de expansión) se seleccionan las partes que forman la consulta sin expandir, para ello se puede marcar una selección manual de los términos o una selección automática indicando que se use, todos los términos de los tags.

A los términos seleccionados se les puede aplicar un filtro o no según lo escogido en la selección Stop List. En ella indicaremos si no se aplicará ningún tipo de filtrado, o bien si se eliminarán aquellas palabras consideradas como muy comunes por las herramientas de recuperación de información Smart. La última opción Tipo de expansión de los términos permite escoger el tipo de expansión pudiéndose expandir mediante términos de los synsets escogidos como correctos.

Al terminar de cumplimentar el formulario, pulsando el botón Seleccionar crearemos el ejercicio del que nos aparecerá un breve resumen con sus datos y que ya estará disponible para ser realizado. Este formulario será el mismo para todos los tópicos que haya que procesar en el ejercicio. En primer lugar, la página sitúa al usuario en los tópicos que se van a procesar, identificándolos con sus números y mostrando los campos del título, descripción y narrativa. Esto se hace así independientemente de que la selección se efectúe a partir de un solo campo del tópico. A partir de esta información el usuario podrá conocer el significado de las palabras seleccionadas dentro del contexto del tópico. Debajo de estos campos aparece una lista con estas palabras seleccionadas y sus posibles significados y sinónimos obtenidos de WordNet. Según el tipo de ejercicio, el usuario podrá seleccionar automáticamente el significado correcto para cada palabra.

5. Experimentos y resultados

Para la evaluación de la expansión se ha utilizado el sistema de recuperación de información Lemur [11] y la colección de referencia TREC. Concretamente, en todos los experimentos realizados que se comentarán a continuación, se tomaron los tópicos TREC [12] numerados desde 401 a 450 para obtener cincuenta consultas de pruebas en cada experimento. En cuanto a la colección de documentos se usó la colección Small Web de TREC-8 “WT2g” que contiene 250.000 documentos aproximadamente, un conjunto de documentos relevantes proporcionados por especialistas, y además se utilizaron las medidas de recall y precisión para determinar los resultados de la recuperación.

El objetivo de definir el experimento fue probar si la expansión de la consulta original y la selección del significado correcto producen mejoría en la recuperación de la información. Como se puede observar en la figura 5.1, los resultados de la expansión de consulta con la selección del significado correcto fueron mejor con una precisión media no interpolada de 0.205 que la expansión de la consulta sin la selección del significado correcto con una precisión media no interpolada de 0.164. Por otro lado la consulta expandida con la selección del significado correcto mejoró ligeramente los resultados de la consulta original con una precisión media no interpolada de 0.243.

Los experimentos realizados consistieron en seleccionar automáticamente los términos relacionados para cada una de las palabras de las consultas originales y efectuar dos expansiones, una con los sinónimos para el significado correcto de cada palabra y otra con los sinónimos para los dos tipos de significados. Los términos relacionados se añaden, en este y en todos los experimentos, directamente al término original y no como términos independientes al final de la consulta. Es decir, dada una consulta $\langle A, B, C \rangle$ el término A_1 relacionado con A y el término B_1 relacionado con B la consulta expandida, quedará del siguiente modo $\langle A, A_1, B, B_1, C \rangle$. Los términos relacionados seleccionados fueron los sinónimos para

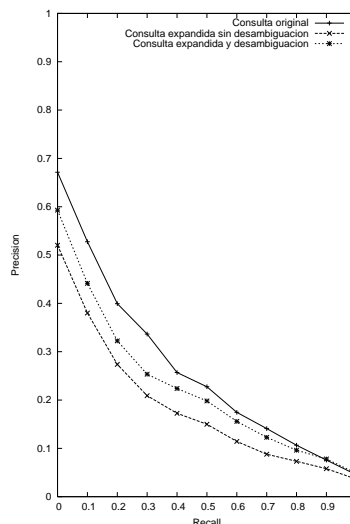


Figura 2: Gráfico de Recall-Precisión de las consultas original, expandida sin desambiguación del significado de la palabra y expandida con desambiguación del significado.

el significado correcto de cada palabra; el método desambiguación utilizado fue el basado en el diccionario, utilizando el recurso lingüístico WordNet bajo el siguiente algoritmo:

- Adquirir desde WordNet los conjuntos de sinónimos de las palabras a desambiguar.
- Determinar la coincidencia entre el contexto de las palabras a desambiguar y los conjuntos de sinónimos.
- Escoger el significado correcto de las palabras en un texto, dado el contexto de forma automática.

Este algoritmo aumentó la eficacia de 20% con relación a la consulta expandida sin la selección del significado correcto y 16% con relación a la consulta original.

6. Conclusiones

Con este trabajo hemos mostrado una aplicación que permite hacer la experimentación con expansión de consultas en el ámbito de la

recuperación de información mediante el uso de recurso lingüístico WordNet y bajo el modelo del usuario, utilizando una metodología de evaluación basada en la simulación. Para implementar estas funcionalidades de formulación y expansión de consultas se ha accedido a la información almacenada en WordNet. Una vez obtenidas las consultas expandidas de los distintos experimentos contemplados se ha procedido a su evaluación y la comparación entre ellas y con las consulta sin expandir. Los resultados de la evaluación son positivos con relación a la consulta expandida sin la selección del significado correcto y han permitido extraer algunas conclusiones importantes:

- Por un lado utilizando diferentes técnicas de expansión los resultados mejoraron ligeramente frente a la consulta original.
- Por otro lado los resultados demuestran que es preciso hacer uso de la desambiguación del significado correcto de la palabra para sacar el máximo partido a la expansión de la consulta con WordNet.

Referencias

- [1] A. Miller, R. Beckwith, C. Fellbaum, D. Gross y K. J. Miller, Introduction to WordNet: An on-line lexical database, en *International Journal of Lexicography*, Vol.3, No.4, 1990, pp.235-312.
- [2] D. Wolfram, A. Spink, B. J. Janses y T. Saracevic, Vox populi: The public searching of the web, *Journal of the American Society for Information Science and Technology* 52(12), 1073-1074 (2001).
- [3] E. M. Voorhees (1994): Query expansion using Lexical -Semantic Relations, *Proceedings of ACM-SIGIR' 94*, 61-69.
- [4] G. Salton, M. E. Lesk (1971): Computer evaluation of indexing and text processing, *The Smart Retrieval System: Experiments in Automatic Document Processing*, 143-180, Prentice-Hall, Inc. Englewood Cliffs, New Jersey.
- [5] M.F. Porter (1997): An algorithm for suffix stripping, *Readings in information retrieval*, Morgan Kaufman Publishers, 313-316.
- [6] Miller, R. Beckwith, C. Fellbaum, D. Gross y K. J. Miller: Introduction to WordNet: An on-line lexical database, en *International Journal of Lexicography*, vol. 3, No. 4, 1990, pp. 235-312.
- [7] R. Mandala, T. Tokunaga, H. Tanaka (1999): Combining multiple evidence from different types of thesaurus, *Proceedings of ACM-SIGIR' 99*, 191-197. (5)
- [8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman
- [9] : Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, 41(6), 391-407.
- [10] S. M. Harabagiu, G. A. Miller y D.I. Moldovan (1999): WordNet 2 -A Morphologically and Semantical Enhanced Resource, en *Proceedings of the SIGLEX Workshop*.
- [11] Sitio web del LEMUR: <http://www-2.cs.cmu.edu/~lemur>
- [12] Sitio web del TREC NIST: <http://trec.nist.gov>
- [13] Yin-Chen Wang, J. Vandendorpe, M. Evens (1985): Relational thesauri in information retrieval, *Journal of the American Society for Information Science*, 36(1): 15-27.