Dynamic User Modeling in a System for Personalization of Web Contents*

Alberto Díaz[1] and Pablo Gervás[2]

[1] CES Felipe II – Universidad Complutense de Madrid
C/ Capitán 39, 28300 Aranjuez, Madrid
adiaz@cesfelipesegundo.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

Abstract. This paper presents a system for personalization of web contents based on a user model that stores long term and short term interests. Long term interests are modeled through the selection of categories and keywords for which the user need information. However, user needs change over time as a result of his interaction with received information. For this reason, the user model must be capable of adapting to those shifts in interest. In our case, this adaptation or dynamic modeling is performed by a short term model obtained from user provided feedback. The experiments that have been carried out determine that the combined use of long and short term models performs best when both categories and keywords are used for the long term model.

1 Introduction

Web content appears in many forms over different domains of application, but in most cases the form of presentation is the same for all users. The contents are static in the sense that they are not adapted to each user. Content personalization is a technique that tries to avoid information overload through the adaptation of web contents to each type of user.

A personalization system is based on 3 main functionalities: content selection, user model adaptation, and content generation. For these functionalities to be carried out in a personalized manner, they must be based on information related to the user that must be reflected in his user profile or user model (Mizarro&Tasso, 2002).

Content selection refers to the choice of the particular subset of all available documents that will be more relevant for a given user, as represented in his user profile or model. In order to effect this choice one must have a representation of the documents, a representation of the user profile, and a similarity function that computes the level of adequacy of one to the other.

* This research has been partially funded by the Ministerio de Ciencia y Tecnología (TIC2002-01961).

R. Conejo et al. (Eds.): CAEPIA-TTIA 2003, LNAI 3040, pp. 281–290, 2004.

© Springer-Verlag Berlin Heidelberg 2004

User model adaptation is necessary because user needs change over time as a result of his interaction with information (Billsus&Pazzani, 2000). For this reason the user model must be capable of adapting to those interest changes, it must be dynamic. This adaptation is built upon the interaction of the user with the system, which provides the feedback information used to evolve the profile.

In our case, content generation involves generating a new result web document that contains, for each selected document, its title, its relevance as computed by the system, a summary, and a link to the full document.

In this paper we focus on user model adaptation and the various possible combinations of modeling alternatives for this process. The aim is to identify which is the best way of carrying out the user model adaptation process to improve content selection.

2 Available Methods and Techniques

Existing literature provides different techniques for defining user interests: keywords, stereotypes, semantic networks, neural networks, etc. A particular set of proposals (Chiu&Webb, 1998; Billsus&Pazzani, 2000) model users by combining long term and short term interests: the short term model represents the most recent user preferences and the long term model represents those expressed over a longer period of time. To determine whether a document is relevant for a given user the short term user model is used wherever it can provide a satisfactory answer. The long term model is used only as a backup solution for cases in which the short term model fails to provide an answer.

The representation of the text content of the documents is usually achieved by means of techniques based on term weight vectors (Salton, 1989). The vector associated with a document can be obtained by eliminating the words contained in a stop list and extracting the stems of the remaining words by means of a stemmer. Weights are usually calculated by means of the tf · idf formula, based on frequency of occurrence of terms (Salton, 1989).

Various classification algorithms are available for carrying out content selection depending on the particular representation chosen for user models and documents: cosine formula, rules associated to stereotypes, neural networks, nearest neighbour, naive Bayes classifier, etc.

The feedback techniques needed to achieve a dynamic modeling of the user are based on feedback given by the user with respect to the information elements selected according to his profile. The information obtained in this way can be used to update accordingly the user models in representation had been chosen: term weights, semantic networks, rules associated to stereotypes, etc.

In particular, a system based on intelligent agents is applied in (Nakashima &Nakamura97) to a digital newspaper. The user model stores "conscious" information about the user as terms with an associated weight and "unconscious" information as terms associated with aspects such as age, sex, occupation, marital status, city, etc. A selection is computed using a combination of both types of information. For the first one, relevance is based on "conscious" user terms appearing in the document, with additional relevance accorded to terms appearing in the title. For the second, a similar computation is applied over the terms associated with each aspect of the "unconscious" part of the model.

The next process must be carried out each day for each user u to obtain / update the user terms associated to the "conscious" part of the model:

Two set of documents are distinguished according to the feedback provided by the user: $R_u(+)$, is the set of documents for which the user has provided positive feedback, $R_u(-)$, is the set of documents for which no feedback has been provided. The set of all documents is R_u.

The access value for term t in document d for user u is defined as:

$$a_{tdu} = \begin{cases} P \cdot (T \cdot title_{td} + body_{td}) & si\, d \in R_u(+) \\ -N \cdot (T \cdot title_{td} + body_{td}) & si\, d \in R_u(-) \end{cases} \tag{1}$$

where $title_{td}$ is the frequency of appearance of term t in the title of document d, $body_{td}$ is the frequency of appearance of term t in the body of document d, P is the weight applied to positive feedback, N is the weight for no-feedback and T is the weight applied to the title. The particular values chosen are: P = 0.9, N = 0.9 and T = 2.

In this way, a term will have a high access value if it appears frequently in titles and bodies of documents with positive feedback, and it will have a low access value if it appears in documents for which no feedback is provided. This value computes the representativity of terms as a function of user feedback.

The update rate of a term t for a user u is computed as:

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

In this way, the access values for all the terms are added together and normalised to ensure that the term with highest update rate has value 1, and the rest take values between 0 and 1.

The new interest value for term t for user u is obtained with the following formula:

$$N_{tu} = \begin{cases} O_{tu} + ((1 - O_{tu}) \cdot S \cdot p_{tu}) & si\, p_{tu} \geq 0 \\ O_{tu} - (O_{tu} \cdot S \cdot |p_{tu}|) & si\, p_{tu} < 0 \end{cases} \tag{3}$$

where O_{tu} indicates the old interest value for term t for user u and S indicates the speed of change of the degree of interest of a term. The higher the value of S, the faster the degree of interest will change, in the sense that there will be more difference between its initial value and the new value. The value chosen for S is 0.8.

3 Our Proposal

We propose a browsable user model or user profile that represents user interests from three different points of view (Amato&Straccia, 1999). The user model stores three types of information: personal information, information concerning the format in which information is to be received, and specific information about user interests according to various reference systems that will be used to carry out the personalization.

When a user accesses an information filtering system, he defines a more or less static set of interests that are stored in his user profile. For Web personalization we can have a similar situation in which the user has a set of fixed reference interests about which he wants to receive information on a regular basis. These interests will make up the long term model. However, user needs change over time as a result of the interaction with information (Bilssus&Pazzani, 2000). For this reason, it is probable that the interests of a user will not remain static but will in the short term suffer temporary oscillations around this initial reference. The interests associated with these oscillations will constitute the short term model. Our proposal is based on the combination of both models to represent user's information needs.

Long term user interests are modelled with respect to two reference frameworks: one based on a domain specific system of classification, and another based on the content of the documents.

A basic reference system is the classification system specific to the particular domain under consideration - for instance, in a digital newspaper, this system will be based on the set of sections used by the newspaper -. This system is composed of a set of first level categories that represent different types of information - for instance, examples of sections of digital newspapers would be: national, international, sport, etc. Each web document belongs to a category of that classification system. Information concerning these categories is stored as a matrix where rows correspond to categories and columns correspond to users. Users may assign a weight to each category to indicate their interest in them (C_{cu}).

The other system of reference is based on the content of documents. The user can enter a number of keywords to characterise his interests. The appearance of these keywords in the documents will be taken to indicate that the document may be interesting to the user. For each keyword the user introduces a weight that indicates its importance to him. These keywords are stored, for each user u, as a term weight vector (k_u).

Short term interests are represented by means of feedback terms. These terms are obtained from user provided feedback over the documents he receives. That is, the user provides positive or negative feedback over the documents he receives, and a set of representative terms is extracted from them. This information is handled by the user model adaptation process, which returns a term weight vector (t_u) for each user. This term weight vector is taken to represent the current short term interests of that user. Short terms interests tend to correspond to temporary information needs whose interest to the user wanes after a short period of time. Therefore their weight must be progressively decreased over time.

Documents are downloaded from the web as HTML documents. For each document, title, category, URL and text are extracted and stored for ulterior processing. Term weight vector representations are obtained by application of stop lists, stemmer, and the tf · idf formula for computing actual weights.

The only restrictions that must be fulfilled by a domain for the proposed model to be applicable are that there exist textual information associated with web documents and that a domain specific classification exist to classify the documents.

4 Content Selection

Content selection refers to the choice of those among the available documents that are particularly relevant for a user, according to his profile. Once particular representations have been fixed for documents and user model, it becomes feasible to establish which documents are more adequate for each user.

Since we have different reference frameworks in the user model we will indicate how content selection is performed with respect to each one of them, and later we will explore different possible combinations of the resulting selections. Combinations will be based on the relevance obtained for each document within each particular reference framework, and the relative weight used for each reference framework in a particular combination. For all combinations, the final result is a ranking of the set of documents according to the computed overall relevance.

4.1 Selection with Respect to the Long Term Model

As each web document has a preassigned category, selection with respect to this reference framework is immediate. Each document is assigned the weight associated with the corresponding category in the particular user model. The relevance between a document d, belonging to a category c, and a user model u is directly the value assigned to category c by user u:

$$r_{du}^{c} = C_{cu} \tag{4}$$

The relevance between a document d and the keywords of a user model is computed using the cosine formula for similarity within the vector space model (Salton, 1989):

$$r_{du}^{k} = sim(d_d, k_u) \tag{5}$$

When all documents have been ordered with respect to the various reference frameworks, the results are integrated using a particular combination of reference frameworks. Therefore, the total relevance between a document d and a user model u is computed with the following foyrmula:

$$r_{du}^{l} = \frac{\alpha r_{du}^{c} + \beta r_{du}^{k}}{\alpha + \beta} \tag{6}$$

where Greek letters α and β represent the importance assigned to each reference framework (α, for categories and β, for keywords). For this combination to be significant, relevance obtained for each framework must be normalised with respect to the best results for the document collection under consideration.

5 User Model Adaptation

Adaptation of the user model involves obtaining / updating a short term model of the user from the feedback information provided by the user. This model can be used to improve the process of selection in the personalization system.

5.1 Obtaining the Short Term Model

The short term model is obtained as a result of the process of adaptation of the user model. The user receives a web document that contains an automatically generated summary (Acero et al. 2001) for each of the 10 web documents that the system has found more relevant according to his user profile. With respect to this information the user may interact with the system by giving positive or negative feedback - refraining from providing feedback is interpreted as a contribution as well, taken to imply indifference - for each of the information elements that he has received. The feedback terms of the short term model are obtained from the news items for which either positive or negative feedback has been provided.

Because these terms represent an interest of the user over a short period of time, an algorithm is used to decrement their value over time: each day the starting value of the new weights is obtained by subtracting 0.1 from the previous day's value. Terms that reach a weight less or equal to 0 are eliminated from the model.

To select / update the new feedback terms all documents are preprocessed in the same way as was done for the selection process: stop lists and stemmer are applied. The starting point for the adaptation process are the terms of the representation of the documents, with their associated frequency (tf).

The algorithm in (Nakashima&Nakamura, 1997) is then applied to obtain the feedback terms. The feedback process for the "conscious" part of their model is used to obtain the short term model of our proposal. As an innovation, the set $R_u(-)$ is taken to be the set of documents for which the user has provided negative feedback. Also the set R_u is now the set of all documents for which feedback of some kind has been provided.

The final result of this process is a set of terms ordered according to their new interest value. A subset of them is selected - the 10 most relevant ones - to obtain / update the feedback terms of the short term model.

5.2 Selection with Respect to the Short Term Model

Relevance between a document d and a short term user model u is computed in the same way used for the keywords of the long term model, but using the term weight vector obtained in the process of adaptation of the user model:

$$r_{du}^{s} = r_{du}^{t} = sim(d_d, t_u) \tag{7}$$

5.3 Selection with Respect to the Combined Long Term - Short Term Model

When all documents have been ordered with respect to the different sources of relevance, the results are integrated using a particular combination of reference frameworks. Therefore, the total relevance between a document d and a user model u is computed with the following formula:

$$r_{du} = \frac{\chi r_{du}^{c} + \delta r_{du}^{k} + \varepsilon r_{du}^{t}}{\chi + \delta + \varepsilon} \tag{8}$$

where Greek letters χ, δ, and ε represent the importance assigned to each of the reference frameworks -χ, for categories, δ, for keywords, ε, for feedback terms. For this combination to be significant, the relevance obtained from each reference framework must be normalised with respect to the best results over the document collection being used.

6 Evaluation

As an example of web documents for experimentation we have chosen the web pages of the digital edition of a Spanish newspaper[1]. Experiments are evaluated over data collected for 11 users and the news items corresponding to 5 consecutive days - Monday to Friday - of the digital edition of the ABC Spanish newspaper. These days correspond to the period 6th -10th May 2002. The number of news items per day is respectively 128, 104, 87, 98 and 102.

To carry out the evaluation, judgements from the user are required as to which news items are relevant or not for each of the days of the experiment. To obtain these judgements users were requested to check the complete set of news items for each day, stating for each one whether it was considered interesting or not. Users were explicitly asked not to confine their judgements on interest to relevance with respect to the initial user profiles they had constructed on first accessing the system, but rather to include any news items that they found interesting on discovery, regardless of their similarity with respect to their initial description of their interest. It is hoped that enough information to cover these rogue items will be captured automatically and progressively by the system through the feedback adaptation process.

6.1 Metrics

Since our experimental set up combines a binary relevance judgement from the users and a ranking of news items provided by the system, it was decided to use normalised precision (Salton, 1989; Mizarro, 2001) as our evaluation metric. In addition, with respect to equal relevance values for consecutive positions of the ranking, the average ranking of the whole set of conflicting positions has been taken as ranking for each

[1] This provides a consistent format, which simplifies systematic processing.

and all of them. This adjustment avoids the problem of ordering items at random within the ranking when they have equal relevance.

Normalised precision is computed using the following formula:

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

where REL is the number of relevant documents, $\mathrm{RANK_i}$ is the ranking of document i, and N is the total number of documents.

6.2 Statistical Significance

Data are considered statistically significant if they pass the *sign-test*, with paired samples, at a level of significance of 5% ($p \leq 0.05$). This decision is based on the fact that no specific assumption is made concerning the distribution of data, and that due to the different normalisation processes carried out, it is more convenient to consider relative values instead of absolute values (Salton, 1989).

6.3 Experiments

The following experiments have been carried out to check the validity of the proposed model. Each experiment combines different possibilities for long term modeling - only categories, only keywords, and categories and keywords together - either acting on their own or in combination with the short term model. This implies giving different values to the parameters χ, δ and ε of formula (8).

6.3.1 Experiment 1

This experiment compares the long term model using only keywords L(Ke) (χ=0, δ=1, ε=0), with the short term model S (χ=0, δ=0, ε=1) and with a combination of both models L(Ke)S (χ=0, δ=1, ε=1).

Table 1. Relative increments in normalised precision between different combinations of L(Ke) and S, L(Ca) and S, and L(Ca,Ke) and S.

Experiment 1	Pr	Experiment 2	Pr	Experiment 3	Pr
L(Ke)S > L(Ke)	26.9	L(Ca)S > L(Ca)	26.9	L(Ca,Ke)S > L(Ca,Ke)	8.5
L(Ke)S > S	16.1	L(Ca)S > S	29.0	L(Ca,Ke)S > S	32.9
S > L(Ke)	8.5	L(Ca) > S	10.9	L(Ca,Ke) > S	22.4

The only statistically significant result (Table 1) is that L(Ke)S > L(Ke). This means that combining the long and short term models, is better than using only the long term model. There is also a relative improvement of the combination with respect to the short term model, but it is not statistically significant. The short term model performs better than the long term model, but again not significantly.

6.3.2 Experiment 2

This experiment compares the long term model using only categories L(Ca) (χ=1, δ=0, ε=0), with the short term model S (χ=0, δ=0, ε=1) and with the combination of both models L(Ca)S (χ=1, δ=0, ε=1).

The statistically significant results (Table 1) are that L(Ca)S > S and L(Ca)S > L(Ca). This means that the combination is always better than using each model separately. The long term model performs better than the short term, but without significance.

6.3.3 Experiment 3

This experiment compares the long term model using both categories and keywords L(Ca,Ke) (χ=1, δ=1, ε=0), with the short term model S (χ=0, δ=0, ε=1) and with the combination of both models L(Ca,Ke)S (χ=1, δ=1, ε=1).

All results are statistically significant (Table 1). This means that the combination performs better than either model on its own, and the long term model is better than the short term model.

6.3.4 Experiment 4

This experiment compares the best performing combinations of previous experiments - long and short term models used together - when the long term model is built using only keywords L(Ke)S (χ=0, δ=1, ε=1), only categories L(Ca)S (χ=1, δ=0, ε=1) and a combination of both L(Ca,Ke)S (χ=1, δ=1, ε=1).

Table 2. Relative increments in normalised precision between different combinations of L and S together.

	Pr
L(Ca,Ke)S > L(Ca)S	2.9
L(Ca,Ke)S > L(Ke)S	12.6
L(Ca)S > L(Ke)S	11.2

All results are statistically significant (Table 2). This means that the long term / short term combination that uses categories and keywords in the long term model is better than the combinations that use either only categories or only keywords for the long term model. Using categories only for the long term model is better than using keywords only.

7 Conclusions

This paper presents the improvement in personalisation achieved by the inclusion of a process of user model adaptation, due to the fact that the selection that is obtained by combining the long term and short term profiles performs better than the one obtained by using the long term model on its own.

The results show that using a combination of a long term model based on categories and keywords, together with a short term model, improves the adaptation to the user because values of normalised precision increase.

The only restrictions for this model to be applicable to a particular domain are that there exist textual information associated to each web document, and that a domain dependent classification be available to classify the documents to be considered.

References

1. Amato, G. & Straccia, U., 1999. "User Profile Modeling and Applications to Digital Libraries". Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99), Springer-Verlab LNCS 1696, pp. 184-197.
2. 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.
3. 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.
4. Chiu, B. & Webb, G., 1998. "Using decision trees for agent modeling: improving prediction performance", User Modeling and User-Adapted Interaction (8), pp. 131-152.
5. 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.
6. Mizarro, S. & Tasso, C., 2002. "Ephemeral and Persistent Personalization in Adaptive Information Access to Scholarly Publications on the Web". 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, Málaga, España, Mayo 2002.
7. 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.
8. Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley Publishing, Reading, Massachusets, 1989.