Developing a recommendation system for e-newspaper articles through personalizing digital contents

Sung Ho Ha, Jae-Shin Yi

School of Business Administration, Kyungpook National University,

Sangyeok-dong, Buk-gu, Daegu, Korea, 702-701, hsh@bh.knu.ac.kr

Abstract

This study presented a personalization system that adopted a methodology which is applicable for digital content recommendation and executed by the Internet service providers. The system made a recommendation to the users on the basis of their preferences, while most techniques for recommending digital content have focused on considering the similarity of content. In addition, it developed a method of evaluation to determine the priority of recommendations and adopted measures when selecting a set of recommendations. To experiment the feasibility and effectiveness of the presented methodology, a prototype system was developed and was applied to an English newspaper on the Internet.

Keywords: Personalization, recommender system, data mining, digital content

1. Introduction

With the fast development of the Internet and Web technologies recently, enterprises can achieve their business goals through lower cost and they can provide information to their customers faster and easier than before.

Rapidly changing business conditions, however, have brought a difficult time to both enterprises and customers: Every enterprise faces intense competition for survival and

needs innovative strategies to gain a competitive advantage over the competition; It is getting difficult for customers to select products and services effectively, due to the excess of information.

On the Web, where large amounts of information and a large number of users exist, personalization that aims to offer suitable information to the user is an essential tool in trying to overcome these difficulties. Even portal sites attracting many users supply personalized content to sustain their advantage. For example, Amazon.com recommends its goods to users following their purchase background. Yahoo! provides "My Yahoo!" service, encouraging users to create their own personalized pages from extensive lists presented by the site (Manber et al., 2000)

Most Web content providers, however, still offer all users equal content and yet do not satisfy individual user's needs. They should be able to offer users suitable content on time. To do so, they must be able to identify the customers, predict and understand their preferences and interests, identify appropriate content, and deliver it directly to customers in a personalized format during their online sessions.

Therefore, this study develops a system that recommends Web content, such as Internet news articles, based on a user's preference, when he or she visits an Internet newspaper site and reads the published articles. This content recommender system allows to form a one-to-one relationship between user and a content provider, heightens user's satisfaction, and raises the degree of loyalty to the content provider.

2. Content personalization

Usually, online service providers, including Internet newspapers, maintain significant

interest in marketing activities that attract and keep customers (Kohavi and Provost, 2001). One-to-one marketing is one of these online activities and it aims to preserve continuous relationships between an enterprise and individual customers by raising their degree of loyalty (Peppard, 2000). A major advantage of one-to-one marketing is its ability to customize products and services at a reasonable cost. This is referred to as personalization which offers the most suitable products and services to customers regarding their background information which was collected either implicitly or explicitly (Langheinrich, 1999; Sarwar et al., 2001).

Basic types of content personalization that are currently available fall into three categories: User-controlled, rules-based, and information-driven (Roberts, 2003). User-controlled personalization allows a user to choose the content elements to be displayed. It remains unchanged until the user decides to modify it. Rules-based personalization delivers content on the basis of decision rules made from the user profiles. It can also be static and rules must be established in advance.

Information-driven personalization can be relatively dynamic and usually adopts three types of filtering techniques: Content-based, collaborative, and non-intrusive (Linoff and Berry, 2001). Content-based filtering recommends content which should be similar to what a user has liked previously (Balabanovic and Shoham, 1997; Aggarwal and Yu, 2000). Collaborative filtering selects content based on the opinions of other users with similar preferences (Resnick et al., 1994). Although it is said to be the most successful personalization technology it still has several problems, such as reliance on subject user ratings, early rater difficulty, sparsity problems, and gray sheep (Claypool et al., 1999; Sarwar et al., 2000).

Nonintrusive-filtering incorporates Web usage data to discover navigation patterns

from them and predicts user behavior while the user interacts with the Web. It can create dynamic user profiles from Web usage data and maintain up-to-date personal preferences (CACM).

To date, typical content personalization systems often involve such functions as a news article classification or document search. SMART is a representative information retrieval system that expresses electronic documents by vector. It classifies similar documents and stores them so as to improve retrieval performance (Salton and McGill, 1997). Personal Webwatcher supports the monitoring of a user's Web experience using browsers, learns about the user's interests to obtain profiles, and delivers the appropriate Web documents to the user based on the profiles (Mladenic and Grobelnik, 2003). InfoFinder, unlike Personal Webwatcher, uses supervised learning to generate user profiles through the receiving of a user's direct input for document information of interest (Balabanovic and Shoham, 1997). Webby is a Web agent offering preferred documents to a user who tries to get necessary information using a browser. When the user refers to the same document again, the system provides the user with an immediate approach to the document. The system can find out areas where a user expresses interest by the number of visits for documents within those areas (Ivory et al., 2001). NewT is an agent system that helps a user select articles from a continuous stream of news. It examines the information stream and finds articles of interest to the user. As more and more information is available on the Internet, it has become one of the more useful agents (Maes, 1997).

3. Methodology for personalizing information content

A digital content recommender system presented here can be viewed as having two environments: The development environment and the online use environment. The development environment performs content analysis and user analysis, and builds several components used in the online environment, such as an Article-keywords index, a Keyword-articles index database, and a User preference rule base, as shown in Figure 1.

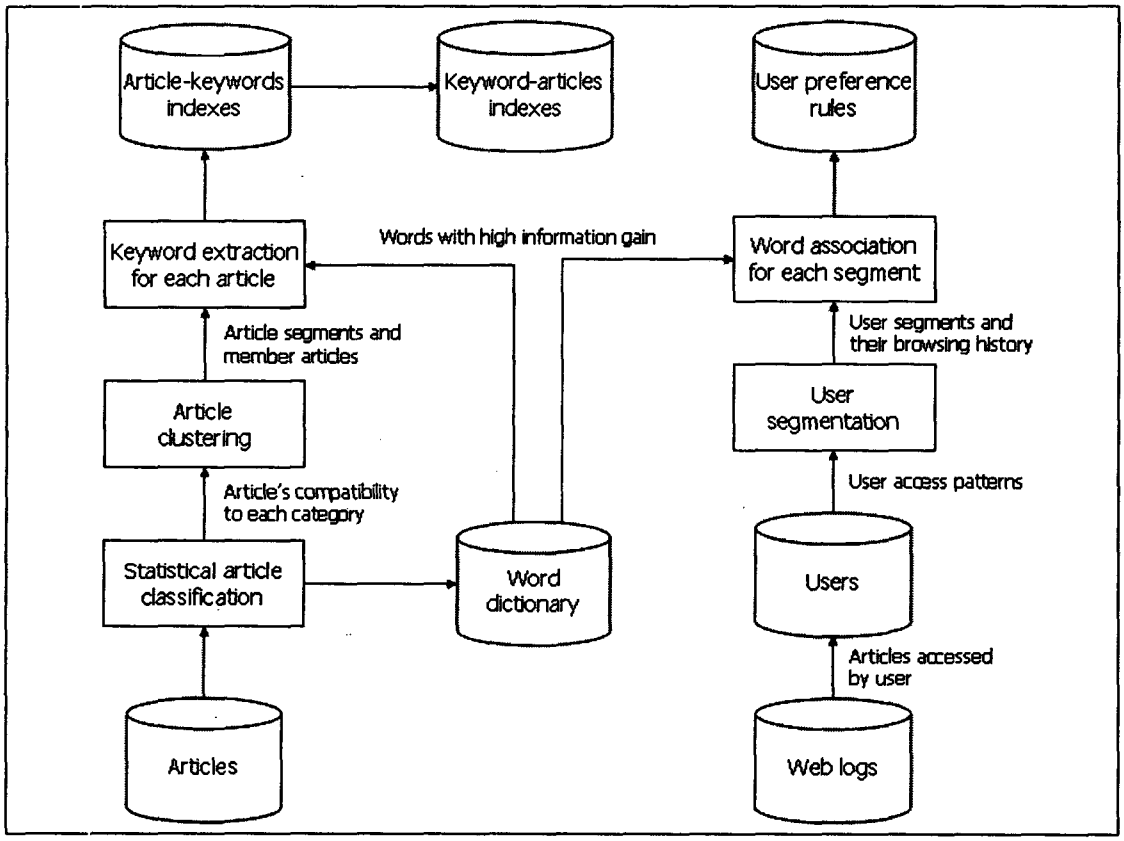

Figure 1. Overall architecture for developing the recommender system

3.1. Development environment

Content analysis classifies digital contents and prepares them for recommendation. User analysis keeps track of changes in users' preferences over time.

3.1.1. Content analysis

Most information content providers supply their content on the Web, classifying them according to pre-classified categories. Since some content, however, may belong to several categories, establishing a classification scheme by a provider is not simple.

In this article, a text mining technique such as the naïve Bayesian classifier, reclassifies information content with respect to their vocabulary and assigns statistical probabilities that measure the degree of compatibility to each content category.

The naïve Bayesian classifier is a learning method based on the Bayesian theorem. In order to classify information content, it simply assumes that given the target category, the words in a document are independent and identically distributed, and the probability of word occurrence is independent of the position within the text document. The Bayes classification maximizes the probability of observing words that were actually found in the document, subject to the usual naïve Bayes independence assumption.

Despite the inaccuracy of this independence assumption, much research shows that it performs very well in text document classification, when compared with other learning algorithms such as a neural network and a decision tree (Mitchell, 1997).

The following expression describes the Bayes classifier:

$$\upsilon_{CO} = \arg\max_{\upsilon_j \in V} P(\upsilon_j) \prod_i P(a_i \backslash \upsilon_j) \tag{1}$$

where υ_j denotes each category of content, V is a set of the categories, $P(\upsilon_j)$ is the prior probability per category, $P(a_i \backslash \upsilon_j)$ is the probability that a randomly drawn word from a document in category υ_j will be the word a_i, and υ_{CO} is the category output

by the classifier. To calculate the probability terms in equation (1), the following estimates (2-1) and (2-2) are adopted:

$$P(\upsilon_j)=\frac{|S_j|}{|S_{total}|} \tag{2-1}$$

$$P(a_i \setminus \upsilon_j)=\frac{|S_{ji}|}{|S_j|} \tag{2-2}$$

where S_j is the number of training samples belonging to category υ_j, S_{total} is the total number of training samples, S_{ji} is the number of training samples of category υ_j which have the word a_i.

Also, information gain measures are used to reduce the number of features and thus improve the effectiveness of classifying text documents. The greater the increase in information gain, the more distinguished the words become in classifying documents. Information gain is equivalent to the expected cross entropy that shows the reduction in entropy caused by partitioning the examples according to the feature *F*, as shown in expression (3).

$$InfoGain(F)=I(C,F)=H(C)-H(C \setminus F) \tag{3}$$

where $I(C,F)$ is the expected cross entropy, $H(C)$ is Shannon's entropy of class *C* and calculated from $\sum_i P(C_i)\log_2 P(C_i)$, and $H(C/F)$ is the conditional entropy of class *C* and calculated from $\sum_j P(F_j)\sum_i (C_i \setminus F_j)\log_2 P(C_i \setminus F_j)$.

Distinguished words, along with information gain measures, organize a *word*

dictionary. These words are later used to build an Article-keywords index database, a Keyword-articles index database, and a User preference rule base. The size of the dictionary varies according to the set threshold of the information gain, and it determines the amount of work needed to build the Article-keywords indexes. Therefore, it is important to decide on an appropriate word dictionary size in order to improve the performance of developing the recommender system. If a word dictionary becomes bigger by lowering the threshold, the recommendation accuracy may increase, since the dictionary contains more words which are extracted from news articles.

Using the posterior probability for each content category calculated by the Bayes classifier, a self-organizing map (SOM)—a neural network using an unsupervised learning scheme—divides this content into numerous segments in order to identify hidden categories, such as compound ones that hold multiple kinds of content simultaneously. Each article in a particular segment tends to have similar content, whereas articles in a different segment have dissimilar content.

A SOM tries to uncover patterns in the input fields set and clusters the data set into distinct groups without a target field (Han and Kamber, 2001). The SOM algorithm uses competitive learning. When an input pattern is imposed on the neural network, the algorithm uses an equation (4) and selects the output node with the smallest Euclidean distance between the presented input pattern vector ($\hat{X}$) and its weight vector ($\hat{W}_j$).

$$\max_j \left(\hat{X}^t \hat{W}_j \right) \tag{4}$$

Only this winning neuron generates an output signal from the output layer; all other

neurons in the layer have an output signal of zero. Since learning involves a weight vector adjustment, only the neurons in the winning neuron's neighborhood can learn with this particular input pattern. They do this by adjusting their weights closer to the input vector according to the equation (5).

$$w_j(n+1) = \begin{cases} w_j(n) + \eta(n)[x(n) - w_j(n)], & j \in N(n) \\ w_j(n), & otherwise \end{cases} \quad (5)$$

where η is the learning rate and N is the neighborhood function.

The neighborhood's size initially includes all units in the output layer. As learning proceeds, however, it shrinks progressively to a predefined limit, and fewer neurons adjust their weights closer to the input vector.

Once the clustering of articles has been finished, keywords are then extracted from each article. When extracting words, the recommender explores the word dictionary which contains eminent words in terms of the information gain. An article identifier, category information, and keyword lists constitute an Article-keywords index database.

A Keyword-articles index database is also constructed to expedite the recommendation of articles to users. It mirrors the Article-keywords index database, but it consists of keywords and article identifiers which contain these keywords. Both the Article-keywords index and Keyword-articles index databases need to include up-to-date information in order to maintain accurate and current recommendations.

3.1.2. User analysis

Successful content providers offer a bundle of customized content which satisfies a user's needs. To maximize the effectiveness of a recommendation, a Web site needs to

be able to provide proper content to the customer audience. It should identify users' habits and interests, which can be discovered by mining their Web page navigation patterns.

When extracting a user's access history, on which the mining algorithms can be run, several data preprocessing issues—cleaning data, identifying unique users, sessions, and transactions—have to be addressed (Srivastava et al., 2000; Kosala and Blockeel, 2000). Preprocessed user information forms a User database.

Once users and their access histories have been identified, segmenting the users follows. A SOM is chosen to break users into peer groups with similar Web page navigation behavior. Segmentation by traversal history uses a segmentation variable on the basis of the *visit RFM* values: *Visit recency* describes how long it has been since a user visited a category; *visit frequency* is how many times a user visits each category; *visit monetary* value calculates the number of visited pages per category.

Given the user segments and articles which users within these segments read, association mining finds all word affinities that are among the articles for segmented users based on the interests of the peer group members. Word affinity information guides users to articles they are not aware that they need. As consulting the word dictionary, the recommender generates word associations for each word distinguished in the articles. Word associations assume the form of IF-THEN, in which the left-hand side is an antecedent of a user preference rule and the right-hand side is a consequent. They are kept in a User preference rule base. Since it is obtained from dynamic user patterns, the User preference rule base can be adapted over time.

3.2. Online use environment

Once the recommender system accomplishes the development tasks, it provides users dynamic recommendations based on their browsing history and monitors its recommendation performance over time in an online use environment. Figure 2 shows some components of which the online environment is comprised.

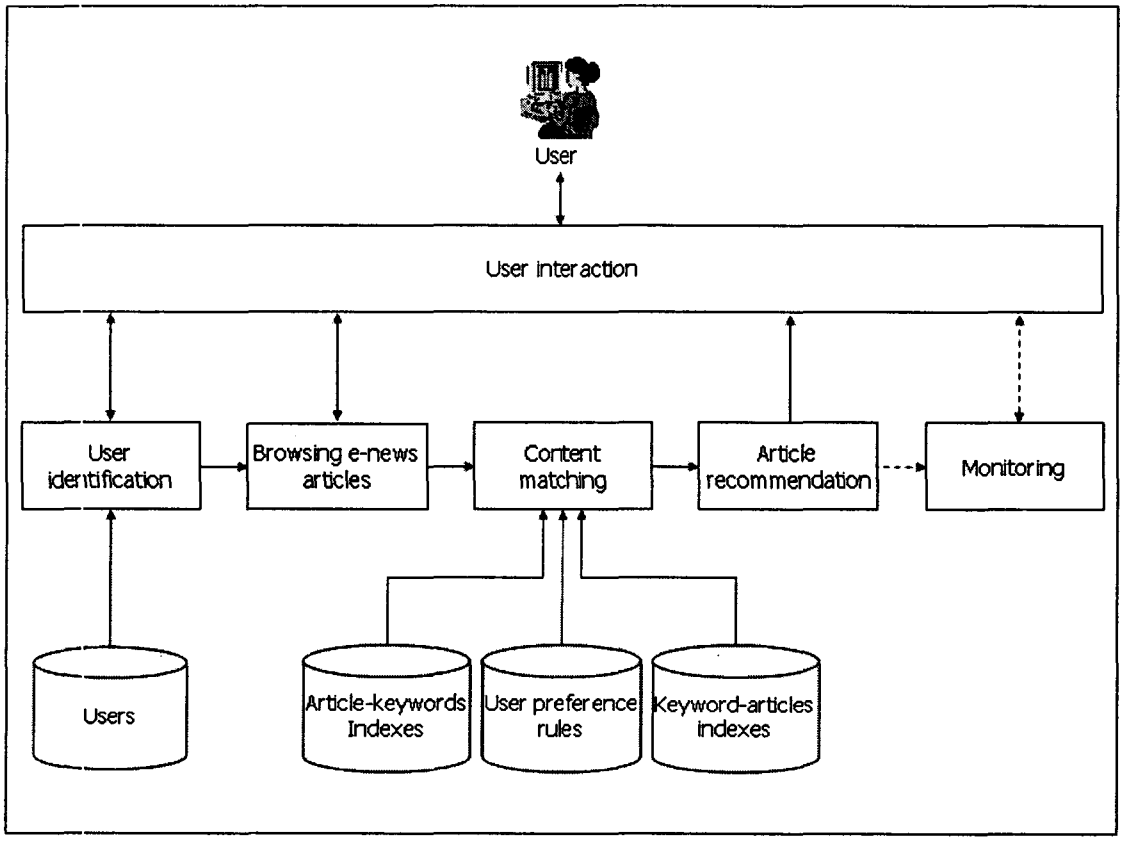

Figure 2. Components of online use environment

Whenever a user visits an electronic news site and browses articles, the recommender identifies a user segment to which the user belongs and determines the preference rules for the user. By looking up the Article-keywords index database, the recommender grasps keywords from the current-browsing article. It applies the keywords to antecedents of user preference rules and identifies the consequents of the rules executed

for only those antecedents which are true. Using these consequents, the recommender consults the Keyword-articles index database to extract a set of recommended articles for the user and then sends them to the user browser. The *User Preference Score* (UPS) is developed to prioritize those recommended articles according to their relevance to the user. This leads to finding the most applicable content from a variety of article segments.

User interests and needs, however, continuously change. The recommender should capture these changes. To do so, it should continue to monitor user navigational activities which reflect their interests before recommendations. As a result, it can decide on its success or failure after the recommendations. If the user follows a page link to a recommended page, it signifies that a recommendation has succeeded.

The recommender may use a metric, such as a click-through ratio, to quantitatively represent a chance of success. Using such a measure allows the determining of a predictive performance of user preference rules, user segments, or content classification in order to produce better recommendations. Whenever the measure decreases below a set threshold, the recommender can alert the management if new preference rules will be needed in order to catch changes in the user's interests. If alarms continue, the system can automatically start deriving new rules on behalf of the management.

4. News article personalizing system

To test the feasibility and effectiveness of the presented methodology, a prototype system was developed and applied to an English newspaper on the Internet (with established headquarters in South Korea). An experimental Articles database housed electronic news during a seven-month period between Dec. 2002 and Jun. 2003. This amounted to 8,968 articles randomly selected from six main sections, such as business

(*Biz*: 1,645), culture (*Cul*: 1,040), nation (*Nat*: 2,605), opinion (*Opi*: 1,434), sports (*Spo*: 956), and technology (*Tec*: 1,288).

4.1. Information content classification

Web pages containing news are usually organized by Hypertext Markup Language (HTML). HTML is helpful in communicating with each other through the Internet but, in itself, contains no semantic meanings. This can make it difficult to analyze content. Eliminating the HTML tags in Web pages accomplishes data reduction for classification and results in exposing clear semantics. The prototype system also removes irrelevant items including Web ads, writers, and positions. News without the HTML tags is maintained in the Newspaper article database.

The Bayesian classifier assorts news articles into one of the aforementioned categories according to their vocabulary, as shown in Figure 3.

Article ID	Article title	Biz	Cul	Nat	Opi	Spo	Tec
10603	WB Staff to Visit for Job Interview	0.8087	0	0.1913	0	0	0
10613	WTO Urged to Remove Non-Tariff Barriers	0.1524	0	0	0	0	0.8476
20159	Snowboarding has become a popular activity here since 1998	0.0448	0.0337	0.9132	0.0001	0	0.0082
20309	Multiplex to Open at Kimpo Airport	0.0002	0.4503	0.5495	0	0	0
30409	Kimhae Airport Sees Increase in Int'l Flights	0.0043	0	0.4541	0	0	0.5415
31925	[Campus Life] Cheerleading Squad Like a Family	0	0	0.7742	0.002	0.2239	0
40493	North Korea's Brinkmanship Goes Nowhere	0	0	0.7544	0.2456	0	0
41399	China's Accession to WTO	0.7655	0	0	0.2345	0	0
50159	English Premiership	0	0	0.1157	0	0.8842	0
50448	Choi In Hot Pursuit	0.0001	0.0001	0.5394	0	0.4603	0
60941	Woori's CEO club	0.8334	0	0.1648	0	0	0.0018
61237	Cooker or robot?	0.1008	0	0.0964	0	0.0001	0.8027

Figure 3. Article classification using the Bayesian classifier

This study organized a word dictionary which contained approximately 1,500 words, scoring a high information gain. Sample words with the information gain in the word dictionary are described as follows: (nation, 0.41863), (finance, 0.41807), (technology, 0.32586), (sports, 0.29558), (market, 0.12263), (league, 0.10312), (political, 0.08719), (government, 0.07660), (economic, 0.07370), and (editorial, 0.06967).

At the same time, the content recommender divides news articles into segments on the basis of classification probabilities as shown in Figure 3, and assigns each article to the resulting news category. Figure 4 illustrates 14 article categories derived by a four by four (4 × 4) SOM and the number of articles within each category.

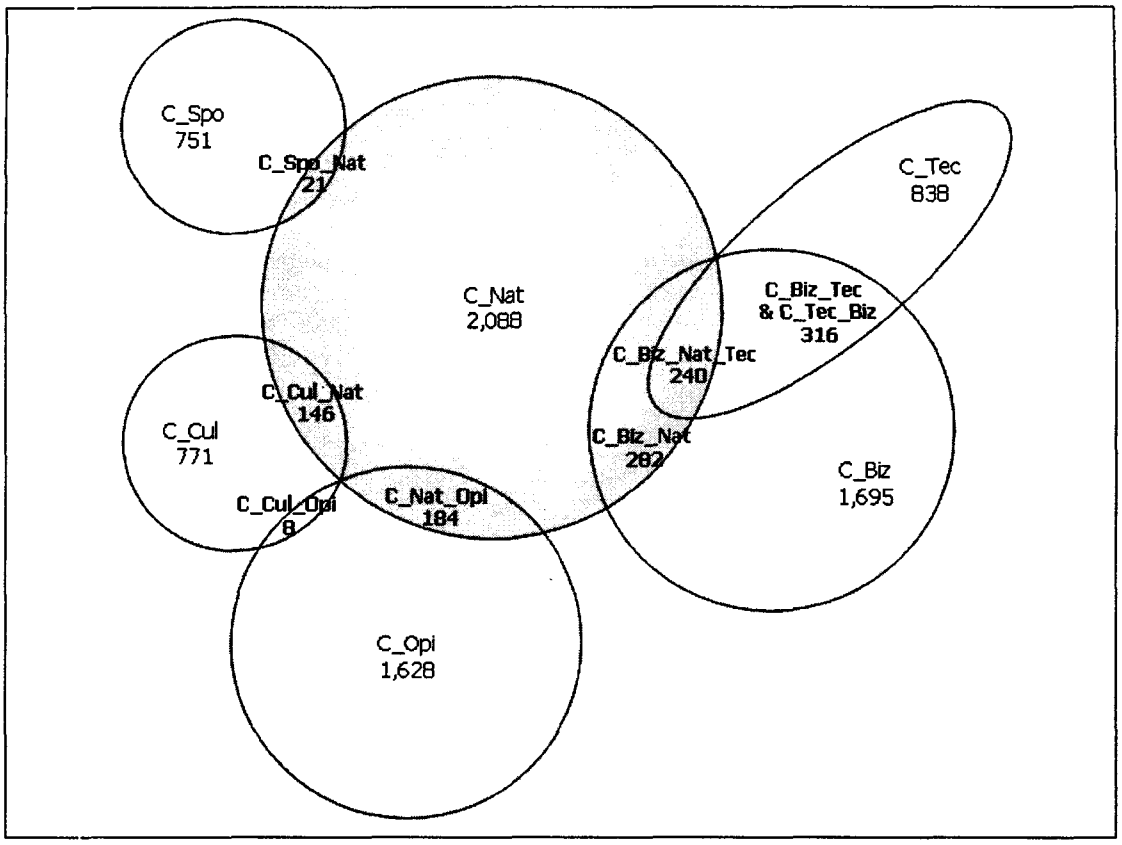

Figure 4. News categories and the number of articles derived by a SOM

A SOM derived compound categories that hold multiple kinds of content at the same time. They are *C_Biz_Tec*, *C_Tec_Biz*, *C_Biz_Nat*, *C_Biz_Nat_Tec*, *C_Nat_Opi*, *C_Cul_Nat*, *C_Spo_Nat*, and *C_Cul_Opi*. For example, category C_Biz_Tec contains those news articles which have both business and technological content. C_Biz_Tec and C_Tec_Biz, however, differ primarily in content composition. Business content is more dominant than technological content in C_Biz_Tec, while technological content is more prominent in C_Tec_Biz. Some articles change their membership after a SOM has been applied. Categories C_Spo_Nat and C_Cul_Opi are ruled out in the following analysis, since they contain only a small number of articles.

Referring to the word dictionary, the recommender extracts key words from all

articles in each category and stores them in the Article-keywords index database. Figure 5 shows a sample of an Article-keywords index.

Article ID	Category	Keywords
10603	C_Biz	applicant, bank, candidate, college, comment, economist, environment, finance, ...
10613	C_Tec	agenda, barrier, competition, consultation, country, custom, intellectual, settlement, ...
20159	C_Biz_Nat	activity, back, comment, list, popular, printer-edition, reader, snowboarding, view, ...
20309	C_Cul_Nat	airport, area, cinema, comment, complex, domestic, entertainment, international, multiplex, theater, ...
30409	C_Biz_Nat_Tec	accommodate, airstrips, domestic, factor, passenger, refurbishment, route, shanghai, transporting, ...
31925	C_Spo	childcare, children, dawson, enthusiastic, festival, fresh, graduating, lifestyle, role, ...
40493	C_Nat_Opi	abandon, approach, authority, brinkmanship, bush, complaint, decision, dmz, organization, pyongyang, ...
41399	C_Biz	accession, certificate, commitment, decision-making, economy, equity, ministerial, requirement, ...
50159	C_Spo	arsenal, block, clash, goal, heavyweight, manchester, premiership, shot, team, united, won, ...
50448	C_Spo	carding, comment, distance, golfer, Korean, kyoung-ju, list, Nissan, open, striking, ...
60941	C_Biz	agriculture, applicant, cost, globalization, local, lower, machinery, opportunity, overseas, ...
61237	C_Tec_Biz	appliance, blender, changwon, electronic, province, robot, south, store, view, ...

Figure 5. Sample articles and their keywords in the Article-keywords index database

4.2. User analysis and rule generation

The recommender performs a clustering analysis of Web page traversal history which users form as they navigate news on the Internet site. The system mines server log files, extracts a user's surfing history per news category in order to express it as a vector of categories that the users approach. A SOM then splits users into different segments on the basis of their similar browsing patterns for each category. Traversal histories for segmentation purposes display four features: A user identifier and the visit RFM values

per category.

A four by four (4 × 4) SOM derived 12 user segments, such as *U_Biz*, *U_Cul*, *U_Nat*, *U_Opi*, *U_Spo*, *U_Tec*, *U_Biz_Nat*, *U_Biz_Nat_Tec*, *U_Biz_Tec*, *U_Nat_Opi*, *U_Cul_Nat*, and *U_Tec_Biz*. Users within the U_Biz segment mainly have an interest in business-related news articles, and users in U_Biz_Nat have an interest in articles related to both business and nation.

After assigning users to peer groups based on the visit RFM values, the recommender discovers word affinities for segmented users based on their surfing history extracted from the User database. Word affinities are limited to those words in the word dictionary. Table 1 shows a subset of association rules for the user segment U_Tec_Biz.

Table 1. Sample of user preference rules for the user segment U_Tec_Biz

Antecedent	Consequent	Transactions	Support (%)	Confidence (%)	Lift
won	billion	13	26	81.25	2.26
billion	won	13	26	72.22	2.26
million	market	12	24	75.00	1.56
local	market	11	22	78.57	1.64
trade	market	10	20	71.43	1.49
local	south	10	20	71.43	2.1
foreign	market	9	18	75.00	1.56
company	billion	9	18	75.00	2.08
global	market	9	18	81.82	1.7

Notice that because users can move from one segment to another, new user segments

arise and old segments disappear. Also, word affinities for each segment may change over time.

When *Minimum Support*, *Minimum Confidence*, and *Lift* are set to 10%, 70%, and 1, respectively, the number of association rules for each user segment is summarized as follows: U_Biz (126), U_Cul (162), U_Nat (92), U_Opi (194), U_Spo (173), U_Tec (165), U_Biz_Nat (76), U_Biz_Nat_Tec (56), U_Biz_Tec (149), U_Nat_Opi (132), U_Cul_Nat (93), and U_Tec_Biz (111).

Those thresholds were chosen to limit rules after different values on these parameters were tested and the results were inspected. Since word affinities within an article are less likely to appear in all news articles, Minimum Support can be set relatively low. Minimum Confidence can be set relatively high, since there is a high possibility of semantically associative words appearing together within an article. Lift is of significance if it is above one. Generally, the number of rules for a user segment increases as the number of articles, which users within the segment have seen, increases.

4.3. Online recommendation

Assume that a user who belongs to the user segment U_Tec_Biz is browsing a certain article, i.e. 60828, which belongs to content segment C_Tec_Biz. The article is titled "US Preliminary Ruling on Hynix to Create Int'l Trade Dispute."

The recommender searches for the article in the Article-keywords index database and finds keywords within the article. It then locates the user preference rules for the user segment U_Tec_Biz, whose antecedents match the keywords found. The consequents of the rules imply the user's preferences. Articles containing the consequents become candidates for recommendations to the user, which will be found in the Keyword-

articles index database.

The number of candidates for recommendation by content category is as follows: C_Biz (152), C_Cul (20), C_Nat (98), C_Opi (388), C_Spo (1), C_Tec (90), C_Biz_Nat (8), C_Biz_Nat_Tec (3), C_Biz_Tec (8), C_Nat_Opi (0), C_Cul_Nat (27), and C_Tec_Biz (5). As the number indicates, many candidates (i.e. 800 articles) exist for recommendation, since there are many articles that contain at least one consequent of the rules. Limiting the number of consequents can be a way of controlling the number of candidates.

When the recommender determines a set of recommendations from the candidates, a criterion, UPS, prioritizes the recommendations using the following equation (6):

$$UPS = \sum_{i=1}^{n} (avgSup_i + avgConf_i)C_i \quad (6)$$

where n is the number of consequents, $avgSup_i$ is the average support for the i^{th} consequent, $avgConf_i$ is the average confidence for the i^{th} consequent, and C_i is the number of times a candidate article contains the i^{th} consequent.

Table 2 represents a process for obtaining the UPS measures for sample candidates. It shows consequents, average support and average confidence for each consequent, and the UPS values of each candidate.

Table 2. Calculating UPS measures of sample candidates

Consequent	avgSup (%)	avgConf (%)	Candidates for recommendation											
			11254	20967	60785	60053	50252	60192	11179	41169	60587	41232	11606	10229
won	25.76	73.91	18.94	7.97	0	0	1.99	0	3.99	1.99	3.99	0	1.99	0
bank	18.51	76.62	24.73	0	0	0	0	0	0	0.95	0	0	0	0
financial	17.82	73.96	7.34	0	0	0	0	0	0	2.75	0	0	0	0
market	18.18	70.59	1.78	0	23.08	10.65	0	0.89	0	0.89	3.55	0.89	0	18.64
Samsung	12.74	76.6	0.89	0	0	0	0	0	0	0	0	0.89	0	0
government	19.7	81.25	0	0	0	5.05	0	0	0	9.09	5.05	1.01	0	3.03
company	16.0	80.0	0	0	0	1.92	0	0	0	4.80	0.96	4.80	7.68	4.80
industry	16.0	88.89	2.10	1.05	0	4.20	0	1.05	0	1.05	2.10	3.15	0	9.44
south	22.0	100	0	0	6.10	0	0	0	1.22	0	12.2	0	0	3.66
electronics	15.15	76.92	0	0	0	0	0	0	0	0	0	0.92	0	0
million	14.0	77.78	1.84	0	0	0	0	10.10	2.75	0	6.42	1.84	0.92	0
commerce	12.12	80.0	0	0	0	0	0	0.92	0	0	0	0	0	0
ministry	12.12	80.0	0	0	2.76	0.92	0	1.84	0	0.92	0	0	0	2.76
north	16.82	97.66	1.14	0	0	0	0	0	0	1.14	0	0	0	0
party	17.67	91.08	0	1.09	0	0	0	0	0	0	0	0	0	3.26
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
	UPS		81.5	17.87	49.74	46.03	8.67	51.0	20.5	34.6	52.87	26.74	17.79	74.15

Table 3 summarizes the articles which received the best UPS in each content category and lists them in descending order of the UPS. It shows that article 11254 recorded the highest UPS measure and had the highest recommendation priority among all others. The recommender, therefore, provides article 11254 first to the user who is reading article 60828.

Table 3. The best UPS Article in each category and their characteristics

Seq.	Article ID	Title	Content category	UPS measure
1	11254	S&P Upgrades Woori Bank Credit Rating	C_Biz	81.5
2	10229	Korea Lukewarm to Further Opening of Audiovisual Market	C_Tec_Biz	74.15
3	60587	FTA Leads Nation Toward Global Market	C_Biz_Tec	52.87
4	60192	Computer Exports to Exceed $13 Bil. This Year	C_Tec	51.0
5	60785	Seoul to Facilitate Limited Education Market Opening	C_Nat	49.74
6	60053	Policymakers Gather Public Opinion for Doha Round	C_Opi	46.03
7	41169	Growing Economic Woes/Cut Expenses for Corporate Lobbying	C_Biz_Nat_Tec	34.6

8	41232	SARS and Korean Economy	C_Nat_Opi	26.74
9	11179	Foreign Workers Face Trouble in Raising Kosians	C_Biz_Nat	20.5
10	20967	Grass Looks Greener at Samyang Ranch	C_Cul	17.87
11	11606	English Website on Tax Service Attracts Foreign Investors	C_Cul_Nat	17.79
12	50252	Soft-Spoken Han Ready to Make Noise	C_Spo	8.67

5. UPS versus similarity measures of information retrieval

Determining whether a user shows an interest in the recommendations is to understand the effectiveness of the recommender system. Before applying the system to a real-world environment, it goes through the evaluation of the recommendations from the perspective of information retrieval using vector models such as *Term Frequency Inverse Document Frequency* (TFIDF) or *Latent Semantic Index* (LSI) measure (Sullivan, 2001). Although evaluating the recommendations, based on user preferences using similarity measures, is to some degree not appropriate, this kind of evaluation gives us an opportunity to compare the recommender system using UPS with ones using other evaluation measures.

When a user looks at a news article, such as 60828, an evaluation procedure is outlined as follows: 1) for article 60828, find the highest UPS article in each content category; 2) calculate the similarity measures of the recommendations found (see the appendix for information on TFIDF); and 3) compare the UPSs with the similarity

measures. Figure 6 illustrates the results of the comparison.

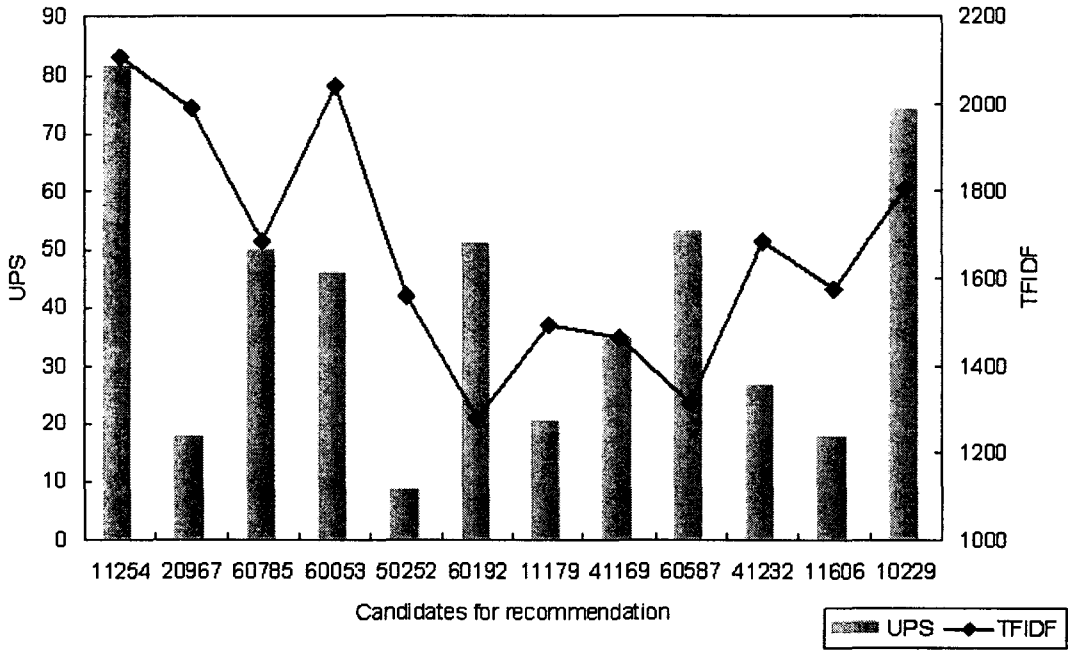

Figure 6. The comparison of UPS with TFIDF

In Figure 6, the principal Y-axis displays the UPS values and the secondary represents the TFIDF scores. The TFIDF axis maps the Euclidean distance between the top UPS article in each content category and article 60828. Therefore, it is more similar document as the TFIDF value is smaller. The UPS value of article 11254 is the largest among all others (i.e. 81.5) and so is the TFIDF value (i.e. 2110). That is, the user preference for a document which is the least similar is the highest.

From the information retrieval perspective, a vivid method that can satisfy users is to offer documents that contain the most similar information with search keywords that the users enter. This displays only similar documents but does not show the user preferences. It only reflects a user's short-term interests or preferences when a search is

conducted. Since similarity measures between documents are determined regardless of the number of articles or user segments, they cannot react to dynamic changes in user preferences. That is, they cannot effectively reflect users' mid- and long-term preferences. The system developed here recommends content based on homogenous user group's preferences, captures dynamic changes of preferences over time, and consequently complements shortcomings of similarity-based systems.

Over time, however, things may change. Additional experiments are tracking changes in the recommendations over time, when a user within the user group U_Tec_Biz reads article 60828. For convenience sake, these experiments only allow users, user segments, and user preference rules to be changed. Other factors, including the Article database, remain unchanged. These experiments are performed three times at regular intervals (i.e. two weeks). Figure 7 illustrates the results of monitoring changes in the number of rules and recommended articles over time. The best UPS article selected from all content categories is also displayed.

As a result, the best UPS articles discovered by the experiment over time are summarized as follows:

- **Observation 1**: Article 11169 titled "Footwear Trade Turns to Deficit For First Time Ever" in content category C_Tec
- **Observation 2**: Article 10291 titled "Ulsan Posts Highest Per Capita Output in 2001" in content category C_Biz
- **Observation 3**: Article 60306 titled "Annual Exports to Hit Record High This Year" in content category C_Tec

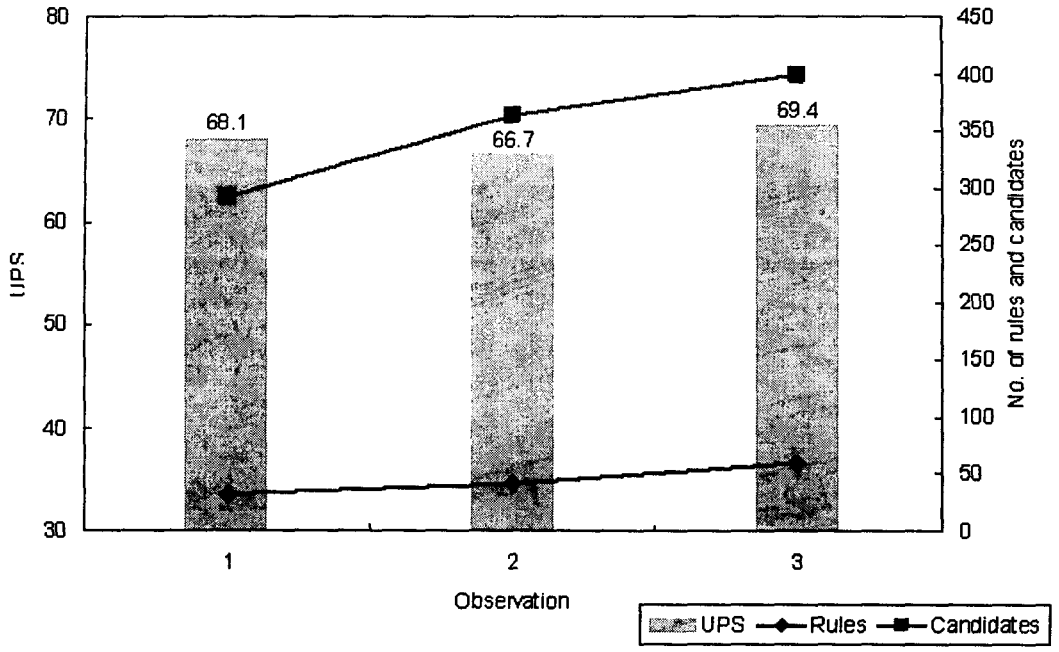

Figure 7. Additional experiments tracking changes in the recommendations over time

6. Conclusions and further research

These days, several big Web sites, such as Yahoo!, Netscape, and Amazon.com, have begun providing personalized content. Most of the online content providers, however, supply the same content for all users, so that they have received poor responses from customers. In addition, most personalization techniques until now have focused on developing recommender systems for physical products. In terms of digital content personalization, very few applications have been reported to date. They have limited effectiveness in recommending digital content, since they are only considering the similarity of content. Therefore, the recommender system described here offers several advantages which overcome the limitations of the traditional recommender systems.

- The recommender system adopts a framework applicable for digital content recommendation which can be executed by the Internet service providers, including the electronic newspaper. It enables the system to translate its performance into customer satisfaction.
- The system discovers word affinities from the user's surfing history for each user segment and makes a recommendation to the user on the basis of their preferences the word affinities manifest. It differs from other personalization systems that use the traditional information retrieval techniques.
- The recommender reveals document-level affinities through mining word-level affinities which reside in news articles. Word-level affinities can discover more accurate decision rules suitable to a user's preference which may be unseen at the document level.
- The recommender develops an evaluation measure, UPS, and adopts it to determine priority of articles that are appropriate for recommendation. UPS is calculated by using the support and confidence measures of the applied association rules. UPS can capture changes in a user's preference over time, which are expressed as changes in association rules.

Future research can extend the work this study presents in several ways: First, when matching content, this study uses the association rules algorithm. Although it has several advantages, including simple rules which are easy to use, a variety of matching techniques need to be developed and their performances need to be compared.

Second, the system may heighten the accuracy of rules and therefore improve the accuracy of recommendations if the word dictionary is organized more efficiently.

Further research including text processing needs to be performed.

References

[1] Aggarwal, C.C. and P.S. Yu, "Data Mining Techniques for Personalization," *IEEE Data Engineering Bulletin*, Vol.23, No.1(2000), pp.4-9.

[2] Baeza-Yates, R. and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley Longman, New York, 1999.

[3] Balabanovic, M. and Y. Shoham, "Fab: Content-based Collaborative Recommendation," *Communications of the ACM*, Vol.40, No.3(1997), pp.66-72.

[4] Claypool, M., A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, "Combining Content-Based and Collaborative Filters in an Online Newspaper," *Proceedings of ACM SIGIR Workshop on Recommender Systems*, Aug. 19, 1999, Berkeley, CA.

[5] Han, J. and M. Kamber, *Data mining: concepts and techniques*, San Diego, CA, 2001.

[6] Ivory, M.Y., R.R. Sinha, and M.A. Hearst, "Empirically Validated Web Page Design Metrics," *Proceedings of ACM SIGCHI'01 Conference on Human Factors in Computing Systems*, Mar. 31-Apr. 4, 2001, Seattle, WA, pp.53-60.

[7] Kohavi, R. and F. Provost, *Applications of Data Mining to Electronic Commerce - Special issue of Data Mining and Knowledge Discovery*, Kluwer Academic Publishers, 2001.

[8] Kosala, R. and H. Blockeel, "Web Mining Research: A Survey," *SIGKDD Explorations*, Vol.2, No.1(2000), pp.1–15.

[9] Langheinrich, M., A. Nakamura, N. Abe, T. Kamba, and Y. Koseki, "Unintrusive Customization Techniques for Web Advertising," *Computer Networks*, Vol.31, No.11-16(1999), pp.1259-1272.

[10] Linoff, G.S. and M.J.A. Berry, *Mining the Web: Transforming Customer Data into Customer Value*, John Wiley & Sons, New York, 2001.

[11] Maes, P., "Agents that Reduce Work and Information Overload," Bradshaw, J.M (ed.). *Software Agent*, AAAI Press/MIT Press, CA, 1997.

[12] Manber, U., A. Patel, and J. Robinson, "Experience with personalization on Yahoo!," *Communications of the ACM*, Vol.43, No.8(2000), pp.35-39.

[13] Mitchell, T.M., *Machine Learning*, McGraw-Hill, Singapore, 1997.

[14] Mladenic, D. and M. Grobelnik, "Feature selection on hierarchy of web documents," *Decision Support Systems*, Vol.35, No.1(2003), pp.45-87.

[15] Peppard, J., "Customer Relationship Management (CRM) in Financial Services," *European Management Journal*, Vol.18, No.3(2000), pp.312–327.

[16] Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, Oct. 22-26, 1994, Chapel Hill, NC, pp.175-186.

[17] Roberts, M.L., *Internet marketing: integrating online and offline strategies*, McGraw-Hill, New York, 2003.

[18] Salton, G. and M.J. McGill, "The SMART and SIRE Experimental Retrieval Systems," *Readings in Information Retrieval (K.S. Jones and P. Willett, eds.)*, Morgan Kaufmann, 1997, pp.381-399.

[19] Sarwar, B., G. Karypis, J. Konstan, and J. Riedl, "Analysis of Recommendation Algorithms for E-Commerce," *Proceedings of ACM conference on E-Commerce 2000*, Oct. 17-20, 2000, Minneapolis, MN, pp.158-167.

[20] Sarwar, B., G. Karypis, J. Konstan, and J. Riedl, "Item-based Collaborative

Filtering Recommendation Algorithms," *Proceedings of the Tenth International World Wide Web Conference*, May 1-5, 2001, Hong Kong, pp.285-295.

[21] Srivastava, J., R. Cooley, M. Deshpande, and P.-N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKDD Explorations*, Vol.1, No.2(2000), pp.12–23.

[22] Sullivan, D., *Document Warehousing and Text Mining*, John Wiley & Sons, New York, 2001.

Appendix

TFIDF (Term Frequency/Inverse Document Frequency) word weighting scheme is described as follows (Baeza-Yates and Ribeiro-Neto, 1999): Each occurring term has to be assigned a unique number that corresponds to a dimension of the term space. This can be done in several different ways; one could simply take ordinal of the term in the alphabets or the order of appearance.

After that, the TFIDF weights for the terms are calculated as:

$$TFIDF = TF(w,d) \times IDF(w)$$

where *TF(w, d)* is the number of times word *w* occurs in a document *d* and *DF(w)* is the number of documents in which the word *w* occurs at least once. The inverse document frequency is calculated as:

$$IDF(w) = \log(\frac{|D|}{DF(w)})$$

where $|D|$ is the total number of documents.