KeyConcept: Un motor de búsqueda conceptual

Juan Manuel Madrid Molina
Departamento de Redes y Comunicaciones
Universidad Icesi-I2T
Cali, Colombia
jmadrid@icesi.edu.co

Susan Gauch
EECS Department
University of Kansas
Lawrence, KS 66045 (USA)
sgauch@ittc.ku.edu

RESUMEN

A medida que el número de páginas web disponible crece, los usuarios experimentan dificultades para hallar documentos que les sean de interés. Una de las razones subyacentes de este problema es que la mayoría de los motores de búsqueda encuentran documentos basándose solamente en palabras claves, sin fijarse en los significados de dichas palabras. Para brindar al usuario información más útil, se requiere de un sistema que incluya en la búsqueda información acerca del marco conceptual de la consulta, además de las palabras claves. Este es el objetivo de KeyConcept, un motor de búsqueda que recupera documentos usando una combinación de conceptos y palabras claves. Los documentos se clasifican automáticamente para determinar sus conceptos asociados. Los conceptos relacionados con la consulta son introducidos manualmente por el usuario, o determinados automáticamente mediante una pequeña descripción textual de la consulta. Este artículo describe la arquitectura del sistema, el entrenamiento del clasificador, y los resultados de nuestros experimentos de evaluación de desempeño del sistema. Se demuestra que KeyConcept incrementa en forma significativa la precisión de las búsquedas mediante el uso de la recuperación conceptual de información.

PALABRAS CLAVES

Búsqueda conceptual, clasificación de texto, ontologías.

Clasificación: A

ABSTRACT

As the number of available Web pages grows, users experience increasing difficulty finding documents relevant to their interests. One of the underlying reasons for this is that most search engines find matches based on keywords, regardless of their meanings. To provide the user with more useful information, we need a system that includes information about the conceptual frame of the queries as well as its keywords. This is the goal of KeyConcept, a search engine that retrieves documents based on a combination of keyword and conceptual matching. Documents are automatically classified to determine the concepts to which they belong. Query concepts are determined automatically from a small description of the query or explicitly entered by the user. This paper describes the system architecture, the training of the classifier, and the results of our experiments evaluating system performance. KeyConcept is shown to significantly improve search result precision through its use of conceptual retrieval.

KEYWORDS

Conceptual search, text classification, ontologies.

INTRODUCCIÓN

La Web ha experimentado un crecimiento sostenido desde su creación. En marzo de 2002, el motor de búsqueda más grande contenía aproximadamente 968 millones de páginas indexadas en su base de datos [SES 02]. Encontrar la información correcta en una colección de documentos de tal tamaño es extremadamente difícil. Una de las razones principales para obtener resultados insatisfactorios en las búsquedas es que muchas palabras poseen múltiples significados [Krovetz 92]. Por ejemplo, dos personas que efectúan una búsqueda usando la palabra clave "jaguar" pueden estar buscando cosas completamente diferentes (animales salvajes y automóviles), y sin embargo obtendrán exactamente los mismos resultados. Esta dificultad se presenta porque la mayoría de los motores de búsqueda usan un algoritmo de búsqueda de frases que regresa como resultados todos los documentos que contengan una ocurrencia exacta de los términos usados en la consulta, sin importar su significado.

Para resolver este problema estamos desarrollando KeyConcept, un motor de búsqueda que, además de las palabras claves, toma en cuenta los tópicos o conceptos relacionados con la consulta. Los documentos se indexan por palabras claves y por conceptos que se escogen automáticamente de la ontología Open Directory. En el proceso de recuperación, además de las palabras claves, el motor de búsqueda recibe uno o más identificadores de concepto que se usan para restringir el dominio de la búsqueda. Los tópicos de la consulta pueden ser especificados manualmente por el usuario, o se pueden hallar automáticamente pasando una pequeña descripción de la consulta por un clasificador. Este artículo reporta los resultados de los experimentos que se realizaron para evaluar la importancia relativa dada a la búsqueda por conceptos y por palabras claves en el proceso de recuperación.

El resto del artículo comienza con una presentación de investigación previa relacionada con el tema en la sección 2. La arquitectura de KeyConcept se describe en la sección 3. La sección 4 describe los experimentos de entrenamiento del clasificador y desempeño del sistema, junto con la discusión de los resultados. Finalmente, se concluye con algunas observaciones e ideas para trabajo futuro en la sección 5.

INVESTIGACIÓN PREVIA RELACIONADA

Clasificación de texto

Los algoritmos de clasificación de texto organizan la información, asociando un documento con los conceptos que más se relacionan con su contenido; dichos conceptos se escogen de un conjunto predefinido. Se han desarrollado varios métodos para clasificación de texto, cada uno con diferentes métodos para confrontar un documento nuevo con el conjunto de referencia, tales como representaciones vectoriales de los documentos (Support Vector Machines, k-Nearest Neighbor, Linear Least-Squares Fit, TF-IDF), uso de probabilidades conjuntas de que las palabras estén en el mismo documento (Naïve Bayesian), árboles de decisión y redes neuronales. Se puede encontrar un resumen muy completo y una comparación de dichos métodos en [Yang 99], [Pazzani 96] y [Ruiz 99].

De acuerdo con [Chekuri 97] y [Matsuda 99], una de las aplicaciones prin-

cipales de la clasificación de texto es restringir el dominio de búsqueda. En [Chekuri 97], los usuarios tienen la opción de especificar algunos conceptos de interés cuando introducen una consulta. El sistema busca documentos únicamente en los conceptos que se especifiquen. [Matsuda 99] extiende esta idea clasificando los documentos usando, además del contenido, otros atributos tales como tamaño, número de imágenes, presencia o ausencia de ciertas etiquetas. En este sistema, el usuario tiene la opción de especificar el tipo de documento que está buscando (por ejemplo un catálogo, un FAQ) conjuntamente con las palabras claves de la búsqueda.

En el proyecto Obiwan [Pretschner 99] se emplean ontologías para representar perfiles de usuario. Las consultas se efectúan por medio de un motor de búsqueda tradicional, y los resultados se clasifican en los conceptos de la ontología usando los resúmenes de documento que se incluyen en la página de resultados. Luego, el conjunto de resultados es reorganizado con base en el cruce de los conceptos relacionados con cada resumen y los conceptos con más peso en el perfil de usuario. A pesar de que este método permitió mejorar el orden de los resultados, no fue capaz de encontrar más información para los usuarios debido a que este esquema trabaja con los resultados de un motor de búsqueda convencional. KeyConcept mejora este proceso efectuando directamente la recuperación conceptual desde el índice.

Ontologías

Una ontología es un arreglo de conceptos que representa una visión del mundo [Chaffee 2000]; dicho arreglo puede emplearse para estructurar información. Las ontologías se pueden construir especificando las relaciones semánticas entre los términos de un diccionario. Un ejemplo de una ontología de este tipo es Sensus [Knight 94], una taxonomía de más de 70.000 nodos. El sistema OntoSeek [Guarino 99] emplea esta ontología para recuperar información desde catálogos de productos.

Las ontologías también se pueden derivar de colecciones jerárquicas de documentos, tales como Yahoo! [YHO 02] y el proyecto Open Directory [ODP 02]. [Labrou 99] reporta el uso del clasificador TellTale [Pearce 97] para introducir documentos nuevos en la ontología Yahoo! en forma automática. Esto se logra entrenando el clasificador con documentos de cada uno de los conceptos de la ontología, y luego encontrando el concepto que tenga la similitud más grande con el documento nuevo. Aún más, una ontología se puede emplear para permitir a los usuarios navegar por la Web y hacer búsquedas usando una jerarquía de conceptos personalizada. El proyecto Obiwan [Chaffee 00, Zhu 99] logra esto permitiendo que cada usuario defina su propia jerarquía de conceptos, y haciendo luego la equivalencia de esta ontología personal con una ontología de referencia (Lycos en este caso).

Los autores también pueden incluir información ontológica en sus documentos. SHOE (Simple HTML Ontology Extensions) [Heflin 2000] es un lenguaje diseñado para este propósito, que permite la creación de nuevas ontologías personalizadas y la extensión de ontologías existentes.

ARQUITECTURA DEL SISTEMA

Durante el proceso de indización, KeyConcept incluye en el índice información acerca de los conceptos con los que se relaciona cada documento. Para lograr esto, se extendió la estructura tradicional de archivo invertido del índice para incluir las relaciones entre conceptos y documentos. El proceso de recuperación utiliza este índice extendido y soporta consultas que empleen sólo palabras claves, sólo conceptos, o una combinación de los dos. El usuario puede seleccionar la importancia relativa de cada criterio (cruce por palabras o cruce por conceptos).

Indización

El proceso de indización se compone de dos fases: Entrenamiento del clasificador e indización de la colección. Durante el entrenamiento del clasificador se reúne y combina un número fijo de documentos de muestra para cada concepto, y los super-documentos resultantes son preprocesados e indexados utilizando el algoritmo *tf * idf*. Cada concepto queda entonces representado por el centroide del conjunto de documentos de entrenamiento para dicha categoría. Durante la indización de la colección, se indexan los documentos nuevos usando el método vector-espacio para crear un índice tradicional basado en palabras claves. Luego, se clasifica el documento calculando la similitud de su vector con el centroide de cada concepto. Los valores de similitud calculados se almacenan entonces en el índice de conceptos.

Recuperación

Durante la recuperación, el usuario introduce una consulta que puede contener palabras claves, identificadores de concepto, o ambos. En esta versión inicial de KeyConcept el usuario introduce también un número entre 0 y 1 (llamado factor α) que especifica la importancia relativa del cruce de conceptos y el cruce de palabras claves. Si α es 1, solamente se trabaja con cruce de conceptos; si es 0, solamente se considera el cruce de palabras. Cuando α es 0.5, el cruce de palabras y el de conceptos contri-

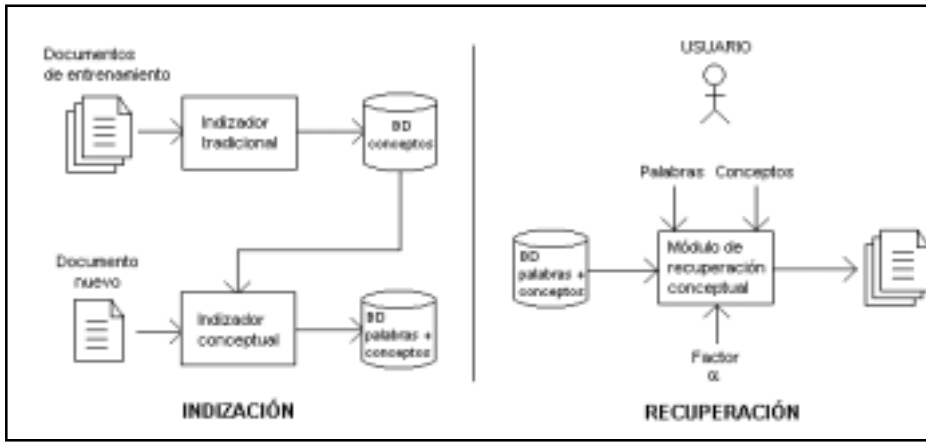

Figura 1. Funcionamiento del motor de búsqueda conceptual

buyen por igual. En un futuro, se espera que α posea un valor por omisión, de tal modo que el usuario no tenga que introducirlo; sin embargo se permite ajustarlo en esta versión inicial para propósitos de evaluación. Después de recibir los datos del usuario, el motor efectúa la búsqueda y almacena los resultados del cruce de palabras claves y conceptos en acumuladores separados. El puntaje final para cada documento se calcula mediante la siguiente fórmula:

Puntaje documento = (α *x puntaje conceptual*) + ((1-α) x puntaje palabras claves)

DESCRIPCIÓN DE LOS EXPERIMENTOS Y RESULTADOS

Se llevaron a cabo dos series de experimentos para evaluar y afinar el motor de búsqueda conceptual. La meta de la primera serie de experimentos fue determinar el tamaño de la colección de documentos que sería usada para entrenar el clasificador; más específicamente el número de conceptos y la cantidad de documentos de entrenamiento por concepto. La segunda serie de experimentos fue diseñada para determinar el grado de contribución que debería tener el cruce de conceptos en el proceso de recuperación.

Elección del conjunto de datos de entrenamiento

Se eligió usar el proyecto Open Directory [ODP 02] como fuente para los datos de entrenamiento, porque esta jerarquía está disponible en la Web en un formato compacto fácil de procesar; además de que se está convirtiendo en un estándar *de facto* para este tipo de experimentos. En abril de 2002, Open Directory poseía más de 378.000 conceptos. Debido al gran volumen de datos que involucra, el entrenamiento del clasificador hubiera sido una tarea larga y difícil si se hubiera empleado el conjunto completo de conceptos. Adicionalmente, existen diferencias sutiles entre conceptos que son obvias para un ser humano, pero podrían ser imposibles de detectar para un algoritmo de clasificación. Por esa razón se eligió emplear los conceptos de los tres primeros niveles de la jerarquía del Open Directory. El conjunto inicial de entrenamiento constó de 2.991 conceptos y aproximadamente 125.000 documentos.

Experimentos de entrenamiento del clasificador

Para determinar cuántos documentos se requerían para "entrenar" cada concepto, se diseñaron tres experimentos:

Experimento 1: Determinación de una cota superior para el número de documentos

Se deseaba probar la hipótesis de que, después de un cierto número de documentos de entrenamiento por concepto, la precisión del clasificador no aumenta, y antes puede disminuir. En este experimento se entrenó el clasificador usando los conceptos que tenían al menos setenta documentos asociados (633 de los 2.991 conceptos). De dichos documentos, dos se reservaron para pruebas y el resto se usó en el entrenamiento. El clasificador arroja los diez mejores resultados para cada caso de prueba: esto permite computar la precisión para coincidencia exacta (cuando el concepto correcto del documento de prueba aparece de pri-

mero en los resultados del clasificador) y la precisión sobre los dos primeros y los diez primeros (cuando el concepto correcto aparece entre los lugares 2 y 10 de la lista).

En este experimento se entrenó cuatro veces el clasificador, iniciando con cuarenta documentos por concepto en la primera corrida y agregando diez documentos en cada corrida subsiguiente. El Gráfico 1 muestra que, con cuarenta documentos, el clasificador posee una precisión de coincidencia exacta del 48.2%. Para cincuenta documentos se alcanza un pico del 50.9%. Con sesenta documentos, la precisión de coincidencia exacta cae al 50.2%; sin embargo el resto de los valores de precisión todavía experimenta un leve incremento. Finalmente, la precisión de coincidencia exacta cae al 34.2% cuando se usan todos los setenta documentos. Esto prueba que nuestra hipótesis es verdadera, y que la cota superior para el número de documentos de entrenamiento es 60.

Experimento 2: Determinación de una cota inferior para el número de documentos

El propósito de este experimento es determinar el número mínimo de documentos por concepto necesarios para entrenar el clasificador. Como muchos conceptos de la jerarquía Open Directory tienen menos de sesenta documentos, se podría incrementar el número de conceptos a incluir en el índice si se puede obtener una precisión aceptable usando menos documentos de entrenamiento por concepto. Se empleó el mismo procedimiento del experimento anterior, pero en esta ocasión se disminuyó el número de documentos de entrenamiento por concepto y se midió el efec-

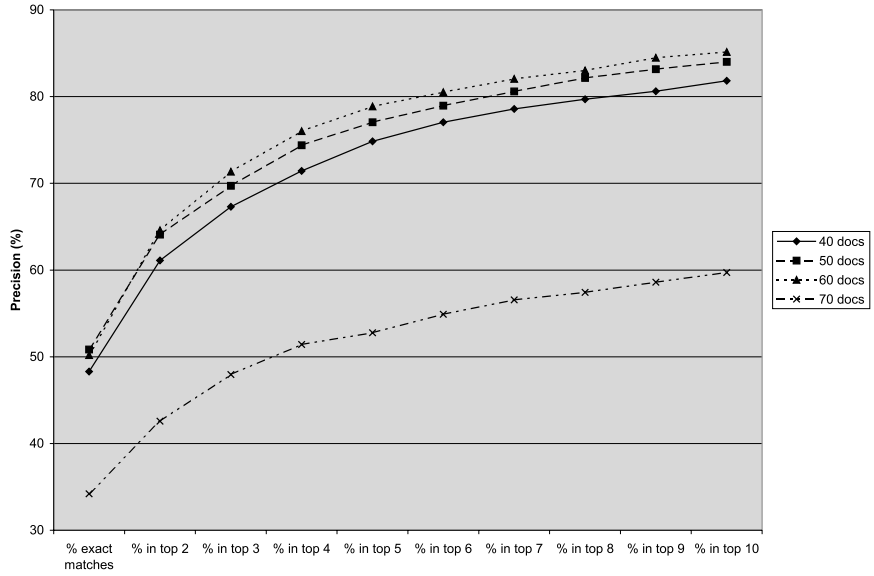

Gráfico 1. Determinación de la cota superior del número de documentos por concepto.

to sobre la precisión. Debe notarse que el número de conceptos involucrados en este experimento es mayor, pues se usaron todos los conceptos que poseían al menos cincuenta documentos de entrenamiento (941 de los 2.991 conceptos).

En este experimento se entrenó el clasificador seis veces, comenzando con cincuenta documentos por concepto y usando diez menos en cada corrida subsiguiente, a excepción de la última corrida en la que tan sólo se emplearon cinco documentos por concepto. Como se muestra en el Gráfico 2, los valores de precisión para 50, 40 y 30 documentos casi coinciden. Al usar veinte documentos se observa una disminución en la precisión (comenzando con la precisión calculada sobre los cinco primeros resultados), y para diez documentos toda la curva de precisión se desplaza hacia abajo. Esto permite concluir que la cota inferior para el número de documentos de entrenamiento por concepto es treinta. Después de analizar los datos, se encontró que usar más de treinta documentos por concepto no incrementa la precisión en forma significativa (valor p de la prueba $X^2 \approx 1$). En conclusión, se decidió usar treinta documentos por concepto para entrenar el clasificador en todos los experimentos subsiguientes. Esto permitió usar una base de datos de entrenamiento de 46.920 documentos escogidos de los 1.564 conceptos que tenían al menos treinta documentos de muestra. Usando treinta documentos de entrenamiento se logra una precisión del 51% en coincidencias exactas, 77.05% de precisión sobre los cinco primeros resultados y 84% sobre los diez primeros.

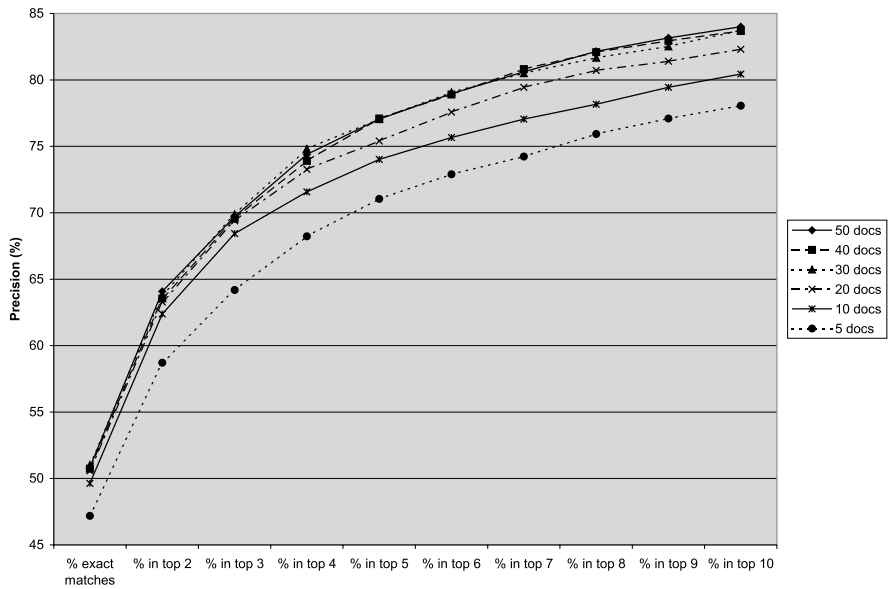

Gráfico 2. Determinación de la cota inferior del número de documentos por concepto.

Experimento 3:

Relación entre número de conceptos y precisión (Escalabilidad)

Este experimento intenta determinar si la precisión del clasificador depende o no del número de conceptos entre los cuales debe decidir. También se evalúa la dependencia de la precisión con respecto a un conjunto de conceptos en particular. Para lograr esto, se entrena el clasificador con subconjuntos diferentes de los 1.564 conceptos que poseen treinta documentos de muestra. Se escogen primero cien conceptos al azar, se entrena el clasificador con ellos, luego se clasifican los casos de prueba y se determina el número de coincidencias exactas. A continuación se escogen otros cien conceptos, el clasificador se entrena con doscientos conceptos (los cien nuevos más los cien anteriores), y se clasifican los doscientos casos de prueba. El proceso se repite hasta que todos los conceptos han sido procesados. Luego, se repite el experimento varias veces, de suerte que los conceptos se escojan en un orden diferente. Si la precisión del clasificador es independiente del número de conceptos, la precisión debería permanecer constante a medida que se incrementa el número de conceptos. Asimismo, si la precisión del clasificador es independiente del conjunto de conceptos que se empleó en cada corrida, la precisión de coincidencias exactas no debe cambiar entre corridas.

Después de entrenar el clasificador usando la base de datos de entrenamiento del experimento anterior, se crearon diez secuencias aleatorias de conceptos, se corrió el experimento 3 para cada una de ellas y por último se promediaron los resultados. El Gráfico 3 muestra los resultados para las diez corridas, y el puntaje promedio en una línea más oscura. El puntaje pro-

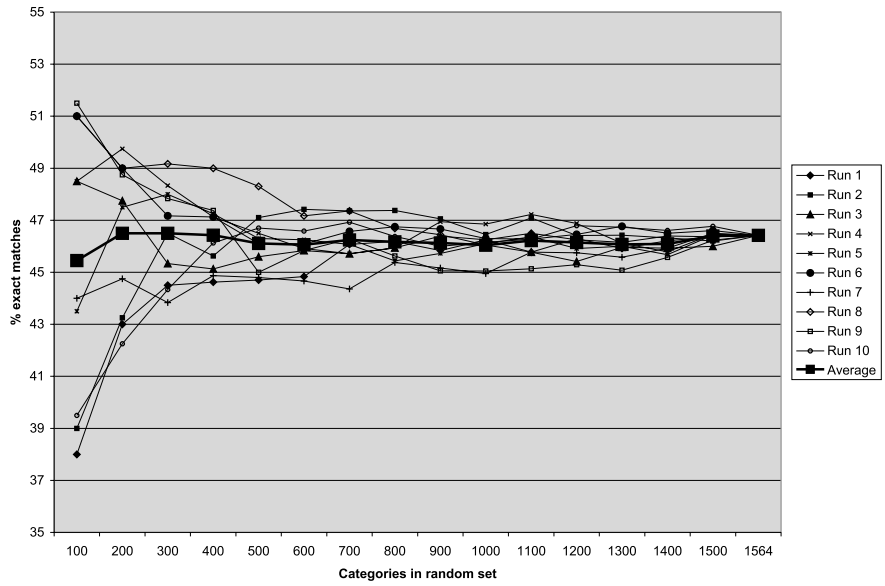

Gráfico 3. Relación entre número de conceptos y precisión

medio de precisión se mantiene aproximadamente constante (alrededor de 46%), demostrando que la precisión del clasificador no sufre a medida que se incrementa el número de conceptos. Adicionalmente, cuando se emplean únicamente cien conceptos la precisión fluctúa entre 38% y 51%; sin embargo, cuando se usan al menos seiscientos conceptos, los valores de precisión se mantienen entre 45 y 47%, con lo que se concluye que el conjunto de conceptos escogido no afecta los resultados, siempre y cuando dicho conjunto sea grande. Esto ilustra la independencia entre el número de conceptos y la precisión del clasificador.

Discusión

Los resultados de los experimentos muestran que el uso de treinta documentos de entrenamiento por concepto permite lograr un buen balance entre la cantidad de datos de entrenamiento y la precisión del clasificador. Si se emplean más documentos por concepto, menos conceptos cumplirían con el requerimiento de cantidad disponible de datos de entrenamiento, haciendo que la clasificación sea demasiado general. Por otro lado, si se usan muy pocos documentos por concepto, cada concepto se vuelve demasiado específico y no quedaría representado por un número significativo de documentos; en consecuencia, a medida que el número de conceptos crece, las diferencias entre ellos se vuelven demasiado sutiles como para que un clasificador automático las pueda detectar.

Experimentos de precisión en la recuperación de información

Después del entrenamiento del clasificador, se diseñaron dos experimentos para estudiar si un motor de búsqueda que combine búsqueda conceptual y de palabras claves se desempeña mejor que uno basado únicamente en palabras claves. Para este propósito se construyó un conjunto de documentos de prueba consistente en 100.000 documentos de la colección WT2g [Hawking 99]. El conjunto de prueba incluye los 2.279 documentos de la colección que tienen juicios de relevancia positivos, y 97.721 documentos seleccionados al azar. Debido a que el título de cada uno de los cincuenta tópicos de la colección WT2g se parece a una consulta típica (de dos a tres palabras) [Zien 01], se empleó dicha sección de título como las palabras claves a usar en cada consulta. Luego se corrieron los párrafos de narración y descripción (cincuenta palabras en promedio) incluidos con cada tópico por el clasificador, y se emplearon los identificadores de concepto obtenidos como la entrada conceptual para cada consulta.

Debe notarse que también se determinaron manualmente los identificadores de concepto que mejor se correspondían con cada tópico. Sin embargo, la diferencia entre los resultados obtenidos usando clasificación automática y manual no es significativa (valor p de la prueba $X^2 \approx 1$).

Experimento 4: Efecto de α en la precisión de la búsqueda

En este experimento se trató de encontrar un buen valor para α, el factor que balancea las contribuciones de la búsqueda conceptual y la de palabras claves. Se variaron dos factores en el experimento: (1) el factor α, y (2) el número de identificadores de concepto correspondientes a cada

tópico que se emplean en la consulta. Luego se calculó la precisión sobre diez primeros resultados y veinte primeros resultados, contando los resultados relevantes que aparecieron entre los primeros diez y veinte resultados de la consulta, y se promedió el puntaje de todas las consultas.

La línea de base para este experimento es α=0, es decir, solamente se usa recuperación por palabras claves. En ese caso, la precisión es del 42.8% (sobre los diez primeros resultados) y 33.5% (sobre los veinte primeros resultados).

Como se muestra en el Gráfico 4, la mejor precisión sobre diez primeros resultados (44.4%) se obtiene cuando se emplean los tres mejores conceptos para cada consulta, con α = 0.2. Los puntajes de precisión caen abruptamente para valores de α mayores que 0.2, queriendo decir que la recuperación conceptual por sí sola no es suficiente para obtener una buena precisión en las búsquedas, y que las palabras claves son el factor más importante en la identificación de documentos relevantes. La precisión también cae ligeramente cuando se usan más de tres conceptos por consulta, indicando que los conceptos adicionales introducen ruido en la búsqueda, subiendo el puntaje de documentos no relevantes.

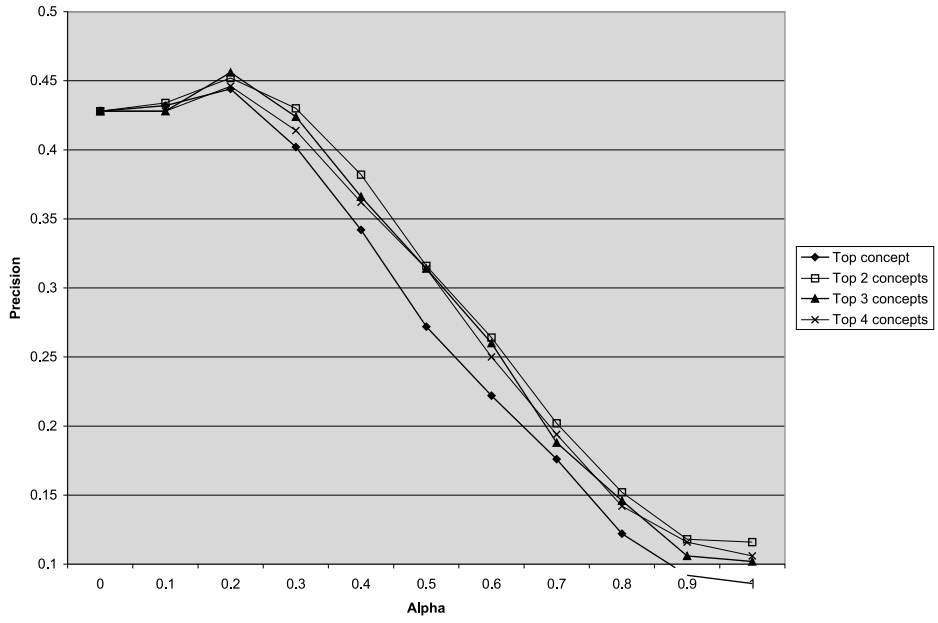

Gráfico 4. Efecto del factor α en la precisión calculada sobre los diez primeros resultados

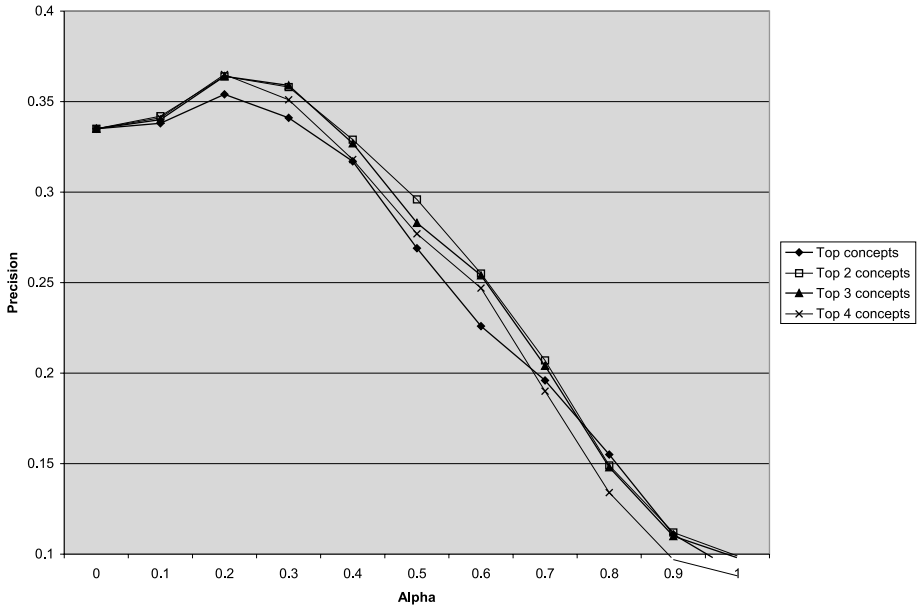

Gráfico 5. Efecto del factor α en la precisión calculada sobre los 20 primeros resultados

El Gráfico 5 muestra los resultados del experimento, considerando la precisión calculada sobre los veinte primeros resultados. La precisión también alcanza un pico (36.5%) para α=0.2, cuando se usan los tres mejores conceptos para cada consulta.

Experimento 5: Comparación de precisión por consulta

El propósito de este experimento es comparar la precisión que se obtiene con cada uno de los tópicos de la WT2g cuando se ejecuta la búsqueda con α=0 (solamente palabras claves) y α=0.2 (mejor combinación de conceptos y palabras claves). En las búsquedas combinadas se usaron los tres mejores identificadores de concepto para cada consulta.

Los resultados de este experimento se muestran en el Gráfico 6. De las cincuenta consultas de la colección, doce experimentaron un incremento de precisión cuando se empleó el cruce combinado de conceptos y palabras claves, una experimentó una disminución de puntaje, y no hubo variación en 37. A pesar de ser pequeña, la diferencia entre recuperación de información basada únicamente en palabras claves y la basada en conceptos y palabras claves es significativa (valor p de la prueba $X^2 \approx 10^{-6}$). Para los consultas en las que no se encontraron documentos relevantes nuevos, el uso de la recuperación combinada no tuvo efectos significativos en el ordenamiento de los resultados.

Discusión

Se encontró que un valor de 0.2 para el factor α incrementa la precisión de las búsquedas propuestas. Esto indica que el cruce de palabras claves es

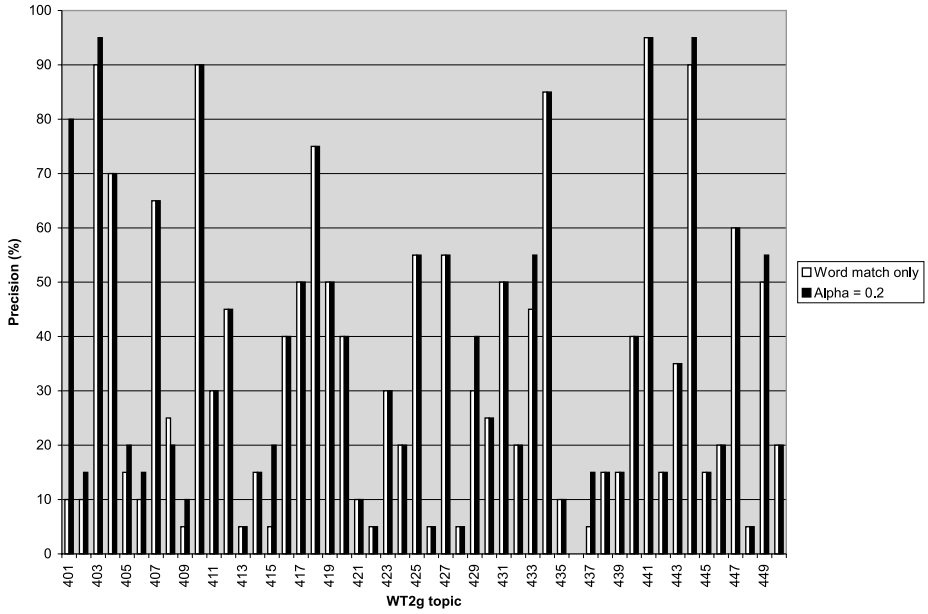

Gráfico 6. Comparación por consulta de precisión sobre los veinte primeros resultados

mucho más importante que el cruce de conceptos, pero la precisión de la búsqueda se puede elevar en forma significativa si se incluye el cruce conceptual. Los conceptos para una consulta dada pueden determinarse en forma automática usando un texto descriptivo asociado, y el desempeño es el mejor cuando se emplean en la consulta los tres mejores conceptos relacionados con la misma. El hecho de que se necesiten varios conceptos relacionados con la consulta en lugar de uno solo (el mejor) se desprende del experimento 2: El concepto correcto para cierto documento solamente aparece en el primer lugar el 50% de las veces, pero aparece entre los tres primeros el 70% de las veces.

Creemos que la razón para el incremento más bien modesto de precisión registrado en nuestros experimentos está en las características de las consultas de la colección WT2g. La definición de relevancia para dichas consultas es muy restrictiva. Un documento que sea válido tanto por palabras claves como por conceptos puede ser clasificado como irrelevante por algún detalle semántico o de contenido que puede resultar obvio para un evaluador humano pero indetectable para un clasificador automático. Esto ocasiona que el número de documentos relevantes por consulta sea muy pequeño (40 en promedio) y que muchos documentos que son del tema sean considerados irrelevantes.

CONCLUSIONES Y TRABAJO FUTURO

Este artículo presentó la idea de un motor de búsqueda conceptual, que almacena en su índice información acerca de los conceptos con los que se relaciona cada documento. Cuando se

efectúa una consulta, se usa la información conceptual, además de las palabras claves, con el fin de proporcionar al usuario resultados más relevantes.

Se mostró que es posible mejorar la precisión de las búsquedas cuando se emplea información conceptual en combinación con las palabras claves. Se empleó un subconjunto de la jerarquía Open Directory para construir la base de datos de conceptos, y se determinó que era adecuado usar 1.564 conceptos del nivel 3 de la jerarquía, con treinta documentos de entrenamiento por cada concepto. En los experimentos de recuperación se obtuvo una precisión calculada sobre los veinte primeros resultados del 36.5%, que resultó ser un incremento significativo con respecto a la búsqueda por palabras claves únicamente. Nuestra intención es efectuar experimentos usando juicios de relevancia más amplios, para ver si es posible un incremento mayor en precisión y capacidad de recuerdo.

Actualmente, los conceptos relacionados con cada consulta deben ser introducidos manualmente o corriendo a través del clasificador un texto corto relacionado con la consulta. La mayoría de los usuarios ven esto como una tarea pesada y que consume mucho tiempo; de ahí que determinar los conceptos de la consulta de manera completamente automática sería muy beneficioso. En un futuro cercano se planea hallar los conceptos de la consulta usando tecnología de perfiles de usuario. El motor compararía la consulta con un perfil de usuario y extraería de dicho perfil los conceptos que más relacionados estén con la consulta. De esta manera, los usuarios recibirán resultados más relacionados con sus intereses y actividades actuales.

REFERENCIAS

[Chaffee 2000] Jason Chaffee, Susan Gauch. *Personal Ontologies For Web Navigation*. En Proceedings of the 9th International Conference On Information Knowledge Management (CIKM), 2000, pp. 227-234.

[Chekuri 97] Chandra Chekuri, Michael H. Goldwasser, Prabhakar Raghavan, Eli Upfal. *Web Search Using Automatic Classification*. En Proceedings of the 6th International WWW Conference. Santa Clara (CA), USA, 1997. http://www.scope.gmd.de/info/www6/posters/725/web_search.html

[Guarino 99] N. Guarino, C. Masolo, and G. Vetere. *OntoSeek: Content-Based Access to the Web*. IEEE Intelligent Systems, 14(3), Mayo 1999, pp. 70-80.

[Hawking 99] David Hawking, Ellen Voorhees, Nick Craswell, y Peter Bailey. *Overview of the TREC8 Web Track*. En Eighth Text Retrieval Conference (TREC-7), Noviembre 1999. http://citeseer.nj.nec.com/article/hawking99overview.html

[Heflin 2000] Jeff Heflin, James Hendler. *Dynamic Ontologies on the Web*. En Proceedings of 17th National Conference on Artificial Intelligence (AAAI), 2000. http://www.cs.umd.edu/projects/plus/SHOE/pubs/aaai2000.pdf

[Knight 94] K. Knight y S.K. Luk. *Building a Large-Scale Knowledge Base for Machine Translation*. En Proceedings of the 12th National Conference on Artificial Intelligence (AAAI), 1994, volumen 1, pp. 773-778.

[Krovetz 92] Robert Krovetz y Bruce W. Croft. *Lexical Ambiguity and Information Retrieval.* ACM Transactions on Information Systems, 10(2), Abril 1992, pp. 115-141.

[Labrou 99] Yannis Labrou, Tim Finin. *Yahoo! As An Ontology – Using Yahoo! Categories To Describe Documents.* En Proceedings of the 8th International Conference On Information Knowledge Management (CIKM), 1999, pp. 180-187.

[Matsuda 99] Katsushi Matsuda, Toshikazu Fukushima. *Task-Oriented World Wide Web Retrieval By Document Type Classification.* En Proceedings of the 8th International Conference On Information Knowledge Management (CIKM), 1999, pp. 109-113.

[ODP 02] *Open Directory Project.* http://dmoz.org

[Pazzani 96] Michael Pazzani, Jack Muramatsu, Daniel Billsus. *Syskill & Webert: Identifying Interesting Web Sites.* En Proceedings of the 13th National Conference On Artificial Intelligence, 1996, pp. 54-61.

[Pearce 97] Claudia Pearce, Ethan Miller. *The TellTale dynamic hypertext environment: Approaches to scalability.* En Advances in Intelligent Hypertext, Lecture Notes in Computer Science. Springer-Verlag, 1997.

[Pretschner 99] Alexander Pretschner, Susan Gauch. *Ontology Based Personalized Search.* En Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Noviembre 1999, pp. 391-398.

[Ruiz 99] Miguel Ruiz, Padmini Srinivasan. *Hierarchical Neural Networks For Text Categorization.* En Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Agosto 1999, pp. 281-282.

[SES 02] Search Engine Showdown: Size statistics. http://www.searchengineshowdown.com/stats/sizeest.shtml

[Yang 99] Yiming Yang, Xin Liu. *A Re-Examination Of Text Categorization Methods.* En Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Agosto 1999, pp. 42-49.

[YHO 02] Yahoo! http://www.yahoo.com

[Zhu 99] Xiaolan Zhu, Susan Gauch, Lutz Gerhard, Nicholas Kral, Alexander Pretschner. *Ontology-Based Web Site Mapping For Information Exploration.* En Proceedings of the 8th International Conference On Information Knowledge Management (CIKM), 1999, pp. 188-194.

[Zien 01] Jason Zien, Jörg Meyer, John Tomlin, Joy Liu. *Web Query Characteristics And Their Implications On Search Engines.* En Proceedings of the 10th International WWW Conference. Hong Kong, China, 2001. http://www10.org.hk/cdrom/posters/1077.pdf

CURRÍCULOS

Juan Manuel Madrid Molina es ingeniero de sistemas de la Universidad Icesi (1995), especialista en gerencia de informática con concentración en redes y comunicaciones de la misma Universidad (1999) y candidato a doctor en Ciencias de la Computación de la Universidad de Kansas, con la disertación "Aspectos temporales de perfiles para búsqueda en la Web". Ha estado vinculado laboralmente con la universidad Icesi desde 1994, y desempeñó hasta 1999 funciones de soporte técnico a sistemas, diseño, puesta en marcha y administración de la red institucional. En la actualidad es profesor de tiempo completo del Departamento de Redes y Comunicaciones y director del programa de Ingeniería Telemática.

Susan Gauch recibió títulos de pregrado y máster en Ciencias de la Computación en Queen's University (Ontario, Canada) y es doctora en Ciencias de la Computación de la Universidad de North Carolina-Chapel Hill (1990). Actualmente trabaja como profesora-investigadora de tiempo completo en el departamento de Ciencias de la Computación de la Universidad de Kansas. Su investigación está enfocada hacia el desarrollo de agentes inteligentes para búsqueda y uso de información. Ha publicado una gran cantidad de artículos sobre el tema en revistas especializadas y congresos.