Generation of User Profiles for Information Filtering – Research Agenda

Tsvi Kuflik and Peretz Shoval
Information Systems Program, Department of Industrial Engineering & Management
Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel
E-mail: {tsvikak, shoval}@bgumail.bgu.ac.il

ABSTRACT

In information filtering (IF) systems, user long-term needs are expressed as user profiles. The quality of a user profile has a major impact on the performance of IF systems. The focus of the proposed research is on the study of user profile generation and update. The paper introduces methods for user profile generation, and proposes a research agenda for their comparison and evaluation.

KEYWORDS: Information filtering, user profile; content-based filtering; rule-based filtering.

INTRODUCTION

Information Filtering (IF) is a research area that offer tools for discriminating between relevant and irrelevant information by providing personalized assistance for continuous retrieval of information. IF is needed in situations of information overflow in general, and for digital libraries or the Internet in particular [4]. This area combines tools from the field of artificial intelligence (AI), such as intelligent agents or software robots ("softbots"), guided by user profiles, with information retrieval (IR) methods, geared to representing, indexing and retrieving of content [2,6]. IF differs from traditional IR in that the users have long term interests (information needs) that are described by means of user profiles, rather than ad-hoc needs that are expressed as queries posed to some IR system.

Agent technology provides the framework for automated information gathering over the Internet in general or any large information repositories, such as digital libraries. IF systems can be implemented as "intelligent agents". Applications of agent technology include passive filtering of incoming messages, like email and Usenet data, and active information seeking, like interesting Web-site detection, browsing assistance and digital libraries search [3,7,8,9,10,15]. The heart of such an agent is "user profile", a representation of the user's needs.

The quality of the user profile has a major impact on the performance (effectiveness) of the IF system. Problems related with user profile are how to generate an initial profile for a new user, and how to update an existing profile over time. An improper user profile causes poor filtering performance (namely, the user may be overloaded with irrelevant information, or not get relevant information that has been erroneously filtered out). The focus of this research is on the study of methods for user profile generation and update.

CONCEPTS IN INFORMATION RETRIEVAL AND FILTERING

Information retrieval (IR) may be characterized as "leading the user to those documents that will best enable him/her to satisfy his/her need for information" [13]. This definition (among many others) can be described in a model of information retrieval where the user seeks, by the use of queries, relevant information in some data space (e.g. a database of documents).

The vector space model [14], according to which a document is represented by a vector of terms, is perhaps the most commonly used model for text representation. The vector of terms, which may be weighted, represents the document content. Similarly, the user information interests (i.e. queries) can be represented as a vector of keywords [1,4,12]. The main task of IR is to correlate the vector that represents the user query with the vectors that represent the data items, and based on that provide the user with relevant data items, i.e. those data items that are mostly correlated with the query.

There are several different IR methods for determination of the weights of term in documents or queries. For example, TF-IDF [14] is a well-known statistical method that assigns a weight to a term in proportion to the number of its occurrences in the document, and in inverse proportion to the number of documents in which it occurs at least once.

In IF systems, user needs are expressed as user profiles. A profile represents the user's long-term information needs. There are two main distinct types of user profiles [1,12]:

a) Content-based profile: the user profile is represented similar to a query, i.e. as a vector of terms.
b) Collaborative profile: this approach is based on the rating patterns of similar users. It is assumed that people with similar rating patterns seem to like the same kind of information – "like minded people". Hence, a collaborative profile may be expressed as a list of similar users.

Another way to express user information needs in IF systems is by filtering-rules, which may be used in addition to the content-based profile. A rule can be expressed in the form of "if <condition> then <value>". The <condition> may be related to attributes of the data-items (for example, its source, author, publisher, date, length, etc.), and to demographic and social characteristics of the user (e.g. profession, education, position, age, etc.), because different users may judge the relevance of a certain data element differently. Here is an example for a possible filtering rule that applies to a database of documents:

If document-publisher = "ABC" or XYZ" or …,
and user-profession = "researcher" or "student",
then relevance-rule = 0.8 (on a 0..1 scale)

In order to support user needs, any type of user profile should be adaptable according to feedback from user reaction to information provided to him/her, since user interests tend to change over time. This calls for incorporating learning mechanisms into user profiling [5]. The user reacts to the information provided by the IF system (expressing level of relevancy). This reaction is used to adapt the user profile.

METHODS FOR USER-PROFILE CREATION AND UPDATE

The focus of this research is on the creation and update of user profiles. Therefore we describe some of such methods:

User-Created Profile

This is the most simple and natural approach. The user specifies his/her area(s) of interest by a list of (possibly weighted) terms. The specified terms are used to guide the filtering process.

System-Created Profile by Automatic Indexing

A set of data items which have already been judged by the user as relevant, are analyzed by software (using stemming algorithms), in order to identify the most frequent and meaningful terms in the text. Those terms, weighted according to the frequency of their appearance, constitute the user profile.

System- plus User-Created Profile

This is a combination of the above two approaches. First, an initial profile is created automatically (by automatic indexing). Then, the user reviews the proposed profile and updates it (by adding or deleting terms, and changing their weights).

System-Created Profile based on Learning by Artificial Neural-Network (ANN)

Based on a sample set of data items that have already been judged relevant by the user, an ANN may be trained. The inputs of the ANN are the meaningful terms, and the outputs are the relevance judgments of the users. An algorithm can calculate a Causal Index that gives the relative magnitude (and sign) of the influence of each input on each output. After training, the ANN may serve as the user profile for future filtering [11,16].

User-Profile Inherited from a User-Stereotype

This method assumes that the IF system has pre-defined user-stereotypes. A user-stereotype is represented as a content-based profile, i.e. a weighted-vector of terms that represents a set of (virtual) users who have common information usage and filtering behavior. A user-stereotype is also represented by a set of demographic and social attributes that are common to those users. (It is beyond the scope of this paper to describe how user-stereotypes and their profiles are created.) A new user is attached to a predefined stereotype to which he/she is most close with respect to the demographic and social attributes. The user inherits from his stereotype its content-based profile [17].

Rule-based Filtering

All previous methods deal with the creation of a content-based profile. Contrarily, a rule-based profile consists of a set of filtering rules. Questioning the user on his/her information usage and filtering behavior can generate such rules. An alternative method for creating a rule-based profile for a user is to inherit filtering rules from user-stereotypes, similar to the inheritance of a content-based profile. As before, this method assumes that the IF system has pre-defined user-stereotypes, but in this case a user-stereotype is represented by a set of filtering-rules that are common to the users who belong to the stereotype. As in [17], user-stereotype is also represented by a set of demographic and social attributes that are common to those users. A new user is attached to a predefined stereotype to which he/she is most close with respect to those attributes, and inherits from that stereotype filtering rules.

RESEARCH AGENDA

We plan to evaluate and compare various methods for creation of initial user profiles and for updating them over time. The research will be conducted in a research center, with real users (researchers and other consumers of information) who access various research databases and the Internet. We plan to develop an IF system that employs content-based and rule-based filtering techniques. We will generate several initial user profiles for each participant, according to the various profile creation methods, and measure the performance of the IF system according to each method. Later on we will update the profiles according to user feedback, based on actual filtering results. We will compare the performance of the system according to the different profile creation and update methods. Eventually, we will propose a model for the most effective method or combination of methods for user profile creation and update. Here are some more details about the research issues:

Comparing the Accuracy of Various User-Profile Creation Methods

The above mentioned approaches will be used in order to create initial user profiles for each of the participants. The participants will read data items (documents) obtained from certain databases and evaluate their relevancy. Then, the IF system will filter the same data items for each user, according to his/her various profiles, and compute their relevancy. System-computed relevancy will be compared to the user evaluations. This will enable determining the accuracy of the various user profiles. The following approaches will be compared:

- User (manual) created profile: the user defines a set of weighted-terms that represent his/her area of interest.
- Systems (automatic) created profile: The profile is created by applying a classical IR text analysis (TF-IDF) algorithm [Salton & McGill 83] on a sample of "interesting", rated data items provided by the user.
- Combination of the above: the user is enabled to modify and adjust the system-created profile.
- Automatic profile generation using ANN: Profile is created by applying a neural net learning mechanism, on a sample of "interesting", rated data items provided by the user.

Stabilizing the Process of User-Profile Creation:

An initial user-profile can be in various levels of detail (i.e. include many or few terms or filtering rules). Then, after initial generation, and based on user feedback (in the form of his/her ranking of the relevance of the data-items), the user profile may be updated. We plan to investigate how long it takes to stabilize a user profile, namely how many data items need to be evaluated by the IF system until no more improvement in system performance is achieved. This will also enable determining which combination of initial profile creation and adaptation process is most effective.

For this, a set of predefined user profiles will be used. For each user there will be a set of profiles in different levels of details. Using the different profiles, simulated filtering sessions will be performed, enabling to compare the performance of several levels of detailed profiles. In following steps the profiles will be updated gradually, until the performance of the IF system will stabilize. At this point we will be able to identify the best combination of user-created and training set needed fór the most effective profile.

Combining Content-based and Rule-based Filtering

In addition to the evaluation of each profile creation method individually, we plan to investigate the effect of combining rule-based and content-based filtering. The IF system will combine the two main filtering methods in various ways. For example, a filtering process that starts with user-created content-based profiles, followed by personal filtering-rules, or a filtering process that starts with system-created content-based profiles, followed by rule based filtering where the rules are inherited from proper user-stereotypes.

Personal vs. Stereotype-based Filtering Rules

In an earlier study [17,18] we employed filtering-rules within user-stereotypes, while in this study we plan to employ personal filtering rules. This will also enable us to compare and even combine the two approaches. While it may seem obvious that personal filtering rules are more accurate than filtering rules that are inherited from stereotypes, it must be admitted that personal rules may not be easy to define, in particular if there are many users. Hence, it may be more practical for a new user to inherit filtering rules from an appropriate stereotype, which will later on be adapted according to the specific user's needs, based on feedback. We plan to measure and compare the accuracy of personal filtering rules vs. inherited (stereotype-based) rules, and study how to adapt the inherited rules to the specific user needs. Initial stereotype-based profiles will be generated. Later on, as user feedback is accumulated, the initial sets of rules will be modified and adapted accordingly.

Adaptation of Filtering Method to Area of Application

IF systems may be utilized in various areas of application, e.g. research documents, e-mail lists servers, hotel and tourism information, restaurants, entertainment, etc. In different application areas the data items may have different attributes, and the relevancy of the respective data items to users may depend on different personal sociological factors For example, we think that for a database of research documents, a content-based profile is more important than a rule-based profile, while the opposite is true for tourism information. But this must be proved by research. We plan to investigate this issue by utilizing various profile-creation and filtering methods on a recreational database in addition to the research database

Sociological and demographic data will be gathered from the participating users, together with their preferences regarding recreational activities. That data will be used to create personal filtering rules and content-based profiles. Using the new set of profiles, the users will be asked to rate recreational data items filtered by the system The rated data items will be used in several simulated filtering processes each with different profile (as before). This will enable comparing and evaluating the adaptation of filtering methods to the different areas of application.

EXPECTED BENEFITS

The sought research is important from both theoretical and practical points of view. From the theoretical point of view, the significance of the sought research is in examining and comparing various user-profile generation- and updating methods. From the practical point of view, the results of the research will enable to implement the most appropriate user-profile generation methods in IF systems used on the Internet, digital libraries, Intra-nets, etc. Moreover it will allow a better adaptation of filtering methods to areas of application.

REFERENCES

[1] Aas, K. A Survey on Personalized Information Filtering Systems for the World Wide Web. 1997, *Report No. 922, Norwegian Computing Center.*

[2] Balabanovic', M., Shoham, Y. Learning Information Retrieval Agents: Experiments with Automated Web Browsing. *Spring Symposium on Information Gatheringfrom Heterogenious Distributed Environments.*, AAAI 95, 1995.

[3] Balabanovic', M., Shoham, Y. Fab: Content-Based Collaborative Recommendation. *Comm. of the ACM*, 1997, 40 (3), 66-72.

[4] Belkin, N. J. and Croft, W. B. Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Comm. of the ACM*, 1992, 35 (12), 29-38.

[5] Billsus, D. and Pazzani, M. Learning Collaborative Information Filters. *Proc. of the Int'l Conf. on Machine Learning*, 1998, Morgan Kaufman, Madison, Wisconsin.

[6] Etzioni, O. and Weld, D. Intelligent Agents on the Internet: Fact, Fiction and Forecast. *IEEE Expert*, August 1995 44-49

[7] Joachims, T., Freitag, D. and Mitchell, T. A Tour Guide for the World Wide Web. *Proceedings of IJCAI97*, 1997.

[8] Lieberman, H., Letizia: An Agent that Assist Web Browsing. *Proc. of the Int'l Joint Conference on Artificial Intelligence*, Montreal, 1995.

[9] Lieberman, H. Autonomous Interface Agents. *Proc. of the ACM conference on Computers and Human Interface, CHI 97*, Atlanta, Georgia, 1997.

[10] Maes, P. Agents that Reduce Work and Information Overload. *Comm. of the ACM*, 1994, 37 (7), 31-40.

[11] Moghrabi, C. and Eid, M. S. Modeling Users through an Expert System and a Neural Network. *Computers and Indusrtial Engineering*, 35, (3-4) 583-586, 1998.

[12] Oard, D., W., Marchionini, G. A Conceptual Framework for Text Filtering. *Technical Report* EE-TR-96-25 CAR-TR-830 CLIS-TR-96-02 CS-TR-3643, 1996.

[13] Robertson, S. E. The Methodology of Information Retrieval Experiment. *In (Ed) K. Sparks Jones*, Chap. 1, 9-31. Butterworths, 1981.

[14] Salton, G., McGill, W. J. *Introduction to Modern Information Retrieval*, McGraw-Hill. New York, 1983

[15] Sanchez, J. A., and Leggett J. J., Agent services for users of digital libraries,. *Journal of Network and Computer Applications*, 20, 45-58, 1997.

[16] Schutze, H., Hull, D. A., and Pedersen, J. O. A comparision of classifiers and document repreentations for the routing problem. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 229-237, 1995.

[17] Shapira, B., Shoval, P. and Hanani, U. Stereotypes in Information Filtering Systems. *Information Processing & Management*, 33 (3), 273-287, 1997.

[18] Shapira, B., Shoval, P. and Hanani, U. Experimentation with an Information Filtering System that Combines Cognitive and Sociological Filtering Integrated with User Stereotypes", *Decision Support Systems*, 27, 5-24, 1999.