Agentes Inteligentes para Búsqueda de Información

Daniela Godoy[1], Analía Amandi

ISISTAN - Facultad de Ciencias Exactas - UNICEN
Tandil, Bs. As., Argentina
[1] Facultad de Ingeniería, Universidad Nacional de La Pampa
General Pico, La Pampa, Argentina
{dgodoy, amandi}@exa.unicen.edu.ar

1. Introducción

La gran cantidad de información disponible en forma on-line a través de Internet puso de manifiesto la necesidad de proveer a los usuarios herramientas que faciliten la navegación y búsqueda en este vasto espacio de información. Las primeras soluciones en este sentido la constituyen los motores de búsqueda que brindan, mediante una interface simple basada generalmente en palabras clave, el acceso a un gran número de documentos Web. Sin embargo, esta simplicidad para expresar requerimientos de información usualmente conlleva pobres niveles de precisión en los resultados obtenidos a partir de ellos.

Esto implica que un usuario debe dedicar una considerable cantidad de tiempo y esfuerzo en revisar o navegar a través de lista ordenada de documentos, donde normalmente varios de ellos no son de su interés, antes de encontrar información verdaderamente relevante.

Una alternativa a este problema la constituyen los agentes inteligentes que asisten activamente al usuario proveyéndole información personaliza mientras navega o realiza sus actividades normales en la Web. Recientes desarrollos en este sentido son PersonalWebWatcher [3], Letizia [2], Syskill&Webert [4].

Para lograr esto un agente debe contar con un perfil de usuario que deberá construir necesariamente a partir del análisis de las interacciones del usuario con el sistema. Un componente de aprendizaje permitirá que este modelo evolucione ante cambios en las preferencias del usuario, utilizando el feedback de relevancia como una fuente importante de información.

En este trabajo se presenta un agente inteligente que observa el comportamiento de un usuario durante su actividad normal en la Web y realiza un análisis del contenido de las páginas a las que él accede con el objeto de deducir los temas en los que se encuentra interesado. Se propone la utilización de *Razonamiento Basado en Casos* (CBR) para, por un lado, descubrir dinámicamente el tema de un documento con el suficiente grado de detalle que permita ser distinguido de temas similares y, por otro, decidir la conveniencia o no de recomendar un documento particular. La utilización de esta técnica se describe con más detalle en las sección 2, mientras que en la sección 3 se explica como se construye una jerarquía de temas de interés para posteriormente clasificar documentos como se indica en la sección 4. Finalmente, en la sección 5 se presentan las conclusiones.

2. Utilización de CBR para el filtrado de Documentos

La técnica de Razonamiento basado en Casos (CBR) consiste en resolver un nuevo problema recordando una situación previa similar y reusando la información y conocimiento obtenido de ella en la resolución de la nueva situación [1].

Las lecturas de un usuario se almacenan en forma de *casos* dentro de su perfil. Nuevos documentos serán recomendados si son similares a aquellos encontrados en el perfil de usuario en un momento dado, bajo la suposición de que pertenecen a un mismo tema específico.

De esta manera, supongamos que el usuario realiza una consulta utilizando *agentes* como palabra clave, probablemente los resultados en un motor de búsqueda tradicional devolverá documentos que traten tanto de *agentes inteligentes,* como de *agentes de viajes, agentes de seguros*, etc. Si, en cambio, se utilizan los *casos* almacenados en el perfil de usuario para contextualizar la consulta, y ellos demuestran una preferencia del usuario por lecturas acerca de *agentes inteligentes,* los documentos que traten este tema serán recomendados descartando el resto de ellos.

Las características que mejor describen un documento Web para ser representado en forma de caso la constituyen un subconjunto de las palabras contenidas en el texto del documento. Particularmente se seleccionan aquellas más significativas para la determinación del tema considerando los sustantivos, palabras en mayúscula o en partes preponderantes del documento y eliminando *stop-words* (palabras que no aportan ninguna información para la identificación de un tema, como artículos, conectivas, etc.). A cada uno de estos términos se le asocia una grado de importancia en el documento de acuerdo a un conjunto de criterios como son: la frecuencia con la que aparece en el documento, la ubicación (título, encabezado, texto normal, etc.), el estilo de letra en el que aparece (mayúsculas o minúscula, negrita, subrayado) y tamaño de letra.

La totalidad de los documentos obtenidos como resultado de enviar la consulta del usuario a algunos de los más populares motores de búsqueda en la Web son analizados por el agente para establecer su similitud con los casos existentes. La siguiente función permite establecer este grado de similitud:

$$\frac{\sum w_i * sim_i(f_{i1}^I, f_{i2}^R)}{\sum w_i}$$

Donde w_i es la importancia de cada dimensión i (la principal dimensión es la lista palabras relevantes del documento), *sim* es la función de similitud y f_i^I, f_i^R son los valores para la característica f_i en los casos de entrada y los casos recuperados respectivamente. Si un documento se encuentra similar a otro ya existente se le asigna el mismo tema específico y se recomienda al usuario para su lectura.

3. Construcción de una Jerarquía Temática

Los casos que conforman el perfil de usuario no sólo se encuentran agrupados de acuerdo a su tema específico sino que se crea una jerarquía temática de especificidad creciente conformando una *red de características compartidas* [1]. Este tipo de organización jerárquica, no sólo es una forma natural de organización de documentos textuales, sino que ha sido adoptada en forma exitosa por los servicios de directorios de Internet (como Yahoo, Infoseek, etc.) para categorizar el contenido de la Web.

La jerarquía se conforma de tal manera que los casos que comparten muchas características pertenecen a un mismo grupo. Cada nodo interno de la red describe las características que poseen los casos que se encuentran debajo de él (en la forma de un clasificador para la categoría como se explica más abajo), mientras que las hojas son los casos mismos.

Cuando un nuevo problema requiere ser analizado en una base de casos organizada en forma jerárquica sólo se analiza la similitud de la nueva situación con un subconjunto de los casos de la base, aquellos que pertenecen a su misma categoría, a diferencia de una base de casos plana donde la totalidad de los casos deben ser recuperados para ser analizados. Esto conlleva una disminución del tiempo implicado en la comparación a la vez que introduce la necesidad de clasificar los nuevos casos, tanto para ser almacenados en la base como para ser recuperados luego para su análisis.

Las características que conforman los nodos internos del árbol se obtienen a partir de un conjunto de casos agrupados de acuerdo a un mismo tema específico. Mas tarde estas palabras serán utilizadas para discriminar nuevos documentos como pertenecientes al tema o no. De esta manera, el agente aprende a clasificar nuevos documentos para determinar su temática general.

4. Clasificación de Documentos dentro de la Jerarquía

Dada la organización de la base de casos el problema de clasificación global se descompone en un conjunto de problemas más simples en cada nodo del árbol. De esta manera, un clasificador puede distinguir documentos relacionados al tema *computación* de los de *política*.

Para cada categoría (o tema) c_i se construye automáticamente un clasificador observando las características del conjunto de casos agrupados debajo de ella y extrayendo las características que debe tener un nuevo documento para ser clasificado bajo c_i.

Un clasificador para una categoría $c_i \in C$ consta tanto de una función de evaluación, F_i: D $\rightarrow$[0,1], representando la evidencia para el hecho de que d_j debe ser clasificada bajo c_i; como de un valor umbral, τ_i, tal que $F_i(d_j) > \tau_i$, se interpreta como la decisión de clasificar d_j bajo c_i, y $F_i(d_j) < \tau_i$, la decisión de no clasificar d_j bajo c_i [5].

Cada problema de clasificación que debe ser resuelto en cada nivel del árbol es considerablemente más simple que el problema total de clasificación ya que sólo requiere distinguir entre un pequeño número de términos. Por ejemplo, las palabras *computer*, *science*, *systems*, etc., presentes normalmente en un texto referido a *computación* son útiles para clasificarlo como tal, pero resultan intrascendentes para clasificarlo en otras jerarquías del árbol como *economía* o *política*, e incluso para clasificarlo en sus propios subtemas para los cuales se requerirá la presencia de términos más específicos.

Los clasificadores utilizados en este trabajo son del tipo *lineal* permitiendo representar la categoría en términos de un vector $c_i = \langle w_{1i}, \ldots, w_{ri} \rangle$ tal que el valor de la función F es el producto interno entre el vector que representa el documento (tomando sólo las características que posee el clasificador) y el vector de la categoría, que luego de ser normalizado resulta en la medida de similaridad del coseno:

$$S(c_i, d_j) = \cos(\alpha) = \frac{\sum_{k=1}^{r} w_{ki} * w_{jk}}{\sum_{k=1}^{r} w_{ki}^2 * \sum_{k=1}^{r} w_{kj}^2}$$

Cuando un nuevo caso requiere ser clasificado se evalúan los clasificadores comenzando en el nivel superior de la jerarquía, hasta obtener el tema general del documento, luego se establece su tema específico dada la similaridad con otros casos que traten el mismo tema.

5. Conclusiones

En este trabajo se presentó un agente capaz de observar al usuario en sus distintas actividades en la Web, con el fin de crear y actualizar un perfil de usuario. En base a este perfil, y dada una consulta particular del usuario, el agente filtra los resultados obtenidos de un conjunto de motores de búsqueda para sugerir la lectura de un número reducido de documentos.

La principal contribución de este trabajo es el uso de Razonamiento Basado en Casos, tanto para el filtrado de documentos Web, como para su clasificación dentro de una jerarquía temática construida incrementalmente para cada usuario. Un conjunto de pruebas con usuarios reales permitió observar que las recomendaciones realizadas por el agente concuerdan con las preferencias de los usuario. En futuros trabajos se pretende evaluar el comportamiento del agente con diferentes tipos de clasificadores.

6. Referencias

[1] Kolodner, J. *Case-Based Reasoning*. Morgan Kaufmann, 1993.

[2] Lieberman, H. *Letizia: An agent that assists web browsing*, Proceedings of the International Joint Conference on Artificial Intelligence. Montreal. 1995.

[3] Mladenic, D. *Machine learning used by Personal WebWatcher*. Proceedings of ACAI-99 Workshop on Machine Learning and Intelligent Agents. Chania, Crete. 1999.

[4] Pazzani, M.;Billsus, D. *Learning and Revising User Profiles: the Identification of Interesting Web Sites*. Machine Learning, 1997.

[5] Sebastiani, F. *Machine Learning in Automated Text Categorisation*. Technical Report. Consiglio Nazionale delle Ricerche. 1999.