PERGAMON

Expert Systems with Applications 21 (2001) 131–137

Expert Systems with Applications

www.elsevier.com/locate/eswa

Web personalization expert with combining collaborative filtering and association rule mining technique

C.-H. Lee[a], Y.-H. Kim[b,*], P.-K. Rhee[b]

[a]BI Part IT Solution Development Team, Information Technology Center, SK Telecom, 267 NamDaeMunRo 5-ga, ChungGu, Seoul 100 711, South Korea
[b]Intelligent Media Lab, Department of Computer Science and Engineering, Inha University, 253 Yong-Hyun Dong, NamGu, Incheon 402 751, South Korea

Abstract

Web personalization has been providing electronic businesses with ways to keep existing customers and to obtain new ones. There are two approaches for providing personalized service: a content-based approach and a collaborative filtering approach. In the content-based approach, it is not easily applied to web objects (pages, images, sounds, etc) which are represented by multimedia data type information. Collaborative filtering approaches have cold-start problem. More serious weakness of collaborative filtering is that rating schemes can only be applied to homogenous domain information. In this paper, we present a framework of personalization expert by combining collaborative filtering method and association rule mining technique to overcome problems that traditional personalized systems have. Since multimedia data type web object cannot be easily analyzed, we adopted a collaborative filtering method that considers each object as an item, and attempts a personalized service. Similar users of each domain object are found as the result of the collaborative filtering method. These similar users' web object access data is used by apriori algorithm to discover object association rules. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Web personalization expert; Collaborative filtering; Association rule mining

1. Introduction

As the WWW and its related technologies have been explored, they provide web users with ways to access necessary information without any restriction. In addition to this, electronic businesses are trying to apply web personalization techniques to their web sites for obtaining new customers and retaining existing ones. Web personalization is what service providers and merchants want to do to tailor the web objects (pages, product, image, sound, etc.) to web users based on their past behavior and inference from other similar users.

Most web personalization systems adopt two types of techniques: a content-based approach and a collaborative filtering approach. In the content-based approach, it recommends web objects that are similar to what the user has been interested in the past (Lang, 1995). In the collaborative filtering approach, it finds other users that have shown similar tendency to the given users and recommends what they have liked (Shardanand & Maes, 1995). However, most web objects are represented by multimedia type information so it is difficult to analyze web objects to attempt content-based method. Also, the collaborative filtering method can only be applied within a homogenous type information domain and suffers from cold-start problem (Philip, 1999).

In this paper, we present a framework that combines the collaborative filtering method with data mining technique, especially association rule mining technique to develop the improved personalized system to overcome mentioned drawbacks of traditional systems.

2. Related work

The content-based approach to the personalization system has been adopted in the information retrieval (IR) society, and employs diverse techniques. Text documents are recommended by comparing between their contents and user profiles. The weights for the words extracted from the document are added to the weights for the corresponding words in the user profile, if the user is interested in a page. Examples of such systems are WebWatcher (Joachims, Freitag & Mitchell, 1994), NewsWeeder (Lang, 1995), InfoFinder (Krulwich & Burkey, 1996), and client side agent Letizia (Leiberman, 1995). These approaches have several shortcomings. Generally, in some domains, the items are not applicable to any useful feature extraction methods with current technology, such as motion pictures, music, etc.

* Corresponding author. Tel.: +82-32-860-7448; fax: +82-32-875-0742.
E-mail address: yhkim@im.inha.ac.kr (Y.-H. Kim).

0957-4174/01/$ - see front matter © 2001 Elsevier Science Ltd. All rights reserved.
PII: S0957-4174(01)00034-3

Another problem is that a user is restricted to seeing items similar to those already rated, since the system can only recommend items scoring highly against the users' profile (Shardanand & Maes, 1995).

The collaborative filtering method usually requires users to explicitly input ratings about pieces of information. These ratings are then used to compute pairwise correlation coefficients among existing users. The correlation coefficient is the measure of how similar two users are. The system can make predictions or recommendations based on the correlation coefficients. This approach does not consider any analysis of the items. This characteristic enables the collaborative filtering-based system to naturally apply to such domains as music, image, sound, etc. Examples of systems taking this approach include GroupLens (Konstan, Miller, Maltz, Herlocker, Gordon & Rield, 1997), and Ringo (Shardanand & Maes, 1995).

3. Web object personalization expert system

Since multimedia data type web object's content cannot be easily analyzed, the content-based filtering method that has to analyze content of the item is not suitable for current web environment. For this reason, we adopt a collaborative filtering method that considers each object as an item, and attempts personalized service. The collaborative filtering task is performed for each domain by using users' web object access information.

The collaborative filtering task generates the similarity information among users that is valid within one specific domain and predicted rating value information of the objects. Given user similarity information of each domain, user similarity information over the all domains is gained by performing linear combination task. Using this information, the nearest neighbors (i.e. similar users) of one certain user can be found and these users' web object access information is to be the input data of apriori algorithm to discover object association rules. Discovered association rules among web objects are the information source that predicts whether certain web object is current user's favorite object or not (Fig. 1).

3.1. Data preprocess

The web log file gives information of who accessed the web site, what pages were requested and in what order, and how long each page was viewed. We transform from the web log data to a certain data format for effective personalization. We get the information of who visited what pages and how long he/she stayed in those pages by analyzing web server log file. We can build web user transaction database using the information. Given the preprocessing steps outlined above, we can assume that there is a set of n unique object URLs appearing in the preprocessed log:

$$\mathrm{URL} = \langle \mathrm{url}_1, \mathrm{url}_2, \ldots, \mathrm{url}_n \rangle$$

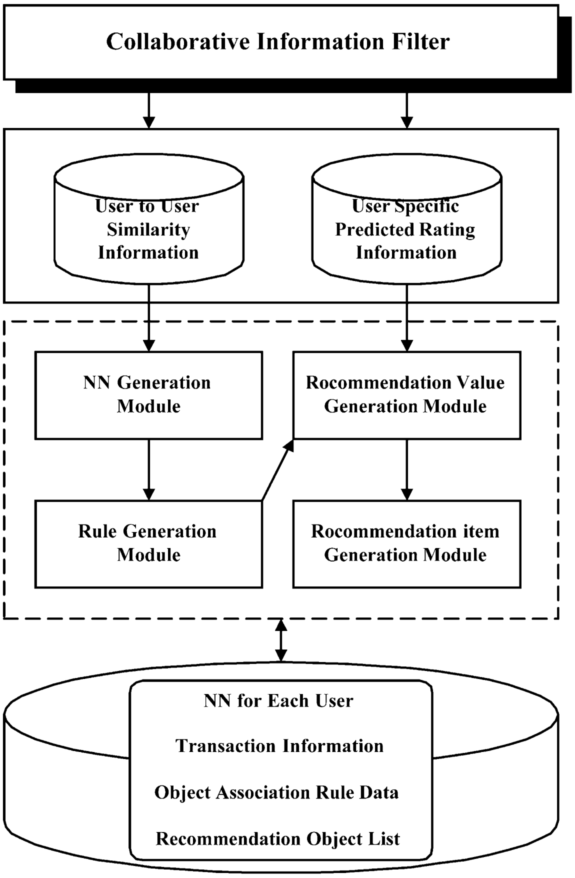
Collaborative Information Filter
User to User Similarity Information
User Specific Predicted Rating Information
NN Generation Module
Rocommendation Value Generation Module
Rule Generation Module
Rocommendation item Generation Module
NN for Each User
Transaction Information
Object Association Rule Data
Recommendation Object List

Fig. 1. Personalized engine architecture.

and a set of m user transactions are

$$T = \{t_1, t_2, t_3, \ldots, t_m\}$$

where each $t_i \in T$ is a non-empty subset of object URLs.

Since one of the purposes of this research is to discover recommendable objects that belong to different domain, we need to extract each user's interests toward certain web objects. Therefore, we identify all web objects that the user has accessed and their domains to construct user profiles that are effective over all domain. In this paper, we assume that the domain of web object is determined by experts, and web objects that belong to domain d_i are defined as,

$$d_i = \langle \mathrm{obj}_{i1}, \mathrm{obj}_{i2}, \ldots, \mathrm{obj}_{in} \rangle$$

and user transaction t within this domain is represented as,

$$t = \langle w_1(t), w_2(t), \ldots, w_n(t) \rangle$$

where $w_i(t)$ is obj_i's weight within the transaction t.

The weight of a certain object is determined by the amount of time a user spent on an object URL. This weight information also can play a role as a rating value for the collaborative filtering engine.

Finally, the transaction database may be filtered to remove web object URLs that are less accessed than a

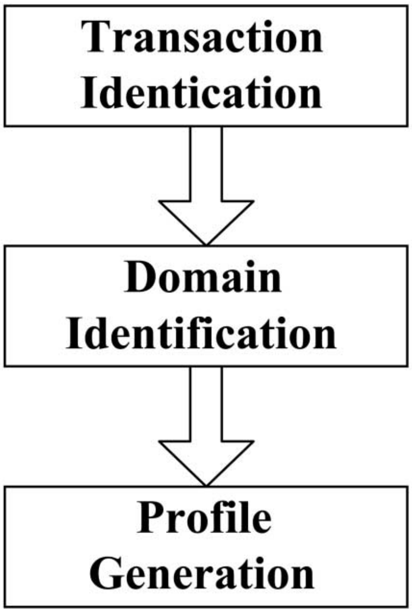

Fig. 2. User profile generation process.

specified number. This type of support filtering should be performed to eliminate noise from the data, and can provide a form of dimensionality reduction while doing collaborative filtering task (Cooley, Mobasher & Srivastava, 1999). The above preprocessed transaction database is used to discover the association rule among the web objects. However, to apply collaborative filtering method, user profile information is needed for generating user similarity information.

3.2. User profile

We adopt the collaborative filtering approach to identify nearest neighbors of each user that are valid over all domains. For this purpose, each user has the same number of user profile as the number of domains. In other words, each user has a certain user profile that is effective within a certain domain. The user profile generation process is explained below (Fig. 2).

The web objects accessed by the user within the each user session belong to different domains. To generate a domain specific user profile, we classify objects according to the domain categorization and record the amount time spent on these objects. We regard this domain specific information (i.e. objects and time length couple) as user profile. User m's user profile (UP_{im}) that is valid domain i is represented as,

$$Up_{im} = \langle (\text{obj}_{i1}, l_{i1}), (\text{obj}_{i2}, l_{i2}), \ldots, (\text{obj}_{in}, l_{in}) \rangle$$

where obj_{ij} is domain's object, and l_{ij} is the amount of time user m spent on obj_{ij}. By adopting this representation, we can map user profile into a multi-dimensional web object space as matrix of object URLs and this information is appropriate for the collaborative filtering method.

3.3. Identifying nearest neighbor within one domain

The collaborative filtering method generally requires users to explicitly input ratings about a pieces of information. These ratings are then used to compute pairwise correlation coefficients among existing users. The correlation coefficient is the measure of how similar two users are. This type of nearest neighborhood-based methods can be separated into three steps (Jonathan, Konstan, Borchers & Riedl, 1999).

1. Weight all users with respect to similarity with the active user.
2. Select a subset of users to use as a set of predictors (possibly for a specific item).
3. Normalize ratings and compute a prediction from a weighted combination of selected neighbors' ratings.

The system can make predictions or recommendations based on the correlation coefficients. The current popular algorithm to compute correlation coefficients is Pearson r Correlation Coefficient. Given two user's list of ratings as $X = [x_1, \ldots x_t]^T$ and $Y = [y_1, \ldots y_t]^T$, the standard Pearson r correlation coefficient is used to measure the similarity (S_r) between two lists of ratings. It is calculated as:

$$S_r = \frac{\sum_{i=1}^{t}(x_i - x_{\text{avg}})(y_i - y_{\text{avg}})}{\sqrt{\left(\sum_{i=1}^{t}(x_i - x_{\text{avg}})^2\right)\left(\sum_{i=1}^{t}(y_i - y_{\text{avg}})^2\right)}}$$

After calculating similarity values (correlation) among all users, neighbors whose similarity value (correlation) is greater than a fixed threshold are selected. A final prediction by performing a weighted average of deviations from the neighbor's mean is calculated as,

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n}(r_{u,i} - \bar{r}_u) \times W_{a,u}}{\sum_{u=1}^{n} W_{a,u}}$$

$P_{a,i}$ represents the prediction for the active user a for item I, n is the number of neighbors and $w_{a,u}$ is the similarity weight between the active user and neighbor u as defined by the Pearson correlation coefficient:

$$W_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i} - \bar{r}_a) \times (r_{u,i} - \bar{r}_u)}{\sigma_a \times \sigma_u}$$

where σ_a, σ_u are variances of user a and u's rating values, respectively (Jonathan, 1999).

Within one domain environment, user's preference toward certain object is obtained by calculating predicted rating value of that object. In this paper, we consider web objects that belong to different domain. Therefore, nearest

Domain 1(d_1)

User	User	Similarity
a	b	S_1(ab)
a	c	S_1(ac)
⋮	⋮	⋮
b	c	S_1(bc)
b	d	S_1(bd)
⋮	⋮	⋮
z	x	S_1(zx)
z	y	S_1(zy)

Domain 2(d_2)

User	User	Similarity
a	b	S_2(ab)
a	c	S_2(ac)
⋮	⋮	⋮
b	c	S_2(bc)
b	d	S_2(bd)
⋮	⋮	⋮
z	x	S_2(zx)
z	y	S_2(zy)

.....

Domain N(d_n)

User	User	Similarity
a	b	S_n(ab)
a	c	S_n(ac)
⋮	⋮	⋮
b	c	S_n(bc)
b	d	S_n(bd)
⋮	⋮	⋮
z	x	S_n(zx)
z	y	S_n(zy)

Fig. 3. User similarity information of multi domain.

neighbors that are valid over the all domains should be found (Fig. 3).

3.4. Identifying nearest neighbor over the all domains

To find the nearest neighbors valid over the all domains, we expand traditional collaborative filtering method.

A similarity value between user A and B of all domains are summed up. This summed value is a similarity value between user A and B over the all domains. After similarity values between user A and C, A and D, and finally A and Z are calculated, users whose similarity value over the all domains is greater than a predefined threshold are selected to become a nearest neighbor of A over the all domains. For each user, above task is repeated to find relevant nearest neighbors of a certain user. Therefore, all users have their own nearest neighbors, and this neighbors' transaction database is to be used to find association rule of web objects that neighbors have accessed.

4. Web object association rule generation

4.1. Aprioi algorithm

The association rule problem was originally proposed in the context of supermarket data to study the relationships of the buying patterns of customers in transaction data, i.e. to find how the items bought in a consumer basket related to each other (Agrawal & Srikant, 1994).

Let $I = \{i_1, i_2, \ldots, i_{,m}\}$ be a set of binary literals called items. Each transaction T is a set of items, such that $T \subseteq I$. This corresponds to the set of items which a consumer may buy in a basket transaction.

An association rule is a condition of the form $X \Rightarrow Y$, where $X \subseteq I$ and $Y \subseteq I$ are two set of items. The idea of an association rule is to develop a systematic method by which a user can figure out how to infer the presence of some sets of items, given the presence of other items in a transaction. Such information is useful in making decisions such as customer targeting, shelving, and sales promotions.

The *support* of a rule $X \Rightarrow Y$ is the fraction of transactions which contain both X and Y.

$$\text{support} = \frac{\text{\# of transactions containing all the items in } X \cup Y}{\text{total \# of transactions in the database}}$$

The *confidence* of a rule $X \Rightarrow Y$ is the fraction of transactions containing X, which also contain Y.

$$\text{confidence} = \frac{\text{\# of transactions that contain both } X \text{ and } Y}{\text{\# of transactions containing } X}$$

Thus, if we say that a rule has 90% confidence then it means that 90% of the tuples containing X also contain Y.

The process of mining association rules is a two phase technique in which all large itemsets are determined, and then these large itemsets are used in order to find the rules (Agrawal, Imielinski & Swami, 1993). The large itemset approach is as follows. Generate all combinations of items that have fractional transaction support above a certain user-defined threshold called *minsupport*. We call all such combinations large itemsets. Given an itemset $S = \{i_1, i_2, \ldots, i_{,m}\}$, we can use it to generate at most k rules of the type $[\mathrm{S} - \{i_r\}] \Rightarrow \{i_r\}$ for each $r \in \{1, \ldots, k\}$. Once these rules have be generated, only those rules above a certain user defined threshold called *minconfidence* may be retained. In order to generate the large itemsets, an iterative approach is used to first generate the set of large 1-itemsets L_1 is, then the set of large itemsets L_2, and so on until for some value of r the set L_r is empty (Agrawal & Srikant, 1994).

4.2. Generating user specific web object association rule

The data mining techniques, especially association rule mining algorithm, have been adopted to target marketing or personalized recommendation service within the E-commerce area. However, to extract hidden and useful knowledge, traditional data mining technique utilizes all user transaction database. This means it never considers each users interests or preferences to provide user specific personalized service. By analyzing web object access pattern using traditional collaborative filtering method, we find users whose interests or preferences are similar each other. Also, we consider these similar users' transaction database to discover web object association rule.

All users have their own nearest neighbors that are valid over the all domains. We use those nearest neighbors' transaction database to generate web object association rule that is specific to each user. As soon as user *i* logins the web site, user *i* specific association rule is activated, and according to this rule and user session information, personalized service is provided.

4.3. Personalized web object generation

Personalized object generation engine is the on-line component of the web object personalization service based on object access information. The task of the personalized object generation engine is to generate web object candidate lists using current user session, current user's association rule and object's predicted rating value. In this paper, we use a fixed-size sliding window method over the active session to generate personalized web object. In other words, only the last *n* visited pages URLs to influence the recommendation value of web objects in the recommendation set.

An efficient method for computing recommendation sets is to directly utilize the frequent itemsets obtained in the offline web object association rule mining stage. After identifying user sessions in the preprocessing steps and then discovering frequent itemsets of URLs, we match the current user session window with itemsets to find candidate URLs for giving recommendations. Given a window size of w, we only consider all itemsets of size $w + 1$ satisfying a specified support threshold and containing the current session window. The recommendation value of each candidate URL is based on the confidence of the corresponding association rule whose consequent is the singleton containing the candidate object. If the rule satisfies a specified confidence threshold requirement, then the candidate object is added to the recommendation set.

Finally, these candidate objects' preference value (predicted rating value) that are the result of the collaborative filtering method, are added to object recommendation value. According to the final recommendation value, web objects are listed to the pop-up windows as recommended web objects.

5. Experimental evaluation

In this section, we report results of the experimental evaluation of our proposed personalized method. We describe the data set used, the experimental methodology, as well as performance measures we consider appropriate for this task.

5.1. MovieLens dataset

We ran experiments using data from the MovieLens recommender system. MovieLens is a web-based research recommender system that was first introduced in September 1997. During the period from September 1997 to April 1998, the database grew to a large size, containing 100,000 ratings records from 943 users on 1682 movies. User ratings were recorded on a numeric five-point scale (from 1 to 5).

All movies that are in the MovieLens database belong to 18 different genre. In this paper, we assume that one genre represents one domain, so a total of 18 domains exist. However, similar domains such as action, thriller, etc. are summed to one representative domain. We only considered users that had rated 200 or more movies since we needed users that had enough rating information to discover similar users that are valid within all domains as well as one domain. The number of users that had rated 200 or more movies is 149. Among 100,000 rating records, 44,122 rating records were generated from those 149 users.

5.2. Experimental methodology

We compared recommended items' real rating value to evaluate the personalization performance. There are two kinds of recommended items. One is the recommended items resulting from the association rule database that using all site users' transaction database. Another is the recommended items resulting from the association rule database that using each user's nearest neighbors' transaction database.

Similarity information between users was discovered by performing collaborative filtering task for each domain. Since we needed nearest neighbors that were valid over the all domains, each domain's user similarity values are summed up to generate similarity information of the all domain. All users whose similarity value over the all domains was greater than 3.5 were selected to become a nearest neighbor of a certain user. Those users' rating transaction data were used to generate user specific rule database. One hundred and forty-nine user specific rule database were obtained by performing above task repeatedly. We then compared user rule database's consequent items' real rating value and site's all users rule database's consequent items' real rating value since those consequent items were the candidate items to be recommended.

Table 1 shows the case of condition items of the all users rule database (AURD) and the condition items of the user

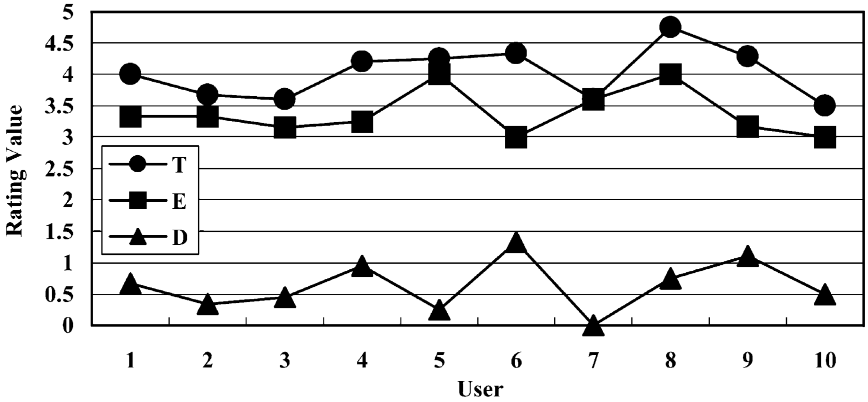

Fig. 4. Comparison results graph of rating value.

Table 1

Rating value comparison information [A, the number of the same conditions that exist in URD *n* and in AURD(R); B, the number of different consequent items (*Q*) that exists in *R* among AURD/the number of different consequent items (*W*) that exists in *R* among URD; C, the average rating value of users for *Q* (*E*)/the average rating value of users for *W* (*T*); D, difference (*T*–*E*)0

	1	2	3	4	5	6	7	8	9	10
A	5	4	5	5	6	4	5	6	7	6
B	4/3	3/3	5/4	4/3	4/2	3/1	4/4	4/3	5/6	4/4
C	4.00/3.33	3.67/3.33	4.20/3.25	3.60/3.15	4.25/4.00	4.33/3.00	3.60/3.60	4.75/4.00	4.28/3.17	3.50/3.00
D	+ 0.67	+ 0.34	+ 0.95	+ 0.45	+ 0.25	+ 1.33	0	+ 0.75	+ 1.11	+ 0.50

rule database (URD) are the same. In other words, in the case that a user is rating the same movies, recommendable items' real rating value from the AURD and recommendable items' real value from the URD were compared.

The number of the same conditions that exist in URD 1 and in AURD is 5. Among 5 rules, the recommendable items (i.e. consequent items) is 4 and 3, respectively. Those items average rating values are 4.00 and 3.33. Therefore, recommendable items from URD 1 were much more favored by the user 1. In this experiment, we discovered that recommendable items from 149 URD had average 0.63 higher rating value than from AURD (Fig. 4).

6. Conclusions and future work

We have presented a framework of web personalization expert based on web mining and collaborative filtering technologies. Our study shows that the proposed framework can provide better recommendation service by analyzing web access pattern of similar users. We adopted collaborative filtering method data mining techniques, especially apriori algorithm, respectively, to find similar users and web object association rule. For each user, similar user group is discovered and those users' web transaction information is utilized to generate web object association rule database. We found those user specific rule database could be better resources for personalization service.

Our future work in this area is to include the content-based filtering technique to our personalized expert system. Since the proposed architecture of personalized expert utilizes only the user behavioral information, it is necessary to consider the object's content information for more effective personalization service.

References

Agrawal R., & Srikant R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th VLDB Conference*.

Agrawal R., Imielinski T., & Swami A. (1993). Mining association rules between sets of items in very large databases. In *Proceedings of the ACM SIGMOD Conference*.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining World Wide Web browsing pattern. *Journal of Knowledge and Information Systems*, *1*(1), 5–32.

Joachims, T., Freitag, D., & Mitchell, T. (1994). Webwatcher: A tour guide for the world wide web. In *Proceedings of the 15th International Conference on Artificial Intelligence, Nagoya, Japan*.

Jonathan L. H., Konstan, J. A., Borchers, A., & Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In *Proceedings of the 1999 Conference on Research and Development in Information Retrieval, August*.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., & Rield, J. (1997). GroupLens: Applying collaborative filtering to usenet news. *Communications of the ACM*, *40*(3), 77–87.

Krulwich, B., & Burkey, C. (1996). Learning user information interests through extraction of semantically significant phrases. In *Proceedings of the AAAI spring Symposium on Machine Learning in Information Access, Standford, California*.

Lang, K. (1995). Newsweeder. Learning to filter netnews. In *Proceedings of the 12th Internationl Conference on Machine Learning, Tahoe City, California*.

Leiberman, H. (1995). Letizia: An agent that assist web browsing. In *Proceeding of the International Joint Conference on Artificial Intelligence, Montreal, Canada*.

Philip, S. Y. (1999). Data mining and personalization technologies. In *Proceedings of the 6th International Conference on DataBase system for Advanced Application, Taiwan*.

Shardanand, U., & Maes, P. (1995). Social information filtering: algorithms for automating 'Word of Mouth'. In *Proceedings of the Conference on Human Factors in Computing Systems-CHI'95*.