A Personalized Recommendation Procedure based on Dimensionality Reduction and Web Mining

Jae Kyeong Kim [1], Jun Yong Xiang [2], Il Young Choi [3], and Yoon Ho Cho [4]
School of Business Administration, Kyung Hee University [1,2,3], jaek@khu.ac.kr [1],
gysang@yahoo.com [2], best102@bcline.com [3]
Department of Internet Information Systems, Dongyang Technical College,
yhcho@dongyang.ac.kr [4]

ABSTRACT

This paper suggests a recommendation methodology based on Web usage mining and dimensionality reduction techniques to address problems of current collaborative filtering-based recommender systems. Web usage mining is used to capture the accurate customer's behavioral intent. The dimensionality reduction techniques are used to reduce the sparsity of ratings and to improve the scalability of searching for neighbors. Experiments on real e-commerce data result that the quality and performance of the proposed methodology are better than those of other collaborative filtering methodologies.

INTRODUCTION

In various e-commerce applications, personalization has become an important business issue (Peppers, et al., 1993.). One of the successful personalization technologies is collaborative filtering (CF). But this system has some limitations such as sparsity and scalability. In this paper, we propose a new methodology for personalized recommendations to address these problems in online stores. The key characteristics in our suggested methodology are (1) Web usage mining is used to capture customer's preference and relations between products in online stores. (2) Dimensionality reduction techniques using singular value decomposition (SVD) and adjusted product taxonomy are introduced and compared to address sparsity and scalability problem. We also provide experimental evaluation to compare the techniques with real e-commerce data, and some guidelines are discussed to be applied in real field.

BACKGROUND

Collaborative filtering (CF) is a technique to identify customers whose interests are similar to those of a given customer and recommend products they have liked (Pazzani, et al., 1996). However, widespread use of CF systems exposed limitations such as sparsity and scalability problems. To address these problems, various techniques are developed and used.

First, Web usage mining is used to capture the accurate customer preference. Web usage mining is the application of data mining technologies to huge Web data repositories to discover and analyze user's usage pattern (Srivastava, et al., 2000). The result of Web usage mining can be used to perform the Web personalization. Lee et al (2001) divided e-shopper's behavior according to the following four shopping steps; product impression, clickthrough, basket placement and purchase. A part of Lee et al's model was adopted for our research, because they focus the online retailer that is also our consideration.

Second, the dimensionality reduction techniques such as adjusted product taxonomy and SVD technique are used to overcome both problems. SVD is a matrix factorization technique

used for producing low-rank approximation of the original space. It is possible for us to construct a low dimensional matrix by reducing the singular matrix (Sarwar, et al., 2000a). A taxonomy reflects the relationship among different groups of elements. Choosing the right levels of the product hierarchy may lead to improve the results of the analysis.

RESEARCH METHODOLOGY

We divide the entire procedure of CF-based recommendation into four phase; Profile Creation, Dimensionality Reduction, Similarity Computation and Recommendation generation.

Profile Creation: In e-commerce sites, Web data or purchase data are available to look into customer's behavioral intention. Profile creation based on purchased data uses the purchase database of customers.

First, the customer profile based on Web data is constructed based the following three general shopping steps (click-through, basket placement, and purchase) in online stores modified from works of Lee et al. (2001). Let p^c_{ij}, p^b_{ij} and p^p_{ij} be total number of occurrence of click-throughs, basket placements and purchases of customer i for a product j, respectively. This is represented as a $m \times n$ customer-product matrix P, where each entry $p_{i,j}$ is defined as follows:

$$p_{i,j} = \begin{cases} \dfrac{P^c_{i,j}}{\sum_{i=1}^{m} P^c_{i,j}} + \dfrac{P^b_{i,j}}{\sum_{i=1}^{m} P^b_{i,j}} + \dfrac{P^p_{i,j}}{\sum_{i=1}^{m} P^p_{i,j}} & \text{if } p^c_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $i=1,\cdots,m$ (number of customers), $j=1,\cdots,n$ (number of products)

Second, profile using purchase data is collections of historical purchasing transaction of m customer on n products. It is usually represented as a $m \times n$ customer-product matrix P, such that $p_{i,j}$ is one if the *ith* customer has purchased the *jth* product, and zero, otherwise (Sarwar, et al., 2000).

Dimensionality Reduction: To solve sparsity and scalability problem of CF systems, the following dimensionality reduction techniques are used.

First, the product taxonomy is used for identifying similar products and grouping them together, by specifying the level of aggregation in the product taxonomy which is provided by the marketer or domain experts. Such specification is called a grain specification, which is similar to the concept of "cut" proposed by Adomavicius and Tuzhilin (2001). We can consider several examples of specifying the grain as shown in Figure 1 where grains are denoted by shaded region. However although product taxonomy can be provided by marketers or experts, the provided product taxonomy is often necessary to reorganize or adjust the existing product taxonomy as a set of nodes with relatively even data distribution. Accordingly, we provide an algorithm for grain specification using the adjusted product taxonomy.

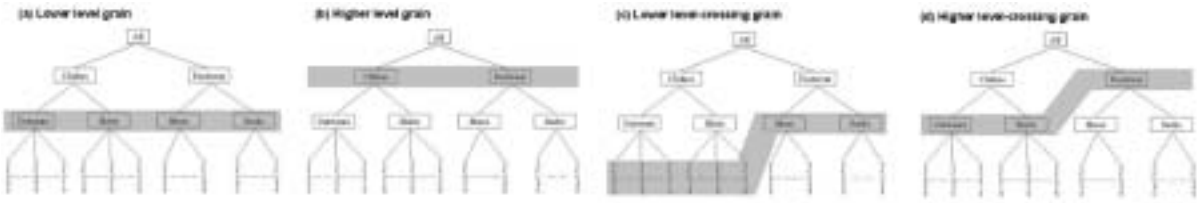

Figure 1. Various types of grain

Second, SVD is a matrix factorization technique commonly used for producing low-rank approximation. The SVD of $m \times n$ matrix P with rank r is its factorization into a product of three matrices as U, S and V. U and V are orthogonal matrices and the diagonal entries of S are the singular values of P sorted in descending order. Accordingly, it is possible to reduce dimensionality by choosing the k most significant dimensions from the factor space which is then used for estimating the rank by evaluating the number of singular value. If we keep k dimensions, matrix, we obtain $U_k S_k^{1/2}$ represented ratings of m users in k-dimensional space. Please refer Sarwar, et al.(2000) for more detail about this.

Similarity Computation: This phase computes the similarity between two customers. The key motivation is that items that are similar or related to the items that he/she has already purchased will be purchased more likely. Given the customer-product matrix P, the similarity between two customers a and b, denoted by $sim(a,b)$, is usually measured using either the correlation or the cosine measure. This paper compute similarity using the following cosine measure, which results to find latent semantic among the product:

$$sim(a,b) = \cos(\vec{a},\vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\| \vec{a} \|^2 \cdot \| \vec{b} \|^2}$$

where $\vec{a}$ and $\vec{b}$ are row vectors of the customer profile for two customers a and b, respectively.

Recommendation Generation: The final step of the recommender system is to derive the top-N recommendations from the neighborhood of customers. Two techniques are used for the top-N recommendations. One is the recommendation of the most frequently purchased product (MFP). This technique, adopted from of the study of Sarwar, et al. (2000), looks into the neighborhood N and for each neighbor, scans through a sales database and counts the purchase frequency of the products. The other is recommendation of the most frequently referred product (MFR) based on purchase frequencies of all neighbors. This technique sorts the products according to their reference frequencies (Cho, 2002).

PEFORMANCE EVALUATION

In this chapter, a case application to the method proposed in chapter 3 is illustrated. This case is based on the real W Internet shopping mall.

Data Set And Evaluation Metrics: Data set for our experiments is a collection of Web log files from 1st May 2001 to 30th May 2001. This data set contains click through, basket-placement and purchase information of 102,284 customers on 3,158 products. The product taxonomy of W shopping mall consists of three levels of hierarchy except the root, “All”. Training period is set between 1st May 2001 and 24th May 2001 and test period is set between 25th May 2001 and 30th May 2001.

We employ two evaluation metrics for evaluating our methodology in terms of quality and performance requirements. First, generally recall and precision are used to evaluate the quality of top-*N* recommendation and are defined as followings; Recall = size of hit set/size of test set, Precision = size of hit set/size of top-*N* set. But increasing the number *N* tends to increase recall but decreases precision, and vice verse. Therefore, to evaluate the quality of top-*N* recommendation we use the standard *F1* metric as the following equation:

$$F1 = \frac{2 \times \mathrm{Re}\,call \times \mathrm{Pr}\,ecision}{\mathrm{Re}\,call + \mathrm{Pr}\,ecision}$$

Second, the response time or the throughput was employed to measure the system performance. The response time is defined as the amount of time required to compute all the recommendations for the training set, and the throughput denotes the rate at which the recommendations are computed in terms of recommendations per second.

Experimental Results: In this section, we present an experimental result for the phases of our recommendation methodology and compare their performance to that of original CF. Before we obtain top-*N* recommendation result, we run experiments as follows:

First, we determined the sensitivity of the neighborhood size. The size of the neighborhood has significant impact on the recommendation quality (Sarwar, et al., 2000b). To determine the sensitivity of neighborhood size, we performed experiments in which we varied the number of neighbors and computed the corresponding *F1* metric. Over experimental results, we can see that the *F1* value of Web data is always higher than that of purchase data and the size of the neighborhood does affect the quality of top-*N* recommendations. In case of Web data it reaches its peak at 50, whereas in case of purchase data it reached at 60.

Second in order to evaluate the impact of grain specification on the recommendation quality, we performed experiments with five types of grains; the lower level grain (labeled T2; Figure 1(a)), the higher level grain (labeled T1; Figure 1(b)), the lower level crossing-grain (labeled TC2; Figure 1(c)), the higher level crossing-grain (labeled TC1; Figure 1(d)), and grain on adjusted taxonomy (labeled TA). Figure 2 shows the comparative results obtained from these five grains for Web data and purchase data, respectively. Looking into the results of these figures, we can see that TA, the grain with an even distribution among products leads to the better quality of recommendations regardless of data type.

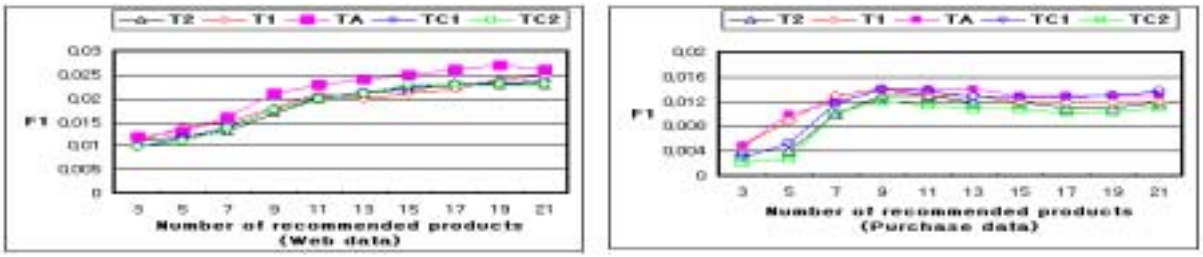

Figure 2. Impact of grain specification on recommendation quality

Finally when using SVD as a dimensionality reduction methodology, the number of dimension has an impact on the recommendation quality. We performed the experiments with large number of dimensions enough for avoiding over-fitting error. In case of Web data, recommendation quality reaches its peak quickly in the rank of lower dimensional space, whereas in case of purchase data, recommendation quality continues to improve to 35

dimensions.

Through the experiments we obtained top-*N* recommendation results of our proposed methodology as shown in Table 1. High dimensional implies no dimensionality reduction is performed, and low dimensional means dimensionality reduction performed by grain adjustment or SVD.

Table 1. Top-*N* comparison of methodologies

Experimental Data set / Dimensionality Reduction		Web data		Purchase data	
		Quality (F1)	Performance (Second)	Quality (F1)	Performance (Second)
Low dimensional	Grain	0.0213	24.2	0.0121	11.1
	SVD	0.0178	22.5	0.012	13.2
High dimensional		0.01762	286.4	0.0087	263.4

We experimentally evaluated our proposed methodology and reached the following conclusions through the result of experiments. First, the performance of CF using the Web data is better than that of CF using the purchase data. Second, the recommendation quality and scalability-related performance of CF using dimensionality reduction algorithm are better than those of CF using original sparse data set. Third, using adjusted product taxonomy results better or almost the same performance than using the SVD. Fourth, experiments with diverse grain specification show that our suggested algorithm for grain specification using adjusting product taxonomy results better performance than other grains. This implies that grain specification is determined without the burden of marketers.

CONCLUSION

In this paper we identified the problems of collaborative filtering techniques and showed how these problems can be addressed through our methodology. Our results show that the performance of adjusted product hierarchy is better than that of other methodologies. However these results are based on experiments limited to the particular e-commerce site that has not enough customers, products, and transactions. Therefore, it is required to evaluate our methodologies in more detail using data sets from a variety of large e-commerce sets.

REFERENCES

Adomavicius, G., Tuzhilin, A., “Expert-driven validation of rule-based user models in personalization applications,” *Data Mining and Knowledge Discovery*, 5 (1-2), 2001

Cho, Y. H., Kim, J,K. Kim,S,H., “A Personalized Recommender System based on Web Usage Mining and Decision Tree Induction,” *Expert Systems With Applications*, 2002

D.Peppers and M. Rogers, “The One-to-One Future,” Doubleday, 1993

Lee, J., Podlaseck, M., Schonberg, E., Hoch, R., “Visualization and analysis of clickstream data of online stores for understanding web merchandising,” *Data Mining and Knowledge Discovery*, 5 (1-2), 2001

Pazzani, M., Muramatsu, J., Billsus, D., “ Syskill & Webert: identifying interesting web site,” In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1996

Sarwar, B. M., Karypis, G., Konstan, J. A., & Riedl, J., “Application of Dimensionality Reduction in Recommender System - A Case Study,” In *Proceedings of the ACM WebKDD-2000 Workshop*, 2000

Srivastava, J., Cooley, R., Deshpande, M., & Tan P., “Web usage mining: discovery and applications of usage patterns from web data,” *SIGKDD Explorations*, 1 (2), 2000