User Model-Based Information Filtering

Fabio A. Asnicar, Massimo Di Fant, and Carlo Tasso

Department of Mathematics and Computer Science - University of Udine
Via delle Scienze, 206 - 33100 UDINE (Italy)
e-mail: tasso@dimi.uniud.it

Abstract The IFT (Information Filtering Tool) project has the goal of developing new approaches to information filtering which are based on user modeling techniques for building and managing the representation of the user information preferences. In this paper we describe three prototypes which have been developed and evaluated within the project. All of them are dealing with textual semistructured documents and exploit a semantic network representation of user preferences: the first two prototypes (IFTool and PIFT) are characterized by two different matching algorithms utilized for assessing the relevance of an incoming document against the user model, whereas the third (ifWeb) concerns an application of IFTool to the navigation and filtering of documents in the INTERNET. The three prototypes have been evaluated in order to compare their performance with similar systems presented in the literature. The results achieved show that information filtering can positively profit from user modeling techniques, and point out interesting challenges for future investigations.

1 Introduction

The recent development of communication networks and multimedia systems provide potential users with the availability of a huge amount of information, making worse and worse the problem of information overload ([12]). This situation has favoured the development of systems capable of automatically identifying the subset of the available information, which is potentially relevant to the user information needs. More specifically, filtering systems have been proposed ([3]), which interface the information source to the user, and are aimed at automatically evaluating the potential relevance of incoming information on the basis of an explicit description of the user information interests (user profile). However, while the need for these systems has been widely recognized ([10], [14], [4]) and adequate techniques for their implementation have emerged — we mainly refer to the intelligent agent technology ([17]) — two basic problems still remain open and need further investigation: (i) the mechanisms for learning, representing and updating the user's information preferences, and (ii) the processing algorithms to be adopted to extract the information content of the incoming documents and the matching algorithms to be exploited to assess their potential relevance.

The *IFT* (*Information Filtering Tool*) project has the goal of developing and evaluating new approaches to information filtering, where the potential

relevance of incoming information is computed by comparison with an explicit and dynamic user model, which represents user information needs and preferences. More specifically, we have investigated new representation and matching paradigms for classifying incoming documents and evaluating their potential relevance. Another major point of investigation has been the development of suitable tools capable of learning and tracking over time the information preferences of the user. The proposed approaches have been experimented in two main applications, namely SDI services and information gathering on the WWW.

The goal of this paper is to illustrate two specific prototype systems, where (i) the user model includes an explicit representation of the co-occurrence relationship ([7]) between pairs of terms appearing simultaneously in the documents, and (ii) two specific mechanisms (relevance feedback and rent) are exploited in order to manage the temporal evolution of the content of the user model.

In section 2, we present the *IFTool* (*Information Filtering Tool*) prototype, which is based on a matching algorithm specifically designed for taking advantage of the co-occurrence relationship. In section 3 we introduce the *PIFT* (*Probabilistic Information Filtering Tool*) prototype, characterized by a probabilistic approach to filtering. For both prototypes we show also the results of a systematic evaluation activity, which support the claim that a specific user modeling component improves the performance with respect to other filtering system reported in the literature. In section 4 we illustrate the *ifWeb* (*Information Filtering Web*) prototype, which extends the proposed approach to the navigation and filtering within the WWW, providing also an account of its evaluation and a comparision with commercial search engines. Section 5 concludes the paper.

2 IFTool

IFTool ([11]) is a prototype of an information filtering system devoted to semistructured textual documents, which exploits the *UMT* (*User Modeling Tool*) ([5]) shell for building and managing the user model devoted to represent the user information preferences.

2.1 The User Model

The model includes information which represents both the interests and the 'not interests' of the user. More precisely, it is constituted by a weighted semantic network whose nodes correspond to terms (concepts) found in documents and where arcs link together terms which co-occurred in some document. Each node has an associated weight, which is positive (negative) if the corresponding term has been extracted from a document which has been judged "interesting" ("not interesting") by the user. Each arc is characterized

by a weight which represents the frequency of co-occurrence of the two terms in the previously analyzed documents. The specific method described above has been proposed to overcome the polisemy problem ([9]) which stems from the use of keywords for representing the information preferences of the user: the co-occurrence relationships allows to associate to each term a 'pragmatic context' which helps in the disambiguation of the meaning of the term.

The content of the user model is acquired and managed by the Feedback Handler Module (see next section). In this way the main technique available to the user for providing information about his/her preferences is constituted by relevance feedback ([18]). Direct inspection and modification of the user model by the user is also possible through a specialized interface.

2.2 The Architecture

The main modules of IFTool (cfr Figure 1(a)) are illustrated in the following.

The *Document Representation Module* analyses the incoming documents and produces an internal representation, containing information about their content. In particular, this is constituted by a weighted vector of terms which is obtained trough standard techniques (such as segmentation, stop list deletion, stemming and weighting) and a specific algorithm which is devoted to identify the best terms that represent the content of a document (compression).

The *Document Classifier Module* receives in input the internal representation of a document and the current user model, and produces in output a classification of the document with respect to its potential relevance for the user. The algorithm exploited includes two phases: the comparison of the internal representation of the document with the user model and, later, the classification of the document on the basis of the results of the previous phase. The comparison phase produces two numerical values representing the similarity between the current document and (a) the information preferences and (b) the 'not interests' of the user. More specifically, the comparison phase exploits in an original way the co-occurrence relationship. In standard keyword matching, a simple count is performed of the terms which are simultaneously present in the document representation and in the user model. IFTool, on the other hand, considers also a contribution obtained by the pairs of terms included in the document which have already co-occurred in previous documents (information represented by arcs in the semantic network): this allows to add more evidence to the classification process. The classification of the document is then performed by means of a suitable criteria which considers the two matching values produced by the comparison phase and results in the final classification ("interesting", "not interesting" or "indifferent").

The *Feedback Handler Module* receives in input the relevance judgements possibly provided by the user and the internal representation of the corresponding document and updates the user model accordingly. Updates are constituted by (i) insertions of new nodes with corresponding weights in

the semantic network included in the user model and/or (ii) changes in the weights associated to nodes and arcs of the semantic network, taking into account the frequency of occurrence of terms and of co-occurrence of pair of terms in the considered document. The Feedback Handler Module allows IFTool to better follow the temporal evolution of the user information preferences and this is obtained by means of an increase (decrease) of the weights associated to nodes and arcs. The module is also used to initialize the content of a new user model.

The *Rent Handler Module* is devoted to delete from the user model the terms which have been inserted accidentally in it through the feedback operation. This allow to highlight the terms that better represent the user preferences. The module is activated periodically on the user model and it identifies the changes which have to be performed on the model: these are constituted by a decrease of the weight associated to the nodes, so as to results as the payment of a rent ([1]). After the rent has been paid, all the terms which have a weight lower than a given threshold are deleted from the model.

The *User Modeling Subsystem* has been obtained by extending the original representational capabilities of the UMT shell (basically attribute-value pairs and frame-based stereotypes) with semantic networks.

The *User Interface* is devoted to manage the interaction with the user.

Finally, the *Document Data Base* stores the original version of the documents and their internal representation.

2.3 Experiments and Evaluation

The goal of the experimental activity performed on IFTool has been the evaluation of the performance of the prototype. Several experiments have been performed. The most significant one has been concerned with the filtering operation of 2000 abstracts of technical report on the basis of the information preferences of four subjects. We have simulated the following situation: IFTool receives 20 documents per day (globally 100 sessions) which are classified according to the user model and then are shown to the user. Every day the Rent Handler Module is activated. For each set of 20 documents, each subject provides a relevance ranking which is then utilized for computing normalized precision and recall ([18]) for both "interesting" and "not interesting" classifications. In Figure 1(b) the evolution of the normalized precision of the interesting documents is reported. At the beginning of the 100 session sequence, the user model, initially empty, is incrementally acquired through relevance feedback. The performance increases until a saturation value is reached (94% for precision, 93% for recall). The dotted line refers to a simple keyword matching algorithm, whereas the continuous line refers to the new algorithm exploited in IFTool: the obtained results show an average improvement of the precision of 18% and an average improvement of the recall of 30%. The performance of IFTool have also shown better values than those reported in [9].

A specific experiment has been performed to evaluate the behaviour of IFTool when the user preferences drastically change. The results obtained show that, after an initial strong decrease of the performance, the prototype can learn and adapt quickly to the new information preferences of the user.

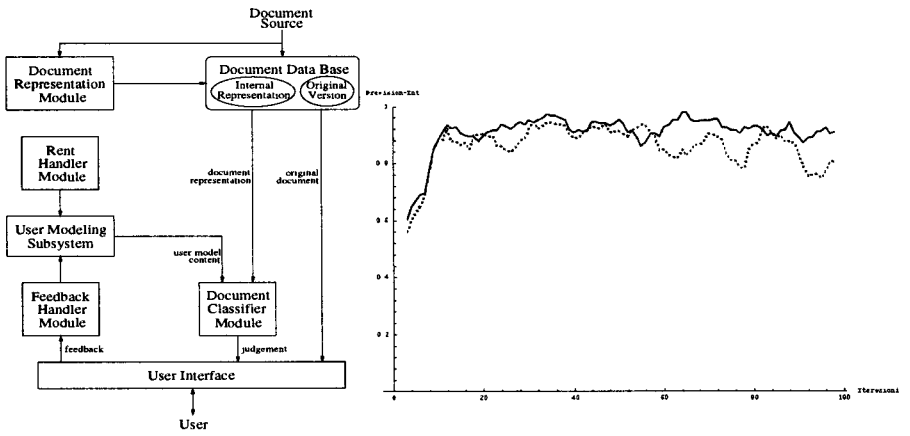


Figure1. *(a) The architecture of the IFTool prototype. (b) Normalized precision of interesting documents over 100 sessions.*

3 PIFT

PIFT is a prototype of a filtering system for semistructured documents, where their potential relevance is computed by means of two bayesian networks ([16]). These are built through statistical analysis of the terms which occurred in the documents judged as "interesting" and, respectively, "not interesting" by the user.

Several similarities characterize PIFT and the InRoute ([6]) system, essentially due to the same approach based on bayesian networks. However, PIFT is characterized by an explicit and dynamic user modeling activity which is absent in InRoute.

3.1 The Architecture

PIFT has basically the same architecture of IFTool. However, for supporting the statistical analysis of the documents, a specific data base (*Term Dictionary*) has been introduced, together with a corresponding *Term Dictionary*

Handler Module. Moreover, the User Modeling Subsystem is not based on the UMT shell, but a new reduced version of it has been ad hoc developed. Another difference between the two prototypes concerns the co-occurrence relationship, that in the PIFT prototype is not computed by exploiting the entire document, but only a part of it.

The two main modules of PIFT are described in the following.

The *User Modeling Subsystem* is devoted to represent the information preferences of the user. Analogously to IFTool, the user information preferences are represented by means of two semantic networks, one for the interests and one for the 'not interests'. The co-occurrence between words is verified within a *contextual window* of size m: two words are considered to co-occur if they appear together in the same document and the number of words between them is not greater than m. The exploitation of contextual windows is aimed at reducing inconsistencies in the model, since in long documents the co-occurrence relationship between a term at the beginning and a term at the end of the document may mean nothing ([20]). Each node of the semantic network is associated to a weight which is computed by means of tf.idf weighting ([18]), where tf represents the sum of values proportional to the absolute frequency of occurrence of the term, and idf represent the reciprocal of the number of documents where the term has appeared. Analogously, each arc of the semantic network is associated to a numerical value which is a function of the frequency of co-occurrence of the linked terms.

The *Document Classifier Module* builds two bayesian networks on the basis of the internal representation of the current document, of the information stored in the Term Dictionary, and of the content of the user model. The first network is devoted to compute the probability that the current document satisfies the user interests, and the second one is devoted to the user 'not interests'. The networks are constituted by $n+2$ nodes $S, T_1, \ldots, T_n, Q_I$, where n is the number of terms present in the semantic network describing the interests (or not interests) of the user. Each node is associated with a proposition: S is associated with the proposition "the document is present in input to the system", each T_i with the proposition "the term t_i is present in the representation of the user interests (not interests)" and Q_I (Q_{NI}) with the proposition "the interests (not interests) of the user are satisfied". There are $2 \cdot n$ arcs $S \rightarrow T_i$ and $T_i \rightarrow Q_I$ ($T_i \rightarrow Q_{NI}$): the first n arcs link S with each of the T_i nodes and the second n arcs link each T_i with the proposition Q_I (Q_{NI}). Each arc is associated with a numerical value which represents the relevance of the term respectively in the representation of the document and in the representation of the user interests (not interests) and it is expressed by means of a generalised tf.idf weighting ([19], [6]). The networks are simplified by means of the conditioning method ([16]) and are evaluated by means of the noisy OR-gate model ([16]). This results in two probability values (one for the interests and one for the 'not interests'), which are exploited in order to compute a final relevance classification for the document.

3.2 Experiments and Comparision with IFTool

We have performed on the PIFT prototype the same experimental evaluation which has been performed on the IFTool prototype. The diagram reported in Figure 2(a) show that the performance of PIFT is only slightly worse than IFTool.

Moreover, in order to better understand which is the optimum size of the contextual window for computing co-occurrence, we have measured the percentage of the documents which PIFT has correctly classified as interesting (not interesting), with a different size of the contextual window. Figure 2(b) reports a diagram which shows the different percentage after the analysis of 2000 documents with a different size of the contextual window. The diagram shows that the performance increases until the size of the contextual window is near 20 and decreases for higher values.

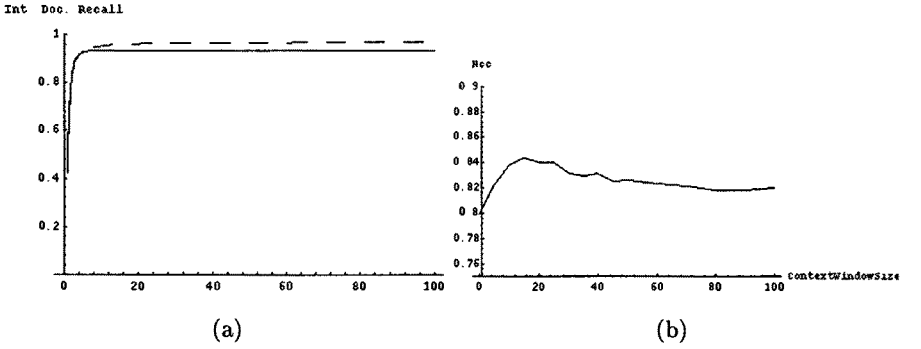


Figure2. *(a) The normalized recall (polynomial interpolation) of the interesting documents in the PIFT prototype (continuous line) and in the IFTool prototype (dotted line). (b) The percentage of the documents correctly classified by PIFT (N_{CC}) with respect to a different size of the contextual window (ContextWindowSize).*

4 ifWeb

ifWeb is a prototype of user model-based intelligent agent capable of supporting the user in the navigation of the WWW, the retrieval and the filtering of documents. Following [3], the filtering process can also be viewed as the process of accessing and retrieving information from remote data bases: in such case the incoming data (document source) is the result of the WWW navigation.

Several tools have been proposed in the literature which are aimed at gathering information from the WWW on the basis of a user profile ([15], [8],

[13], [2]). All of them, however, share some basic limitations: the technique used to represent knowledge in the user profile is based on simple lists of keywords; the types of the considered knowledge are very limited, usually restricted to single words, or to (some) structural characteristic; the learning capabilities are usually very poor, if any; and the navigation strategies, if any, are usually based on very heuristic and limited criteria. ifWeb is aimed at overcoming the above limitations by exploiting the possibilities of the information filtering approach proposed in the IFT project.

ifWeb is characterized by two modes of operation. The first one is called *navigation support*: from a specific WWW document pointed out by the user, ifWeb starts an autonomous navigation, it collects WWW documents, it analyses and classifies them and, as a result, it shows graphically to the user the structure of the hypertextual links present in the documents which have been accessed. The second mode of operation is called *document search*: from a specific WWW document pointed out by the user, the system autonomously performs an extended navigation in the WWW, retrieves and classifies documents. As a result, the system shows to the user the set of the documents which have been classified as most relevant, ordered downward from the most interesting one.

4.1 The Architecture

Figure 3(a) shows the functional architecture of ifWeb which includes the following modules: the *ifWeb Interface Agency* which manages all the interaction with the user; the *ifWeb Agency* which performs the specific function of navigation support and document search; the *IFTool Agency* based on the IFTool prototype, which is devoted to classify incoming documents on the basis of the content of the user model.

The overall operation is the following. The ifWeb Agency requires from the Network through an URL the retrieval of a specific document; when the document is retrieved, information about its content is extracted and is then exploited in order to build the document internal representation.

ifWeb considers only HTML documents. The analysis of these documents is performed by a syntax direct parser for the HTML (Data Type Definition) format and, moreover, it includes some basic processing (segmentation, stop list deletion, stemming, contextual weighting, and compression). The document internal representation produced is then sent to the IFTool Agency for comparison with the user model. The result of the comparison allows the classification of the document in one of the three categories "interesting", "not interesting" or "indifferent". Then, the ifWeb Agency manages the choice about accessing or not the documents mentioned in the links (URLs) specified in the currently analyzed document. In case of a positive decision it forwards the request for access, and the operation continues in an analogous way.

The strategies for the autonomous navigation performed by ifWeb are based on the evaluation of how much a link (URL) is considered *promising* for accessing other documents which can be relevant to the user interests. The computation of the *degree of promise* of a link is executed by means of two parameters: the first one is called *expectation rate*, and it is a function of the values obtained from the comparison carried out with the user model and the consequent classification performed on the currently analyzed document (i.e., the document containing the URLs whose potential is considered for further navigation). The second parameter, called *confidence rate*, is a function of the values of the degree of promise of the documents previously accessed on the path which concludes with the currently analyzed document. In navigation support mode, all the promising link are considered, whereas in document search mode an hybrid best/breadth-first search is adopted. Figure 3(b) shows the user interface: it includes a normal browser window

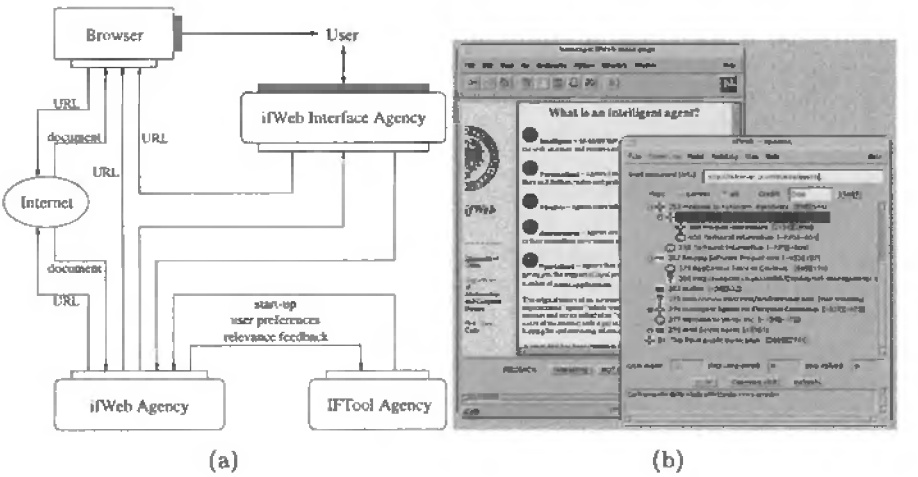


Figure3. *(a) Functional architecture of ifWeb. (b) User interface of ifWeb.*

and a specific window managed by the ifWeb Interface Agency. This window is used for showing to the user the intermediate status and the results of the various analyses and for allowing the user to modify some system parameter.

The documents are displayed in a tree-like structure, where the arcs correspond to hypertextual links. The various icons represent the result of the classification performed on the documents: '+' means "interesting", '=' means "indifferent", '-' means "not interesting", '?' means 'analysis not performed' (i.e. document not available), 'STOP' means that the ifWeb Agency has decided not to continue the analysis from that document. The user can easily modify the order of analysis, request the access to links which where consid-

ered not promising by ifWeb, exclude some document from navigation, and ask for display (through the browser) of a specific document in its original form. In the browser window, two specific buttons allow the user to provide relevance feedback on a document.

4.2 Experiments and Evaluation

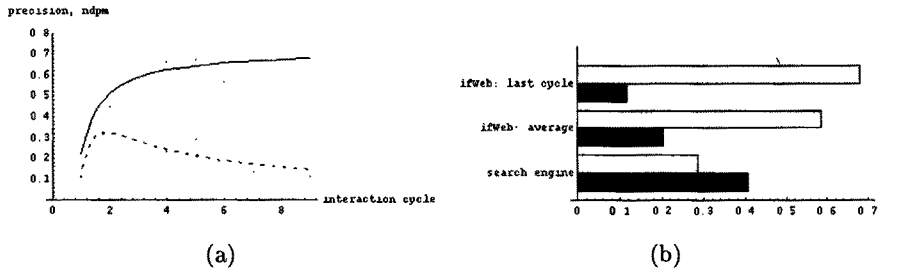
A specific evaluation activity has been carried out on ifWeb. The main goals of the evaluation were: (i) assessing the quality of the learning capability and the quality of the results of the classification performed on the documents and (ii) comparing ifWeb with some traditional search engines.

The evaluation activity has been performed through real-time access to INTERNET: four subjects where using ifWeb having in mind a specific field of interest (usually their specific field of expertise). Each subject was carrying out 9 sessions with ifWeb. ifWeb was started with a user model obtained through (positive and negative) relevance feedback on a limited set of 4–6 documents. The model was later incrementally refined by ifWeb, thank to further relevance feedback provided by the user. During each session ifWeb was working autonomously and, at the end of the session, it displayed the results to the subject. The subject was then requested to provide relevance feedback, and to order the results according to his/her relevance judgement. The data collected at the end of each session were then used for computing two performance figures: ndpm ([21]) (a measure of the capability to order correctly the documents from interesting to not interesting) and standard precision ([18]).

Figure 4(a) shows the results of this experimentation. After the initial session where ifWeb has a too limited knowledge of user preferences, both the computed figures show that the system progressively improves its performance through the user feedback, both in terms of overall precision of classification (upper continuous line) and of the capability to order correctly the documents according to their potential interest for the user (lower dotted line). Comparisons with similar tools found in the literature are difficult, because those systems have been scarcely evaluated and cannot be well compared one to the other.

Moreover, we have performed a comparision with the standard search engines AltaVista and ExCite. More specifically, the most relevant keywords included in the user model of ifWeb at the end of the 9 session experiment described above, were submitted to the two search engines as query. On the outcome, ndpm and precision were computed, according to the relevance ranking provided by the subjects.

Figure 4(b) illustrates the results: they show that ifWeb provides more precise results and better ordering.



(a) (b)

Figure4. *(a) Precision (continuous line) and ndpm (dotted line) over the 9 sessions (polynomial interpolation). (b) ifWeb vs. standard search engines (precision in grey, ndpm in black).*

5 Conclusions

In this paper we have described the IFTool, PIFT, and ifWeb prototypes, three user model-based information filtering systems developed and evaluated within the IFT project. All of them deal with textual semistructured documents: the first two prototypes are characterized by two different matching algorithms utilized for assessing the relevance of the documents, whereas the third prototype concerns an application of the IFTool prototype to the navigation and filtering of documents in the INTERNET.

The evaluation activity carried out on the three prototypes allows to support the claim that the use of sophisticated user modeling techniques can improve the performance of intelligent agents for information filtering. Specific results in this direction have been produced in the IFT project and are reported in this paper.

A basic problem whose importance has emerged from our research and which needs further effort, is constituted by the methodologies which are exploited for the evaluation of these systems. The availabilty of such methodologies — frequently not reported in the literature — is strongly needed in order to compare the various approaches. The performance measures and the experimental procedures adopted in the IFT project can be considered a first step towards this goal.

References

1. P. E. Baclace. Competitive Agents for Information Filtering. *CACM* 35(12), p. 50, Dec. 1992.
2. M. Balabanovic. An Adaptive Web Page Recommendation Service. In *Proc. of the 1st Int.l Conf. on Autonomous Agents*, Marina del Rey CA, Feb. 1997.
3. N. J. Belkin, W. B. Croft. Information Filtering and Information Retrieval: Two Sides of the Same Coin? *CACM* 35(12), pp. 29–38, Dec. 1992.

4. T. A. Bell, A. Moffat. The Design of a High Performance Information Filtering System. In *Proc. of the 19th Int.l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 12–20, Zurich, CH, Aug. 1996.
5. G. Brajnik, C. Tasso. A shell for developing non-monotonic user modeling systems. *Int. J. of Human-Computer Studies* 40, pp. 31–62, 1994.
6. J. Callan. Document Filtering with Inference Networks. In *Proc. of the 19th Int.l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 262–269, Zurich, CH, Aug. 1996.
7. W. B. Croft. Effective Text Retrieval Based on Combining Evidence from the Corpus and Users. *IEEE Expert*, pp. 59–63, Dec. 1995.
8. P. Edwards, D. Bayer, C. L. Green, T. R. Payne. Experience with Learning Agents which Manage Internet-Based Information. In *Proc. of the AAAI Spring Symposium on Machine Learning in Information Access*, Stanford, Mar. 1996.
9. P. W. Foltz. Using Latent Semantic Indexing for Information Filtering. In *Proc. of the ACM SIGOS Conf. on Office Information Systems*, pp. 40–47, Boston, MA, 1990.
10. M. Höfferer, B. Knaus, W. Winiwarter. An Evolutionary Approach to Cognitive Information Filtering. In *Proc. of the 18th Int.l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 1–15, Seattle, WA, July 1995.
11. M. Minio, C. Tasso. IFT: un'interfaccia intelligente per il filtraggio di informazioni basato su modellizzazione di utente. *AI*IA Notizie* IX(3), pp. 21–25, Sep. 1996.
12. M. Morita, Y. Shinoda. Information Filtering Based on User Behavior: Analysis and Best Match Text Retrieval. In *Proc. of the 17th Int.l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 272–281, Dublin, IR, June 1994.
13. A. Moukas. Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem. In *Proc. PAAM96, The Practical Application of Intelligent Agents and Multi-Agent Technology*, London, UK, Apr. 1996.
14. S. Mukhopadhyay, J. Mostafa, M. Palakal, W. Lam, L. Xue, A. Hudli. An Adaptive Multi-level Information Filtering System. In *Proc. of the 5th Int.l Conf. on User Modeling*, pp. 21–28, Kailua-Kona, Hawaii, Jan. 1996.
15. M. Pazzani, J. Muramatsu, D. Billsus. Syskill & Webert: Identifying interesting web sites. In *Proc. of the 13th National Conf. on Artificial Intelligence*, pp. 54–61, Portland, OR, Aug. 1996.
16. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan-Kauffman, 1988.
17. C. J. Petrie. Agent-Based Engineering, the Web, and Intelligence. *IEEE Expert*, pp. 24–29, Dec. 1996.
18. G. Salton, M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, 1983.
19. H. R. Turtle, W. B. Croft. Evaluation of an Inference Network-Based Retrieval Model. *ACM TIS* 9(3), pp. 188–222, July 1991.
20. J. Xu, W. B. Croft. Query Expansion Using Local and Global Document Analysis. In *Proc. of the 19th Int.l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 4–11, Zurich, CH, Aug. 1996.
21. Y. Y. Yao. Measuring Retrieval Effectiveness Based on User Preference of Documents. *Journal of the American Society for Information Science* 46(2), pp. 133-145, 1995.