NBM and WNBM: Algorithms and Evaluation for personalizing information retrieval in METIORE

David Bueno, Ricardo Conejo, Amos David, Cristina Carmona

NBM and WNBM: Algorithms and Evaluation for personalizing information retrieval in METIORE

David Bueno[1], Ricardo Conejo[1], Amos A. David[2], Cristina Carmona[1]

[1] Department of Languages and Computer Science, University of Málaga,
29071, Málaga, Spain.
bueno@lcc.uma.es
[2] LORIA, BP 239, 54506 Vandoeuvre, France
adavid@loria.fr

Abstract. The current Information Retrieval Systems return hundreds or thousands of documents in response to a query. Users consider only the first 20 or 30, but documents are often sorted according to the query and the results are perhaps not relevant to the users' needs. This situation is even more problematic when non expert users have difficulties in expressing their information requirements. One solution to this problem could be the use of a user model which would complement the query in order to find the best solutions for the user. Most personalized retrieval systems have a single user model for each user which works on the basis that this user will have always the same information needs. Our proposal however is a personalization method based on the current objective of the user. To this end, we have developed two probabilistic algorithms NBM and WNBM which support different kinds of multi parameter databases and in this paper we present new experiments with these algorithms in METIORE[1] to validate our proposal.

1. Introduction

The Information Retrieval System (IRS) has arisen from the necessity to organize information contained in libraries in order to locate the documents they contain. Currently everyone who connects to the Web uses systems such as *Altavista*, *Google* or *Yahoo*. The problem with these systems is that they sometimes return hundreds or thousands documents in response to a query, and yet the user will consider only the first 20 or 30. Sorting is done according to the query or even according to economic criteria. A company or an individual can pay to be on the top of a list for queries containing some keywords, but this does not necessarily mean that their content will interest the users. Also the novice user may experience difficulties in expressing his/her information needs in the query language of the system.

One possible solution to the information retrieval problem comes from the application of a recommender system which can offer a personalized response according to the preferences and needs of each user. Very good reviews can be found

[1]METIORE is available in: http://www.lcc.uma.es/metiorew/

in [1] [2]. The most important aspects to take into consideration will be related with sorting the query results.

In this article we present results related to the study of personalization in information retrieval systems. We offer a new vision of personalization oriented to the user objectives as opposed to other visions centered on query refinement. Most of the systems that offer some kind of personalization service have a general vision of the user's interests and have only one user model. Consequently, when the user is interested in a specific objective, the system will offer him only documents related to that objective. We believe however that the same user can have various and apparently unrelated interests.

In this paper, we present probabilistic algorithms which allow personalization based on user objectives relating to different documental/multimedia databases. The algorithms take into account multiple parameters which can be inferred from document descriptions.

METIORE is an information research system which incorporates all the proposals outlined here. It has been evaluated in different environments (conventional applications/ the Web) and has also been applied to different databases.

2. Personalization Centered on Objectives

2.1 Objectives

In this section we present the main characteristics of our proposal. To solve the problem of managing the different information needs of the user, we will center the personalization on the concept of Objective:

"*The user's objective is the expression in natural language representing the user's information need when he uses the information retrieval system*"

In some systems, the user expresses his/her information request using one or more queries, but these must be written using a specific language. This can limit the user's capacity of expression and in many cases the user may not know exactly either what he/she wishes to find or how to interact with the system in order to obtain the best results. Some of these difficulties are presented in [3] .

The question is: How can this concept help users to acquire a more relevant and personalized response? The objective is used as an identifier which groups a set of queries, concepts and decisions made by the user with this objective in mind. An easy example to understand the necessity of objective could be: "*Sometimes I'd like my search tool, which has my profile/model, to help me to find documents related to "user modeling or information retrieval"*. In another case perhaps *I'm looking for information on "games programming"*. In this case my search objective is different and the conclusions of the system should be managed separately according to the individual objective.

This concept is fundamental in our work and is not applied in most personalized information retrieval systems or recommender systems. The easiest way to solve this example could be to have one user model with common information relating to the user and some specific parts for different objectives. In METIORE the user can create

a new objective or select an existing one. The recommendations are associated to the selected objective and the user can explore his previous evaluations for each objective in the history.

2.2 Obtaining User Feedback

In order to acquire the user information necessary to produce his/her model, different strategies can be used. In our proposal we use explicit feedback after the evaluation of each document presented to the user. Usually systems use only positive feedback [4] to modify the model. We propose a different way of evaluating documents which takes into account both positive and negative feedback, and attempts to obtain the reasons that led the user to evaluate in this way.

2.3 Personalization Algorithms

2.3.1 The NBM Algorithm

The goal of these algorithms is to predict the evaluation of a user for each document according to his/her current objective. NBM (Naïve Bayes Metiore) [5] is based on the user's objective and not only on their queries as is the case in other related works such as [6] where they use probabilities and Markov models to predict the user's preferences. In systems which do not integrate any kind of personalization technique, the list of solutions to a query is presented in the same way for different users on the basis of the similarity between these documents and the words used the original query. NBM however is combined with typical retrieval algorithms to obtain documents related to the query, and subsequently these are further sorted, combining firstly the relevance of each document with the user model and secondly, to establish if different documents have similar relevance to the user model using their similarity with the query.

The algorithm can be applied to different evaluations. For example (*ok(2), ok(1), , wrong(1), wrong(2)*) or (*ok, known, ?, wrong)*. If the user evaluates a document as *ok* all the parameters that are part of the document (keywords, author, year,…) will be incremented/included in the model for this type of evaluation.

The difficulties of associating a document with a particular evaluation can be seen as a classification problem. The goal is to put a document in the most probable class according to the user model and then within this classification to rank it among the other similar documents.

Many studies have been done to compare different classification methods: Neural Networks, ID3 and Bayesian methods. Most of these involve complex calculations which can be a limitation to using them in systems requiring a fast response. However, the *Naïve Bayes* algorithm, in addition to its simplicity, and the assumption of independence between parameters, is able to provide similar results which are sometimes better than the other algorithms with a simpler calculus. Some of these comparisons are in[7] [8] [9] [10] and [11].

The algorithm we propose is inspired by the probabilistic theory of Bayes. For each document the algorithm gives its relevance in relation to each kind of evaluation

of the current objective. The proposal adapts the original Naïve Bayes [8] as shown in equations (1) and (2).

In Equations (1) and (2) is one of the possible classes of evaluations (*ok, known, ?, wrong* for example). $V_{i,Ji}$ is a boolean variable that is 1 if the current instance has a value J_i and P(C) is the probability of evaluating the class C. The application of this equation will take the values that describe a document (J_i), and then return the probable evaluation of the users for each type. The class with the biggest probability will be selected.

$$P'(C/V_{1,J_1},...,V_{n,J_n}) = P(C)\frac{\sum_{i=1}^{n} Q_i(C,J_i)}{n} \quad \textbf{(1)}$$

Where

$$Q_i(C,J_i) = \frac{P(V_{i,J_i}/C)}{P(V_{i,J_i})} \quad \textbf{(2)}$$

2.3.2 Advantages of NBM vs. Naïve Bayes

The classical Naïve Bayes algorithm (NB) has a problem when the number of documents evaluated is few. It makes calculates a product of probabilities, and if a document has *n* different attributes and one of them has not been evaluated for a class, the probability of belonging to this class is zero. To avoid this problem a modification has been proposed. In [12] (Naïve Bayes Cestnik- NBCestnik) this problem is solved by giving an initial probability to all the possible attributes. In [13] the NB, NBCestnik and NBM are compared to see how they sort 1000 documents according to the user preferences. For the three lists generated we calculated the correlation factor to see how similarly they sorted the documents. In Table 1 we can see that the results are indeed interesting. Our algorithm gives similar results to the others with a similarity of 93% to the original NB, but this is closer to NBCestnik. This means an improvement on NB. These results only reflect the similarity with the other algorithms but in the following section it will be possible to see the real improvement.

NB-NBCestnik	*NB-NBM*	*NBM-NBCestnik*
0,922	0,930	0,937

Table 1. Comparison of the three variants of Naïve Bayes

2.3.3 The WNBM Algorithm

The previous results show that NBM gives similar solutions to the classical Naïve Bayes classifier, but what else does it offer? Let us look at an example; In Fig. 1 an

article with different parameters is presented. To personalize, we opted to use author, *year* and *keywords*. For this document, there are 2 authors, 1 year, and 10 keywords.

Title: METIORE: A Personalized Information Retrieval System Authors: Bueno, D; David, A Publisher in: 8th International Conference on User Modeling.UM'2001 Pages: 168-177 Year: 2001 Keywords: information retrieval, user modeling, METIORE, Personalization…

Fig. 1 Example of the parameters of an article

NB will calculate the probability for a class of evaluation in a simplified way as shown in Equation **(3)**. The problem is that as there are many more keywords than dates, and the relevance of the date to the final evaluation of the document is low. For this reason it would be interesting to have some kind of weighting in order to give more importance to the parameters with few attributes such as the year.

$$P(C/V_{1,J_1},...,V_{n,J_n}) = P(C)\prod_{i=1}^{n} Qi(C,J_i) = \quad \textbf{(3)}$$

$$P(C)\big(Q(C,bueno)Q(C,david)Q(C,2001)Q(C,usermodel)Q(C,infretrieval)...\big)$$

The NBM algorithm can be easily adapted to give appropriate importance to a given parameter. This extension is called WNBM *(Weighted Naive Bayes Metiore)* and is shown in equation (4):

$$P(C/V_{1,J_1},...,V_{n,J_n}) = P(C)\sum_{p=1}^{m}\omega_p \frac{\sum_{i=1}^{n} Q_{pi}(C,J_{pi})}{N_p} \quad \textbf{(4)}$$

In equation (4) the addition with index *p* adds the partial values for the *m* parameters analyzed (in our example m=3: year, keywords and author). N_p is the total number of elements associated with the parameter p (in the example, for author N_p=2 and for the year N_p=1). The factor ω_p shows the importance for each parameter (for example 1/5 for the year, 2/5 for the author, and 2/5 for the keywords. If it is appropriate to give the same importance for all the parameters ω_p=1/m. To have a real weighted mean, the addition of all ω_p must be 1.

With WNBM the problem of NB and NBCestnik is solved, and all the attributes are treated in the same way.

2.4 Application of the Algorithms in METIORE

In this section we describe the details of the prototype where this proposal has been implemented. The METIORE system (***M**ultimedia coop**E**rative Informa**TIO**n **R**etrieval Syst**E**m*) is an environment where different personalized information retrieval systems can be implemented. Some of its characteristics are: it gets data from XML, it works internally in the same way as an object oriented database, it

offers the possibility of graphical data analysis of the queries, as well as personalization, cooperation, multilingual interface and multiplatform possibilities. The user can work using two different interfaces: the first is as an application (executable file) and the second is from the Web[2]. Both connect to the main module of METIORE that will do all the calculi of the application and will manage databases and user models. Another important module is the one that transforms data into its original format (such as XML) in order to adapt it to the internal format of METIORE.

The *search/analysis module* receives the user queries for simple searches, or for complex queries with data analysis. Data are normalized and sent to the *Generator of response module.* This, using the queries, will look for possible results in the database, taking into account the context of the search determined by the current user objective. This means that results will be shown according to the *user model manager* which helps to show the results in the most helpful and clearest way possible. The user model manager will also receive the users' evaluation of documents which after analysis will be used to update the models.

3 Experimentation

In previous sections, a proposal for personalized information retrieval has been proposed, based on a personalization algorithm implemented in the METIORE system. Here the experiments carried out to evaluate this proposal are described. Firstly some experiments compare the way a list of documents are sorted using different algorithms. This shows that in similar conditions, the Naïve Bayes and Naïve Bayes Cestnik algorithms produce similar results to the algorithm proposed here, giving NBM results but with fewer evaluations. Secondly, we present the experiments performed with the METIORE application and with a controlled group of users, Researchers of the laboratory LORIA of Nancy (France), where the advantages of the application were summarized in [5]. These experiments made it possible to check that the results generated by the algorithm were satisfactory for the users. The new experiment analyzes the results of METIORE in the Web with an unsupervised group of users. With this group, it was possible to show that non instructed users were able to obtain good results for the systems and it was also possible to show the viability of using the algorithm with different databases.

3.1 Experimentation with the Web Prototype

The characteristics of an experiment performed in the Web are a little different from one performed in a controlled environment. The motivations of the Web user may be different from those of other users. The first may be curious to see how this system works. If the content is interesting he could use it as any other web browser. Only in

[2] http://www.lcc.uma.es/metiorew/

the case where the user is interested in benefiting from the personalization characteristics, he will evaluate the documents to get personalized results.

In order to carry out the Web experiments, it was not enough simply to have the system and any database, it was necessary to have a database that would be interesting for a considerable number of users. In order to achieve this, we compiled scientific publications about Adaptive Hypermedia (AH) selected from different places: the home page of adaptive hypermedia[3], Citeseer[4], from a general web crawler such as Google[5] and from related conferences. Each publication had to be indexed, extracting typical data such as the title, author, conference/journal, year etc... Furthermore it was necessary to extract keywords from the document. The characteristics of this version of METIORE are described in [14].

Once the database was created, the analysis has been done by analyzing the user interactions. From the registry, we can ascertain that the users are mainly researchers, university professors or students from at least 25 countries. When the data was analyzed, a total of 146 users had registered.

	ok	?	wrong
pok	60,00%	10,00%	30,00%
pnormal	77,27%	18,18%	4,55%
pwrong	33,33%	16,67%	50,00%

Fig. 2. Results of the evaluations of the users of METIORE-AH

In Fig. 2 are shown the results of the predictions of METIORE for the evaluations obtained in the AH experiments. Although it would have been very interesting to have a greater number of evaluations done on the total number of documents examined by users, the results of the evaluations are very interesting. In the following paragraphs the most relevant data that can be obtained from Fig. 2 are commented upon:

- 60% of the system predicted that a document would interest the user (*pok-prediction o f interest)*) the user subsequently evaluated as interesting.
- The fact that a high percentage (77,27%) of documents without any prediction (*normal),* are evaluated as interesting, indicates that in the initial case when the system does not have enough information to know the user interests, the criteria of sorting the documents according to the relevance of the query works fine.
- It is interesting to note that also the prediction of not interest(pwrong) were right in a 50%.

[3] Adaptive Hypermedia. http://wwwis.win.tue.nl/ah
[4] http://citeseer.com
[5] http://www.google.com

4 Conclusions

Until the beginning of the 90s computer systems in general were very traditional and were developed in a "one size fits all" basis,, or in other words systems worked in the same way for all users. With the advent of the WWW and up until recently, this philosophy has continued to be the case in the general web search engines such as Altavista, Yahoo or Google.

The tendency nowadays however is to create systems that try to adapt progressively to the users. We can see examples in the latest version of Windows Operating System, where the desktop is personalized so that when the user wants to execute an application, the more recently used applications appear more easily. In information retrieval systems, the objective is also to limit the hundreds of possible documents because the user will ultimately examine only a few. For this reason the personalization of responses is an interesting area of research. In this work we have proposed some methods and algorithms to personalize and recommend applications with satisfactory results as justified in the experiments.

- We have developed original algorithms inspired by others already existing such as Naïve Bayes, which show the results in a personalized way. Our proposal improves the existing algorithms in the sense that it can group data describing objects with multiple parameters, weighting the importance of each parameter as necessary.
- In the analysis of the existing recommenders, we found that the main problem has been that they have a single user model to recommend documents in any situation. The vision we propose is to center the user model on the concept of objective. Therefore, depending on the current objective, the recommendations to the same user will be different.
- We already have an unpublished prototype (Kraken[6]) which personalizes web searches using the Google web services.
- These techniques can also be applied to other contexts such as personalized newspaper, personal shops (My Amazon) or TV Recommenders. Newspapers offer lots of news that is not necessarily interesting to all users. Using either the user evaluation or implicit evaluation about what s/he reads, it is possible to generate a model where user preferences will be kept in order for example to show him/her the most interesting news on the first page or to generate an email with this information. The other possibility is that of the personal web shops. Previous purchases can be used to create a profile and to recommend related products. Both proposals can be implemented with the algorithms we propose: NBM or WNBM.

[6] http://pluton.lcc.uma.es/kraken/

5 References

1. G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 6, pp. 734-749, 2005.

2. R. Burke, "Hybrid Recommender Systems: Survey and Experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331-370, Nov.2002.

3. J. Vassileva, "A Practical Architecture for User Modeling in a Hypermedia-Based Information System," Proceedings of the 4-th International Conference on User Modeling: 1994, pp. 115-120.

4. Schwab I. and Pohl W., "Learning Information Interest from Positive Examples," User Modeling (UM'99): 1999.

5. D. Bueno and A. A. David, "METIORE: A personalized information retrieval system," *User Modeling 2001, Proceedings*, vol. 2109, pp. 168-177, 2001.

6. I. Zukerman, D. W. Albrecht, and A. E. Nicholson, "Predicting Users' Request on the WWW," 1999.

7. E. Keogh and M. Pazzani, "Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches.," Ft. Lauderdale, Florida: 1999, pp. 225-230.

8. I. Kononenko, "Comparison of Inductive and Naive Bayesian Learning Approaches to Automatic Knowledge Acquisition," *Current Trends in Knowledge Adquisition*, pp. 190-197, 1990.

9. T. M. Mitchell, *Machine Learning* The McGraw-Hill Companies, Inc., 1997.

10. M. Singh and G. M. Provan, "Efficient learning of selective Bayesian network classifiers," 1996.

11. L. Versteegen, "The Simple Bayesian Classifier as a Classification Algorithm," 2000.

12. B. Cestnik, "Estimating probabilities: A crucial task in machine learning"," *Proceedings of the Ninth european Conference on Artificial Intelligence* London: 1990, pp. 147-149.

13. D. Bueno, "Recomendación Personalizada de documentos en sistemas de recuperación de la información basada en objetivos." Ph.D. Universidad de Málaga, 2003.

14. D. Bueno, R. Conejo, C. Carmona, and David A.A., "METIORE: A Publication Reference for Adaptive Hypermedia Community," *Adaptive Hypermedia and Adaptive Web-Based Systems.AH'2002* Málaga: 2002.