NewsAgent : un agente que genera diarios digitales personalizados

D. Cordero, P. Roldán, S. Schiaffino, A. Amandi

Instituto de Sistemas ISISTAN – Facultad de Ciencias Exactas
Universidad Nacional del Centro de la Pcia. de Buenos Aires
Campus Universitario Paraje Arroyo Seco
(7000) – Tandil – Bs. As., Argentina
Email: {dcordero, proldan, sschia, amandi}@exa.unicen.edu.ar

Resumen

Debido a la gran cantidad de información disponible en Internet, un usuario debe dedicar gran parte de su tiempo al análisis de una cantidad considerable de documentos digitales en busca de aquellos que son de su interés. NewsAgent es un agente inteligente que tiene la capacidad de generar diarios digitales personalizados basándose en un perfil de preferencias que construye para cada usuario. Este perfil se obtiene a partir de la observación del comportamiento del usuario y del feedback que éste brinda ante los documentos presentados. El agente utiliza la técnica de Razonamiento Basado en Casos para realizar una clasificación específica de los documentos que son relevantes para el usuario.

1. Introducción

Internet es un medio que provee gran cantidad de información sobre diferentes temas. Si bien a primera vista esta característica parece ser una de sus mayores ventajas, representa una dificultad para las personas que desean acceder rápidamente sólo a aquellos documentos que son de su interés. Además, debido a la naturaleza dinámica de este espacio, generalmente se accede sólo a una parte de la gran cantidad de información disponible sobre un tema en particular o tal vez no se encuentra la información específica deseada.

Existen diferentes soluciones a estos problemas. Un usuario puede utilizar servicios personalizados, realizar consultas en los motores de búsqueda, como Altavista, Yahoo, Lycos, o directamente navegar para obtener la información que le interesa.

Los servicios personalizados requieren que el usuario ingrese sus preferencias previamente. La limitación de este enfoque es que generalmente las personas no saben precisamente el tipo de información que están buscando. Sin embargo, una vez que encuentran los documentos que le son relevantes, desean obtener más documentos que traten ese tema en particular o relacionado con él.

Las búsquedas de información realizadas por medio de los motores de búsqueda se basan en una o más palabras claves ingresadas por el usuario. Este mecanismo presupone que el usuario es capaz de formular el conjunto adecuado de palabras claves para recuperar la información deseada. Si se realizan las consultas con las palabras claves erróneas puede recuperarse información irrelevante o perderse documentos de interés.

La navegación a través de Internet obliga al usuario a acceder a los diferentes vínculos en busca de los documentos relevantes.

Los inconvenientes que presentan estos métodos de recuperación de información es que requieren una importante dedicación de tiempo por parte de los usuarios para analizar qué documentos contienen información relevante.

Otro problema que se presenta frecuentemente al recuperar información es que los usuarios acceden a la Web sin saber precisamente qué están buscando. Una vez que encuentran la información que les resulta relevante, desearían obtener documentos similares que traten del tema en particular o relacionado con éste.

Por otro lado, los motores de búsqueda son independientes del dominio. Es decir, reúnen la información requerida sin considerar el dominio específico involucrado.

Una solución totalmente diferente para tratar estos problemas y recuperar información relevante es utilizar agentes inteligentes. Un agente inteligente es un programa que asiste a un usuario en la realización de una actividad o tarea compleja.

Los agentes crean su propia base de conocimiento sobre las fuentes de información disponibles en Internet, la cual se va actualizando y expandiendo después de cada búsqueda.

Los agentes son capaces de buscar información basada en un contexto. El contexto se deduce a partir de la información obtenida de los usuarios. Un agente se puede adaptar a las preferencias y deseos de cada usuario. Esto lleva a que los agentes se adapten cada vez mejor a los intereses y deseos de los usuarios, a lo que generalmente buscan aprendiendo de las tareas que realizan y a la forma en la cual los usuarios reaccionan a los resultados que ellos brindan.

Los agentes inteligentes no sólo se utilizan para ayudar a buscar y filtrar información relevante sino también para categorizar, priorizar y seleccionar documentos. Ayudan a los usuarios durante la navegación en Internet.

Un agente es capaz de buscar información relevante para un usuario y luego sugerirle un pequeño conjunto de páginas interesantes para ser leídas. De esta manera evita que el usuario dedique una excesiva cantidad de tiempo en la búsqueda de información interesante. En este trabajo se presenta un agente inteligente que genera un diario personal conteniendo solamente la información relevante para un usuario.

NewsAgent es un agente capaz de obtener las noticias del día de un conjunto de diarios digitales y filtrar aquellas que son de interés para el usuario. Se eligió el dominio de los diarios digitales como dominio de estudio debido a las características cambiantes de los temas que tratan los documentos que se pueden encontrar en estos sitios. Este agente tiene la capacidad de detectar las preferencias del usuario mediante la observación de su comportamiento. Utiliza la técnica de Razonamiento Basado en Casos para establecer la similitud entre documentos. El almacenamiento de las características de las noticias preferidas permite distinguir las noticias del día que son relevantes para el usuario.

Cómo obtener los intereses del usuario, cómo detectar las alteraciones en tales intereses, cómo analizar la semejanza de documentos para recuperar documentos relevantes y cómo asistir al usuario sin transformarse en una carga, son algunos de los objetivos que guiaron el desarrollo de NewsAgent. Este artículo presenta las soluciones propuestas para la implementación de un agente con tales características. A continuación se resume la estructura de este trabajo. En la sección 2 se presentan las técnicas utilizadas para la adquisición del conocimiento acerca de las preferencias de un usuario. En la sección 3 se describe el análisis de los documentos leídos por el usuario. En la sección 4 se propone la construcción de un modelo de filtrado para la recuperación de información relevante. En la sección 5 se describe la construcción del diario personal y se explica como se utiliza la evaluación del usuario para refinar el modelo de filtrado. Luego se presentan los trabajos relacionados y las conclusiones.

2. Adquisición de conocimiento acerca de las preferencias de un usuario

Para adquirir conocimiento acerca de las preferencias de un usuario, e ir construyendo de esta forma el estado mental del agente, se utilizan tres técnicas diferentes.

La primera de ellas se basa en la interacción directa con el usuario. El usuario informa sus intereses voluntariamente. La segunda técnica utilizada para obtener conocimiento sobre los intereses del usuario le provee un método para que éste informe si considera relevante la información contenida en un documento particular. Adicionalmente, se utiliza una tercera técnica para la determinación de las preferencias, que no requiere ningún tipo de interacción con el usuario. Esta técnica se basa en la observación del comportamiento del usuario. Es decir, si un usuario no proporciona información sobre sus preferencias explícitamente, éstas se pueden obtener observando el comportamiento del usuario durante la lectura de los documentos.

2.1 Conocimiento provisto directamente por el usuario

El método más directo para obtener conocimiento sobre las preferencias de un usuario es mediante la interacción directa con el mismo. El usuario provee voluntariamente los temas que considera relevantes. En la figura 1 se presenta la interfaz utilizada para obtener la información

mediante la interacción con el usuario. En ella se muestra el conjunto de opciones disponibles para que el usuario provea sus preferencias. Las opciones pueden ser combinadas con el fin de presentar las características generales que un documento debe contener para ser considerado relevante. Las preferencias obtenidas directamente de cada usuario se presentan como el conjunto de las características generales de un documento que el usuario considera de interés.

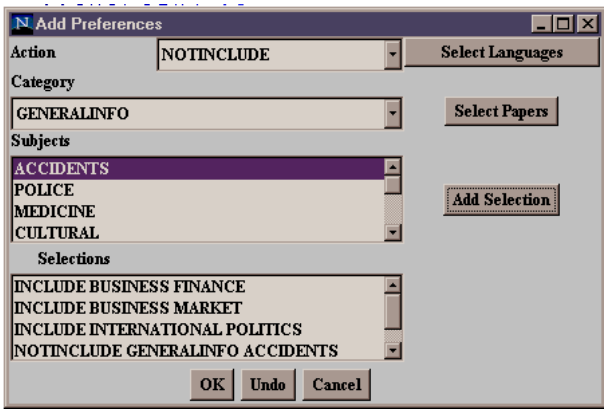


Figura 1. Interfaz utilizada para obtener información del usuario

2.2 Conocimiento provisto indirectamente por el usuario

El segundo método que se presenta para obtener conocimiento acerca de las preferencias de un usuario requiere también de interacción con el mismo. Si un usuario desea indicar explícitamente su preferencia por el documento mientras está navegando, puede hacerlo mediante la selección de un icono provisto por la aplicación. Esto se puede observar en la figura 2.

La aceptación explícita de un documento hace posible inferir las características que cada usuario desea encontrar en los documentos. Debido a que el usuario no indica explícitamente los tópicos que le resultan interesantes, estos pueden deducirse a partir de las propiedades encontradas en los documentos que han sido calificados como interesantes.

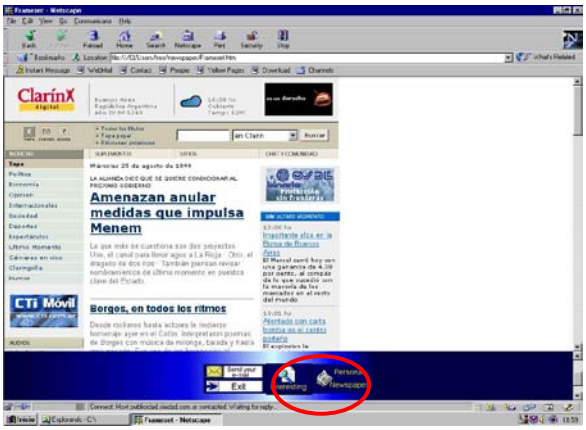


Figura 2. Forma de indicar interés explícito sobre el documento.

2.3 Conocimiento inferido a partir del comportamiento del usuario

La observación de las lecturas sin interacción con el usuario implica que todo el conocimiento acerca de las preferencias de éste se debe deducir a partir de su comportamiento [Lieberman 95]. Se definen patrones de comportamiento para identificar la conducta del usuario ante un documento relevante.

El patrón que se utiliza en este trabajo se basa en el tiempo de permanencia en el documento que se está analizando. Un usuario permanece más tiempo en aquellos documentos que le resultan de

interés, debido a que el tiempo de lectura es proporcional al grado de interés. Sin embargo, si el tiempo de lectura excede limites que se consideran razonables para la lectura de una página, las características de este documento no son incorporadas a las preferencias del usuario.

A partir de la observación del usuario, se deducen las características preferidas analizando el contenido del documento. Estas características son incluidas como preferencias del usuario.

3. Análisis de la información

Como se mencionó en la sección anterior, si un usuario no proporciona información sobre sus preferencias explícitamente, éstas se pueden obtener observando el comportamiento del usuario durante la lectura de los documentos.

Para poder determinar cuáles de los documentos leídos le interesan al usuario, se necesita información que represente los documentos que lee habitualmente. Para ello se extrae y analiza información de las distintas páginas por las que navega un usuario. Las características que se registran de cada página son: el sitio, la categoría a la que pertenece, el tema, la fecha y el tiempo que dedica el usuario para leer el documento.
En la figura 3 se resume esta información.

Sitio: `Clarín`
Categoría: `Economía`
Tema: `Bolsa`
Tiempo: `180000`
Fecha:`12/06/99`

Figura 3. Ejemplo de preferencias inferidas a partir del comportamiento del usuario

3.1 Análisis de los documentos leídos por un usuario

El agente NewsAgent deduce los temas de interés de un usuario a partir del análisis de información de las páginas por las que éste navega. El sitio, la categoría a la que pertenece un documento, su tema, la fecha y el tiempo que dedica el usuario para leerlo son las características registradas por el agente para inferir las preferencias de lectura de un usuario.

El sitio al cual pertenece un documento brinda información general acerca de las preferencias de lectura. Se infiere que un lector está interesado en los documentos que se publican en aquellos sitios que lee habitualmente. Por ejemplo, un usuario puede preferir las notas del sitio de Clarín por sobre las de La Nación.

Para determinar el sitio al que pertenece una página se analiza su URL (Universal Resourse Locator). Por ejemplo, del URL:

`http://www.clarin.com/diario/hoy/o-02601d.htm`

se determina que el documento pertenece al sitio del diario Clarín.

Una vez determinado el sitio al que pertenece la página se clasifica el documento ubicándolo en uno de los niveles de una jerarquía de temas estática. Esta jerarquía puede ser construida para cualquier dominio en el que se trabaje. Es necesario realizar un análisis de la información que se puede encontrar en los sitios de interés y organizarla en forma jerárquica.

El primer nivel de la jerarquía permite ubicar al documento en categorías generales del dominio. El segundo nivel está compuesto por diferentes temas que se pueden hallar dentro de cada una de las categorías. El tercer nivel de la jerarquía está conformado por los subtemas que puedan hallarse dentro de cada tema del nivel anterior, en caso de poder realizar esta subclasificación.

En el dominio de los diarios digitales el primer nivel de la jerarquía corresponde a las distintas secciones que posee un diario, es decir política, economía, internacionales, deportes, espectáculos o información general. Esta primera clasificación nos da información relevante acerca del tema general que le interesa leer al usuario.

La categoría a la que pertenece el documento, en el caso de los diarios digitales, se obtiene a partir del URL de la página. Mediante un análisis de los diferentes sitios de los diarios, se observó que cada uno de estos sitios posee una estructura determinada que les permite organizar las páginas que los componen. Por ejemplo, una página cuyo URL es `http://www.clarin.com/diario/hoy/pol_sum.htm` indica que pertenece a la sección de política del diario Clarín.

Cada una de las secciones de un diario contiene noticias de temas variados. Con los distintos temas identificados para cada una de las secciones, se construye el segundo nivel de la jerarquía. Por ejemplo, un documento que se encuentre en la sección de economía puede referirse a la bolsa, a las importaciones o exportaciones, a las finanzas.

Los diferentes subtemas para cada uno de los temas de las distintas secciones, conforman el último nivel de la jerarquía estática. Por ejemplo, en el tema automovilismo de la sección de deportes se pueden identificar los subtemas rally, turismo carretera, fórmula 1.

En la figura 4 se muestran algunos de los temas que conforman la base de temas estática utilizada para clasificar documentos en el dominio de los diarios digitales.

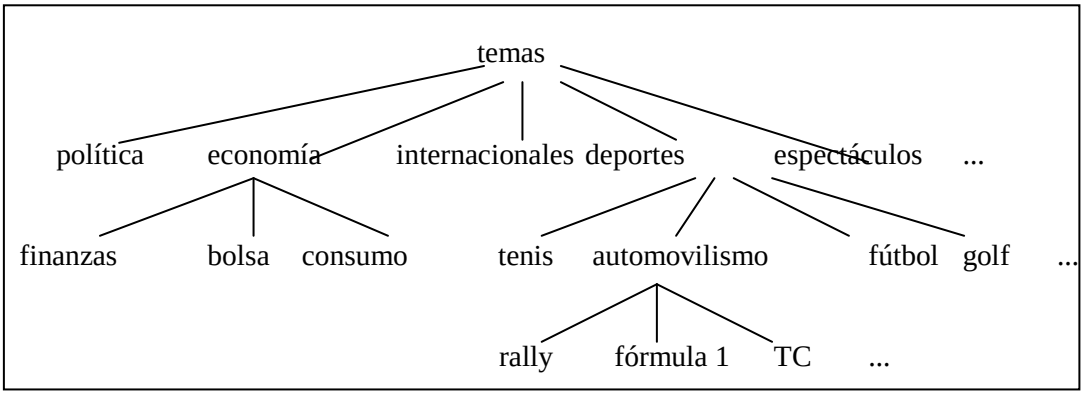


Figura 4. Jerarquía de temas estática

Cada uno de los temas y subtemas que conforman la jerarquía se encuentra definido por un conjunto de palabras. Estas palabras son aquellas que determinan al tema o subtema en cuestión. Por ejemplo algunas de las palabras que definen al tema bolsa de la sección de economía son MerVal, alza, operadores, caída, bonex, índice, acciones, cotización, tasa, Alpargatas, etc.

Una vez ubicado el primer nivel de la jerarquía, se determina el tema del documento analizado. Para ello se establece qué porcentaje de las palabras del documento se corresponden con las palabras que definen a cada uno de los temas dentro de esa categoría.

Sin embargo, no todas las palabras del documento se utilizan para determinar el tema, pues no todas representan información relevante. En primer lugar, dado que los documentos digitales se encuentran en formato HTML, se identifica la parte del documento que contiene el texto de la nota, eliminando todos los tags que se utilizan para darle formato a la página.

Una vez recuperado el texto de la noticia, se descartan los artículos, los conectivos, las preposiciones, los pronombres, ya que no brindan información propia de un determinado tema, sino que pueden encontrarse en cualquier documento.

Del total de palabras restantes se establece que proporción se corresponden con las palabras que definen a un tema determinado. Una vez obtenido este porcentaje, se compara con un valor umbral prefijado para cada uno de los temas de la jerarquía. Si el porcentaje supera al valor umbral, podemos decir que la página se refiere al tema en cuestión. De otro modo, no se le asigna ningún tema.

Una vez que se clasificó en base al segundo nivel de la jerarquía, se realiza un proceso similar al anterior para determinar a que subtema particular pertenece la página, de existir este tercer nivel en la jerarquía.

Luego de clasificar el documento, si el usuario no brinda información explícita acerca de sus intereses, se necesita algún método para inferir el nivel de relevancia de este documento. Para ello se registra el tiempo de permanencia en la página para deducir el grado de interés que el usuario muestra por la misma.

El tiempo de permanencia en cada una de las páginas se obtiene registrando el instante en el cual el usuario comienza a leer una página y el tiempo de finalización. Para determinar si el tiempo de permanencia en una página es suficiente para que el usuario haya leído la página detenidamente se lo compara con dos umbrales, uno mínimo y otro máximo. En el umbral mínimo se contempla el tiempo que puede tardar en bajar la página en el navegador. Además se considera el tiempo necesario para que el usuario observe de qué trata la nota en función del tamaño de la misma. El valor umbral máximo se tiene en cuenta pues un tiempo demasiado grande tal vez indique que el usuario no está utilizando la aplicación en ese momento o hubo algún inconveniente por el cual no se pudo cargar la página. En ese caso, la lectura no se tiene en cuenta.

Con el fin de mantener actualizada la información, se registra la fecha en la que el usuario está navegando y en el momento de tener en cuenta la información de navegación para determinar el perfil de lectura del usuario, se descartan aquellas lecturas que se consideran demasiado antiguas.

3.2 Consideraciones al clasificar documentos

La base de temas estática descripta en las sección anterior se encuentra en español. Cada uno de los temas y subtemas que componen la jerarquía están definidos con palabras en español. Cuando el usuario accede a diarios digitales en otros idiomas, se realiza un proceso de traducción para poder realizar el análisis de los distintos documentos. Por ejemplo, si un usuario está leyendo una página del diario Le Monde, una vez extraído el texto de la noticia, se traducen las palabras de francés a español. El análisis de estas palabras, una vez traducidas, se realiza de la forma explicada anteriormente.

Cabe destacar que no todos los diarios nombran a sus secciones de la misma manera. Por ejemplo, para referirse a la sección de noticias internacionales, el diario Clarín utiliza la denominación 'Internacionales' pero el diario La Nación la llama 'Exterior'. Ante esta situación se traducen los nombres reales de las secciones a los nombres asignados a las categorías correspondientes en la base estática de temas, utilizando alias.

Un caso particular es el diario El Cronista, que por ser un diario con orientación económica, posee secciones diferentes a los demás diarios. Mientras que La Nación y El Clarín poseen solamente la sección 'Economía', el diario El Cronista contiene además de ésta, las secciones 'Mercados', 'Mercosur' y 'Negocios'. Para homogeneizar el análisis de las páginas de todos los diarios, las páginas que corresponden a estas secciones específicas se categorizan como pertenecientes a la categoría economía y el nombre de la sección se utiliza para ubicar al documento en el segundo nivel de la jerarquía, es decir se toma el nombre de la sección como tema.

3.3 Clasificación específica de la información

Una base fija de temas sólo permite clasificar los documentos en categorías muy generales. Si se desea realizar una categorización más específica de un determinado documento, esta clasificación posiblemente resultaría insuficiente y podría volverse obsoleta rápidamente debido a la naturaleza dinámica de la información existente en Internet.

Se necesitan técnicas que permitan analizar los documentos en un contexto específico. Este análisis requiere de algún modo establecer el significado de los documentos correspondientes. Se considera que la técnica de Razonamiento Basado en Casos es adecuada porque permite integrar conocimiento semántico. El significado de un documento incluye el tema específico del que trata, así como también el contexto en el que se encuentra [Lenz 98].

La utilización de Razonamiento Basado en Casos permite realizar una clasificación específica haciendo uso de la información registrada de documentos considerados de interés para el usuario. Esta técnica constituye un paradigma para construir sistemas inteligentes que se basa en observaciones previas. Básicamente resuelve un nuevo problema recordando una situación previa similar y reusando la información y el conocimiento de esa situación [Kolodner 93].

En este trabajo se utiliza una jerarquía estática de temas para obtener las características generales del contenido de un documento. En forma complementaria se realiza una clasificación dinámica y específica utilizando la técnica de Razonamiento Basado en Casos.

Para modelar este problema se considera que cada lectura de una nota es un caso, donde cada dimensión del caso es un par atributo-valor que registra la información relevante del documento leído.

El contenido de un documento se describe utilizando un conjunto de palabras representativas del mismo. Se propone extraer de los documentos los sustantivos destacando entre ellos los sustantivos propios. Estos brindan información sobre el tema específico al que se refiere el documento, ya que representan nombres de personas, países, empresas. Los sustantivos comunes se consideran palabras relevantes porque definen el contexto en el cual se utilizan estas palabras en mayúscula. Para extraer los sustantivos se realiza un análisis gramatical del documento. Se realiza un análisis gramatical diferente para cada uno de los idiomas considerados.

Con el fin de clasificar un documento en forma más específica se le asigna un tema codificado basándose en las palabras recuperadas que representan su contenido y en los documentos anteriormente clasificados.

Ante un nuevo documento a clasificar, se recuperan de la base de casos aquellos casos que representan documentos similares. Para ello se utilizan como índices de recuperación la categoría a la cual pertenece el documento y el tema asignado al mismo a partir de la jerarquía de temas estática. De los casos obtenidos se selecciona mediante el proceso de matching aquel que más se asemeje a la nueva situación. Este proceso establece las correspondencias entre las características que describen ambos casos y halla el grado de similitud entre ellos. Encontrar las correspondencias entre casos tiene por objetivo determinar qué características de la nueva situación se corresponden con qué características de una situación almacenada.

Cada uno de los casos está compuesto por un conjunto de descriptores que representan los documentos leídos. Como se observa en la figura 5 las dimensiones de cada caso, en el dominio de los diarios digitales, se corresponden al diario al que pertenece la página, la sección, el tema, el tema más específico, el conjunto de palabras representativas del documento, el total de palabras y el total de palabras representativas.

Problema: Analizar Información de páginas de Internet en el dominio de diarios digitales.
Objetivo: Determinar el tema
Situación:
diario (La Nación)
sección (Deportes)
tema (automovilismo)
temaMasEspecífico(Fórmula Uno)
palabrasRelevantes
<Formula,2,up.2><Uno,2,up,2><Hakkinen,2,up,2>
<McLaren,5,up,2><finlandés,2.low,1><carrera,2.low,1>
<campeonato,2.low,1><piloto,1.low,1><Ferrari,2,up,2>
<camino,1.low,1><equipos,1.low,1><Coulthard,1,up,2>
<promedio,1.low,1><ensayos,1.low,1><Schumacher,1,up,2>
<Irvine,1,up,2><radiador,1.low,1>...
cantidadTotalPalabras (424)
cantidadPRs (101)
Solución:
subtema (COD1)
Evaluación:
éxito(si_no)
Feedback:
utilizado para modificar el peso de las palabras relevantes

Figura 5. Representación de un documento del sitio La Nación.

Los casos se comparan unos con otros, dimensión por dimensión, teniendo en cuenta la importancia de cada dimensión en el matching. La importancia asociada con cada dimensión indica el grado de atención que se debe prestar a las similitudes y diferencias de los valores en dicha dimensión al calcular la similitud entre dos casos.

Para comparar dos casos, se obtiene el grado de similitud de un caso con otro, combinando los puntajes de matching dimensionales individuales. Para ello se utiliza una función de evaluación numérica que combina el grado de similitud de cada dimensión con un valor que representa la importancia de la dimensión en el caso.

La función de evaluación utilizada para calcular el grado de similitud total se muestra a continuación.

$$SIM(N,F) = \sum_{D_{iN} \in N,\ D_{iF} \in F} sim_i(D_{iN}, D_{iF}) * W_i$$

Donde N es el caso nuevo, F es el caso recuperado, w_i es la importancia de la dimensión i, sim_i es la función de similitud local de la dimensión i y D_i es el valor correspondiente al i-ésimo descriptor.

El valor obtenido a partir de la función de evaluación numérica es comparado con un umbral preestablecido. Si el valor es mayor a este umbral los casos son similares y se le asigna al nuevo caso el código de tema puntual del caso recuperado. De no ser así se le asigna un código nuevo de tema puntual. Finalmente este nuevo caso se almacena en la base de casos para ser tenido en cuenta en clasificaciones posteriores.

4. Modelo de filtrado

El perfil del usuario es utilizado para filtrar la información relevante para el mismo, a partir de la información disponible.

La evaluación que el usuario realiza de la calidad de los documentos recuperados también se incorpora como parte de la tarea de aprendizaje. La información proveniente de las evaluaciones sobre las predicciones del agente se utiliza para refinar el perfil del usuario. También es posible mejorar la calidad de los conjuntos de palabras que definen las preferencias en cada dominio.

Utilizando el perfil del usuario como modelo de filtrado, el agente puede obtener el conjunto de documentos que podrían ser de interés para el usuario, a partir de la información accesible. La calidad de los documentos recuperados es mejorada conforme se incrementa el conocimiento de las características específicas que el usuario prefiere. Este conocimiento se incrementa con la observación de sucesivas navegaciones y la evaluación del usuario sobre las predicciones del agente. El usuario puede evaluar cada documento presentado e informar cuando un documento es relevante. El perfil del usuario se modifica teniendo en cuenta este criterio.

La construcción del perfil se realiza a partir del conocimiento que el usuario provee explícitamente y de las creencias que se deducen a partir de la observación del comportamiento del usuario. La colección de preferencias de un usuario se utiliza para determinar su perfil de intereses. El perfil de un usuario brinda la información necesaria para elaborar un criterio de relevancia. Este criterio es utilizado para realizar la tarea de selección y recuperación automática de documentos desde las fuentes de información que de otra forma deberían ser constantemente monitoreadas por el usuario.

La información obtenida acerca de las características preferidas por cada usuario es combinada de acuerdo al nivel de detalle del tema de interés. Cuando la información acerca de las preferencias se obtiene mediante la interacción directa con el usuario, sólo se tiene la categoría general y el nombre del tema. Por otro lado, a partir de la observación de la conducta previa del usuario es posible obtener preferencias con un nivel más específico de detalle. Este nivel está dado por el tema puntual acerca del cuál trata el documento obtenido automáticamente utilizando Razonamiento Basado en Casos. Estas preferencias, que incluyen temas puntuales, tienen una validez limitada. De esta forma es posible mantener el perfil de preferencias actualizado, sin que esto le demande esfuerzo al usuario.

Las diferencias que se presentan en el nivel de detalle del conocimiento adquirido se utilizan como base para desagregar el perfil de cada usuario. Se construye un perfil de características generales a partir de las preferencias cuyos temas fueron provistos por el usuario explícitamente y un perfil de características específicas sobre la base de los temas puntuales obtenidos automáticamente. A continuación se describe como se obtiene cada una de las partes del perfil.

4.1 Construcción del perfil de características generales

El perfil general de un usuario, contiene las características más generales de los documentos que considera relevantes. Estas preferencias incluyen los sitios en los cuales debe ser buscada la información, junto con la categoría y el tema que deben contener los documentos recuperados. El

perfil general se construye exclusivamente basándose en las preferencias ingresadas directamente por el usuario.

En este trabajo se utilizan un conjunto limitado de sitios, que pertenecen al dominio de los diarios digitales. Debido a que la selección de los sitios preferidos no es obligatoria, las combinaciones de categorías y temas pueden asociarse con el nombre de un sitio predeterminado. Es decir, si un usuario no selecciona el sitio del cual privilegia la recuperación de documentos, se utiliza un sitio ya establecido.

4.2 Construcción del perfil de características específicas

La información mantenida en el perfil de características específicas del usuario es similar a la registrada en el perfil general. Las preferencias almacenadas incluyen el sitio en el cual deben buscarse estos documentos, junto con la categoría y el tema que el usuario considera relevantes en el contenido de los mismos. La diferencia entre las características almacenadas en ambos conjuntos es el nivel de detalle del tema seleccionado.

El perfil específico se construye a partir de las creencias que el agente mantiene sobre las preferencias del usuario. Este perfil puede corregirse a partir de las evaluaciones que el mismo usuario realiza sobre la calidad de los documentos recuperados. La evaluación de los resultados puede ser explícita o implícita. Si el usuario no provee tal información la evaluación de las predicciones puede ser deducida a partir de su comportamiento ante los documentos presentados.

Para generar el perfil específico de un usuario se tiene en cuenta la observación del comportamiento que el mismo ha presentado anteriormente junto con la información que ha brindado en forma indirecta. El usuario no manipula esta información directamente, por lo tanto es necesario que el agente se encargue de mantener tal información actualizada. Para asegurar que las preferencias que se consideran al construir el perfil no son obsoletas, se eliminan aquellas que no fueron obtenidas recientemente. Las preferencias que describen lecturas realizadas en un período mayor a un mes atrás son consideradas obsoletas y por lo tanto deben ser eliminadas. El agente es también el encargado de mantener la consistencia de la información que se mantiene sobre las preferencias del usuario. Esta tarea consiste en eliminar de las preferencias inferidas las características que el usuario no considera relevantes. Esta información se obtiene directamente a través de la interfaz, o indirectamente mediante la evaluación que el usuario realiza de los documentos presentados.

La primera aproximación que se realiza al perfil de características específicas se basa en las observaciones del comportamiento del usuario. Se mantiene un registro de las características de las lecturas realizadas por el usuario que se consideraron relevantes. El primer paso en la construcción del perfil específico es la agregación de las lecturas realizadas que tienen las mismas características. Se genera un nuevo conjunto conteniendo una única instancia de cada combinación sitio, categoría y tema. Se registra el tiempo de lectura acumulado para cada agregación de preferencias. El enfoque utilizado para determinar el grado de relevancia de cada conjunto de características se basa en la suposición de la existencia de una relación directa entre el tiempo de permanencia en un documento y el grado de relevancia que este documento presenta para el usuario. Por lo tanto, al ordenar las preferencias agregadas por el tiempo de lectura acumulada, es posible obtener un orden inicial de la relevancia que cada conjunto de características presenta para el usuario. Las preferencias son ordenadas en forma descendente. Las características que el usuario privilegia se encuentran en las primeras posiciones del perfil del usuario. Si no existe información adicional acerca de las características preferidas por un usuario, el perfil obtenido es utilizado para el filtrado de la información.

Sin embargo, es posible contar con información provista indirectamente por el usuario. Si el usuario ha señalado su interés por documentos particulares, las características obtenidas al analizar estos documentos se utilizan para corregir el perfil generado. Las características de los documentos seleccionados explícitamente han sido registradas para ampliar el conocimiento acerca de las preferencias del usuario. Estas características mantienen el mismo nivel de detalle que las preferencias inferidas a partir del comportamiento del usuario. Es decir la definición del tema contiene un nombre de tema codificado.

Es posible comparar las preferencias inferidas con aquellas que han sido obtenidas indirectamente, en busca de valores idénticos de combinaciones de sitio, categoría y tema. Todas las características de los documentos seleccionados son evaluadas.

Si un conjunto de características del vector de preferencias seleccionadas se encuentra entre las que están presentes en el perfil preliminar, la importancia de este conjunto de características es incrementado. El motivo de este incremento se debe a que el usuario ha indicado explícitamente su interés, el cual queda reflejado en un aumento de la relevancia de esas características. La certeza del interés del usuario en un conjunto de características incrementa su importancia.

Si existen conjuntos de características en el vector de preferencias seleccionadas que no se encuentran en la primera versión del perfil deben ser ahora incorporados. Estos conjuntos se incorporan con un nivel de relevancia preestablecido, debido a la certeza de que tales conjuntos son de interés para el usuario. Este nuevo conjunto obtenido se encuentra ordenado de acuerdo a la relevancia que posee para el usuario.

Adicionalmente puede existir información acerca de características generales que ha sido provista previamente por el usuario. Si bien esta información ha sido utilizada para la construcción del perfil de carácter general, tales características son utilizadas para mejorar la calidad del perfil específico que describe los intereses del usuario. El conocimiento provisto directamente por el usuario es utilizado para mejorar la calidad del perfil construido. Las características preferidas por el usuario son ponderadas teniendo en cuenta si pertenecen a las preferencias ingresadas explícitamente por el usuario. De ser así, el nivel de relevancia de esta preferencia es incrementado. En forma complementaria, si alguna característica no es de interés del usuario, el grado de relevancia de las preferencias que la incluye se disminuye.

5. Construcción del diario digital

La tarea principal involucrada en la recuperación de la información relevante es el análisis de la información disponible y el filtrado de la información relevante. Este filtrado involucra el estudio de cada artículo, evaluando si su contenido es relevante o irrelevante para cada usuario sobre la base de su perfil.

La información recuperada de los sitios de interés del usuario es analizada utilizando este perfil del usuario como modelo de filtrado, de tal forma que para cada usuario sólo se presenta el conjunto de documentos con información relevante. El resultado de tales predicciones, puede ser evaluado por el usuario para mejorar la calidad de futuros resultados.

Para obtener los documentos relevantes para cada usuario, se contrastan las características de todos los documentos recuperados con las propiedades en las que el usuario está interesado. Aquellos documentos cuyas características se corresponden con las propiedades que el usuario considera de interés, son recuperados y posteriormente clasificados de acuerdo al grado de relevancia que tales características presentan para ese usuario. En forma análoga, los documentos que no presentan las características de interés para el usuario son descartados.

5.1 Recuperación de la información relevante

La información relevante para el usuario, debe ser recuperada analizando los documentos que se encuentran en los sitios que el usuario considera de interés. En el contexto de los diarios digitales, las noticias del día son los documentos que se analizan para encontrar la información relevante. Para obtener las páginas que contienen las notas de cada día es necesario realizar una conexión con el sitio del diario digital del cual se desea obtener tales documentos.

Si algunas de las preferencias del usuario indican que ciertos documentos relevantes pertenecen a una categoría determinada, se busca la información de interés en la sección equivalente del diario. Por ejemplo, si un usuario está interesado en documentos que contengan información que se clasifica dentro de la categoría de economía, esta información debe ser buscada a través de todas las notas pertenecientes a la sección de economía del diario.

Dado que en el dominio de los diarios digitales todas las notas que corresponden a una misma categoría están agrupadas en la misma sección, para realizar la búsqueda de documentos con esa característica, es suficiente con obtener la página índice de la sección en cuestión. Una página índice contiene los vínculos que posibilitan el acceso a la información que debe ser analizada. Utilizando estos vínculos se obtienen los documentos que contienen las notas de las secciones preferidas por el

usuario. Una vez obtenido el documento que corresponde a cada noticia, se analiza su contenido para determinar si la información que provee es relevante para el usuario.

El análisis de los documentos es el que marca la diferencia entre las notas que deben ser recuperadas para cubrir los requerimientos del perfil general y el específico.

El perfil de características específicas contiene una combinación de sitio, categoría y tema, que un documento debe poseer para ser considerado relevante. La definición del tema es la propiedad que diferencia los dos conjuntos de preferencias. En el perfil de características específicas el tema se define con un código.

Cada uno de los documentos recuperados para una categoría determinada es analizado teniendo en cuenta estas propiedades. Una noticia se considera relevante si cubre todos los requerimientos de alguna preferencia. Se parte de la base de que la noticia pertenece al sitio y a la categoría que incluye la preferencia que se trata de cubrir, debido a que tales características del perfil de preferencias se utilizaron para recuperar los documentos que se analizan. Un documento es considerado como relevante sólo si su nombre de tema, nombre de subtema y código de tema, tienen el mismo valor que las propiedades incluidas entre las características preferidas por el usuario.

El perfil de características generales incluye también una combinación de sitio, categoría y tema que un documento debe cubrir para ser considerado relevante. A diferencia de las características del perfil específico, las características del perfil general no contienen una definición específica del tema. Un tema definido en el contexto del perfil general, sólo incluye el nombre de un tema perteneciente a la jerarquía estática de temas. Un documento se presenta al usuario si a partir del análisis de su contenido es posible asignarle el mismo nombre de tema que el incluido en la preferencia que se trata de satisfacer.

En ambos conjuntos de preferencias, se analizan aquellas que se encuentran en las posiciones más altas de los perfiles. Sin embargo, si no es posible encontrar documentos que cubran los requerimientos de las primeras posiciones se debe continuar el análisis esperando conseguir un número de documentos previamente determinado.

5.2 Presentación del diario personalizado

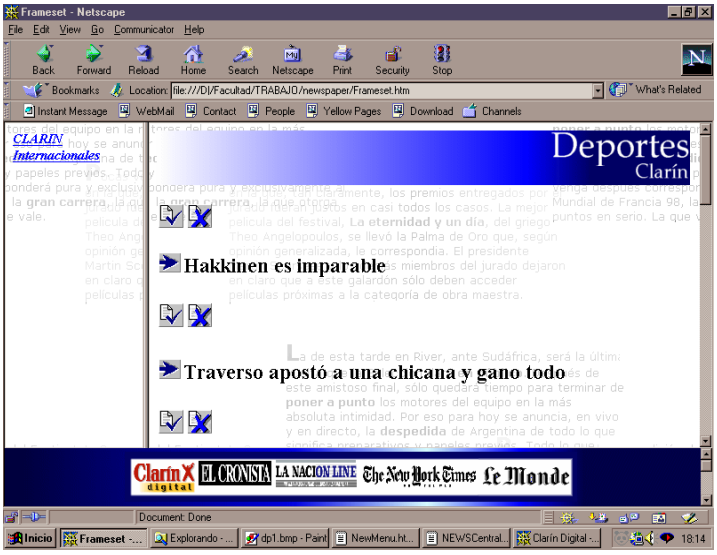
La recuperación de la información desde los diarios digitales, y el posterior análisis de los documentos recuperados, son las principales tareas para la construcción de un diario digital personalizado. La presentación de la información al usuario, es una tarea complementaria a las anteriores.

El análisis de la información contenida en los documentos pertenecientes a los diarios digitales presenta como resultado el conjunto de noticias del día que presentan algún grado de relevancia. Estos documentos, se encuentran agrupados en dos subconjuntos disjuntos. Esta división se realiza teniendo en cuenta si el documento recuperado cubre los requerimientos del perfil de características específicas, o en su defecto, los requerimientos del perfil de características generales.

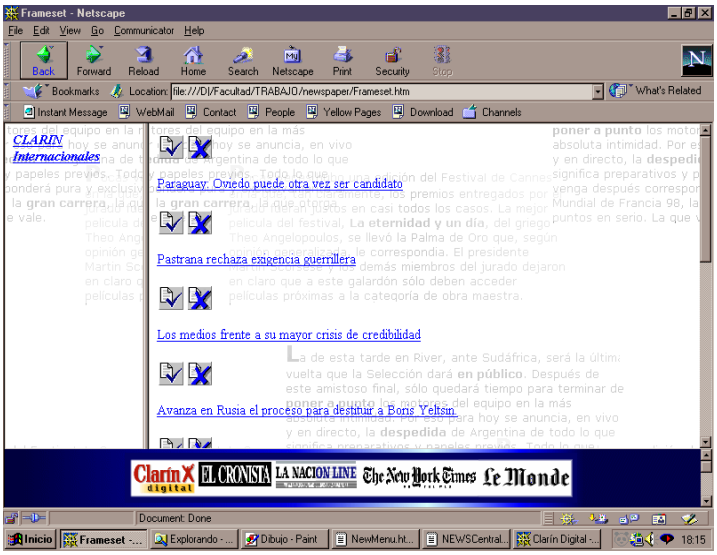
Los vínculos para acceder a las noticias que cubren los requerimientos del perfil específico se presentan en la página principal del diario digital generado. Las noticias se agrupan de acuerdo a la sección y al diario al que pertenecen. Si la noticia incluye alguna imagen, ésta se presenta junto con el titular de la misma en la tapa principal.

Los documentos que contienen información con las propiedades exigidas por el perfil de características generales, son presentados en grupos de acuerdo al diario y la sección a la que pertenecen. Se construye una página para cada una de estas combinaciones. Junto a la página principal se encuentran los vínculos que posibilitan el acceso a estas páginas.

En la figura 6.a se muestra una imagen del diario personalizado generado para un usuario. En la figura 6.b se presenta un ejemplo de una página que contiene los titulares correspondientes a una combinación de diario y sección que satisface el perfil de características generales.



(a)



(b)

Figura 6. (a) Página central del diario personalizado con las noticias que pertenecen al perfil específico (b)Pagina perteneciente a una combinación de diario y session perteneciente al perfil general

5.3 Evaluación del diario digital presentado.

Una de las técnicas utilizadas por la tecnología actual de recuperación de información es la utilización de la evaluación que el usuario provee acerca de la relevancia de los documentos presentados [Chen 97]. En este proceso el usuario identifica los documentos relevantes a partir de la lista de documentos mostrados. La evaluación que el usuario realiza sobre las predicciones del agente se denomina feedback de relevancia.

La evaluación de los resultados puede ser explícita o implícita. El usuario puede señalar sus intereses explícitamente evaluando la lista de documentos recuperados. Con este fin, junto a cada titular que se presenta al usuario, se incluyen dos iconos para que este provea una evaluación acerca de los documentos presentados. Si la información recuperada es de interés para el usuario, puede indicarlo explícitamente a través del icono correspondiente. De lo contrario, puede mostrar que la evaluación ha sido negativa, utilizando el icono incluido para tal fin. La evaluación que el usuario realiza es utilizada para mejorar el conocimiento del agente acerca de las preferencias del usuario.

Si el usuario no provee tal información la evaluación de las predicciones puede ser deducida a partir de su comportamiento ante los documentos presentados. Es decir, se infiere la relevancia de

cada documento para el usuario en base al comportamiento que éste manifiesta frente a los documentos presentados. Para poder inferir una evaluación acerca de la calidad de las predicciones realizadas por el agente para un usuario en particular, se considera el tiempo que éste habitualmente dedica para la lectura de diarios digitales. Si el usuario dedica un tiempo distinto al tiempo promedio de lectura no es posible sacar conclusiones a partir de su comportamiento. En cambio, si el usuario dedica un tiempo comparable con el promedio se puede deducir la relevancia de los documentos recuperados. Si un documento no es visitado por el usuario o su permanencia en él es inferior a la esperada, se concluye que el documento no es de su interés. Si se accede al documento y el tiempo de permanencia se encuentra dentro de los límites esperados, se infiere que el documento es relevante.

El feedback explícito, provisto por el usuario, se utiliza para refinar el conjunto de palabras que definen el tema puntual de su preferencia registrado entre las características de su perfil. La depuración del conjunto de palabras se consigue modificando la importancia de cada una. La importancia de una palabra está determinada por el peso que ésta posee en la definición de un tema preferido por un usuario.

La evaluación obtenida a partir del usuario, ya sea en forma explícita o deducida por el agente, se utiliza para mejorar la calidad de las preferencias del usuario.

6. Trabajos relacionados

NewsAgent es un agente capaz de obtener la información relevante para un usuario a partir de los documentos disponibles en los sitios de los diarios digitales. Los gustos y preferencias del usuario son utilizados para distinguir la información relevante. WebWatcher [Joachims 97] y Syskell & Webert [Pazzani 96] son sistemas desarrollados para asistir al usuario durante la navegación a través de la Web. La información es evaluada utilizando los gustos e intereses provistos explícitamente por el usuario. NewsAgent utiliza este tipo de información y cuenta con la capacidad adicional de deducir los intereses del usuario a partir de su comportamiento.

Letizia [Lieberman 95] es un asistente personal que monitorea el comportamiento del usuario y trata de inferir las preferencias de éste. Utiliza un conjunto de heurísticas para inferir aproximaciones útiles sobre las preferencias de los usuarios.

Estos agentes asisten al usuario en la búsqueda de información relevante a medida que el usuario navega. Por el contrario, NewsAgent obtiene un conjunto de documentos que podrían ser de interés para el usuario y los presenta en una página. Lira [Balabanovic 95] es un sistema de recuperación de información que también trabaja off-line retornando un conjunto de páginas que coinciden con los intereses del usuario.

Estos asistentes personales utilizan diferentes técnicas de recuperación de información para distinguir los documentos relevantes. Una técnica de recuperación de información es utilizada para determinar qué características del documento deben ser registradas. NewsAgent utiliza Razonamiento Basado en Casos para determinar la relevancia de un documento calculando su similitud con documentos previamente calificados como interesantes. SPIRE [Daniels 97] es un sistema híbrido que combina Razonamiento Basado en Casos y técnicas de recuperación de información para la recuperación de texto a partir de documentos completos. La utilización de este enfoque híbrido permite producir resultados o funcionalidades que no pueden ser alcanzados con cada técnica individualmente. BROADWAY [Jaczynski 97] es un asistente de navegación que utiliza el paradigma de Razonamiento Basado en Casos para aprender a partir de las navegaciones de los usuarios el conjunto de casos relevantes. El uso de Razonamiento Basado en Casos se basa en las siguientes hipótesis: si dos usuarios van a través de una secuencia similar de documentos ellos intentarán navegar en forma similar. Es posible aconsejar a un usuario en base a los documentos evaluados como relevantes para el otro usuario.

7. Conclusiones

En este trabajo se presenta un agente capaz de generar diarios personales a partir de las preferencias de un usuario. El agente tiene la capacidad de observar el comportamiento del usuario y deducir sus intereses a partir de esta conducta. El modelo de filtrado usado para la recuperación de

información relevante se basa en estas preferencias y en los gustos ingresados explícitamente por el usuario. Se utiliza la técnica de Razonamiento Basado en Casos para realizar una clasificación específica de los documentos interesantes a partir de similitudes con documentos previamente registrados. Se propone un análisis gramatical para obtener las características que describen el tema específico del documento y el contexto en el cual se encuentra.

El agente está siendo utilizado por diferentes usuarios y las pruebas realizadas hasta el momento sugieren que las noticias presentadas por el agente concuerdan con los intereses del usuario.

Este proyecto está siendo continuado sin poner restricciones en el dominio de la información.

Referencias

[Balabanovic 95]
Marko Balabanovic, Yoav Shaam. 1995, Learning Information Retrieval Agents: Expriments with Automated Web Browsing. *Proceedings of the AAAI Spring Symposium Series on Information Gathering from Heterogeneous, DistributedEnviroment*

[Daniels 97]
J. J. Daniels and E. L. Rissland. What You Saw Is What You Want: Using Cases to Seed Information. *In Case-Based Reasoning Research and Development* ,1997

[Jaczynski 97]
M. Jaczynski and Brigitte Trosse BRODWAY: A World Wide Web Browsing Advisor Reusing Past Naviations from a Group of Users. *In Proceedings of the Third UK Case-Based Reasoning Workshop. Manchester, UK,* September 9, 1997

[Joachims 97]
Thorsten Joachims, Dayne Freitag, Tom Mitchell 1997. WebWatcher: Atour Guide for the World Widw Web. *Proceedings of IJCA,* August 1997.

[Kolodner 93]
Janet Kolodner. *Case Based Reasoning* . 1993

[Lenz 98]
Mario Lenz, André hübner, and Miriam Kunze. *In proceeding of International Conference on Flexible Query Answering Systems,* May 1998, Roskilde, Denmark

[Lieberman 95]
Henry Lieberman 1995, Letizia: An agente that assists web browsing. *In International Joint Conference of Artificial Intelligence*.

[Lieberman 97]
Henry Lieberman 1997, Autounomous Interface Agent. *In Proceedings of the ACM Conference on Computers and Human Interface.* March, 1997.

[M. Pazzani 96]
M. Pazzani, J Muramatsu, D. Billsus. 1996, Syskill & Webert: Identifying interesting web sites. *In AAAI Conference, Portland* 1996.