Caracterización Formal y Análisis Empírico de Mecanismos Incrementales de Búsqueda basados en Contexto

Carlos M. Lorenzetti*

Director: Guillermo R. Simari
Codirector: Ana G. Maguitman

Universidad Nacional del Sur

18 de marzo de 2011

Resumen

La Web se ha vuelto un recurso potencialmente infinito de información, transformándose además en una herramienta imprescindible para muchas tareas de la vida diaria. Esto provocó un aumento en la cantidad de información existente en el contexto de los usuarios, que no es tenida en cuenta por los sistemas de recuperación de información actuales. En esta tesis se propone una técnica semisupervisada de recuperación de información que ayuda al usuario a recuperar información relevante para su contexto actual. El objetivo de la misma es contrarrestar la diferencia de vocabulario que pudiera existir entre el conocimiento que tiene el usuario sobre un tema y los documentos relevantes que se encuentran en la Web. Se presenta un método de aprendizaje de nuevos términos asociados a un contexto temático, a través de la identificación de términos que sean buenos descriptores y términos que sean buenos discriminadores del tópico del contexto actual del usuario. Para la evaluación del método propuesto se desarrolló un marco teórico de evaluación de mecanismos de búsqueda y, a partir de este, se implementó una plataforma de evaluación, que además permitió comparar las técnicas desarrolladas en esta tesis con otras técnicas existentes en la literatura. La evidencia experimental muestra que las mejoras alcanzadas son significativas respecto de otros trabajos publicados. Dentro de este marco se desarrollaron asimismo nuevas métricas de evaluación que benefician al material novedoso y que incorporan una medida de relación semántica entre documentos. Los algoritmos desarrollados a la largo de esta tesis evolucionan consultas de alta calidad, permitiendo recuperar recursos relevantes al contexto del usuario, e impactan positivamente en la forma en la que éste interactúa con los recursos que tiene disponibles.

*Laboratorio de Investigación y Desarrollo en Inteligencia Artificial, Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Av. Alem 1253, Bahía Blanca, Argentina, cml@cs.uns.edu.ar

La tesis completa puede accederse en **http://cs.uns.edu.ar/~cml/tesisdoctoral.html**.

1. Introducción

La Recuperación de Información (IR[1]) web es un área de investigación relativamente nueva, que se popularizó desde la aparición de la Internet a principios de los '90s y trata de afrontar los desafíos de la IR en la Internet. La investigación de la IR con la ayuda de computadoras data de los '50s, cuando el esfuerzo estaba enfocado en la resolución de problemas de IR en colecciones de documentos pequeñas, con consultas descriptivas, en un dominio acotado y con usuarios particulares. Las características del nuevo entorno que resultó la World Wide Web (Web), hicieron que la tarea fuera algo diferente de la IR tradicional. La Web es un recurso prácticamente ilimitado, con información heterogénea, con usuarios dotados de distintas habilidades y con gran variedad de requisitos, buscando información que satisfaga sus necesidades. Estos necesitan que la Web sea accesible a través de sistemas de recuperación de información efectivos y eficientes. El *tamaño*, la *heterogeneidad*, el *dinamismo* y la *estructura* de la Web, junto con la *diversidad* en los comportamientos de búsqueda de los usuarios, son las principales características que hacen que la IR tradicional tenga grandes desafíos en la Internet.

Los motores de búsqueda comerciales, que son los sistemas de IR más populares, han resuelto parcialmente los desafíos con los que se enfrenta la IR en la Web, ofreciendo una herramienta para la búsqueda de información relevante. En efecto, los usuarios actuales esperan ser capaces de encontrar la información que buscan en la Web, de forma rápida y fácil. La IR en la Web, sin embargo, continúa siendo un área con muchas cuestiones por resolver, probablemente con muchas aplicaciones por descubrir. En la actualidad sigue existiendo la necesidad de desarrollar métodos novedosos para facilitar el acceso eficiente a la información relevante en la Web. Algunos problemas de investigación van desde comprender mejor las necesidades del usuario, al procesamiento de enormes cantidades de información para brindar mejores métodos de ordenamiento, que hagan uso de la estructura y las características de la Web.

2. Motivación

La omnipresencia de las computadoras personales, unida a la conectividad de la Internet han cambiado para siempre el rol de la información en la computación. Los recursos de información ya no están más relacionados con una única ubicación ni son accedidos sólo por profesionales. Los sistemas de IR están disponibles para los usuarios de Internet cada día, desde el confort de su propia computadora personal. Estos repositorios de información se acceden de la misma forma en la que se escriben artículos, se leen diarios y se navegan sitios de la Web. Desafortunadamente, los sistemas de IR tradicionales resultaron difíciles de usar para usuarios nuevos, lo que impulsó el desarrollo de una gran cantidad de sistemas para buscar, filtrar y organizar la gran cantidad de información que se tenía disponible. Se desarrollaron sistemas de IR para aplicaciones que van desde la clasificación y organización de correo electrónico [30, 18], el filtrado de noticias [24], sistemas para responder consultas basados en las FAQ[2] de Usenet[3] [21], y la búsqueda en la Web [32, 6]. También se han desarrollado algunas aplicaciones para organizar la información del usuario, como pueden ser archivos de notas, diarios y calendarios [22, 23].

Sin embargo, la mayoría de estos sistemas, que se han convertido en la piedra angular del acceso a la información, sólo se han concentrado en la generación de consultas para recuperar información por demanda, lo que significa que el usuario tiene que invocarlos explícitamente, interrumpiendo el proceso normal de navegación y esperando ocioso por los resultados de la búsqueda. Tales sistemas no pueden ayudar a un usuario cuando éste no está suficientemente

[1] del inglés, Information Retrieval.
[2] del inglés, Frequently Asked Questions, Preguntas frecuentes.
[3] del inglés, USErs NETwork.

familiarizado con el tema en cuestión, o desconoce el vocabulario exacto con el que debe formular las consultas para acceder a los recursos de interés.

Este escenario trae nuevos desafíos y oportunidades a los diseñadores de tales sistemas, tanto para crear sistemas accesibles como para aprovechar por completo este nuevo espacio de información oculta. El crecimiento explosivo que ha tenido la Web y otras fuentes de información on-line han hecho crítica la necesidad de alguna clase de asistencia inteligente para el usuario que está buscando información relevante. Al desarrollarse computadoras de escritorio cada vez más potentes, la mayor parte del tiempo de CPU de éstas se desperdicia esperando que el usuario presione la siguiente tecla, lea la siguiente página o se cargue el siguiente paquete de la red. No hay razón para que esos ciclos de CPU desperdiciados no puedan ser usados constructivamente para realizar búsquedas de información útil para el contexto actual del usuario. Por ejemplo, mientras un ingeniero lee un correo electrónico sobre un proyecto, un agente puede recordarle la planificación, los reportes de avance u otros recursos relacionados con ese proyecto. Cuando el ingeniero no lee más el correo y, por ejemplo, comienza a editar un archivo, el agente cambiaría automáticamente sus recomendaciones para adecuarse a la nueva tarea.

Para los diseñadores de interfaces de exploración de información también se presentan problemas interesantes, ya que la forma en la que un usuario genera una consulta depende de su conocimiento previo y de su entendimiento del tema. Algunas preguntas que surgen son: ¿cómo les presentamos a los usuarios las posibles acciones que pueden tomar teniendo en cuenta su entendimiento actual?, ¿cómo podemos ayudar a los usuarios a tener un mejor entendimiento de estas referencias?, y ¿cómo podemos ayudar a los usuarios a volver a sitios visitados con anterioridad en la exploración, una vez que se ganó una nueva perspectiva?

La motivación para las investigaciones presentadas en esta tesis es desarrollar una herramienta que ayude y asista al usuario de un sistema de IR en la tarea que está realizando, brindándole información relevante y basada en el contexto en el cual está trabajando.

3. Contribuciones

Este trabajo de investigación propone una técnica de IR novedosa que incrementalmente aprende nuevos términos que pueden ayudar a reducir la distancia existente entre el vocabulario empleado en las consultas formuladas por un usuario y el vocabulario utilizado para indexar los documentos relevantes para dicho usuario. Es decir, las principales contribuciones de esta tesis son:

1. Un *Algoritmo semisupervisado* que utiliza una estrategia de recuperación incremental de documentos web para el ajuste de la importancia de los términos utilizados en la generación de consultas, de forma tal que éstos reflejen mejor su valor como descriptores y discriminadores del tópico del contexto del usuario. El vocabulario enriquecido de esta forma permite la generación de consultas para una búsqueda más efectiva.

2. Una *Plataforma de evaluación* de nuevos métodos y algoritmos desarrollados para la IR. Una plataforma de evaluación es algo fundamental en el desarrollo de nuevos métodos en IR, permitiendo la comparación con las técnicas existentes. También se proponen nuevos métodos de evaluación sustentados en una métrica de *similitud semántica* para la comparación de documentos.

3.1. Método incremental de recuperación de información basado en contexto

Este trabajo presenta técnicas generales para aprender incrementalmente términos relevantes asociados a un contexto temático. Específicamente se estudian tres preguntas:

1. ¿Puede el contexto del usuario explotarse satisfactoriamente para acceder a material relevante en la Web?

2. ¿Puede un conjunto de términos específicos de un contexto ser refinado incrementalmente basándose en el análisis de los resultados de una búsqueda?

3. ¿Los términos específicos de un contexto aprendidos mediante métodos incrementales, son mejores para generar consultas comparados con aquellos encontrados por técnicas clásicas de IR o métodos clásicos de reformulación de consultas?

La contribución de este trabajo es un algoritmo semisupervisado que aprende incrementalmente nuevo vocabulario con el propósito de mejorar consultas. El objetivo es que las consultas reflejen la información contextual y así puedan recuperar efectivamente material relacionado semánticamente. En este trabajo se utilizó una métrica estándar de evaluación del rendimiento y dos métricas ad hoc para descubrir si estas consultas son mejores que las generadas utilizando otros métodos.

La pregunta principal que guió este trabajo es cómo aprender términos específicos a un contexto basándonos en la tarea del usuario y en una colección abierta de documentos web recuperados incrementalmente. Se asume que la tarea del usuario está representada como un conjunto de términos cohesivos que resumen el tópico del contexto del usuario. Consideremos un ejemplo que involucra la *Máquina Virtual de Java*, descripto por los siguientes términos:

`java`	`virtual`	`machine`	`programming`	`language`
`computers`	`netbeans`	`applets`	`ruby`	`code`
`sun`	`technology`	`source`	`jvm`	`jdk`

Los términos específicos a un contexto juegan distintos roles. Por ejemplo, el término *java* es un buen descriptor del tópico para el común de las personas. Por otro lado, términos como *jvm* y *jdk* (acrónimos de "Java Virtual Machine" y "Java Development Kit" respectivamente) pueden no ser buenos descriptores del tópico para esas mismas personas, pero son efectivos recuperando información similar al tópico cuando se los utiliza en una consulta. Luego, *jvm* y *jdk* son buenos discriminadores del tópico.

Para distinguir entre descriptores y discriminadores de tópicos se argumenta que *buenos descriptores de tópicos* pueden encontrarse buscando aquellos términos que aparecen en la mayoría de los documentos relacionados con el tópico deseado. Por otro lado, *buenos discriminadores de tópicos* pueden hallarse buscando términos que sólo aparecen en documentos relacionados con el tópico deseado. Ambos tipos de términos son importantes a la hora de generar consultas. Utilizar términos descriptores del tópico mejora el problema de los resultados falso-negativos porque aparecen frecuentemente en páginas relevantes. De la misma manera, los buenos discriminadores de tópicos ayudan a reducir el problema de los falsos-positivos, ya que aparecen principalmente en páginas relevantes.

En [31] se propone estudiar el poder descriptivo y discriminante de un término basándose en su distribución a través de los tópicos de las páginas recuperadas por un motor de búsqueda. Allí, el espacio de búsqueda es la Web completa y el análisis del poder descriptivo o discriminante de un término está limitado a una pequeña colección de documentos que se va construyendo incrementalmente y que varía en el tiempo. A diferencia de los esquemas de IR tradicionales,

los cuales analizan una colección predefinida de documentos y buscan en ella, los métodos propuestos utilizan una cantidad limitada de información para medir la importancia de los términos y documentos así como también para la toma de decisiones acerca de cuáles términos conservar para análisis futuros, cuáles descartar, y qué consultas adicionales generar.

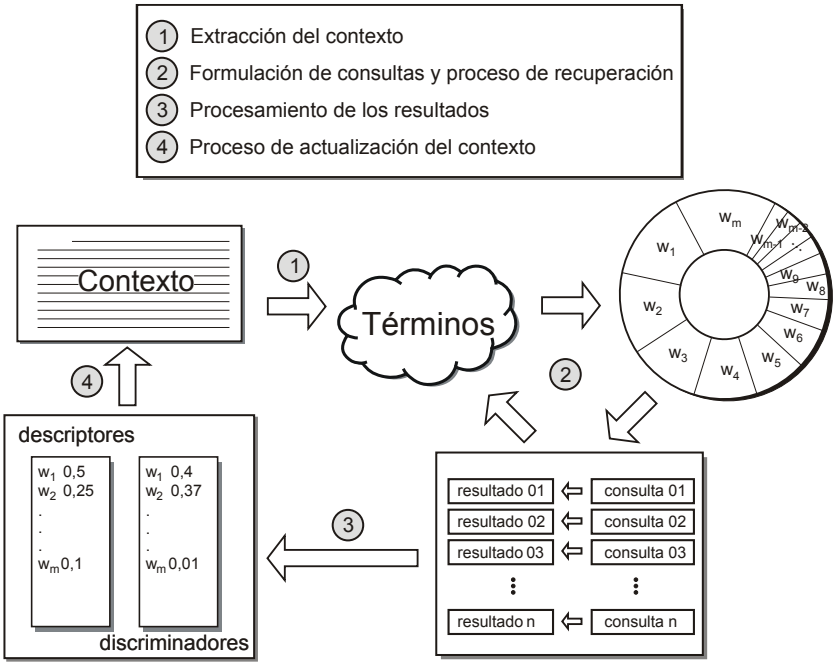

Figura 1: Una representación esquemática del método incremental para el refinamiento de consultas temáticas.

Como parte del trabajo de esta tesis se comenzó formulando un marco teórico [17] que realiza un análisis cualitativo y cuantitativo del contexto del usuario para el mejoramiento de los resultados de una búsqueda. En la Figura 1 puede verse un esquema del método incremental para el refinamiento de consultas basado en un contexto temático. El sistema lleva a cabo una serie de *fases* con el objetivo de aprender mejores descripciones de un contexto temático. En la figura esto está representado por el ciclo de pasos que van desde el paso 1 al paso 4. Al final de cada fase se actualiza la descripción del contexto con el nuevo material aprendido (paso 4).

Continuando con las investigaciones se llegó a la conclusión de que el contexto puede utilizarse para encontrar material relevante, aunque en [29] se mostró que las palabras más frecuentes no siempre son las más útiles. Es por esto que, basándose en los resultados obtenidos, se hizo hincapié en el uso de métodos incrementales para el refinamiento del contexto del usuario, desarrollando una nueva técnica de enriquecimiento del vocabulario [27]. Una versión extendida de este trabajo fue publicada en [28], en donde puede encontrarse un desarrollo más profundo del método incremental y en donde también puede encontrarse su comparación con otros métodos de IR.

Paralelamente a estos trabajos se estudió el impacto de utilizar Algoritmos Genéticos (AG) como alternativa válida para el refinamiento del contexto, dada sus probadas cualidades en problemas de optimización [8]. Continuando con este trabajo, se incluyó un algoritmo basado en AGs en la plataforma propuesta. Se analizaron las razones por las que los AGs son apropiados para la búsqueda Web y se describió el funcionamiento del algoritmo. Las evaluaciones realizadas en [13] mostraron la efectividad de los métodos propuestos y las ventajas que presentan respecto de otros trabajos publicados previamente. Siguiendo con estas investigaciones se decidió analizar el efecto de la variación de distintos parámetros propios de un AG, como son por ejemplo, las

tasas de cruzamiento y mutación [12]. A partir de estos trabajos se fueron publicando distintos resultados a medida que se avanzaba en la investigación [14, 11, 9, 10]. Los resultados completos de estos estudios pueden encontrarse en [15].

3.2. Plataforma de evaluación

Las primeras evaluaciones realizadas a través de esta plataforma se presentan en [25], comparando el método incremental propuesto, basado en la utilización de las nociones de descriptores y discriminadores de tópicos, con un mecanismo simple tomado como referencia. Se introdujo también la noción de similitud novedosa, que es conceptualmente similar a la medida de similitud más conocida en IR, la similitud por coseno. Sin embargo esta medida es capaz de descubrir nuevas relaciones entre los documentos y el contexto del usuario, ya que favorece a aquellos que contienen información relevante y que a su vez contienen términos que no se encontraban en documentos anteriores.

La plataforma incluye actualmente una colección local de documentos que fueron indexados con la plataforma de código abierto Terrier[4], desarrollada por la Universidad de Glasgow. El acceso a este índice se realiza a través de una interfaz que es capaz de aceptar otros tipos de índices e incluso, motores de búsqueda web. En un comienzo, se implementó una interfaz para el servicio web SOAP de Google[5], que luego fuera reemplazado por la empresa por una API AJAX. La utilización del servicio web permitió el desarrollo de las primeras versiones de los algoritmos presentados en esta tesis. Finalmente se optó por un índice local de documentos web debido a las limitaciones que se encontraron en cuanto a los tiempos de ejecución de los algoritmos y al límite impuesto por Google a la cantidad de consultas que se podían realizar por día. Algunos resultados obtenidos mediante el uso de esta nueva plataforma se muestran en [26].

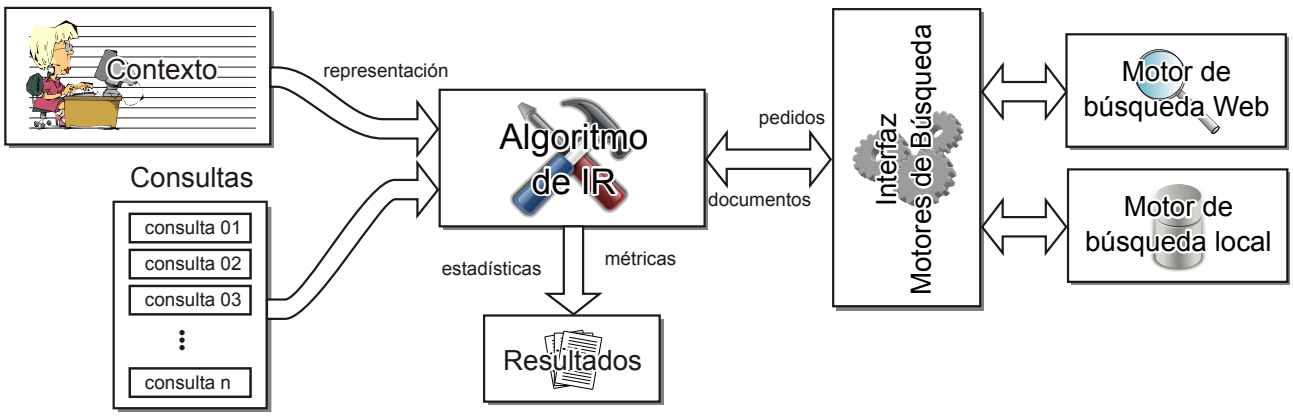

Figura 2: Representación esquemática de la plataforma de evaluación.

Una representación esquemática de la Plataforma de Evaluación se muestra en la Figura 2. Como se puede observar existe una primera parte que se encarga de la representación de las consultas. Estas pueden ingresarse como un conjunto o como un documento, a partir del cual el sistema generará las consultas necesarias. Por otro lado, la plataforma ofrece una interfaz de comunicación con los distintos motores de búsqueda. Como se dijo más arriba, una de las posibilidades es contar con un motor de búsqueda web. También existe un componente dedicado al cálculo de las métricas que guiarán los algoritmos de búsqueda y que también servirán para

[4]http://www.terrier.org
[5]http://www.google.com

su evaluación. En [16] se desarrollan estas métricas y se las compara con otras existentes en la literatura.

Como parte de las contribuciones de esta tesis se desarrollaron dos métricas nuevas para la comparación de documentos en algoritmos de IR. Una de ellas es la Similitud novedosa, una medida de comparación entre documentos que descarta los términos que pudieran introducir un sesgo en la medición, beneficiando a los documentos que incluyen términos nuevos. La otra es la Precisión semántica, una métrica para la comparación de los resultados de un sistema de recuperación de información. Esta medida brinda una noción más rigurosa de la calidad de los documentos recuperados por un algoritmo de IR, al incorporar la noción de relevancia parcial entre tópicos. Este nuevo concepto se basa en una métrica de similitud entre nodos de un grafo arbitrario. En particular se utilizó la ontología creada por el Open Directory Project (ODP[6]), que es un gran directorio de la Web editado por personas, y utilizado por cientos de portales y sitios de búsqueda. El ODP clasifica millones de URLs en una ontología temática. Las ontologías ayudan a darle sentido a un conjunto de objetos y, con esta información, pueden derivarse relaciones semánticas entre esos objetos y, por lo tanto, son una fuente muy útil de donde se pueden obtener medidas de similitud semántica.

4. Conclusiones

A lo largo de esta tesis se desarrolló una herramienta de recuperación de información que ayuda al usuario en la tarea que está realizando, brindándole información relevante y basada en su contexto actual. Para ello se propuso una solución al problema de la sensibilidad semántica, que es la limitación que surge cuando no se puede hallar una relación entre dos documentos similares semánticamente, porque contienen distintos términos en su vocabulario, resultando en un falso-negativo al intentar recuperar material relevante. Además, mediante la identificación de buenos discriminadores de tópicos, la propuesta presentada en esta tesis ayuda a mitigar el problema de falsos-positivos, que aparece cuando el mismo término (p. ej., java) aparece en dos tópicos diferentes. El método enunciado trabaja aprendiendo incrementalmente mejores vocabularios de un gran conjunto de datos como la Web.

A partir de este trabajo se concluye que la información contextual puede ser utilizada con éxito para acceder a material relevante. Sin embargo, los términos más frecuentes en ese contexto no son necesariamente los más útiles. Es por ello que se propone un método incremental para el refinamiento del contexto, que se basa en el análisis de los resultados de las búsquedas y que mostró ser aplicable a cualquier dominio caracterizable por términos.

En este trabajo se demostró que al implementar un método incremental semisupervisado de refinamiento del contexto se puede mejorar el rendimiento alcanzado por un método base, el cual envía consultas generadas directamente a partir del contexto inicial, y mejorar también el rendimiento del método de refinamiento Bo1-DFR [1], el cual no refina las consultas basándose en un contexto. Esto muestra la utilidad de aprovechar simultáneamente los términos existentes en el contexto temático actual y los de un conjunto externo de datos a la hora de aprender mejores vocabularios y de refinar consultas automáticamente.

En esta tesis se implementó una plataforma de evaluación de métodos y técnicas para la recuperación de información. La misma permitió el desarrollo de los algoritmos presentados en este trabajo, proporcionando el soporte necesario para un análisis detallado de los resultados obtenidos. Dentro de esta plataforma también se implementaron las nuevas métricas propuestas en esta tesis.

En la literatura se han propuesto otros métodos basados en corpus para atacar el problema

[6] http://dmoz.org

de la sensibilidad semántica. Por ejemplo, el análisis de la semántica latente [20, 19], o la técnica PMI-IR[7] [34]. Este método de recuperación de información está basado en la información de polaridad mutua, que mide la relación entre dos elementos (p. ej., términos) comparando sus frecuencias observadas con respecto a las esperadas. Estas técnicas se diferencian de la que se propone en que no se basan en un proceso incremental de refinamiento de consultas, sino que utilizan una colección predefinida de documentos para identificar relaciones semánticas. Además, estas técnicas no distinguen las nociones de descriptores y discriminadores de tópicos. Las técnicas para la elección de los términos de las consultas propuestas en este trabajo están inspiradas y motivadas sobre la misma base de otros métodos de expansión y refinamiento de consultas [33, 5]. Sin embargo, los sistemas que aplican estos métodos se diferencian de la plataforma propuesta en que el proceso se realiza a través de consultar o navegar en interfaces que necesitan la intervención explícita del usuario, en lugar de formular consultas automáticamente.

En los sistemas de recuperación proactivos, el uso del contexto juega un rol vital a la hora de seleccionar y filtrar información. Tales sistemas observan las interacciones del usuario e infieren necesidades adicionales de información, buscando documentos relevantes en la Web u otras librerías electrónicas. Aprender mejores vocabularios es una manera de aumentar la percepción y la accesibilidad del material útil. Se propuso un método prometedor para identificar la necesidad detrás de la consulta, lo cual es uno de los principales objetivos para muchos servicios y herramientas web actuales y futuras.

5. Trabajo a futuro

Dentro de las limitaciones encontradas durante el desarrollo de esta tesis, la más importante resultó ser el tiempo de ejecución de los algoritmos presentados. La velocidad es un obstáculo muy grande a la hora de realizar una evaluación con usuarios y es un aspecto a tener en cuenta a futuro. Por otro lado, el tiempo límite de ejecución podría incluirse como un parámetro a ser definido por el usuario, indicando qué tanto está dispuesto a esperar por resultados o si en cambio, desea un determinado número de documentos novedosos sin importar el tiempo de espera. Otro aspecto que no fue abordado dentro de los objetivos y contribuciones de estas tesis es la determinación del contexto actual del usuario, que también es de especial interés al momento de realizar las evaluaciones con usuarios. En lugar de esto, en las evaluaciones presentadas, se utilizó un conjunto de términos extraídos de una página de un tópico dado o la descripción de un tópico realizada por un editor de una ontología temática. En la literatura existen diversos trabajos que abordan el tema del reconocimiento automático del contexto actual de un usuario [2, 4, 3, 7].

Se está trabajando actualmente para aplicar el método propuesto para el aprendizaje de mejores vocabularios en otras tareas de IR, como la clasificación de texto. También se están analizando las distintas estrategias que ayudan a mantener al sistema enfocado en el contexto inicial, luego de que se han llevado a cabo varios pasos incrementales. Por otro lado, se espera adaptar la plataforma propuesta para evaluar otras aplicaciones de recuperación de información, tales como algoritmos de clasificación y clustering.

Se ampliará la plataforma de evaluación presentada en esta tesis con el propósito de ponerla a disponibilidad de la comunidad de IR, lo que resultará de gran utilidad a la comunidad científica del área, proveyéndola de una herramienta que permitirá analizar de manera objetiva la efectividad de nuevos métodos. Entonces, se diseñará un instrumento de evaluación para sistemas de IR basado en un gran número de tópicos y documentos obtenidos a partir de ontologías de tópicos, para luego integrarlo con métodos de evaluación existentes y novedosos.

[7] del inglés, Pointwise Mutual Information – Information Retrieval

En tal sentido será importante el uso de las nociones de similitud semántica y relevancia parcial incorporadas a partir de esta tesis.

La construcción de colecciones de prueba ha merecido especial atención del ámbito de la IR experimental, ya que analizar grandes colecciones de documentos y juzgar su relevancia es una tarea sumamente costosa, especialmente cuando los documentos cubren tópicos diversos.

A la luz de estas necesidades y dificultades, y a partir de ontologías de tópicos editadas por humanos, tales como ODP, hemos desarrollado, y esperamos seguir refinando, un marco de experimentación para la evaluación automática y semi-automática de sistemas de IR, aprovechando el número masivo de relaciones disponibles entre tópicos y documentos.

Agradecimientos

El desarrollo de esta tesis fue financiado en su totalidad por el CONICET, dentro del Laboratorio de Desarrollo y Investigación en Inteligencia Artificial (LIDIA), perteneciente al Departamento de Ciencias y Ingeniería de la Computación (DCIC) de la Universidad Nacional del Sur (UNS). Además parcialmente por los siguientes proyectos: PICT 2005 Nro. 32373, TICs-Sinergia 2008, PGI-UNS 24/ZN13, PGI-UNS 24/N029, PIP N°11220090100863. Algunos de los trabajos publicados se realizaron en conjunto con integrantes del Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC[8]) de la UNS.

Referencias

[1] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[2] M. Balabanović, Y. Shoham, and Y. Yun. An Adaptive Agent for Automated Web Browsing. Technical report, Stanford University, Palo Alto, 1995.

[3] T. Bauer and D. B. Leake. Wordsieve: A Method for Real-Time Context Extraction. En V. Akman, *et. al.*, eds., *Modeling and Using Context*, vol. 2116 de *Lecture Notes in Computer Science*, pp. 30–44. 2001.

[4] K. Bharat. Searchpad: explicit capture of search context to support web search. *Computer Networks*, 33(1-6):493–501, 2000.

[5] B. Billerbeck, F. Scholer, H. E. Williams, and J. Zobel. Query expansion using associated queries. En *Proceedings of the twelfth international CIKM*, pp. 2–9, New York, 2003. ACM.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[7] J. Budzik, S. Sood, K. J. Hammond, and L. Birnbaum. Context transformations for just-in-time retrieval: Adapting the watson system to user needs. Technical Report NWU-EECS-06-21, 2006.

[8] R. L. Cecchini, C. M. Lorenzetti, and A. G. Maguitman. Algoritmos Genéticos para la Búsqueda Web basada en Contextos temáticos. En *IX WICC, ASI*, pp. 6–10, Trelew, Argentina, 2007.

[9] R. L. Cecchini, C. M. Lorenzetti, and A. G. Maguitman. Evolving Disjunctive and Conjunctive Topical Queries based on Multi-objective Optimization Criteria. **Inteligencia Artificial**, 13(44):14–26, 2009.

[10] R. L. Cecchini, C. M. Lorenzetti, and A. G. Maguitman. Multi–objective Query Optimization Using Topic Ontologies. En T. Andreasen, *et. al.*, eds., *Flexible Query Answering Systems, 8th Internat. Conference*, vol. 5822 de **Lecture Notes in Computer Science**, pp. 145–156, Roskilde, Denmark, 2009. Springer.

[11] R. L. Cecchini, C. M. Lorenzetti, and A. G. Maguitman. A Multi-objective Evolutionary Algorithm Approach to Learn Disjunctive and Conjunctive Topical Queries. En *38° Jornadas Argentinas de Informática e Investigación Operativa (JAIIO), ASAI*, pp. 25–36, Mar del Plata, 2009.

[12] R. L. Cecchini, C. M. Lorenzetti, A. G. Maguitman, and N. B. Brignole. Genetic Algorithms for Topical Web Search: A Study of Different Mutation rates. En *XIII CACIC*, pp. 1585–1595, Corrientes, 2007.

[8] http://lidecc.cs.uns.edu.ar

[13] R. L. Cecchini, C. M. Lorenzetti, A. G. Maguitman, and N. B. Brignole. Searching the Web in Context: Genetic Algorithms for Exploring Query space. En *36° Jornadas Argentinas de Informática e Investigación Operativa, SSI*, pp. 183–195, Mar del Plata, 2007.

[14] R. L. Cecchini, C. M. Lorenzetti, A. G. Maguitman, and N. B. Brignole. Using genetic algorithms to evolve a population of topical queries. **Information Processing and Management**, 44(6):1863–1878, 2008.

[15] R. L. Cecchini, C. M. Lorenzetti, A. G. Maguitman, and N. B. Brignole. Multi-objective Evolutionary Algorithms for Context-based Search. **Journal of the American Society for Information Science and Technology**, 61(6):1258–1274, 2010.

[16] R. L. Cecchini, C. M. Lorenzetti, A. G. Maguitman, and F. Menczer. A semantic framework for evaluating topical search methods. **CLEI Electronic Journal**, 14(1):13–27, 2011.

[17] C. I. Chesñevar, C. M. Lorenzetti, A. G. Maguitman, F. M. Sagui, and G. R. Simari. Exploiting User Context and Preferences for Intelligent Web Search. En *Proceedings of the WICC*, Morón, 2006.

[18] W. W. Cohen. Learning rules that classify e-mail. En M. A. Hearst and H. Hirsh, eds., *AAAI Spring Symposium on Machine Learning in Information Access*, pp. 18–25, 1996.

[19] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[20] G. W. Furnas, S. C. Deerwester, S. T. Dumais, T. K. Landauer, *et. al.*. Information retrieval using a singular value decomposition model of latent semantic structure. En *Proceedings of the 11th annual int. ACM SIGIR conf. on Research and development in IR*, pp. 465–480, New York, 1988. ACM.

[21] A case-based approach to knowledge navigation. En U. M. Fayyad, *et. al.*, eds., *Proceedings of the Workshop on KDD*, pp. 383–394. AAAI Press, 1994.

[22] W. P. Jones. On the applied use of human memory models: the memory extender personal filing system. *International Journal of Man-Machine Studies*, 25:191–228, 1986.

[23] M. Lamming and M. Flynn. Forget-me-not: intimate computing in support of human memory. En *FRIEND21: Symposium on Next Generation Human Interfaces*, pp. 125–128, Meguro Gajoen, Japan, 1994.

[24] K. Lang. NewsWeeder: Learning to Filter Netnews. En A. Prieditis and S. J. Russell, eds., *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 331–339. Morgan Kaufmann, 1995.

[25] C. M. Lorenzetti, R. L. Cecchini, and A. G. Maguitman. Intelligent Methods for Information Access in Context: The Role of topic descriptors and discriminators. En *XIII CACIC*, pp. 1608–1619, 2007.

[26] C. M. Lorenzetti and A. G. Maguitman. Learning Better Context Characterizations: An Intelligent Information retrieval approach. En *XXXIV Conferencia Latinoamericana de Informática*, pp. 200–209, 2008.

[27] C. M. Lorenzetti and A. G. Maguitman. Tuning Topical Queries through Context Vocabulary Enrichment: A Corpus-based approach. En R. Meersman, *et. al.*, eds., *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*, vol. 5333 de **Lecture Notes in Computer Science**, pp. 646–655. Springer, 2008.

[28] C. M. Lorenzetti and A. G. Maguitman. A semi-supervised incremental algorithm to automatically formulate topical queries. **Information Sciences**, 179(12):1881–1892, 2009.

[29] C. M. Lorenzetti, F. M. Sagui, A. G. Maguitman, C. I. Chesñevar, and G. R. Simari. Incremental Methods for Context-based Web Retrieval. En *XII CACIC*, pp. 1243–1254, San Luis, 2006.

[30] P. Maes. Agents that reduce work and information overload. *Communicat. of the ACM*, 37(7):30–40, 1994.

[31] A. G. Maguitman, D. B. Leake, T. Reichherzer, and F. Menczer. Dynamic Extraction of Topic Descriptors and Discriminators: Towards Automatic Context-Based Topic Search. En *Proceedings of the Thirteenth CIKM*, pp. 463–472, Washington, 2004. ACM Press.

[32] O. A. McBryan. GENVL and WWWW: Tools for Taming the Web. En *First International Conference on the World Wide Web*, Geneva, Switzerland, 1994. CERN.

[33] F. Scholer and H. E. Williams. Query Association for Effective Retrieval. En *Proceedings of the eleventh international CIKM*, pp. 324–331, New York, 2002. ACM.

[34] P. D. Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. En L. De Raedt, *et. al.*, eds., *Proceedings of the 12th European Conf. on M.L.*, pp. 491–502, London, UK, 2001. Springer-Verlag.