A prototype WWW literature recommendation system for digital libraries

San-Yih Hwang
Wen-Chiang Hsiung and
Wan-Shiou Yang

The authors

San-Yih Hwang, **Wen-Chiang Hsiung** and **Wan-Shiou Yang** work at the Department of Information Management, National Sun Yat-sen University, Taiwan.

Keywords

Digital libraries, Recommendations, Web sites, Cluster analysis, Data mining

Abstract

This article describes a service for providing literature recommendations, which is part of a networked digital library project whose principal goal is to develop technologies for supporting digital services. The proposed literature recommendation system makes use of the Web usage logs of a literature digital library. The recommendation framework consists of three sequential steps: data preparation of the Web usage log, discovery of article associations, and article recommendations. We discuss several design alternatives for conducting these steps. These alternatives are evaluated using the Web logs of our university's electronic thesis and dissertation (ETD) system. The proposed literature recommendation system has been incorporated into our university's ETD system, and is currently operational.

Electronic access

The Emerald Research Register for this journal is available at
http://www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/1468-4527.htm

Online Information Review
Volume 27 · Number 3 · 2003 · pp. 169-182
© MCB UP Limited · ISSN 1468-4527
DOI 10.1108/14684520310481436

Introduction

This is a fascinating period in the history of libraries and publishing. For the first time – with the advances in computing and networking techniques – it is possible to build large-scale services where collections of information are stored in digital formats and accessed over networks (Arms, 2000). Libraries are making good use of digital capacity to provide services such as information seeking and filtering (Furner, 2002; Spink *et al.*, 2002), organising (Arms, 2000), and delivering (Kessler, 1996). Systems intended to provide such digital services are appearing accordingly and are being investigated in digital library environments (Andresen *et al.*, 1995).

This paper reports on a networked digital library project at National Sun Yat-sen University in Taiwan. The project, whose principal goal is to develop technologies for supporting digital services, is a series of three investigations, sponsored by the National Science Council and the National Central Library. The first stage involved the design and construction of a literature recommendation system. The second investigation focuses on the integration of various information sources, and the third investigation addresses the representation and retrieval of multimedia content. The progress of the first investigation is reported in this paper.

The literature recommendation system aims at recommending relevant articles to researchers and library patrons. The system adopted a WWW framework so that subscribers can access the system without time and location constraints, and so that the task of service spreading can be facilitated by a common browsing interface. The core of the literature recommendation system is a recommender mechanism, which analyses literature usage so that publications can be ranked according to the preferences of an active user. Various characteristics of publications and WWW interactions are taken into account, and the endeavour has

Refereed article received 26 February 2003
Accepted for publication 4 April 2003
This work is supported in part by the National Science Council of Taiwan under grant no. NSC91-2413-H-110-004-003 and the National Central Library of Taiwan under grant No. 91-A-015.

resulted in a recommendation system that is particularly suitable for recommending literature in digital library environments.

Related work

Interest in digital libraries has increased tremendously, with several research projects addressing the wealth of challenges in this field. For example, a University of Illinois project has focused on providing integrated access to diverse and distributed collections of scientific literature (Chen *et al.*, 1996). That project deals with heterogeneous interfaces to multiple indices, semantic federation across repositories, and other related issues. A group at the University of California at Berkeley is working on providing work-centred digital information services (Wilensky, 1996). The issues involved include document image analysis, natural language analysis, and computer vision analysis for effective information extraction. Carnegie Mellon University intends to build a large online digital video library featuring full-content and knowledge-based searching and retrieval. The University of California at Santa Barbara has concentrated on geographical information systems, and a Stanford University project addresses the problem of interoperability using CORBA to implement information-access and payment protocols (Baldonado *et al.*, 1997).

The focus of our research reported here is to tackle the problem of information overload. Proposed solutions to this emphasise the need for specialisation in information retrieval services, to help people effectively locate information that meets their individual needs (Bowman *et al.*, 1994). Interest in recommending has increased in the information technology community, and especially in the design of digital libraries (Furner, 2002). The research reported here concentrates on literature recommendations.

The past few years have seen the emergence of many recommendation systems intended to provide personal recommendations for various types of products and services, including:

- news and e-mail messages (see www.netperceptions.com/ for a commercial site, and Goldberg *et al.* (1992), Lang (1995), Konstan *et al.* (1997), and Billsus and Pazzani (1999) for research prototypes);
- Web pages (see http://my.yahoo.com/ for a commercial site, and Balabanovi'c and Shoham (1997), Terveen *et al.* (1997), Pazzani and Billsus (1997), and Armstrong *et al.* (1997) for research prototypes);
- books (see http://www.amazon.com/ for a commercial site, and Mooney and Roy (2000) for a research prototype);
- music (see http://www.CDNow.com/ for a commercial site, and Shardanand and Maes (1995) for a research prototype); and
- movies (see http://movies.eonline.com/ for a commercial site, and Alspector *et al.* (1998), Breese *et al.* (1998), Basu *et al.* (1998), Ansari *et al.* (2000), Pennock *et al.* (2000), and Schafer *et al.* (2001) for research prototypes).

The first type of recommendation technique was called the content-based approach (Loeb and Terry, 1992). A content-based approach characterises recommendable items by a set of content features and represents a user's interests by a similar feature set. Then, the relevance of a given content item to the user's interest profile is measured as the similarity of this recommendable item to the user's interest profile. Content-based approaches select recommendable items that have a high degree of similarity to the user's interest profile.

Another type of recommendation technique, the collaborative approach (sometimes called the social-based approach), takes into account the given user's interests profile and the profiles of other users with similar interests (Shardanand and Maes, 1995). The collaborative approach looks for relevance among users by observing their ratings assigned to products in a training set of limited size. The "nearest-neighbour" users are those that exhibit the strongest similarity to the target user. These users then act as "recommendation partners" for the target user, and collaborative approaches recommend to the target user items that appear in the profiles of these recommendation partners (but not in the target user's profile). It has been observed in several practical settings that the collaborative approach generally achieves more effective recommendations than its content-based counterpart (Alspector *et al.*, 1998; Breese *et al.*, 1998; Mooney and Roy, 2000; Pazzani, 1999).

Pennock *et al.* (2000) proposed using a collaborative approach for recommending articles in CiteSeer (www.citeseer.com/). Their approach implicitly derives users' ratings of articles by observing their actions when viewing an article. Each action is assigned a weight. For example, adding a document to a profile produces a two-point increment, downloading a document a one-point increment, and ignoring a recommendation a one-point decrement in the rating score. However, this approach suffers from the shortage of negative examples, and the method is applicable only to individual members who are willing to identify themselves each time they use the digital library.

We consider that traditional recommendation techniques are not suitable for recommending articles in digital libraries. First, both content-based and collaborative approaches require that users' rating scores on selected items (including both positive and negative instances) are available for analysis. For a typical literature digital library, requiring users to rate some articles before making a recommendation is not realistic. Second, identifying an individual user of a literature digital library is generally not possible, since many literature digital libraries are freely available on the Internet and users can search or browse articles without having to identify themselves. Even for proprietary literature digital libraries, many users gain access via site subscriptions, making it difficult to track an individual's (long term) browsing behaviour.

For the reasons mentioned above, we propose making use of a task-focused approach (Herlocker and Konstan, 2001). In this approach, a task profile (a set of recently accessed items) rather than the long-term interest profile is used to make recommendations. One notable implementation of this approach is Web usage mining, which aims to identify interesting usage patterns of a Web-based system from the Web usage logs that record interactions between users and Web pages (Srivastava *et al.*, 2000). Recently, several approaches have been proposed for recommending Web pages based on the Web page associations discovered by Web-usage mining algorithms (Yan *et al.*, 1996; Mobasher *et al.*, 1999; Pitkow and Pirolli, 1999; Yang *et al.*, 2001). While these approaches vary in their details, they follow the same recommendation framework, which starts with the identification of aggregate usage profiles of Web pages by some data mining method. They then make recommendations by looking into the similarity between the set of recently accessed Web pages of an active user and the collected aggregate usage profiles.

Obviously, literature digital libraries store articles rather than Web pages, and they differ from Web pages in several respects:

- Web pages are more diversified: some serve as index pages, some are content pages, and others have a mixture of indexes and content. On the other hand, since literature articles are more homogeneous in structure they are more likely to have the same set of metadata features.
- A Web site can be viewed as a directed graph whose vertices are Web pages, while a literature digital library is better visualised as a set of articles.
- Literature articles are often retrieved by search queries provided by the system, while Web pages are often browsed through a static site topology.
- Literature articles are incrementally inserted into the digital library at a faster rate than are Web pages inserted into a Web site.

Contributions

The above considerations indicate that literature recommendation services require a different technical approach. Here we describe a recommendation framework for recommending articles in a literature digital library. Several alternatives are proposed for implementing the constituent components of the recommendation framework. These alternatives are compared and analysed by applying the Web usage logs collected from the electronic thesis and dissertation (ETD) system at National Sun Yat-sen University (NSYSU-ETD). We have incorporated the proposed recommendation framework into NSYSU-ETD (www.lib.nsysu.edu.tw/eThesys/english/default_e.htm); the corresponding implementation status is also reported.

This paper is structured as follows. Next the overall architecture is described, followed by detailed design and construction methods. Then implementation experience and evaluation results are presented. The final part summarises this paper and discusses our future research directions.

Architecture

The overall architecture of the literature recommendation system, shown as Figure 1, consists of two basic subsystems, offline and online:

(1) The offline subsystem comprises two sequentially executed tasks: data preparation and log mining. Although Web usage logs are potentially able to provide useful knowledge for making recommendations, the raw log data cannot be used before appropriate pre-processing. Therefore, we first convert raw Web usage logs into a set of user transactions before performing the log mining task. The objectives of log mining tasks include the discovery of article association rules and the derivation of article clusters.

(2) The online subsystem interacts with an active user and provides recommended articles in real time. It keeps track of a set of articles browsed recently by the active user by consulting the current Web usage log provided by the Web server. Then, by comparing the similarities between this set and the article clusters (or associations) produced by the offline subsystem, the online subsystem recommends articles in the clusters (associations) that are highly similar.

We have incorporated the literature recommendation system into NSYSU-ETD. Figure 2 depicts a page view of an article in NSYSU-ETD. The Web page comprises two frames: the left frame displays the metadata of the browsed article and the right frame shows the unseen articles (up to 15) recommended by our literature recommendation system, displayed in the order of relevance. In this example the active user had reviewed two articles, and the literature recommendation system has recommended another seven articles. Once the active user browses another article, the content of the recommendation frame will update accordingly.

Approaches

In this section, we first describe the tasks conducted in the offline subsystem of the literature recommendation system, and then those conducted in the online recommendation subsystem.

Figure 1 Architecture of a literature recommendation system

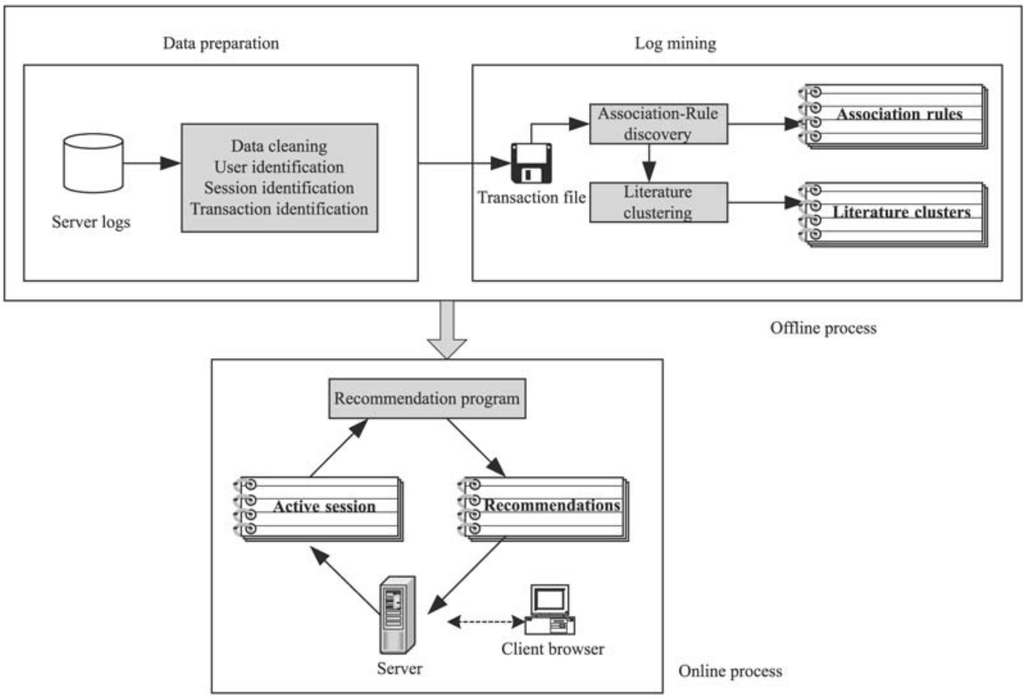
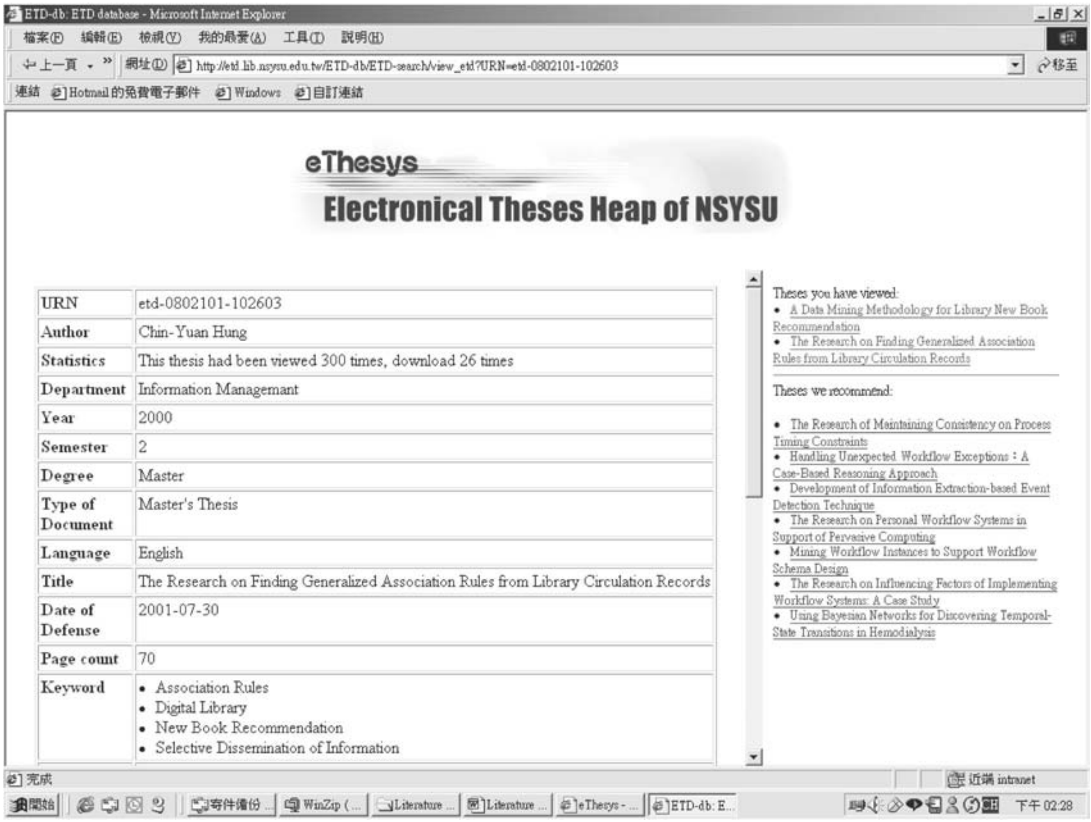


Figure 2 Page view of an article in NSYSU-ETD


ETD-db: ETD database - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
網址(D) http://etd.lib.nsysu.edu.tw/ETD-db/ETD-search/view_etd?URN=etd-0802101-102603
移至
連結 Hotmail 的免費電子郵件 Windows 自訂連結
eThesys
Electronical Theses Heap of NSYSU
完成
近端 intranet
開始
寄件備份 WinZip Literature Literature eThesys ETD-db: E...
下午 02:28

URN	etd-0802101-102603
Author	Chin-Yuan Hung
Statistics	This thesis had been viewed 300 times, download 26 times
Department	Information Managemant
Year	2000
Semester	2
Degree	Master
Type of Document	Master's Thesis
Language	English
Title	The Research on Finding Generalized Association Rules from Library Circulation Records
Date of Defense	2001-07-30
Page count	70
Keyword	• Association Rules • Digital Library • New Book Recommendation • Selective Dissemination of Information

Theses you have viewed:

- A Data Mining Methodology for Library New Book Recommendation
- The Research on Finding Generalized Association Rules from Library Circulation Records

Theses we recommend:

- The Research of Maintaining Consistency on Process Timing Constraints
- Handling Unexpected Workflow Exceptions：A Case-Based Reasoning Approach
- Development of Information Extraction-based Event Detection Technique
- The Research on Personal Workflow Systems in Support of Pervasive Computing
- Mining Workflow Instances to Support Workflow Schema Design
- The Research on Influencing Factors of Implementing Workflow Systems: A Case Study
- Using Bayesian Networks for Discovering Temporal-State Transitions in Hemodialysis

Data preparation for the literature usage log

To prepare data from the Web usage logs of a literature digital library, we basically follow the heuristics adopted by Cooley *et al.* (1999) for processing Web usage logs that involve static Web pages. In their work, the Web usage logs are assumed to be in the extended NCSA format (including referrer and agent fields). The approach contains three sequential steps: data cleansing; user session identification; and transaction identification. The objective of data cleansing is to prune out unwanted Web log records and to add back missing Web log records: some Web log records are surplus as they are accesses to non-HTML pages (e.g. images and other http requests involving no Web page accesses), while other Web log records are missing due to the existence of the local cache, firewalls, and proxy servers. Identifying missing Web log records is especially difficult – several heuristics have been proposed for achieving this. However, we found this difficulty nonexistent when processing the Web usage log of NSYSU-ETD because article Web pages are dynamically generated and are not cacheable. Our university ETD system is database driven in that the theses metadata are stored in a DBMS. Most large-scale digital libraries adopt the same method for maintaining their collections. In the context of a literature digital library, we are concerned with and retain only the Web usage records that involve the following two types of accesses:

(1) *Lookup accesses.* Each lookup access is an execution of a CGI program with searching or browsing conditions specified in the parameters. An example lookup access of NSYSU-ETD is: http://etd.lib.nsysu.edu.tw/ETD-db/ETD-search/search_by_advisor?advisor_name=San-Yih+Hwang, which lists all theses supervised by Professor San-Yih Hwang.
(2) *Article accesses.* Each article access is an execution of a CGI program that displays the detailed metadata on an article. An example lookup access of NSYSU-ETD is: http://etd.lib.nsysu.edu.tw/ETD-db/ETD-search/view_etd?URN=etd-0726100-135739, which shows the metadata of the thesis whose URN is etd-0726100-135739.

Note that article accesses display the detailed metadata of articles and therefore are of primary interest. Lookup accesses provide lookup information to facilitate browsing or searching and can thus be considered auxiliary. Most of the time a user will first execute a lookup access, followed by a selective list of article accesses. We found that some user sessions contained article accesses without prior lookup accesses in the Web usage log of NSYSU-ETD. This is because several information sources had provided hyperlinks directly to articles' page views. In this case, each session can be viewed as a list of queries. Each query (optionally) starts with a lookup access, followed by a list of article accesses related to the articles that the user chose to look at in more detail.

The goal of user session identification is to divide the article accesses of each user into individual sessions. It is reasonable to assume that two records with different IP addresses, browsers, or operating systems belong to two different user sessions. In addition, the time interval between two consecutive requests in a user session should not be too large. As in many commercial products, we use 30 minutes as the default timeout period. When the time interval between the current access and the previous one exceeds this, a new user session is assumed to have started. Some of the identified user sessions are made by Internet robots and hence should not be considered. Some robots have known agent types and/or IP addresses, and can be easily identified. Analysis of the user sessions of these known robots revealed that most of these sessions either have more than 100 article accesses or exhibit a mean adjacent Web page access interval of less than three seconds. User sessions that satisfy this condition are considered as robot sessions, and consequently are removed.

Finally, a user session is further divided into a number of transactions, each of which represents a semantically meaningful unit. However, the various transaction identification approaches proposed in Cooley *et al.* (1999) make use of either the index (auxiliary) pages or the Web site topology. Since neither exists in the context of literature digital libraries, these proposed approaches are not applicable. Our approach identifies transactions by considering the types of accesses, namely lookup and article accesses. In fact, articles listed by the same query must have some degree of similarity in their content (e.g. keyword, title, author, discipline). On the other hand, articles selected in the same user session or query also display some degree of similarity, due to inherent human behaviour. Therefore, we have four methods for defining transactions:

(1) Query-chosen method: the articles selected in a query.
(2) Session-chosen method: the articles selected in a user session.
(3) Query-result method: the articles listed in a query.
(4) Session-result method: the articles listed in queries of a user session.

For the query-chosen and session-chosen methods, article accesses present in the Web logs are grouped into a set of transactions. For the query-result and session-result methods, we construct transactions by reissuing queries to the literature digital library.

As mentioned, the query-chosen and session-chosen methods incorporate knowledge on human selection in making the recommendations. We expect that they will yield more effective recommendations than their counterparts without such knowledge, namely the query-result and session-result methods. We therefore form the following hypothesis:

H1. Users tend to browse the metadata of only the articles they find of interest. Recommendation schemes that consider the Web accesses of these articles will result in more effective recommendations.

Mining the literature usage log

There have been several approaches proposed in the literature (Yan *et al.*, 1996; Mobasher *et al.*, 1999; Pitkow and Pirolli, 1999; Yang *et al.*, 2001) for identifying aggregate usage profiles from Web usage logs. Aggregate usage profiles can be represented in the form of association rules, sequential patterns (or an n-gram Markov model), or clusters of Web pages. In the context of literature digital libraries, however, we decided not to consider sequential patterns because the order of articles in a transaction may not relate to users' preferences. Instead, only association rules and clusters of articles will be discussed.

The problem of finding frequent associations between items in a transaction database, called the association-rule discovery problem, was first introduced by Agrawal *et al.* (1993). Association-rule discovery methods, such as the A priori Algorithm (Agrawal *et al.*, 1993; Agrawal and Srikant, 1994), are typically based on two decompositions: extraction of itemsets and the generation of strong association rules. In the initial extraction phase, the methods find sets of items that frequently occur together. The support of each itemset $I = \{XY\}$, denoted as $\mathrm{Sup}(I)$, is the fraction of transactions containing both X and Y. This itemset I is referred to as a frequent itemset if $\mathrm{Sup}(I)$ exceeds a user-specified minimal support threshold Min_{sup}.

In the second phase, each discovered itemset $I = \{XY\}$ is used to construct association rules in the form $X \Rightarrow Y$. The confidence of each rule, denoted as $\mathrm{Conf}(X \Rightarrow Y)$, is the fraction of transactions containing X that also contain Y. An association rule is said to be strong if it exceeds a user-specified confidence threshold. By applying association-rule discovery algorithms to the transactions of the transformed logs, we can find association rules in the form $\{a_1, a_2, \ldots, a_{\mathrm{m}}\} \Rightarrow a$, which can be used to recommend article a to users who have browsed $\{a_1, a_2, \ldots, a_{\mathrm{m}}\}$ but not a.

The traditional approaches for identifying itemsets with a uniform minimum support threshold, however, cannot be directly applied because articles that arrive later tend to have smaller support even if they are actually more popular. Therefore, a nonuniform support threshold scheme, originally proposed in Liu *et al.* (1999), is adopted. In this scheme, each item is assigned a distinct minimum support value (called the minimum item support, MIS), and the minimum support of an itemset is the minimum of the MIS values of its constituent items. In the literature recommendation system, we view the MIS value of an article as a function of its creation time. That is, articles that are added to the digital library more recently should be assigned smaller MIS values. Let $\mathrm{N}(t)$ be the number of transactions after time t in the Web usage logs. The MIS of an article is defined as follows:

$$MIS(a) = \begin{cases} M(a) \; M(a) > LS \\ LS \text{ Otherwise} \end{cases},$$

$$M(a) = \frac{N(CreationTime(a))}{N(0)} \cdot Min_{\mathrm{sup}},$$

where LS is the lower bound for support values, Min_{sup} is the minimum support threshold based on the entire article browsing log, and $N(0)$ denotes the total number of transactions in the Web usage log. Both LS and Min_{sup} are user-defined constants. After assigning the MIS values of all articles in the literature digital library, the method proposed in Liu *et al.* (1999) can be applied to derive the association rules for articles.

For the clustering technique, we adopt the Association Rule Hypergraph Partitioning (ARHP) approach (Mobasher *et al.*, 1999, 2000) rather than traditional clustering techniques. The main reason for this is that ARHP is more efficient in handling high dimensional data such as those present in literature digital libraries. The dimensions of a transaction are the set of articles, which is huge for a large-scale digital library. This approach starts with the identification of frequent itemsets (as in association-rule discovery methods), each of which contains articles often accessed together in transactions. Each such frequent itemset is then viewed as a hyperedge with a specific weight.

As mentioned, each article has a distinct creation time. We therefore normalise the support values of itemsets as follows before computing their weights:

$$Sup'(a_i) = \frac{N(0)}{N(CreationTime(a_i))} \cdot Sup(a_i)$$

$$Sup'(a_1, \ldots a_k) =$$

$$\frac{N(0)}{\frac{N(\max}{1 \leq i \leq k} \, CreationTime(a_i))} \cdot Sup(a_1, \ldots a_k).$$

There are several ways to define the weight of an itemset, such as using either the support for or the interest in the itemset. The former favours itemsets of smaller size, whereas the latter gives priority to larger itemsets. We define a general weighting formula that covers the broad spectrum between these two extremes. In addition, the supports for or interests in different itemsets can have very diverse values. To prevent itemsets of large weight from dominating the subsequent clustering procedure, we apply the logarithm on the weight. The following is our definition of the weight of an itemset $(a_1, a_2, \ldots, a_k)$:

$$weight(a_1, a_2, \ldots, a_k) =$$

$$\log\left(\frac{Sup'(a_1, a_2, \ldots, a_k)}{[Sup'(a_1) \cdot Sup'(a_2) \cdot \ldots \cdot Sup'(a_k)]^{\alpha}} \times \frac{1}{Min_{\text{sup}}}\right).$$

where $0 \leq \alpha \leq 1$. Note that when $\alpha = 0$ or $\alpha = 1$, this formula is equivalent to using support or interest as the weight, respectively. $1/Min_{\text{sup}}$ is a constant that keeps the weight non-negative. This definition supports our following hypotheses:

H2. A better recommendation effectiveness can be achieved by striking a balance between using support and interest as the weight of an itemset.

H3. A better recommendation effectiveness can be achieved by incorporating the logarithm function in the weight of an itemset.

After deciding the weight of each itemset, the hypergraph partitioning algorithm proposed in Karypis (2002) is applied to partition the set of articles into disjoint clusters of articles. Articles in the same cluster are more "similar" in the sense that they are more likely to be accessed together in the same transaction. To reflect the fact that an article may indeed interest more than one group of users, we adopt the same heuristic as used in Mobasher *et al.* (1999) by adding back articles to clusters, which results in overlapping clusters. Specifically, for a given hyperedge, if the percentage of involved vertices in a cluster is large than a threshold, the other involved vertices are included in the same cluster.

Online recommendations

We propose two recommendation approaches that use the article association rules and article clusters obtained by the methods described above. The goal is to recommend the top-N articles that potentially interest the active user. The first approach makes use of article association rules. The idea is to treat each frequent itemset as the interest profile of a user group and to recommend articles based on the similarity between the current session of the active user and interest profiles of the relevant user groups. Specifically, let s be the active user's current session of length k. We first identify the set of frequent itemsets of size $k+1$ that contain all elements in s and an extra element m (not in s). For each such itemset I, the confidence of the rule $\{I - m\} \Rightarrow \{m\}$ is calculated. These extra elements are then recommended to the user in descending order of confidence value. If these elements are not sufficient (i.e. there are less than N of them), we then search for frequent itemsets of size k that contain $k-1$ elements in s and an extra element (not in s). Again. these extra elements are recommended to the user in descending order of confidence value. This procedure continues until N articles are recommended.

Our other proposed method uses a hypergraph-based approach. In this approach, the recommendation score of each article a is computed by considering the similarity between the current user session and the clusters C to which a belongs, and the coherence weight of a with respect to C. Specifically, each cluster of articles can be viewed as a vector with binary elements, each of which indicates whether an article appears in the cluster. Similarly, the current user session can also be represented as a vector. Then the similarity between the current session s and a cluster C can be defined as a cosine function as follows:

$$match(S, C) = \frac{\sum_k a_k^C \times S_k}{\sqrt{\sum_k (S_k)^2 \times \sum_k (a_K^C)^2}},$$

where S_k is the k'th element in S and a_k^C is the k'th element in C.

The coherence weight of an article a with respect to the cluster C that it belongs to is defined as:

$$weight(a, C) = \frac{\sum_{a \in e, e \subseteq C} weight(e)}{\sum_{e \subseteq C} weight(e)},$$

where *weight*(e) is the weight of a hyperedge e.

The recommendation score Rec(S, a) of an article a with respect to the current user session S is then defined as:

$$\mathrm{Rec}(S, a) =$$

$$\max_{a \in C} \sqrt{weight(a, C) \times match(S, C)}.$$

The top-N articles for recommendation are those with the N highest values in the recommendation score.

The hypergraph-based approach is more carefully designed than the association-rule-based approach, and we expect the former to perform better:

H4. The hypergraph-based approach will result in more effective recommendations and has a quicker response time than the association-rule-based approach.

Empirical evaluations

This section reports our experience in applying the Web usage logs of NSYSU-ETD to the proposed literature recommendation system. The main objective was to test our four hypotheses. NSYSU-ETD runs on PC Solaris 2.7 and uses Apache 1.3.9 as the Web server. Since being commissioned in May 2000, it has been loaded with more than 3,000 electronic theses of National Sun Yat-sen University. Up to February 2003, these theses had been browsed more than 400,000 times and downloaded more than 100,000 times. We analysed the Web usage logs of NSYSU-ETD between February 2002 and May 2002 for our experiments: the data collected from February 1 to April 30 were designated as the training data set, and those collected in May served as the test data set.

We first applied the data cleansing technique on the training data, and obtained 43,349 lookup accesses and 41,627 article accesses. Applying the session identification technique revealed 16,922 user sessions, among which 392 sessions were robot generated, 6,068 sessions contained only one article access, and 5,253 sessions contained no article accesses. We eliminated these trivial user sessions and applied transaction identification techniques, resulting in 5,617 transactions for the query-chosen method, 5,272 transactions for the session-chosen method, and 17,742 transactions for the query-result method. Queries whose results are never chosen by the users are removed from the query-chosen method but remain in the query-result method. The session-result method produced transactions of huge size, each containing thousands of article accesses. We therefore decided not to consider this method in the subsequent experiments.

The two proposed methods for mining literature usage logs both require the identification of frequent itemsets from transactions, which needs the minimum support to be specified. However, the three transaction identification methods under comparison have different numbers of transactions. To be fair, we specify a different minimum support threshold for each method such that the total number of articles involved in large two-item sets of each method – called recommendable articles – is approximately the same. Our recommendation framework recommends an article only if it is associated with other articles a sufficient number of times in Web usage log. Therefore, articles that are not involved in large two-item sets cannot possibly be recommended. Table I shows the specified minimum support threshold and the number of recommendable articles of each method.

To illustrate how we conducted experiments, we define the following notation: let T_{eval} be the set of transactions in the test set, t_{eval} be a transaction in T_{eval}, and $a_t(i)$ be the *i*'th article in t_{eval}. Given a window size W_{size}, we divide each transaction t_{eval} in the test data set into two lists: $t_{\text{eval}}[W]$ and $t_{\text{eval}}[R]$, where $t_{\text{eval}}[W]$ is the first W_{size} article accesses of t_{eval}, and $\text{t}_{\text{eval}}[R]$ is the remaining articles. By treating $t_{\text{eval}}[W]$ as the current session, the recommender system will choose the set t_{pr} of top-N articles for recommendation.

The performance metric we adopted for measuring the quality of recommendation is the precision and recall scheme. The precision is the ratio of the number of recommended articles accessed by a user to the total number of recommended articles, defined as $t_{\text{pr}} \cap t_{\text{eval}}[R]/t_{\text{pr}}$, and recall is the ratio of the number of recommended articles accessed by a user to the total number of articles of interest to the user, defined as $t_{\text{pr}} \cap t_{\text{eval}}[R]/t_{\text{eval}}[R]$. The precision (recall) of a recommendation approach is the average precision (recall) of all transactions in the test set.

Test of *H1*

We first evaluate the performance impact of the three transaction identification methods. In this experiment, α was set to be 0.5 and the logarithm was taken when computing the weight of an itemset. Figure 3(a, b) shows the precisions and recalls, respectively, under association-rule-based recommendation.

Table I Minimum support and number of recommended candidates for each transaction identification method

	Minimum support (per cent)	No. of recommended articles
Session-chosen method	0.16	253
Query-chosen method	0.12	250
Query-result method	3.3	229

The precisions and recalls under hypergraph-based recommendation are shown in Figure 4(a, b), respectively.

Overall, the trends are the same for the association-rule-based and hypergraph-based approaches. It can be clearly seen that both the query-chosen and session-chosen methods outperform the query-result method, in terms of both precision and recall. This implies that the information about articles browsed plays a crucial role in making recommendations – both the query-chosen and session-chosen methods incorporate this information in forming transactions. We therefore accept *H1*. However, the performance difference between the query-chosen and session-chosen methods is not significant.

Test of *H2* and *H3*

We then conducted experiments to shed light on the impact of α and the logarithmic

Figure 3 (a) Precisions of the association-based approach; (b) recalls of the association-based approach

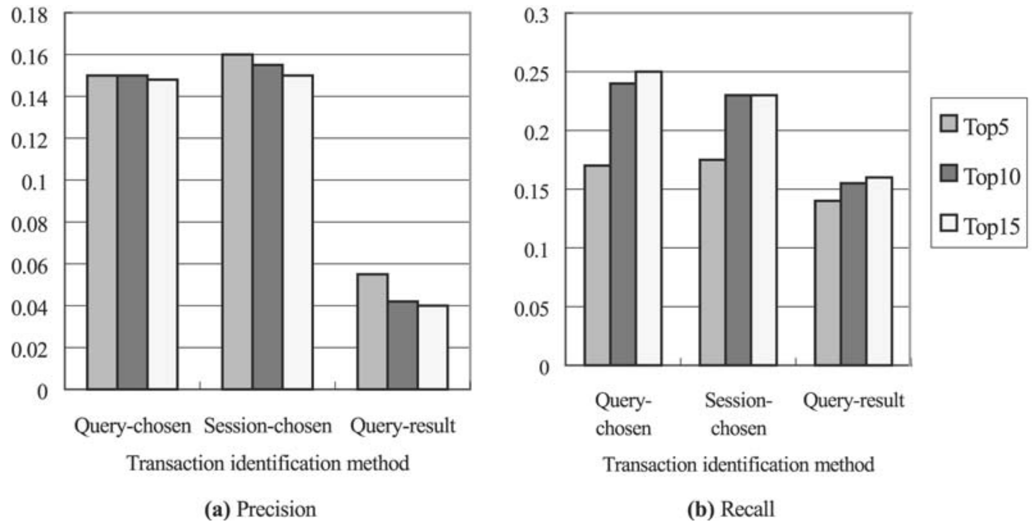
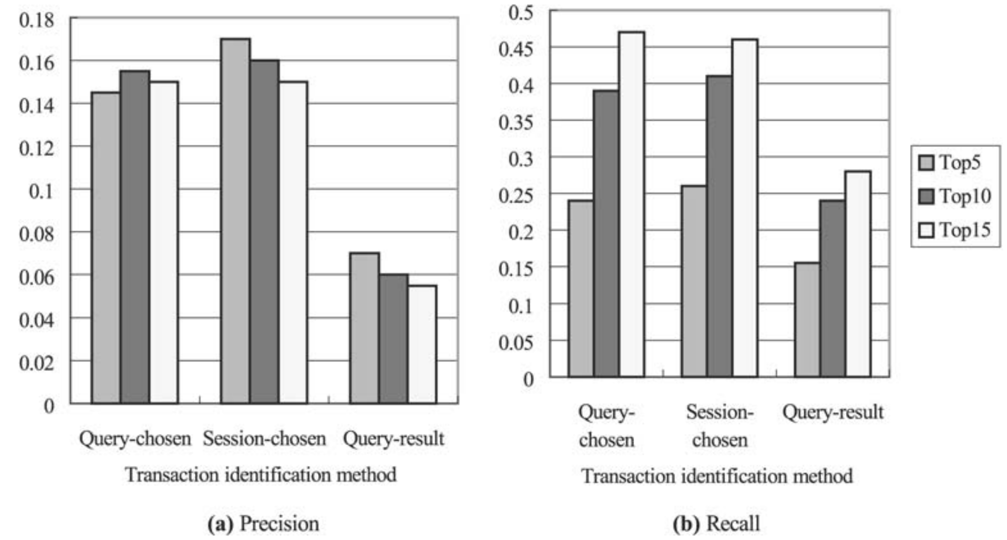


Figure 4 (a) Precisions and (b) recalls of the hypergraph-based approach under different transaction identification methods



function in the weight definition for the hypergraph-based recommendation approach. Figure 5(a, b) shows the precision and recall values when $\alpha = 0$, 0.5, and 1 using the session-chosen method for transaction identification. The window size was set at 2. As can be seen, differences in the precision and recall under different settings of α are very small. We have performed the same experiments for different transaction identification methods and window sizes, and obtained similar results. *H2* is therefore rejected.

We then computed the precisions and recalls with and without application of the logarithmic function on the itemset weight. Figure 6(a, b) shows the resulting precisions and recalls, respectively.

Figure 6 shows that applying the logarithmic function on the itemset weight definition achieves significantly better precision and recall values. This meets our expectation, and *H3* is accepted.

Test of *H4*

Finally, we evaluated the impact of association-rule-based and hypergraph-based recommendation approaches for different window sizes. Figure 7(a, b) shows the precisions and recalls for window sizes of 2, 3, 4, 5, and 6 for a top-15 recommendation. It can be seen that the hypergraph-based approach performs better than the association-rule-based approach, especially for larger window sizes.

Figure 5 (a) Impact of α on precisions and (b) impact of α on recalls of the hypergraph-based approach using the session-chosen method for transaction identification

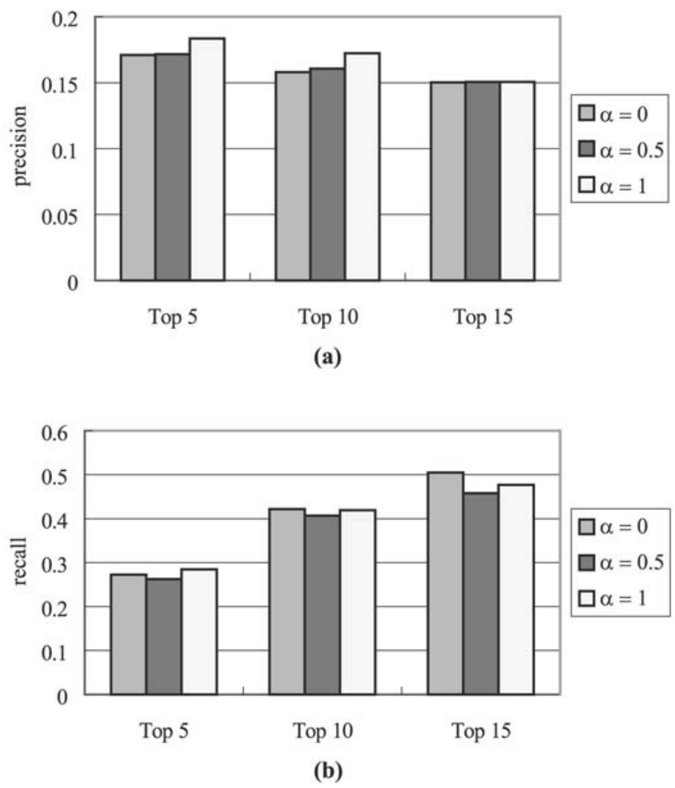
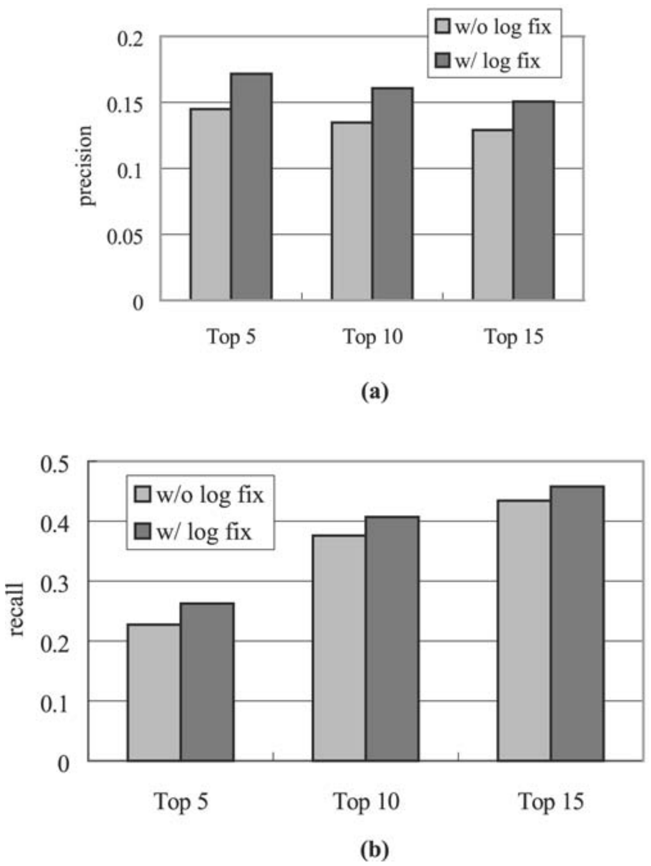


Figure 6 (a) Impact of logarithmic function on precisions and (b) impact of logarithmic function on recalls of the hypergraph-based approach using the session-chosen method for transaction identification



We also compared the running times of both approaches whilst setting different minimum support thresholds. The relative performance of the two approaches under different window sizes are shown in Figure 8(a, b). Overall, the running time of the hypergraph-based approach remained relatively constant, not varying with changes in window size and minimum support threshold. In contrast, the running time of the association-rule-based approach increased with an increase in window size or a decrease in the minimum support. This is because the association-rule-based approach has to search for the frequent itemsets that match the current session. As the number of frequent itemsets increase (as a result of a decrease in the minimum support) or the length of current user session increases (as a

Figure 7 (a) Precisions and (b) recalls of the two recommendation approaches under different window sizes ($Min_{sup} = 0.16$ per cent)

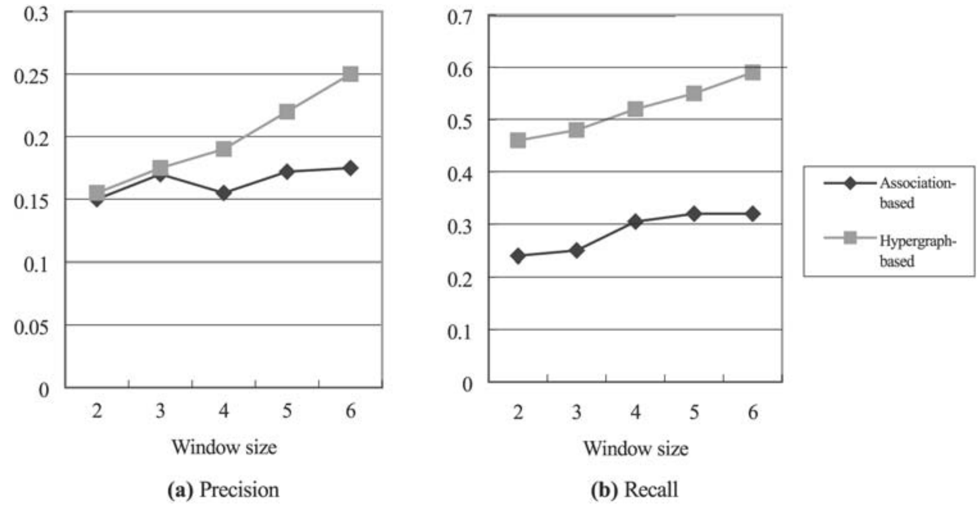
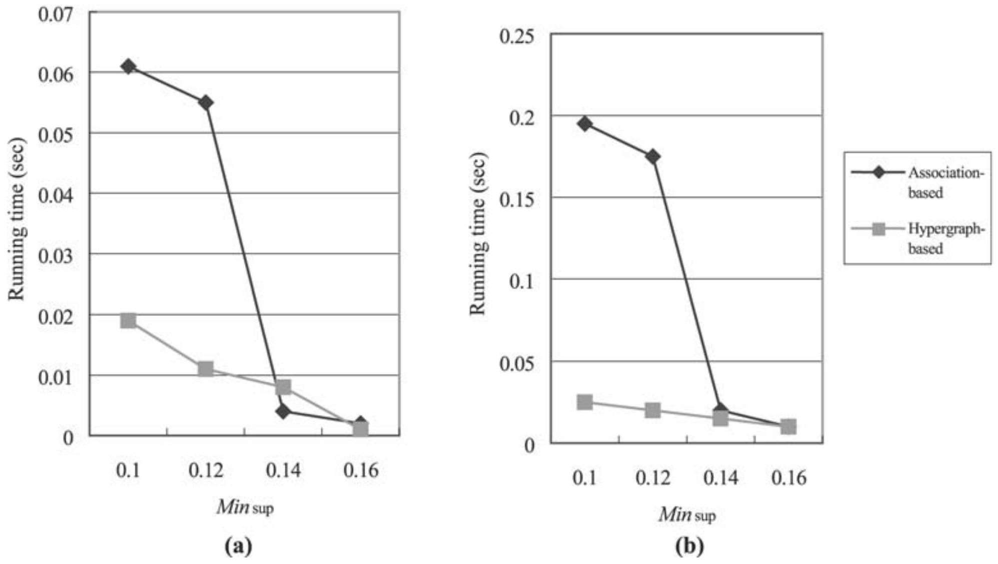


Figure 8 (a) Running times of the two recommendation approaches for a window size of two under different Min_{sup} values and (b) running times of the two recommendation approaches for a window size of six under different Min_{sup} values



result of an increased window size), the association-rule-based approach incurs a larger running time.

Overall, we conclude that the hypergraph-based approach is more attractive since it yields better-quality article recommendation and has a more consistent running time. Thus, *H4* is accepted.

Conclusions

In this paper, we have investigated issues related to the recommendation of articles in a literature digital library. We have developed a literature recommendation system that makes use of the Web usage logs of a literature digital library for making recommendations. The literature recommendation system consists of three sequential steps:

(1) data preparation of the Web logs;
(2) usage log mining; and
(3) generation of article recommendations.

We proposed three alternatives for identifying transactions from Web usage logs and discussed two approaches – association-rule based and hypergraph based – for making recommendations. These alternatives and approaches were evaluated using the Web usage logs of an operational electronic thesis system at National Sun Yat-sen University. It has been found that the query-chosen and session-chosen methods are better for transaction identification, and that the hypergraph-based approach yields better-quality article recommendation and exhibits a more consistent running time, and thus is more scalable.

Our recommendation framework identifies article associations present in the Web usage logs of a digital library. While this approach results in effective recommendations, it fails to recommend those independent articles that are seldom accessed together with others. As evident from our experiments, an analysis of the three-month collection of Web usage logs of NSYSU-ETD (from 1 February to 30 April 2002) showed that only about one-tenth of the total collection are recommendable (see Table I). To extend the scope of recommendable articles, we are currently investigating approaches that make use of multiple sources when making article recommendations in digital libraries. One such source, of course, is the metadata already collected by digital libraries.

References

Agrawal, R., Imielinski, T. and Swami, A. (1993), "Mining association rules between sets of items in large databases", in Buneman, P. and Jajodia, S. (Eds), *Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, DC, May 26-28*, ACM Press, New York, NY, pp. 207-16.

Agrawal, R. and Srikant, R. (1994), "Fast algorithms for mining association rules", in Bocca, J.B., Jarke, M. and Zaniolo, C. (Eds), *Proceedings of the 20th International Conference on Very Large Data Bases, September 12-15, Santiago, Chile*, Morgan Kaufmann, San Francisco, CA, pp. 487-99.

Alspector, J., Kolcz, A. and Karunanithi, N. (1998), "Comparing feature-based and clique-based user models for movie selection", *Proceedings of the 3rd ACM International Conference on Digital Libraries, June 23-26, 1998, Pittsburgh, PA*, ACM Press, New York, NY, pp. 11-18.

Andresen, D., Carver, L., Dolin, R., Fischer, C., Frew, J., Goodchild, M., Ibarra, O., Kothuri, R., Larsgaard, M., Nebert, D., Simpson, J., Smith, T., Yang, T. and Zheng, Q. (1995), "The WWW prototype of the Alexandria digital library", *Proceedings of the International Symposium on Digital Libraries, Tsukuba, Japan, 22-5 August*, pp. 17-27.

Ansari, A., Essegaier, S. and Kohli, R. (2000), "Internet recommendation systems", *Journal of Marketing Research*, Vol. 37 No. 3, pp. 67-85.

Arms, W. (2000), *Digital Libraries*, MIT Press, Cambridge, MA.

Armstrong, R., Freitag, D., Joachims, T. and Mitchell, T. (1997), "WebWatcher: a learning apprentice for the World Wide Web", *AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments*, pp. 6-12.

Balabanovi'c, M. and Shoham, Y. (1997), "Fab: content-based, collaborative recommendation", *Communications of the ACM*, Vol. 40 No. 3, pp. 66-72.

Baldonado, M., Chang, C., Gravano, L. and Paepcke, A. (1997), "Metadata for digital libraries: architecture and design rationale", *Proceedings of the 2nd ACM International Conference on Digital Libraries*, ACM Press, New York, NY, pp. 47-56.

Basu, C., Hirsh, H. and Cohen, W. (1998), "Recommendation as classification: using social and content-based information in recommendation", *Proceedings of the 15th National Conference on Artificial Intelligence, 26-30 July, Madison, WI*, AAAI Press, Menlo Park, CA, pp. 714-20.

Billsus, D. and Pazzani, M. (1999), "A hybrid user model for news story classification", in Kay, J. (Ed.), *Proceedings of the 7th International Conference on User Modelling, Banff, Canada, 20-4 June*, Springer-Verlag, New York, NY, pp. 99-108.

Bowman, C., Manber, P. and Schwartz, U. (1994), "Scalable Internet resources discovery: research problems and approaches", *Communications of the ACM*, Vol. 37 No. 8, pp. 98-107.

Breese, J., Heckerman, D. and Kadie, C. (1998), "Empirical analysis of predictive algorithms for collaborative filtering", *Technical Report MSR-TR-98-12*, Microsoft Research, Seattle, CA.

Chen, H., Schatz, B., Ng, T., Martinez, J., Kirchhoff, A. and Lin, C. (1996), "A parallel computing approach to creating engineering concept spaces for semantic retrieval: the Illinois digital library initiative project", *IEEE Transactions on PAMI*, Vol. 18 No. 8, pp. 17-34.

Cooley, R., Mobasher, B. and Srivastava, J. (1999), "Data preparation for mining World Wide Web browsing patterns", *Journal of Knowledge and Information Systems*, Vol. 1 No. 1, pp. 5-32.

Furner, J. (2002), "On recommending", *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 9, pp. 747-63.

Goldberg, D., Nichols, D., Oki, B. and Terry, D. (1992), "Using collaborative filtering to weave an information tapestry", *Communications of the ACM*, Vol. 35 No. 12, pp. 61-70.

Herlocker, J. and Konstan, J. (2001), "Content-independent task-focused recommendation", *IEEE Internet Computing*, Vol. 5 No. 6, pp. 40-7.

Karypis, G. (2002), "Multilevel hypergraph partitioning", *Tech. Report TR#02-25*, Department of Computer Science, University of Minnesota, MN.

Kessler, J. (1996), *Internet Digital Libraries: The International Dimension*, Artech House Publishers, Norwood, MA.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L, and Riedl, J. (1997), "GroupLens: applying collaborative filtering to Usenet news", *Communications of the ACM*, Vol. 40 No. 3, pp. 77-87.

Lang, K. (1995), "Newsweeder: learning to filter netnews", in Prieditis, A. and Russell, S. (Eds), *Proceedings of the 12th International Conference on Machine Learning, Lake Tahoe*, Morgan Kaufmann, San Francisco, CA, pp. 331-9.

Liu, B., Hsu, W. and Ma, Y. (1999), "Mining association rules with multiple minimum supports", *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 15-18 August, San Diego*, ACM Press, New York, NY, pp. 430-4.

Loeb, S. and Terry, D. (1992), "Information filtering", *Communications of the ACM*, Special Issue on Information Filtering, Vol. 35 No. 12, pp. 26-8.

Mobasher, B., Cooley, R. and Srivastava, J. (1999), "Creating adaptive Web sites through usage-based clustering of URLs", *Proceedings of the IEEE Knowledge and Data Engineering Exchange Workshop*.

Mobasher, B., Dai, H., Luo, T., Nakagawa, M. and Wiltshire, J. (2000), "Discovery of aggregate usage profiles for Web personalisation", P*roceedings of the webKDD Workshop*.

Mooney, R. and Roy, L. (2000), "Content-based book recommending using learning for text categorisation", *Proceedings of the 5th ACM Conference on Digital Libraries, San Antonio*, ACM Press, New York, NY, pp. 195-204.

Pazzani, M. and Billsus, D. (1997), "Learning and revising user profiles: the identification of interesting Web sites", *Machine Learning*, Vol. 27 No. 4, pp. 313-31.

Pazzani, M. (1999), "A framework for collaborative, content-based and demographic filtering", *Artificial Intelligence Review*, Vol. 13 No. 56, pp. 393-408.

Pennock, D., Horvitz, E., Lawrence, S. and Giles, C. (2000), "Collaborative filtering by personality diagnosis: a hybrid memory- and model-based approach", *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, San Francisco, 30 June-3 July*, Morgan Kaufmann, San Francisco, CA, pp. 473-80.

Pitkow, J. and Pirolli, P. (1999), "Mining longest repeating subsequences to predict World Wide Web surfing", *Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems*, Boulder, CO, pp. 139-150.

Schafer, J., Konstan, J. and Riedl, J. (2001), "E-commerce recommendation applications", *Data Mining and Knowledge Discovery*, Vol. 5 No. 1, pp. 10-22.

Shardanand, U. and Maes, P. (1995), "Social information filtering: algorithms for automating 'word of mouth'", *Proceedings of the Conference on Human Factors in Computing Systems, Denver, CO*, ACM Press, New York, NY, pp. 210-17.

Spink, A., Wilson, T., Ford, N., Foster, A. and Ellis, D. (2002), "Information seeking and mediated searching", *Journal of The American Society For Information Science and Technology*, Vol. 53 No. 9, pp. 695-703.

Srivastava, J., Cooley, R., Deshpande, M. and Tang, P. (2000), "Web usage mining: discovery and applications of usage patterns from Web data", *SIGKDD Explorations*, Vol. 1 No. 2, pp. 12-23.

Terveen, L., Hill, W., Amento, B., McDonald, D. and Creter, J. (1997), "PHOAKS: a system for sharing recommendations", *Communications of the ACM*, Vol. 40 No. 3, pp. 59-62.

Wilensky, R. (1996), "Toward work-centred digital information services", *IEEE Computer*, Vol. 29 No. 5, pp. 7-44.

Yan, T., Jacobsen, M., Molina, H. and Dayal, U. (1996), "From user access patterns to dynamic hypertext linking", *Proceedings of the 5th International World Wide Web Conference*, pp. 1007-14.

Yang, Q., Zhang, H.H. and Li, T. (2001), "Mining Web logs for prediction models in WWW caching and prefetching", *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 473-8.