

Modelado dinámico de usuario en un sistema de personalización de contenidos Web*

Alberto Díaz¹ y Pablo Gervás²

¹ CES Felipe II – Universidad Complutense de Madrid
C/ Capitán 39, 28300 Aranjuez, Madrid
adiaz@cesfelipesecondo.com

² Departamento de Sistemas Informáticos y Programación
Facultad de Informática – Universidad Complutense de Madrid
c/ Juan del Rosal, 8, Madrid 28040
pgervas@sip.ucm.es

Resumen. En este artículo se presenta un sistema de personalización de contenidos Web basado en un modelo de usuario que almacena intereses a largo y corto plazo. Los intereses a largo plazo se modelan mediante la elección de las categorías y las palabras clave sobre las que un usuario necesita información. Sin embargo, las necesidades de los usuarios varían con el tiempo como efecto directo de su interacción con la información. Por esta razón el modelo de usuario debe ser capaz de adaptarse a esos cambios de interés. En nuestro caso, esta adaptación o modelado dinámico se realiza mediante el modelo a corto plazo obtenido a partir de la realimentación producida por el usuario. Se han realizado varios experimentos que han determinado que la combinación de los modelos a corto y largo plazo es la más adecuada cuando se utilizan tanto las categorías como las palabras clave para el modelo a largo plazo.

1 Introducción

Los contenidos Web aparecen de muy diversas maneras en distintos dominios de aplicación pero en la mayoría de ellos la forma de presentación de la información es la misma para todos los usuarios, es decir, esos contenidos son estáticos en el sentido de que no se adaptan a cada usuario. La personalización de contenidos es una técnica que trata de eliminar la sobrecarga de información mediante la adaptación de los contenidos Web a cada tipo de usuario.

Un sistema de personalización está basado en 3 funcionalidades principales: selección de contenidos, adaptación del modelo de usuario y generación de contenidos. Para que estas funcionalidades se realicen de manera personalizada deben estar basadas en información relacionada con el usuario que debe estar reflejada en su perfil o modelo de usuario.

* Esta investigación ha sido parcialmente financiada por el Ministerio de Ciencia y Tecnología (TIC2002-01961)

La selección de contenidos se refiere a la elección entre todos los documentos de entrada de aquellos que son más relevantes para un usuario dado, según su perfil o modelo. Para poder realizar esta selección es necesario obtener una representación de los documentos, una representación del modelo de usuario y una función de similitud que calcule el nivel de adecuación entre ambas representaciones.

La adaptación del modelo de usuario es necesaria ya que las necesidades de los usuarios varían con el tiempo como efecto directo de su interacción con la información (Billsus&Pazzani, 2000). Por esta razón el modelo de usuario debe ser capaz de adaptarse a esos cambios de interés, es decir, debe ser dinámico. Esta adaptación se realiza mediante la interacción del usuario con el sistema, a través de la cual se obtiene información para la realimentación del perfil.

En nuestro caso, la generación de contenidos consiste en generar un nuevo documento Web resultado que contenga para cada documento seleccionado, su título, su relevancia, un resumen y un enlace al texto completo.

En este artículo nos centraremos en la parte correspondiente a la adaptación del modelo de usuario y en las distintas posibilidades para realizar este proceso. El objetivo es ver cuál es la mejor forma de realizar el proceso de adaptación del modelo de usuario para mejorar la selección de contenidos.

2 Técnicas y métodos disponibles

Existen en la bibliografía distintas técnicas para definir los intereses de los usuarios: términos, estereotipos, redes semánticas, redes neuronales, etc. En particular hay un conjunto de propuestas (Chiu&Webb, 1998; Billsus&Pazzani, 2000) que modelan a los usuarios combinando intereses a largo y corto plazo: el modelo a corto plazo representa las preferencias del usuario más recientes y el largo plazo las que se han producido en un período de tiempo más largo. Para determinar si un documento es relevante para un usuario se utiliza el modelo a corto plazo siempre que éste dé una respuesta satisfactoria, en caso contrario, se utiliza el modelo a largo plazo.

Para obtener una representación del contenido textual de los documentos se suelen utilizar técnicas basadas en vectores de pesos de términos (Salton, 1989). Los vectores asociados a los documentos se pueden obtener eliminando las palabras vacías almacenadas en una lista de parada y obteniendo las raíces de las palabras aplicando un extractor de raíces. Para calcular los pesos se suele utilizar la fórmula $tf \cdot idf$, basada en frecuencias de aparición de términos (Salton, 1989).

Para realizar la selección de contenidos existen diversos algoritmos de clasificación dependiendo de las representaciones elegidas para el modelo de usuario y los documentos: fórmula del coseno, reglas asociadas a estereotipos, redes neuronales, vecino más cercano, clasificador bayesiano ingenuo, etc.

Las técnicas de realimentación necesarias para poder conseguir un modelado dinámico del usuario se basan en la realimentación del usuario respecto de los elementos de información que se seleccionan según su perfil. La información obtenida se utiliza para actualizar el modelo de usuario de diversas formas según sea la representación elegida: actualización de pesos de términos, actualización de redes semánticas, actualización de reglas asociadas a estereotipos, etc.

En particular, en (Nakashima&Nakamura97) se propone un sistema basado en agentes inteligentes aplicado a un periódico digital. El modelo de usuario almacena información “consciente” del usuario en forma de términos con peso, e información “inconsciente” en forma de términos asociados a aspectos como edad, sexo, ocupación, estado civil, ciudad, etc. La selección se calcula utilizando una combinación de ambas informaciones. Para la primera se tiene en cuenta si los términos del usuario aparecen en el documento, dando una relevancia adicional si el término aparece en el título. Para la segunda se calcula una similitud similar teniendo en cuenta los términos asociados a los distintos aspectos de la parte “inconsciente”.

El siguiente proceso debe realizarse cada día para cada usuario u para obtener / actualizar los términos del usuario asociados a la parte “consciente”:

Se identifican dos conjuntos de documentos sobre los que el usuario u ha efectuado algún tipo de realimentación: $R_u(+)$, es el conjunto de documentos para los que se ha suministrado realimentación positiva, $R_u(-)$, es el conjunto de documentos para los que no se ha suministrado realimentación. El conjunto de todos los documentos para los que ha suministrado algún tipo de realimentación se denota por R_u .

Se define el valor de acceso para el término t en el documento d para el usuario u de la siguiente manera:

$$a_{tdu} = \begin{cases} P \cdot (T \cdot \text{titulo}_{td} + \text{cuerpo}_{td}) & \text{si } d \in R_u(+) \\ -N \cdot (T \cdot \text{titulo}_{td} + \text{cuerpo}_{td}) & \text{si } d \in R_u(-) \end{cases} \quad (1)$$

donde: titulo_{td} es la frecuencia del término t en el título del documento d , cuerpo_{td} es la frecuencia del término t en el cuerpo del documento d , P es el peso para la realimentación positiva, N es el peso para la no-realimentación y T es el peso del título. Los valores elegidos son: $P = 0.9$, $N = 0.9$ y $T = 2$.

De esta manera, un término tendrá valor de acceso alto si aparece con mucha frecuencia en títulos, sobre todo, y en cuerpos de documentos con realimentación positiva, y tendrá valor bajo si aparece en documentos no realimentados. Este valor calcula la representatividad de los términos según la realimentación del usuario.

Se define el porcentaje de actualización de un término t para un usuario u de la siguiente manera:

$$p_{tu} = \frac{\sum_{d \in R_u} a_{tdu}}{\max(|\sum_{d \in R_u} a_{tdu}|)} \quad (2)$$

De esta manera, se suman los valores de acceso de todos los términos y se normalizan de tal forma que el término con mayor porcentaje de actualización tenga valor 1, y el resto tengan valores entre 0 y 1.

El nuevo valor de interés para el término t para el usuario u se obtiene según la siguiente fórmula:

$$N_{tu} = \begin{cases} I_{tu} + ((1 - I_{tu}) \cdot V \cdot p_{tu}) & \text{si } p_{tu} \geq 0 \\ I_{tu} - (I_{tu} \cdot V \cdot |p_{tu}|) & \text{si } p_{tu} < 0 \end{cases} \quad (3)$$

donde V indica la velocidad de cambio del grado de interés de un término, es decir, cuanto mayor sea este valor más deprisa cambiará este valor, en el sentido de que

habrá más diferencia entre el valor inicial y el nuevo valor. El valor elegido para V es de 0.8.

3 Nuestra propuesta

Se propone un modelo o perfil de usuario navegable que representa los intereses de los usuarios desde diferentes puntos de vista. El modelo de usuario almacena 3 tipos de información: información personal, información sobre el formato de la información recibida e información específica sobre los intereses del usuario según varios sistemas de referencia que serán utilizados para realizar la personalización.

Cuando un usuario utiliza un sistema de filtrado de información define unos intereses más o menos estáticos que se almacenan en su modelo de usuario. En el caso de la personalización Web también podemos tener esta situación en la cual el usuario tenga unos intereses fijos de referencia y quiera recibir siempre información del mismo tipo. Estos intereses constituirían el modelo a largo plazo. Sin embargo, las necesidades de los usuarios varían con el tiempo como efecto directo de su interacción con la información (Bilssus&Pazzani, 2000). Por tanto, es bastante probable que los intereses de los usuarios no permanezcan estáticos sino que se produzcan oscilaciones transitorias a corto plazo en torno a esa referencia inicial. Estos intereses asociados a estas combinaciones constituirían el modelo a corto plazo. Nuestra propuesta consiste en la combinación de ambos modelos para la representación de las necesidades de información de los usuarios.

Para modelar los intereses a largo plazo se utilizan dos marcos de referencia: uno basado en el sistema de clasificación dependiente del dominio y otro basado en el contenido de los documentos.

Un sistema básico de referencia es el sistema de clasificación dependiente del dominio (p.ej.: en un periódico electrónico, este sistema estaría basado en las secciones del periódico). Este sistema está compuesto por un conjunto de categorías de primer nivel que representan distintos tipos de información (p.ej.: en periódicos electrónicos serían ejemplos de categorías: nacional, internacional, deportes, etc.). Cada documento Web pertenece a una categoría de dicho sistema de clasificación. Estas categorías se almacenan en una matriz donde las filas son las categorías y las columnas son los usuarios. Los usuarios pueden asignar un peso a cada categoría para indicar su interés en cada una de ellas (C_{cu}).

El otro sistema de referencia es un sistema basado en el contenido de los documentos. El usuario puede introducir un conjunto de palabras clave para definir sus intereses. La aparición de estas palabras en los documentos indicará que estos pueden interesar al usuario. Para cada palabra clave el usuario introduce un peso que indica su importancia. Estas palabras clave se almacenan, para cada usuario u , como un vector de pesos de términos (p_u).

Para representar los intereses a corto plazo se utilizan los términos de realimentación. Estos términos se obtienen de la realimentación del usuario sobre los documentos que recibe. Esto es, el usuario realimenta positiva o negativamente sobre los documentos recibidos y de ellos se extraen una serie de términos representativos. Esta información es manejada por el proceso de adaptación del modelo de usuario el cual

devuelve un vector de pesos de términos para cada usuario (t_u), que representa este interés del usuario. Los intereses a corto plazo suelen ser necesidades de información temporales cuyo interés para el usuario desaparece en poco tiempo, por eso su peso debe decrementarse con el paso del tiempo.

Los documentos se bajan de la Web en forma de documentos HTML. Se extraen el título, categoría, URL y texto de cada documento y se almacenan para ser procesados posteriormente. Para obtener la representación de los documentos se utilizan vectores de pesos de términos aplicando lista de parada y stemmer, y utilizando $tf \cdot idf$ para el cálculo de los pesos.

Las únicas restricciones para que se pueda aplicar el modelo propuesto son que exista información textual asociada a los documentos Web y que se disponga de una clasificación dependiente del dominio para clasificar dichos documentos

4 Selección de contenidos

La selección de contenidos se refiere a la elección entre todos los documentos de entrada de aquellos que son más relevantes para un usuario dado, según su perfil. Una vez fijadas la representación de los documentos y la representación del modelo de usuario se puede establecer qué documentos son más adecuados para cada usuario.

Puesto que tenemos distintos sistemas de referencia en el modelo de usuario se va a indicar como se obtiene la selección de contenidos con respecto a cada uno de ellos y posteriormente se explorarán las distintas combinaciones. Para establecer las combinaciones se tendrá en cuenta la relevancia obtenida con cada sistema de referencia y el peso relativo utilizado en cada una de las combinaciones. En cualquier caso el resultado final será un ranking de los documentos ordenado por relevancia.

4.1 Selección con respecto al modelo a largo plazo

Como cada documento Web tiene una categoría pre-asignada, la selección respecto este marco de referencia es inmediata. Se obtiene para cada documento el valor asignado a la sección a la que pertenece, dentro del modelo de usuario. La relevancia entre un documento d , que pertenece a una categoría c , y un modelo de usuario u es directamente el valor asignado a la categoría c por el usuario u :

$$r_{du}^c = C_{cu} \quad (4)$$

La relevancia entre un documento d y las palabras clave de un modelo de usuario u es calculada mediante la fórmula de la similitud del coseno del modelo del espacio vectorial (Salton, 1989):

$$r_{du}^p = \text{sim}(d_d, p_u) \quad (5)$$

Cuando todos los documentos han sido ordenados con respecto a las diferentes fuentes de relevancia, los resultados son integrados utilizando la combinación particular que es asignada a cada uno de los sistemas de referencia. Por tanto, la relevancia

total entre un documento d y un modelo de usuario u se calcula con la siguiente fórmula:

$$r_{du}^i = \frac{\alpha r_{du}^c + \beta r_{du}^p}{\alpha + \beta} \quad (6)$$

donde las letras griegas α y β muestran la importancia asignada a cada uno de los sistemas de referencia (α , para categorías y β , para palabras clave). Para que esta combinación sea significativa, la relevancia obtenida a partir de cada sistema de referencia debe ser normalizada con respecto a los mejores resultados para la colección de documentos que se esté utilizando.

5 Adaptación del modelo de usuario

La adaptación del modelo de usuario requiere de la obtención / actualización de un modelo a corto plazo del usuario a partir de la realimentación del usuario. Este modelo puede ser utilizado para mejorar la selección en el sistema de personalización.

5.1 Obtención del modelo a corto plazo

El modelo a corto plazo se obtiene como resultado del proceso de adaptación del modelo del usuario. El usuario recibe el documento Web resultado que contiene un resumen generado automáticamente (Acero et al. 2001) para cada uno de los 10 documentos Web que el sistema ha encontrado más relevantes según su modelo de usuario. Con esta información el usuario puede interactuar con el sistema mediante la realimentación, positiva, negativa o nula, sobre cada uno de los elementos de información recibidos. Los términos de realimentación del modelo a corto plazo se obtienen a partir de las noticias sobre las que se ha efectuado algún tipo de realimentación: positiva o negativa.

En primer lugar, puesto que estos términos representan un interés a corto plazo del usuario, se utiliza un algoritmo para decrementar su valor con el paso del tiempo: cada día los nuevos pesos se obtienen restando 0.1 del valor del día anterior. Los términos que llegan a tener peso menor o igual a 0 son eliminados del modelo.

Para la selección / actualización de los nuevos términos de realimentación se pre-procesan primero todos los documentos de la misma forma a como se realiza en el proceso de selección, esto es, se filtran todas las palabras vacías con una lista de parada y después se aplica el extractor de raíces adaptado al español. Es decir, que el punto de partida del proceso de adaptación son los términos de la representación de los documentos, con su frecuencia (tf) asociada.

A continuación se aplica el algoritmo de (Nakashima&Nakamura, 1997) para obtener los términos de realimentación, esto es, el proceso de realimentación de los términos de su modelo “consciente” es utilizado para obtener el modelo a corto plazo en nuestra propuesta. Además, el conjunto $R_u(-)$ son los documentos sobre los que el usuario realimenta negativamente.

El resultado final de este proceso es un conjunto de términos ordenados según su nuevo valor de interés. Se selecciona un subconjunto de ellos, los 10 más relevantes, para obtener / actualizar los términos de realimentación del modelo a corto plazo.

5.2 Selección con respecto al modelo a corto plazo

La relevancia entre un documento d y un modelo de usuario u a corto plazo es calculada de manera similar a la calculada con respecto a las palabras clave del modelo a largo plazo pero utilizando el vector de pesos de términos obtenido en el proceso de adaptación del modelo de usuario:

$$r_{du}^c = r_{du}^t = \text{sim}(d_d, t_u) \quad (7)$$

5.3 Selección respecto a la combinación de largo y corto plazo

Cuando todos los documentos han sido ordenados con respecto a las diferentes fuentes de relevancia, los resultados son integrados utilizando la combinación particular que es asignada a cada uno de los sistemas de referencia. Por tanto, la relevancia total entre un documento d y un modelo de usuario u se calcula con la siguiente fórmula:

$$r_{du} = \frac{\chi r_{du}^c + \delta r_{du}^p + \varepsilon r_{du}^t}{\chi + \delta + \varepsilon} \quad (8)$$

donde las letras griegas χ , δ , y ε muestran la importancia asignada a cada uno de los sistemas de referencia (χ , para categorías, δ , para palabras clave, ε , para términos de realimentación). Para que esta combinación sea significativa, la relevancia obtenida a partir de cada sistema de referencia debe ser normalizada con respecto a los mejores resultados para la colección de documentos que se esté utilizando.

6 Evaluación

Para evaluar los experimentos se dispone de 11 usuarios y de las noticias correspondientes a 5 días consecutivos (de lunes a viernes) de la edición electrónica del periódico ABC. Estos días corresponden al período entre el 6 y el 10 de mayo de 2002. El número de noticias por día es 128, 104, 87, 98 y 102, respectivamente.

Para poder realizar las evaluaciones es necesario conocer los juicios respecto a las noticias relevantes y no relevantes para cada usuario durante cada uno de los 5 días del experimento. Para obtener estos juicios de relevancia los usuarios tuvieron que mirar todas las noticias de todos los días y juzgar si eran de su interés o no. Este interés no se limitaba sólo a si la noticia era relevante respecto su perfil inicial, sino que también captura como relevantes noticias que no estaban relacionadas con su perfil

inicial pero si le interesaban al usuario. Es de esperar que esta información de relevancia se vaya introduciendo en el perfil a través de la realimentación.

6.1 Métricas

Al disponer en nuestro marco de trabajo de una relevancia binaria por parte de los usuarios y de un ranking de noticias seleccionadas proporcionado por el sistema, se ha decidido utilizar como métrica la precisión normalizada (Salton, 1989; Mizarro, 2001). Además en el caso de valores de relevancia iguales para posiciones consecutivas del ranking se ha tomado como *ranking* de todas esas posiciones el ranking medio. Este ajuste evita el problema de la aleatoriedad de tener una noticia antes que otra en el *ranking* aunque tengan la misma relevancia.

La precisión (P) normalizada se calcula a partir de la siguiente fórmula:

$$P = 1 - \frac{\sum_{i=1}^{REL} \log RANK_i - \sum_{i=1}^{REL} \log i}{\log N! / ((N - REL)! REL!)} \quad (9)$$

donde REL es el número de documentos relevantes, RANK_i es el ranking del documento i y N el número total de documentos.

6.2 Significancia estadística

Se consideran los datos como estadísticamente significativos si pasan el *sign-test*, con muestras apareadas, a un nivel de significación del 5% ($p \leq 0.05$). Esta decisión está basada en el hecho de que no se establece ninguna suposición sobre la distribución de los datos y que además, debido a los diferentes procesos de normalización realizados, es más conveniente tener en cuenta los valores relativos en lugar de los valores absolutos (Salton, 1989).

6.3 Experimentos

Se han realizado los siguientes experimentos para comprobar la validez del modelo propuesto. En los distintos experimentos se combinan las distintas posibilidades del largo plazo: categorías sólo, palabras clave sólo, y categorías y palabras clave, combinadas o no con el modelo a corto plazo. Esto supone jugar con los valores de los parámetros χ , δ y ε de la fórmula (8).

6.3.1 Experimento 1

En este experimento se va a comparar el modelo a largo plazo, utilizando sólo las palabras clave L(Pc) ($\chi=0$, $\delta=1$, $\varepsilon=0$), con el modelo a corto plazo C ($\chi=0$, $\delta=0$, $\varepsilon=1$) y con la combinación de ambos modelos L(Pc)C ($\chi=0$, $\delta=1$, $\varepsilon=1$).

Experimento 1	P	Experimento 2	P	Experimento 3	P
L(Pc)C > L(Pc)	26.9	L(Ca)C > L(Ca)	26.9	L(Ca,Pc)C > L(Ca,Pc)	8.5
L(Pc)C > C	16.1	L(Ca)C > C	29.0	L(Ca,Pc)C > C	32.9
C > L(Pc)	8.5	L(Ca) > C	10.9	L(Ca,Pc) > C	22.4

Tabla 1: Incrementos relativos de precisión normalizada entre las distintas combinaciones de L(Pc) y C, L(Ca) y C, y L(Ca,Pc) y C

El único resultado estadísticamente significativo (Tabla 1) que se obtiene es que $L(Pc)C > L(Pc)$, es decir, que es mejor combinar el largo y el corto plazo que utilizar sólo el largo. También hay mejora relativa de la combinación respecto al modelo a corto plazo, pero no es estadísticamente significativa. El modelo a corto plazo supera al modelo a largo plazo, pero no significativamente.

6.3.2 Experimento 2

En este experimento se va a comparar el modelo a largo plazo, utilizando sólo las categorías L(Ca) ($\chi=1, \delta=0, \varepsilon=0$), con el modelo a corto plazo C ($\chi=0, \delta=0, \varepsilon=1$) y con la combinación de ambos modelos L(Ca)C ($\chi=1, \delta=0, \varepsilon=1$).

Los resultados estadísticamente significativos (Tabla 1) son que $L(Ca)C > C$ y $L(Ca)C > L(Ca)$, es decir, que la combinación es siempre mejor que los modelos por separado. El modelo a largo plazo supera al modelo a corto plazo, pero no significativamente.

6.3.3 Experimento 3

En este experimento se va a comparar el modelo a largo plazo, utilizando las categorías y las palabras clave L(Ca,Pc) ($\chi=1, \delta=1, \varepsilon=0$), con el modelo a corto plazo C ($\chi=0, \delta=0, \varepsilon=1$) y con la combinación de ambos modelos L(Ca,Pc)C ($\chi=1, \delta=1, \varepsilon=1$).

Todos los resultados son estadísticamente significativos (Tabla 1), es decir, la combinación supera a ambos modelos en solitario y el largo plazo es mejor que el corto.

6.3.4 Experimento 4

Este último experimento compara las mejores combinaciones de los experimentos anteriores, largo y corto plazo juntos, cuando se utilizan sólo las palabras clave para el largo plazo L(Pc)C ($\chi=0, \delta=1, \varepsilon=1$), sólo las categorías L(Ca)C ($\chi=1, \delta=0, \varepsilon=1$) y la combinación de ambos intereses L(Ca,Pc)C ($\chi=1, \delta=1, \varepsilon=1$).

	Precisión
L(Ca,Pc)C > L(Ca)C	2.9
L(Ca,Pc)C > L(Pc)C	12.6
L(Ca)C > L(Pc)C	11.2

Tabla 2: Incrementos relativos de precisión normalizada entre las distintas combinaciones de L y C juntos

Todos los resultados son estadísticamente significativos (Tabla 2), es decir, que la combinación de largo y corto plazo, cuando se utilizan las categorías y los términos en el modelo a largo plazo, es mejor que la combinación cuando se utilizan sólo categorías o sólo términos. Entre las categorías y los términos la mejor opción es la de utilizar las categorías para el largo plazo.

7 Conclusiones

En el artículo se ha presentado la mejora que supone para la personalización la inclusión de un proceso de adaptación del modelo de usuario, ya que la selección que se obtiene combinando los perfiles a largo y corto plazo del usuario mejora la que se obtiene con la utilización del modelo a largo plazo en solitario.

Los resultados muestran que la utilización de la combinación de un modelo de usuario a largo plazo, formado por categorías y términos, con un modelo a corto plazo, mejora la adaptación del usuario puesto que se incrementan los valores de precisión normalizada.

Las únicas restricciones para que se pueda aplicar el modelo propuesto son que exista información textual asociada a los documentos Web y que se disponga de una clasificación dependiente del dominio para clasificar dichos documentos.

Bibliografía

1. Acero, I., Alcojor, M., Díaz, A., Gómez, J.M., Maña, M., 2001. "Generación automática de resúmenes personalizados". *Procesamiento del Lenguaje Natural*, 27 (2001), pp. 281-290.
2. Billsus, D. & Pazzani, M.J., 2000. "User Modeling for Adaptive News Access", *User Modeling and User-Adapted Interaction Journal* 10(2-3), pp. 147-180.
3. Chiu, B. & Webb, G., 1998. "Using decision trees for agent modeling: improving prediction performance", *User Modeling and User-Adapted Interaction* (8), pp. 131-152.
4. Mizzaro, S., 2001. "A New Measure Of Retrieval Effectiveness (or: What's Wrong With Precision And Recall)". *International Workshop on Information Retrieval (IR'2001)*, Infotech Oulu, pp. 43-52.
5. Nakashima, T. & Nakamura, R., 1997. "Information Filtering for the Newspaper". *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, August 1997. Victoria, B.C., Canada.
6. Salton, G., 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley Publishing, Reading, Massachusetts, 1989.