Journal of Computer Science 1 (1): 40-46, 2005
ISSN 1549-3636

Interest-based Recommendation in Digital Library

Yan Yang and Jian Zhong Li
School of Computer Science and Technology, Harbin Institute of Technology
Harbin 150001, Peoples Republic of China
School of Computer Science and Technology, Heilongjiang University
Harbin 150080, Peoples Republic of China

Abstract: With the huge amount and large variety of information available in a digital library, it's becoming harder and harder for users to identify and get hold of their interested documents. To alleviate the difficulty, personalized recommendation techniques have been developed. Current recommendation techniques rely on similarity between documents. In our work, recommendations are made based on three factors: similarity between documents, *information amount*, and *information novelty*. With the introduction of *degree of interest*, users' interests can be better characterized. Theoretical analysis and experimental evaluations demonstrate that our techniques can improve both the recommendation recall and recommendation precision.

Key words: Digital Library, Personalization, Recommendation, Degree of Interest

INTRODUCTION

With the huge amount and large variety of information available in a digital library, it's becoming harder and harder for users to identify and to get documents they are interested in. Personalized recommendation techniques have been developed as a solution to help users get what they want conveniently and efficiently.

Personalized recommendation techniques have been widely incorporated in systems such as E-commence, Web information retrieval, digital library, and so on. A rich abundance of contents could be stored in these systems, for instance, items in E-commence systems, web pages in Web information retrieval systems, movies, documents and all other media in digital libraries. In this study, we use resource to represent contents in a system where personalized recommendation is expected.

According to their basis of recommendation, personalized recommendation techniques can be categorized as statistics-based, rule-based, content-based, collaborative filtering. Among them, the content-based and collaborative filtering approaches are the most commonly used.

Content-based approaches recommend resources based on the similarity between resource and the user profile. The key problem is to calculate the similarity. Recommender systems taking content-based approaches include Personal Web Watcher [1], CiteSeer [2], IfWeb [3] etc. As an alternative, collaborative filtering approaches give recommendations based on correlation between users. Given a user, these approaches compare his/her profile with that of other users, find the similar users, and provide the resources which they are interested in [4-6]. There are systems which incorporate the above two approaches to obtain more accurate recommendations, called hybrid recommendation techniques. These systems can be grouped into two categories. In the first category, recommendations are generated using the two approaches separately, then the results are simply combined together [7]. While for the second category, the combination of the two approaches occur at a lower level, generating a new representation which encloses both resources and users. Example of systems in this category are GroupLens [8], Fab [9], Graph-based approach [10] etc.

The aforementioned three approaches all depend on results from similarity analysis. They recommend information which is either similar with what the same user was interested in before, or relevant with what other similar users are interested in. Unfortunately, similarity-based techniques cannot always create the accurate recommendations, which could be illustrated by the following examples.

Example 1: In the digital library, there are two copies of the same paper coming from different data sources. Certainly, these two copies have the highest similarity. If a user has read the paper before, the similarity-based approaches will recommend the same paper to the user again. However, this is not appropriate since the user has already seen the information in the study.

Example 2: There are two editions of a book in the digital library, with the second edition having more contents. Still, the similarity between them is pretty high. Suppose the user has downloaded the second edition, existing recommendation approaches, no matter how similarity is defined, will recommend the old edition to the user as well. Obviously, this is not a very useful recommendation.

Example 3: Jack has read a lot of papers about personalized recommendation. Among his unread papers, there are two with the same name. One is a technical report, the other is a short paper in a conference proceeding. If the latter has higher similarity with other papers he read before, the short paper will be recommended by the current similarity-based techniques. While in practice, the user might be more interested in the detailed technical report.

The above examples showed that only considering similarity when making recommendations does not suffice. In addition to similarity, users also care about the amount of information contained in the documents, and the novelty of the information. Thus, a new measurement is needed to reflect the real user interests. This study proposes the concept of *degree of interest*, which indicates the interest of a user to a resource. The degree of interest is determined based on three factors: similarity between resources, *information amount*, and *information novelty*. Theoretical analysis and experimental evaluations demonstrate that our techniques can improve both the recommendation recall and recommendation precision.

The Definition of Degree of Interest: A user's interest to a document is related with how similar a document is with respect to the documents user have accessed before, the amount of information the document contains and the novelty of the document. This section integrates the three factors and defines the degree of interest.

In the rest of the article, *the document set* includes all documents in the recommendation consideration. First, we define the similarity between two documents.

Definition 1: Suppose α and β are vector representations of two documents, the *similarity between α and β*, $RSim_{\alpha\beta}$, is defined as follows:

$$RSim_{\alpha\beta} = \frac{\sum_{k=1}^{n} \alpha_k \beta_k}{|\alpha||\beta|} \qquad (2.1)$$

where:

- * n is the number of distinct words in the document set, that is, dimension of the document vector space
- * α_k is the *k-th* component of vector α,

  $\alpha_k = tf_{\alpha k} \times \log(N / df_k) \times p_{\alpha k}$

  Here, $tf_{\alpha k}$ is the occurrence frequency of the *k-th* term in α. N is the total number of documents in the document set. df_k is the number of documents which contain the *k-th* term, i.e. the document frequency of the *k-th* term. $p_{\alpha k}$ is the position weight of the *k-th* term in α, since the position of a term signals its importance. For example, a term in title is always more important than its counterpart in the body of the document. β_k can be calculated in the same manner.
- * α| and |β| are lengths or norms of vector α and β separately

The amount of information included in a document is also a very important factor affecting a user's interest. The more information a document contains, the more likely a user is interested in it.

Definition 2: For a document *j*, the *amount of information* contained in document *j*, denoted by $Info_j$, is defined as follows:

$$Info_j = \frac{nw_j \times \log(Len_j)}{Nw \times \log(maxlen)} \qquad (2.2)$$

where:

- * Nw is the number of distinct words in the document set
- * nw_j is the number of distinct words in document *j*
- * Len_j is the length of document *j*
- * *maxlen* is the length of the longest document in the document set

Factoring $Info_j$ into the recommendation decision, documents with more information will be recommended first while other conditions are same. As a result, the technical report in Example 3 will be suggested to Jack first. Then, the short paper won't appear in the list, since its novelty is 0 relative to the detailed version, as illustrated in the following part.

To reach the definition of the novelty of a document, the preference and knowledge of a user must be strictly defined.

Definition 3: Suppose *i* is a user and *j* is a document, we define $prefer_{ij}$, the preference of user *i* to document *j*, as follows:

$$prefer_{ij} = \begin{cases} 1, & D_{ij} = 1 \vee t_{ij} > \delta_1 \\ \dfrac{t_{ij}}{Len_j}, & \text{O therwise} \end{cases}$$

where:

- D_{ij} is set to 1 when user i have downloaded document j, otherwise, D_{ij} is set to 0
- t_{ij} is the time(in seconds) spent by user i in browsing document j
- δ_1 is a threshold for browsing time
- Len_j is the length of document j, represented by the number of words in j

To ease the analysis, $prefer_{ij}$ is normalized.

$$prefer_{ij} = \begin{cases} 1, & D_{ij} = 1 \vee t_{ij} > \delta_1 \\ \dfrac{t_{ij}}{Len_j} \Big/ \left(1 + \dfrac{t_{ij}}{Len_j}\right), & \text{Otherwise} \end{cases} \qquad (2.3)$$

Definition 4: Suppose i is a user, the *known knowledge of user i*, $Knowledge_i$, can be presented as follows:

$$Knowledge_i = \{ j \mid prefer_{ij} = 1 \} \qquad (2.4)$$

Intuitively, if a document has been downloaded by i, or was browsed by i for a period of time longer than a certain threshold, we consider the document belongs to i's known knowledge.

The concept of novelty is defined based on the user's knowledge base. For a given document, its relative novelty to each document in user's knowledge base will be computed, and the minimum value will be assigned as its novelty to the user.

Definition 5: The *novelty of document j relative to user i*, denoted by $novu_{ij}$, is represented by:

$$novu_{ij}\square\ \mathbf{min}\ \{novr_{jj`}\},\ \forall j`\in Knowledge_i, \qquad (2.5)$$

where $Knowledge_i$ is the known knowledge of user i.

For two documents j and $j`$, the *novelty of j relative to j`*, represented by $novr_{jj`}$, can be calculated by the following equation:

$$novr_{jj`} = \left(\sum_{k\in j} w_{jk} - \sum_{k\in j \wedge k\in j`} w_{jk}\right) \times e^{-t}$$

where:

- k is a term, $k\in j$ indicates that k is in document j
- t is a time interval, representing how long ago the document was published
- w_{jk} is the weight of term k in document j, which is calculated with the following equation

$$w_{jk} = \frac{tf_{jk}}{maxfre_j} \times \frac{\log(N/df_k)}{\log(N)} \times p_{jk}.$$

Here, tf_{jk} is the occurrence frequency of term k in document j, $maxfre_j$ is the highest term occurrence frequency in document j. p_{jk} is the position weight of term k in document j. N and df_k are already explained in definition 1.

With the introduction of novelty information, the recommendations will be better tailored to user's need. For example, two same papers will only be recommended once, an older edition won't be selected after a newer version has been read or downloaded.

Now we reach the point to integrate the aforementioned three factors together to direct the personalized recommender system.

Definition 6: Suppose i is a user, and j is a document, interest$_{ij}$, the degree of interest of user i to document j, is defined as follows:

$$interest_{ij} = \begin{cases} \max\limits_{j`\in knowledge_i} \{Rsim_{jj`}\} \times Info_j, & novu_{ij} > \eta \\ 0, & \text{Otherwise} \end{cases} \qquad (2.6)$$

where η is the threshold of novelty.

By this definition, it's easy to see that documents that do not contain enough novel information to the user will be filtered out in advance. This will rule out repeated recommendations. Among the rest of documents, those with higher similarity and larger information amount will get recommended with higher priority. To put it in another way, the degree of interest will increase when similarity is enlarged. Similarly, the increment of information amount will also lead to a larger value of the degree of interest.

Theoretical Analysis of Interest-based Approaches vs. Similarity-based Approaches: The theoretical analysis was carried out to compare the Interest- based approaches with Similarity-based approaches. Both recommendation precision and recommendation recall will be studied. In the following, some preliminaries are provided first.

The personalized recommendations for a given user i will be chosen from the document set S. Let the total number of documents in S be n, i.e. $|S| = n$. S can be rewritten as $S = S_1 \cup S_2$, where S_1 is the subset containing all the documents that do not match with the user's interest, and S_2 contains documents in which i is interested. There are various reasons which make a document belong to S_1. Examples include the content of a document is covered by another document; the document is the same as another document; or the document is very similar with another document. Therefore, we further divide S_1 into two parts, $S_1 = S_{11} \cup S_{12}$, where:

$S_{11} = \{\omega | (\exists \upsilon_1 \in S_2)$ (ω is covered by υ_1)$\vee$
$(\exists \upsilon_2 \in S_2)$ $(RSim_{\omega\,\upsilon2} > \lambda) \vee$
$(\exists \upsilon_3 \in S_2)$ (ω and υ_3 are same),
$\omega, \upsilon_1, \upsilon_2, \upsilon_3 \in S$,
λ is a threshold for similarity$\}$,
$S_{12} = S_1 - S_{11.}$

Then, let us be more precise about S_{11}. We can group S_{11} by the document in S_2, that is, a group in S_{11} may only be composed of those documents which are uninteresting to the user because of the same document in S_2. Formally,

$S_{11\upsilon} = \{\omega | (\exists \upsilon \in S_2)$ (ω is covered by υ $\vee$
$RSim_{\omega\,\upsilon} > \lambda \vee \omega$ and υ are same), ω,
$\upsilon \in S$, λ is a threshold for similarity$\}$.

Suppose, there are q such groups in S_{11}, and that on average each group has t documents, then S_{11} contains qt documents in total.

If the total number of interested documents to a user in S is h, the total number of documents in the recommendation set is p, and the number of correct recommendations is x, then the recommendation recall is $\mu = \frac{x}{h}$, and recommendation precision is $\nu = \frac{x}{p}$.

To compare the interest-based and similarity-based approaches, we assumed:

- * *Intr* and *Simr* are the two recommendation sets generated by interest-based approaches and similarity-based approaches, respectively.
- * The size of each set is the same, namely τ
- * And, γ out of n documents in S are of real interest to the user

Then, the recommendation recall and recommendation precision could be calculated as follows:

According to the definition, the recall of similarity-based recommendation approach is:

$$\mu_{simr} = \frac{x_{simr}}{h} = \frac{| Simr \cap S_2 |}{\gamma} \qquad (3.1)$$

To have a closer look, $Simr = Simr \cap S = Simr \cap (S_{11} \cup S_{12} \cup S_2) = (Simr \cap S_{11}) \cup (Simr \cap S_{12}) \cup (Simr \cap S_2)$. Since S_{11}, S_{12}, and S_2 are disjoint, $Simr \cap S_{11}$, $Simr \cap S_{12}$, and $Simr \cap S_2$ are disjoint. Thus,

$$|Simr \cap S_2| = |Simr| - |Simr \cap S_{11}| - |Simr \cap S_{12}| \qquad (3.2)$$

The average number of documents in $Simr \cap S_{11}$ is $\tau \frac{qt}{n}$. Suppose the average number of documents in $Simr \cap S_{12}$ is β. Putting these numbers back into equation (3.2), we get

$$x_{simr} = \tau - \tau \frac{qt}{n} - \beta$$

In the same manner, we can compute

$$\mu_{\text{int}\,r} = \frac{x_{\text{int}\,r}}{h} = \frac{| Intr \cap S_2 |}{\gamma}$$

Here, $Intr \cap S_{11}$ is an empty set, since documents in S_{11} will have a very low novelty and the interest-based approaches can filter them out. Still, we assume $| Intr \cap S_{12} | = \beta$. So $x_{intr} = \tau - \beta$.

It's easy to see that $x_{simr} \le x_{intr}$, so $\mu_{simr} \le \mu_{intr}$. Similarly, $v_{simr} \le v_{intr}$ can be derived. Hence, we can conclude that the interest-based approaches will reflect user's interest better than the similarity-based approaches.

RESULTS AND DISCUSSION

It was reported that a series of experiments designed to i) verify the relationship between document similarity and document novelty, and the relationship between document similarity and the information amount of a document, in order to manifest the necessity of introducing the degree of interest into the recommendation decision; ii) demonstrate that our proposed interest-based approaches outperform the similarity-based approaches and the graph-based approaches in both the recommendation recall and recommendation precision.

The experiments are performed on a PC with 2GHz CPU and 256MRAM, running Windows2000 operating system. The recommender system is implemented on top of the Unlimited Digital Library, which was developed in Harbin Institute of Technology (HIT) in China [11-13]. The system includes modules for extracting the document description information, obtaining user profiles, storing and analyzing query logs, clustering users based on their profiles, and optimizing query processing plans. Supposedly, the interest area of all the recommendations presented in this paper is about computer science related papers written in English. There are 4825 papers and 30 users in the system.

Before we proceed, it's worth mentioning that documents in the digital library are first preprocessed based on user's interest area and the user-to-user similarity. So, after the preprocessing, only a subset of the original digital library is left for recommendation consideration. All the descriptions here are about the procedures after the preprocessing.

Relationships between Document Similarity and Document Novelty and Information Amount of a Document: There are two sets of experiments in this testing. Similarities between different documents are calculated based on equation (2.1). These values are then sorted in decreasing order.

For the first set of experiment, 100 pairs of documents are randomly selected among those with similarity greater than 0.5, and their similarity and novelty values are plotted in Fig. 1. The results show that when the similarity is 1, the novelty is 0; when similarity is decreasing, the novelty is increasing. These documents with higher similarity are either duplicate documents or different versions of the same document coming from different data sources. It's clear that, if we make recommendations only based on similarity, documents with the same or almost same content will be chosen repeatedly.

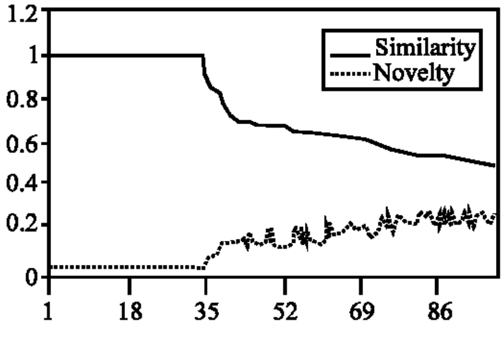

Fig. 1: Relationship between Similarity and Novelty

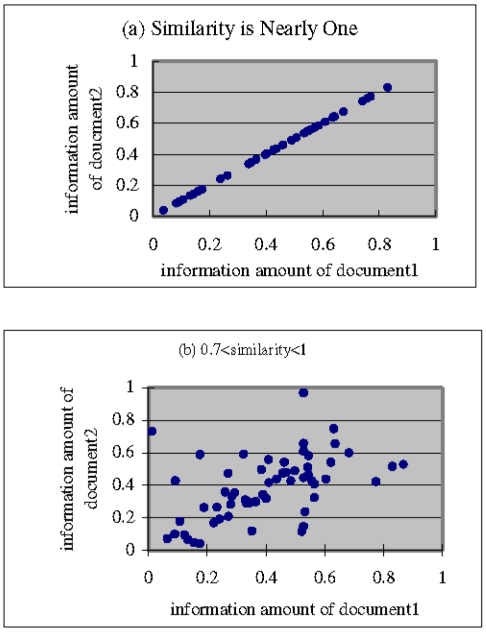

Fig. 2: Relationship between Similarity and Information Amount

The second set of experiment is designed to verify the relationship between document similarity and information amount contained in a document. The top 100 pairs of documents are selected from the sorted list. Fig. 2 presents relationship between similarity and information amount. When the similarity is 1 or close to 1, the amount of information of the two documents is almost the same, shown in Fig. 2a. This could result from the fact that either the same documents are from different data sources, or different versions of the same document. Fig. 2b shows the different amount of information when the similarity is between 1 and 0.7.

To summarize, the above experiments lead to the following conclusions: when two documents are very similar, their relative novelty is very low, and the information amount contained by them are almost the same. Thus, taking only similarity as the recommendation determinant factor cannot give a satisfactory result.

Evaluation of Recommendation Precision and Recommendation Recall: It was compared that our algorithm with two other existing algorithms. One is graph-based approach proposed by Huang[10], the other is a similarity-based approach generated by removing the novelty and information amount factors from our interest-based approach. As mentioned before, there are 30 users in the systems, and we perform two sets of experiments based on the system logs. And the data presented are average over all users.

In the first set, the users are in charge of justifying the recommendation correctness, that is, after getting the recommended documents, each user will tell what are the right recommendations which meet his/her needs. In the experiment, each user is provided with 50 recommended documents. The resulting data are grouped by the number of documents in his/her known knowledgebase. The rationale behind this is, the more the system knows about a user, the more likely a right recommendation will be chosen for the user. Fig. 3 presents this group of data. We can see that our approach can always provide more accurate recommendations than the other two methods. Furthermore, by increasing the number of documents browsed or downloaded by a user, all approaches can predict more accurately.

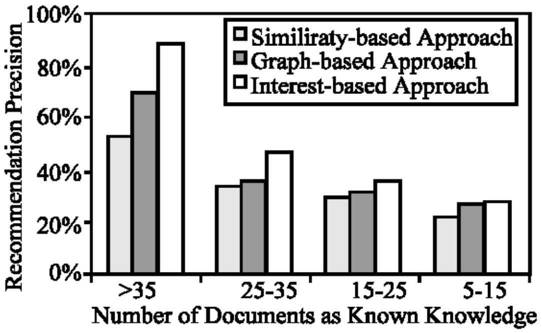

Fig. 3: Comparison of the Recommendation Precision over the Three Approaches

As part of the first set of experiments, we also studied how the recommendation precision is affected by the number or order of recommendations. We calculated the different recommendation precision based on the number of recommended documents, namely 2, 5, 10, 20, 30 and 50, as shown in Fig. 4. Our algorithm can get 30% improvement over the other methods when the recommendation number is 2. And for the recommendation number 5, there are 36% and 40% improvement over the other two, respectively. The reason our method can outperform others is simply because we introduced the novelty and information amount into the recommendation considerations, which can help filter out repetitive recommendations. It's also not hard to tell that users are more likely to be interested in the documents which are closer to the top of the list.

In the second set of experiments, the accessing log of a user is divided into two parts based on time. The earlier happened events are used to generate recommendations, as testing data to deduce the correct and total recommendations for this user. We first recommended 5 documents for each user, then computed the recall and precision for individual user based on the inferred values. The precision and recall shown in Fig. 5 are the averages over those for individuals. Again, the results demonstrate that the interest-based recommendation method proposed in this paper has a better performance than other approaches.

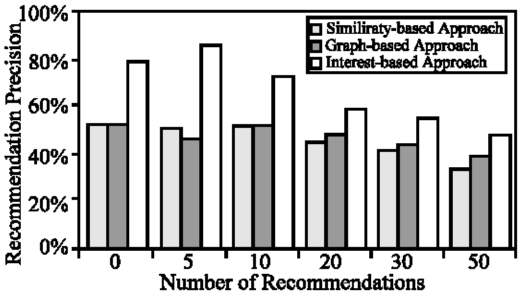

Fig.4: Recommendation Precision is affected by the Number of Recommendations

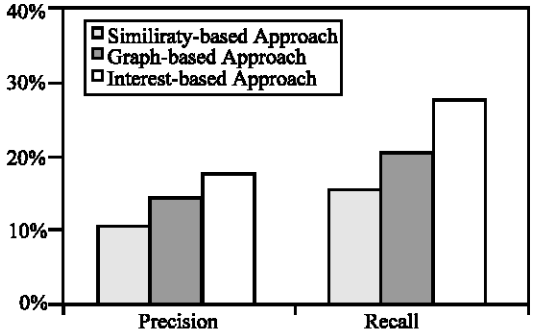

Fig. 5: Comparison of Both Precision and Recall

CONCLUSION

This study presents a novel method for personalized recommendation, namely interest-based approach. We introduce the concept of *degree of interest*, which makes three factors, similarity, novelty and information amount, being integrated together to provide more accurate and complete recommendations. Theoretical analysis and experimental results show that the interest-based recommendation approach can generate more precise and complete recommendations to meet the users' real need.

REFERENCES

1. Mladenic, D., 2000. Machine learning for better Web browsing. AAAI 2000 Spring Symposium Technical Reports on Adaptive User Interfaces.
2. Lee, C. Giles, Kurt D. Bollacker and Steve Lawrence, 1998. CiteSeer: An Automatic Citation Indexing System. In proceeding of the third ACM conference on digital libraries.
3. Asnicar, F. and C. Tasso, 1997. ifWeb: a prototype of user modelbased intelligent agent for documentation filtering and navigation in the World Wide Web. In proceedings of the UM 1997 Workshop on Adaptive Systems and User Modeling on the World Wide Web.
4. Sarwar, B.M., G. Karypis and J.A. Konstan, *et al.* 2000. Analysis of recommendation algorithms for e-commerce. In proceedings of the ACM Conference on Electronic Commerce.
5. Breese, J.S., D. Heckerman and C. Kadie, 1998. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence.
6. Mukund Deshpande and George Karypis, 2003. Item-Based Top-N Recommendation Algorithms. Technical Report. University of Minnesota.
7. Ahmad, M. and A. Wasfi, 1999. Collecting User Access Patterns for Building User Profiles and collaborative Filtering. In Proceedings of the 1999 International Conference on Intelligent User Interfaces.
8. Konstan, J., B. Miller, D. Maltz, J. Herlocker, L. Gordon and J. Riedl, 1997. GroupLens: Applying collaborative filtering to Usenet news. Communications of the ACM, 40: 77-87.
9. Balabanovic, M. and Y. Shoham, 1997. Fab: Content-based, collaborative recommendation. Communications of the ACM, 40: 66-72.
10. Huang, Z., W. Chung, T. Ong and H. Chen, 2002. A Graph-based Recommender System for Digital Library. In proceeding of the second ACM/IEEE-CS joint conference on digital libraries.
11. Yan Yang, Baoliang Liu and Zhaogong zhang, 2003. Partition Based Hierarchical Index for Text Retrieval. In proceedings of the 4-th International Conference of Web-Age Information Management.
12. Yang Yan and Li Jianzhong, 2003. Cluster-based Data Allocation Method of Web Servers in Digital Library. Computer Engineering and Applications, 39: 38-41.
13. Yan Li, Jianzhong Li and Yan Yang, 2002. Query optimization model based on cost in parallel document database. Computer Sci., 29: 255-257.