Acceso a la información mediante exploración de sintagmas

Anselmo Peñas, Julio Gonzalo and Felisa Verdejo

Dpto. Lenguajes y Sistemas Informáticos, UNED
{anselmo,julio,felisa}@lsi.uned.es

Resumen. La sugerencia de sintagmas permite abordar dos problemas en el acceso a la información: precisar necesidades de información y expresarlas en los términos adecuados de la colección. El artículo recorre brevemente algunas aproximaciones existentes y presenta el sistema *Website Term Browser* capaz de sugerir sintagmas presentes en la colección que suponen variaciones morfosintácticas, semánticas y translingües de la consulta[1].

1 Introducción

La búsqueda y recuperación de información textual tienen asociadas una serie de problemas todavía no resueltos satisfactoriamente. Algunos de estos problemas provienen de la dificultad por parte de los usuarios para precisar sus necesidades de información (apartado 2) y para adecuar su expresión al vocabulario del dominio de búsqueda (apartado 3). Estos problemas hacen deseable que los sistemas de acceso a la información superen los presupuestos de los modelos estándar de recuperación que obvian la interacción con el usuario y proporcionen al usuario algún tipo de asistencia mediante niveles intermedios de acceso a la información. El apartado 4 recorre algunas técnicas y sistemas existentes de exploración de términos (palabras y sintagmas) como nivel intermedio de acceso a la información. El apartado 5 presenta el sistema *Website Term Browser* que ofrece al usuario un nivel intermedio de información terminológica que se extrae automáticamente de la colección y que permite al usuario precisar su consulta así como superar algunas barreras del lenguaje.

2 Situaciones de imprecisión

Es frecuente que un usuario no sepa o no pueda expresar de forma concreta el objeto de su búsqueda. En estos casos el usuario sigue una estrategia que no siempre resulta efectiva: iniciar el proceso con una consulta general y poco a poco refinarla a partir de los resultados proporcionados por el sistema. Esto, en los sistemas de Recupera-

[1] Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia y Tecnología a través del proyecto Hermes (TIC2000-0335-C03-01).

ción de Información, se realiza, en general, sin ningún tipo de asistencia y requiere la exploración por parte del usuario de un gran número de documentos. En estas situaciones de búsqueda la consulta varía continuamente, los usuarios se mueven por una variedad de fuentes, y nuevas informaciones conducen a nuevas ideas y nuevas direcciones de búsqueda. Es decir, se trata de situaciones de búsqueda que no se ajustan a los presupuestos de los modelos clásicos de recuperación de información en los que la búsqueda y ordenación por relevancia decreciente de una lista de documentos que se ajustan a la consulta lleva implícito que:

1. Las necesidades de información permanecen estáticas con independencia de los documentos explorados por el usuario.
2. El objetivo es recuperar todos los documentos relevantes y sólo ellos, es decir, maximizar simultáneamente la precisión y la cobertura.
3. El valor se encuentra en el conjunto de documentos obtenidos.

Es decir, los presupuestos implícitos en el modelo tradicional de recuperación de información obvian la interacción del usuario con el sistema, necesaria en situaciones de imprecisión.

3 Barreras del lenguaje en Recuperación de Información

Iniciar el proceso con una consulta general y poco a poco refinarla a partir de los resultados obtenidos es una estrategia que no siempre va a proporcionar buenos resultados cuando nos encontramos con situaciones como:

1. El usuario no conoce la colección, no es experto en el dominio y no conoce la terminología propia del mismo.
2. El usuario, aunque pueda entender textos en otros idiomas, únicamente es capaz de expresar sus necesidades en el idioma propio.

En estos casos, los problemas de ambigüedad léxica, variación terminológica y multilingüismo dificultan la consecución de los objetivos de búsqueda. Se trata de barreras que el usuario se ve obligado a romper por sí mismo sin recibir, tampoco, ayuda por parte del sistema de búsqueda.

Considérese una situación de búsqueda en la que un profesor quiere acceder a *recursos* sobre *educación especial* que le ayuden en el aula. Podría emitir la siguiente consulta en un buscador:

recurso educación especial

El primer documento que recupera uno de los mejores buscadores en Internet es una orden ministerial que contiene:

"... recurso contencioso / administrativo ... educación especial ..."

Ambigüedad léxica. El ejemplo muestra el problema de la ambigüedad del lenguaje natural en recuperación de información: la palabra *recurso* es polisémica, como ocurre con la mayoría de las palabras más frecuentes. En promedio, las palabras que aparecen en textos corrientes tienen unas cinco acepciones diferentes recogidas en el diccionario. El ejemplo muestra también la limitación de las interfaces de búsqueda para expresar objetivos de búsqueda. Esta limitación obliga a que finalmente los usuarios recorran y filtren personalmente los documentos recuperados por el sistema.

Resultaría deseable obtener un nivel intermedio de información que facilitara esta labor.

Variación terminológica. Además de los problemas originados por el carácter inherentemente ambiguo del lenguaje natural existen otros problemas que también afectan a la recuperación de información como la presencia en los textos de expresiones equivalentes (no sólo sinónimos). Si el usuario dispusiera de un interfaz de búsqueda booleana podría indicar al sistema qué expresiones equivalentes (variaciones terminológicas, por ejemplo) debe considerar. El problema es que el usuario debe conocerlas previamente, considerarlas y hacerlas explícitas sin asistencia del sistema de búsqueda. Por ejemplo, la consulta anterior no recupera documentos que contengan expresiones como:

1. *recursos educativos*
2. *medios audiovisuales en la enseñanza*
3. *special needs education*.

La primera expresión evidencia el problema de variación morfosintáctica. Una consulta que contiene "*recursos de educación"* no puede recuperar un documento por contener la expresión "*recursos educativos"* a pesar de que sean expresiones equivalentes. En este caso, la variación morfosintáctica es una *permutación*, pero también pueden darse *inserciones* (e.g. *recursos audiovisuales de educación* es una variación de *recursos de educación*), o variaciones por *coordinación* (e.g. *recursos culturales y educativos* es una variación de *recursos educativos*).

La segunda expresión del ejemplo, *medios audiovisuales en la enseñanza*, además de una variación morfosintáctica de *inserción*, es una variación semántica porque contiene sinónimos de las palabras originales (*recursos de educación*). En los buscadores tradicionales, la palabra *recurso* no puede proporcionar acceso a los documentos que contienen su sinónimo *medio*, ni la palabra *educación* puede recuperar documentos con el sinónimo *enseñanza*, a pesar de que, conceptualmente y a efectos de búsqueda puedan ser equivalentes.

La tercera expresión del ejemplo muestra el mismo problema en el caso translingüe. La consulta "*educación especial*" no puede recuperar un documento que contenga "*special needs education*" aunque contenga traducciones directas de las palabras originales de la consulta.

4 Exploración de términos en el acceso a la información

La consideración de sintagmas no sólo abre la posibilidad de añadir términos a los índices de recuperación, sino que proporciona otra posibilidad en el acceso a la información: en lugar de explorar únicamente los documentos de la colección es posible navegar por su terminología y, a partir de ella, acceder a los documentos que puedan ser relevantes para el usuario. De esta manera será posible abordar los problemas de ambigüedad léxica y variación terminológica.

Existen varias aproximaciones que intentan explotar las posibilidades de interacción mediante la sugerencia o exploración de términos (palabras y sintagmas):

1. De construcción manuales
   - Jerarquías temáticas como la de (Yahoo).
   - Tesauros como el utilizado en (ERIC[2]), etc.
2. De construcción automática
   - Jerarquías extraídas de forma automática mediante la agrupación automática de documentos (clustering) en clases anidadas [5] o mediante relaciones de subsunción entre términos [12].
   - Expansión de la consulta mediante sintagmas [1].
   - Enlaces entre documentos con palabras claves similares extraídas automáticamente [7], [4].
   - Jerarquías de sub-sintagmas [9], [10].

Antes de describir brevemente estas aproximaciones es necesario destacar que ninguna de ellas aborda de forma conjunta los problemas de variación morfosintáctica, semántica y translingüe de los términos.

Jerarquías temáticas. En este tipo de sistemas el usuario navega por una jerarquía de materias hasta el grupo de documentos que son de su interés. El usuario está obligado a identificar la información que le interesa a través de los términos de clasificación. Este tipo de sistemas requiere la clasificación previa de los documentos de acuerdo con el vocabulario, tarea costosa que no siempre resulta sencilla y que en la mayoría de los casos se realiza de forma manual.

Exploración mediante listas y tesauros. Cuando los documentos han sido clasificados de acuerdo con un vocabulario controlado, es posible utilizar ese vocabulario como vía de acceso a la colección. Una de las maneras de navegar por un vocabulario controlado es mediante listas estructuradas de términos predefinidos que cubren todas las ideas y materias que aparecen en textos de un determinado dominio. Los documentos se organizan bajo dichos términos y se puede navegar por la jerarquía de términos a través de relaciones de especificidad. Primero se le pide al usuario que introduzca un término y a continuación se le muestra una lista de los términos semejantes a él que están contemplados en la lista de indexación. El usuario, entonces, elige el término entre los predeterminados y prosigue su interacción:

- añadiendo el término a la consulta,
- navegando por el tesauro hasta encontrar los términos precisos que incluir a la consulta: términos más generales (*Broader Terms, BT*), más específicos (*Narrower Terms, NT*), simplemente relacionados (*Related Terms, RT*), o
- explorando a su vez los términos relacionados con estos otros términos.

De esta forma el usuario va construyendo la consulta. Una vez terminado este proceso se realiza la recuperación y se devuelve la lista de los documentos relevantes para la misma.

La utilización de vocabularios controlados permite abordar parcialmente el problema de variación terminológica y multilingüismo pero requiere su construcción previa generalmente manual y más costosa que en el caso anterior, así como la clasificación de los documentos de acuerdo con el vocabulario. Si bien el uso que los

[2] http://www.ericfacility.net/extra/pub/thessearch.cfm

documentalistas hacen de los tesauros suele ser eficiente, algunos estudios muestran que se requiere cierta experiencia para conseguir buenos resultados, y que los tesauros no suponen una buena estrategia cuando se trata de usuarios poco expertos [6], [2].

Agrupación automática de documentos en clases anidadas. La agrupación automática de documentos (clustering) en clases anidadas que realiza Hearst [5] proporciona como descriptores de una agrupación el conjunto de palabras más características de la misma. El usuario, entonces, puede navegar por la jerarquía de clases en la que la cadena de palabras asociada a cada clase da una idea del contenido de la agrupación. Sin embargo, esta secuencia de palabras se ha obtenido automáticamente y no suele corresponderse con un concepto bien definido por los que su interpretación no siempre resulta sencilla.

Jerarquías de subsunción. Otra manera de navegar por términos es construir automáticamente una jerarquía de los mismos de más generales a más específicos según una relación de subsunción. Sobre las ideas de Forsyth [3], Sanderson [12] define la relación de subsunción de la siguiente manera: "el término *x* subsume al término *y* si el conjunto de documentos que contienen a *y* es un subconjunto de los documentos que contienen a *x*". De esta manera, se proporciona un mecanismo de ordenación de los términos de más generales a más específicos, permitiendo la construcción de jerarquías. Sin embargo, el significado de este tipo de relación jerárquica es difícil de determinar. El objetivo de Sanderson es construir jerarquías conceptuales pero, una vez más, uno de los problemas de esta aproximación es la ambigüedad léxica: ¿dónde situar un término polisémico en la jerarquía? La relación de subsunción no debería aplicarse sobre palabras sino sobre conceptos. Sin embargo, debido a que todavía no se dispone de técnicas suficientemente precisas para anotar semánticamente el sentido de las palabras, Sanderson construye la jerarquía a partir de un subconjunto de documentos donde se espera que los términos se utilicen con el mismo significado: los primeros 500 documentos obtenidos como resultado de una consulta.

Expansión de la consulta mediante sintagmas. Anick [1] explota la tendencia de las palabras que suponen conceptos clave del dominio a participar en familias de compuestos léxicos semánticamente relacionados. La hipótesis de *dispersión léxica* enuncia que los conceptos clave dentro de una colección tienen mayor tendencia a participar en una amplia variedad de compuestos léxicos semánticamente relacionados. La dispersión léxica de una palabra se define como el número de compuestos diferentes en los que aparece dicha palabra dentro de un determinado conjunto de documentos. El sistema *Paraphrase Search Assistant* sugiere al usuario los términos con mayor dispersión léxica en el conjunto de documentos recuperados, como términos para refinar la consulta. La medida de dispersión léxica se ve afectada por documentos aislados con una gran cantidad de términos que no pertenecen al dominio. Por esta razón, los autores enriquecen la medida considerando el número de documentos que contiene el término (*document frequency*) con el fin ajustar los resultados. Sin embargo, el mayor problema de la dispersión léxica es que muchos conceptos clave son a su vez sintagmas lexicalizados que pierden su sentido al considerar sus compo-

nentes aisladas. El concepto de dispersión léxica no permite identificar estos conceptos clave expresados con sintagmas. De esta forma, las palabras identificadas mediante dispersión léxica son demasiado generales, aportan poca información al usuario y tienen poca capacidad de discriminación conceptual.

Navegación por sintagmas clave. *Phrasier* [7] es un sistema que permite navegar entre documentos a través de *sintagmas clave*. Los autores utilizan una herramienta (*KEA*)[4] de extracción automática de *sintagmas clave* para asignar 10 sintagmas a cada documento. Estos sintagmas identificados automáticamente se utilizan como términos de indexación del documento. De esta forma, a partir de un *sintagma clave* es posible acceder a los documentos que lo contienen pero, además, los documentos que comparten *sintagmas clave* quedan enlazados entre sí resultando posible navegar por la colección a través de estos enlaces. Además, se construye un espacio vectorial de *sintagmas clave* en el que las medidas de similitud entre vectores permiten ordenar una lista de los documentos relacionados con otro documento. Este sistema sólo permite navegar y explorar los sintagmas previamente extraídos mediante KEA, no siendo posible relacionar expresiones conceptualmente equivalentes.

Jerarquías de sub-sintagmas. El usuario de un sistema de exploración de subsintagmas (*phrase browsing*) introduce una palabra y el sistema le devuelve todos los sintagmas que contienen esa palabra. A continuación, el usuario puede elegir uno de estos sintagmas para acceder a los documentos que lo contienen o explorar los sintagmas más largos que contienen al subsintagma seleccionado. Uno de los sistemas más representativos de exploración basada en sub-sintagmas es *Phind* de la Universidad de Waikato (Nueva Zelanda) que se utiliza en el paquete de software para bibliotecas digitales *Greenstone* [13] desarrollado por dicha universidad.

El número de sintagmas que pueden contener una palabra puede llegar a ser realmente extenso por lo que resulta necesario imponer criterios de ordenación como, por ejemplo, el número de documentos que contienen al sintagma. Sin conocimiento lingüístico, un sistema de estas características no puede tratar variaciones morfosintácticas, semánticas ni translingües. Algo relativamente sencillo como la exploración de los sintagmas que contienen la palabra *bosque* requiere una sesión diferente a la exploración de los sintagmas que contienen el plural *bosques*.

5 Website Term Browser

Website Term Browser (WTB) [11] es un sistema de acceso a la información basado en la exploración de sintagmas extraídos automáticamente de la colección. De forma similar a los sistemas anteriores, WTB permite la exploración de listas de términos (fundamentalmente sintagmas) organizados jerárquicamente y que el usuario puede recorrer y seleccionar para precisar sus necesidades de información. Sin embargo, a diferencia de los sistemas anteriores, es capaz de presentar al usuario sintagmas que suponen variaciones morfosintácticas, semánticas y translingües de la consulta. De esta manera, WTB aborda el problema del lenguaje ayudando al usuario a expresar sus necesidades de información incluso en idiomas diferentes al de la consulta inicial.

Modelo de indexación. WTB incorpora una extracción automática de sintagmas con criterios terminológicos. Esta extracción se ha realizado sobre la base de patrones morfosintácticos que requieren un etiquetado previo de los textos. El etiquetado es un proceso costoso que puede resultar inviable cuando se procesan cientos de miles de documentos. Para superar esta limitación, WTB implementa un proceso de etiquetado heurístico dirigido a la tarea concreta de extracción de sintagmas terminológicos. Cada palabra tiene una única etiqueta independientemente de su contexto por lo que el proceso de etiquetado resulta suficientemente rápido. La extracción de sintagmas basada en el ajuste de patrones morfosintácticos proporciona una gran cantidad de sintagmas por lo que se realiza un proceso de selección de sintagmas basado en criterios de subsunción y frecuencias de aparición de los sintagmas en la colección.

Modelo de recuperación. Se basa en la traducción y expansión de la consulta con el fin de recuperar el mayor número de sintagmas relacionados. La co-ocurrencia de palabras en un mismo sintagma supone una restricción muy fuerte que desambigua implícitamente categorías gramaticales, sentidos y determina los sinónimos y traducciones más adecuados. De esta manera, los sintagmas recuperados suponen variaciones morfosintácticas, semánticas y translingües de la consulta. Estos sintagmas, al ser extraídos directamente de la colección, quedan asociados a los documentos que los contienen, convirtiéndose en una vía alternativa de acceso a la información.

Interfaz de WTB. Se divide en tres áreas (*Figura 1*): el área de consulta (superior), el área con el ranking tradicional de documentos (inferior derecha), y el área de sintagmas recuperados y organizados de forma jerárquica (inferior izquierda). La selección de sintagmas relevantes puede verse imposibilitada si al usuario se le ofrecen un número excesivo de sintagmas sin orden ni estructura alguna y, por ello, es necesario que el sistema realice una preselección y organización según algún criterio de relevancia. Los criterios utilizados de relevancia, ordenación y selección de sintagmas se basan en la frecuencia de aparición del sintagma en los documentos, en el número de palabras relacionadas con la consulta, y en el origen de la palabra (si son palabras de la consulta original o fruto de expansión o traducción).

Las acciones que puede realizar el usuario son sencillas y directas, requisito fundamental para que los usuarios acepten su uso:

1. Seleccionar un grupo y explorar el conjunto de sintagmas contenidos en él.
2. Seleccionar un término para listar los documentos que lo contienen.
3. Seleccionar un sintagma como nueva consulta al buscador.
4. Explorar un documento.

La *Figura 1* muestra un ejemplo en el que el usuario ha introducido la consulta en inglés *"adult education"*. WTB le ha devuelto una jerarquía de sintagmas relacionados en cada uno de los cinco idiomas contemplados (español, inglés, francés, italiano y catalán). El usuario ha seleccionado el español como lengua destino para la cual hay tres grupos de sintagmas organizados en carpetas: "*formación de adultos", "adultos implicados en el proceso de enseñanza", y "educación de adultos"*. El segundo de ellos no se corresponde con el concepto que subyace a la consulta, pero su discriminación apenas supone coste para el usuario. Un tesauro multilingüe únicamente

habría proporcionado un término de traducción (e.g. *"educación de adultos"*). Sin embargo, WTB es capaz de ofrecer al usuario, además, variaciones morfosintácticas como *"educación de las personas adultas"* y variaciones semánticas como por ejemplo *"formación de adultos"*. En el ejemplo, el usuario ha seleccionado la expresión *"educación de las personas adultas"* y WTB le ha ofrecido la lista de los documentos que contienen a este sintagma. A partir de esta lista, el usuario puede acceder y explorar los documentos.

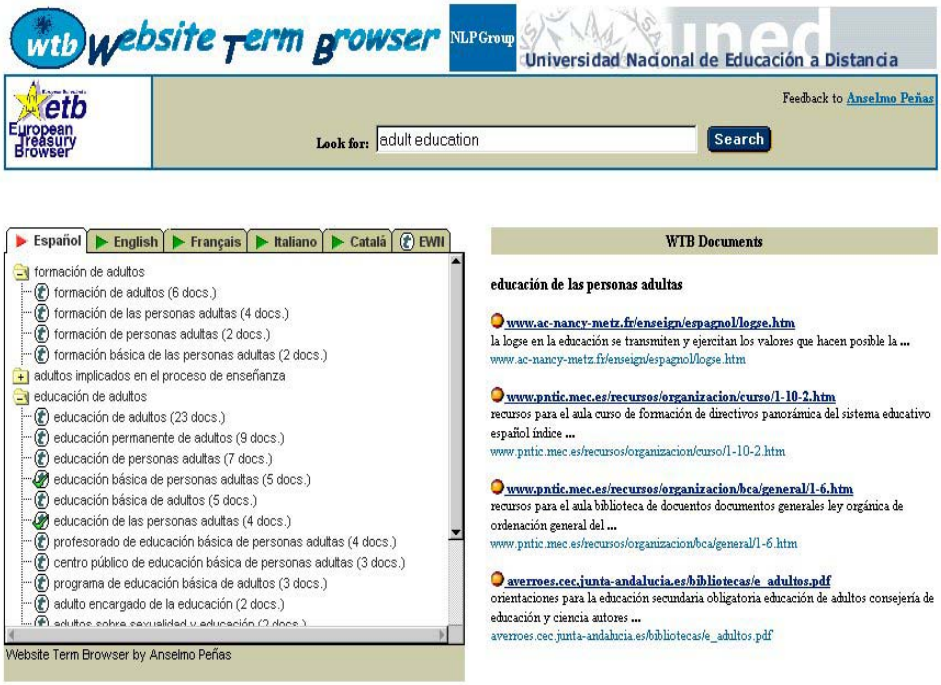

Fig 1. Interfaz de Website Term Browser

Las evaluaciones diseñadas para los sistemas clásicos de recuperación de información no son aplicables a *WTB*. Tampoco resultan apropiadas las evaluaciones diseñadas para los sistemas de búsqueda interactiva, ni para los sistemas que proporcionan interactividad al tratar los problemas de multilingüismo. Por esta razón, se ha diseñado un nuevo marco de evaluación.

Evaluación de la utilidad del área de términos. La utilidad del área de términos se ha evaluado comparando el uso que le dan los usuarios frente al uso que le dan al ranking de documentos proporcionado por uno de los mejores buscadores en Internet[3]. La evaluación se ha realizado en un entorno real de trabajo, registrando las interacciones de usuarios con necesidades reales de información. Para ello, se han regis-

[3] Google: http://www.google.com

trado, almacenado y analizado las interacciones de más de 2000 sesiones de búsqueda en el dominio UNED.es mostrando que los usuarios estiman de utilidad el área de términos y que supone un complemento al ranking tradicional de documentos. La exploración de un término es una acción presente en el 65% de las sesiones con interacción lo cual da una indicación de su uso. La primera acción tras la consulta es mayoritariamente la exploración de un término (60%) frente a la exploración directa de un documento (39%). Esto significa que los términos propuestos por WTB proporcionan mayores expectativas de relevancia que el ranking de Google. Estas expectativas muestran la capacidad de los sintagmas para señalar información de interés. El porcentaje de sesiones que termina con la exploración de un documento a partir del ranking ofrecido por la selección de un sintagma de WTB es del 47% mientras que el porcentaje de las sesiones que terminan con la exploración de un documento ofrecido por Google es del 45.6%. Esto confirma que la información terminológica que proporciona WTB complementa el ranking de documentos proporcionado por los buscadores tradicionales.

Evaluación de la recuperación translingüe de terminología. Como muestra la *Figura 1*, *Website Term Browser* permite realizar una consulta en un idioma y recuperar la terminología relacionada en otro idioma. La capacidad para recuperar términos en otros idiomas depende en gran medida de los recursos lingüísticos disponibles. Esta evaluación de recuperación de terminología en el caso translingüe se ha realizado utilizando como consultas los términos del tesauro multilingüe ETB[4]. La comparación entre los términos recuperados por WTB en otros idiomas con los términos preferidos en el tesauro (descriptores) proporciona unos valores de precisión y cobertura de la recuperación terminológica.

Los resultados evidencian la dependencia del sistema respecto a la disposición de recursos y herramientas de procesamiento lingüístico en cada idioma. Cuando se dispone de ellos, la cobertura alcanza el 55.5% sin contar la recuperación de variaciones terminológicas válidas no contempladas en el tesauro (caso inglés-español).

En cuanto a la precisión, en el mejor caso se ofrece un término coincidente con el tesauro por cada tres términos recuperados por WTB, y uno de cada diez en el peor (sin contar variaciones terminológicas válidas no contempladas en el tesauro). Esto resulta aceptable teniendo en cuenta que es fácil discriminar sintagmas y que WTB los agrupa y organiza jerárquicamente. De hecho, el 70% de los términos relevantes se ofrece en el primer nivel de la jerarquía.

6 Conclusiones

La sugerencia y exploración de sintagmas como vía de acceso a la información es una aproximación que ha despertado el interés tanto de la comunidad de Procesamiento del Lenguaje Natural como la de Recuperación de Información. Esta aproximación se dirige a superar algunas carencias de los sistemas tradicionales de recuperación de documentos introduciendo un nivel intermedio de información e interacción mediante

[4] http://www.eun/org/etb

la exploración de sintagmas. Estos sintagmas ayudan al usuario a precisar su consulta pero, además, y como muestra el sistema WTB, permiten realizar algunas inferencias que ayudan salvar la distancia entre los términos de la consulta y el vocabulario real de la colección, incluso cuando se trata de idiomas diferentes. La evaluación de WTB a partir del registro de más de 2000 sesiones muestra que los usuarios aprecian este nuevo nivel de información suponiendo un buen complemento a los sistemas tradicionales de recuperación y ranking de documentos.

Referencias

1. Anick, P. G. and Tipirneni S. The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking. Proceedings of 22nd ACM SIGIR Conference Research and Development in Information Retrieval. 1999; 153-159.
2. Brem, S. K. and Boyes A. J. Using critical thinking to conduct effective searches of online resources. Practical Assessment, Research & Evaluation. 2000; 7(7).
3. Forsyth R., Rada R. Adding an edge in Machine Learning: applications in Expert Systems and Information Retrieval. Ellis Horwood Ltd. 1986; 198-212.
4. Frank, E. Paynter G. Witten I. Gutwin C. and Nevill-Manning C. Domain-specific keyphrase extraction. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Morgan-Kaufmann. 1999; 668-573.
5. Hearst, M. A. and Pedersen J. O. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. Proceedings of 19th ACM SIGIR Conference on Research and Development in Information Retrieval. 1996.
6. Hertzberg, S. and Rudner L. The quality of researchers' searches of the ERIC Database. Education Policy Analysis Archives. 1999.
7. Jones, S. and Staveley M. S. Phrasier: a System for Interactive Document Retrieval Using Keyphrases. Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval. 1999; 160-167.
8. López-Ostenero, F. Gonzalo J. Peñas A. and Verdejo F. Noun phrase translations for Cross-Language Document Selection. Working Notes for the CLEF 2001 Workshop. 2001; 231-241.
9. Nevill-Manning, C. G. Witten I. H. and Paynter G. W. Lexically-generated subject hierarchies for browsing large collections. International Journal of Digital Libraries. 1999; 2(2/3):111-123.
10. Paynter, G. W Nevill-Manning C. G. and Witten I. H. Phrase hierarchy inference. Proceedings of the JCDL'2001 Workshop on the Technology of Browsing Applications. 2001.
11. Peñas, A. Gonzalo J. and Verdejo F. Cross-Language Information Access through Phrase Browsing. Applications of Natural Language to Information Systems, Proceedings of 6th International Workshop NLDB 2001, Madrid, Lecture Notes in Informatics (LNI), Series of the German Informatics Society (GI-Edition). 2001; P-3:121-130.
12. Sanderson, M. and Croft B. Deriving concept hierarchies from text. Proceedings of 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval. 1999; 206-212.
13. Witten, I. H. et al. Greenstone: a comprehensive open-source digital library software system. Proceedings of ACM Conference on Digital Libraries. 1999; 113-121.