Applying Centrality Analysis to Solve the Cold-Start and Sparsity Problems in Collaborative Filtering

Yoonho Cho
College of Business Administration,
Kookmin University
(www4u@kookmin.ac.kr)

Jounghae Bang
College of Business Administration,
Kookmin University
(bangjh@kookmin.ac.kr)

Collaborative Filtering (CF) suffers from two major problems : sparsity and cold-start recommendation. This paper focuses on the cold-start problem for new customers with no purchase records and the sparsity problem for the customers with very few purchase records. For the purpose, we propose a method for the new customer recommendation by using a combined measure based on three well-used centrality measures to identify the customers who are most likely to become neighbors of the new customer. To alleviate the sparsity problem, we also propose a hybrid approach that applies our method to customers with very few purchase records and CF to the other customers with sufficient purchases. To evaluate the effectiveness of our method, we have conducted several experiments using a data set from a department store in Korea. The experiment results show that the combination of two measures makes better recommendations than not only a single measure but also the best-seller-based method and that the performance is improved when applying the hybrid approach.

Received : August 01, 2011 Revision : August 08, 2011 Accepted : August 16, 2011
Type of Submission : APJITM Special Corresponding author : Yoonho Cho

1. Introduction

Due to the advances of the Internet, the quantity and quality of Web retailers are rapidly increasing. As much, Web retailers have faced severe competitions. For the companies, therefore, it has become critical to have the strategic weapons to obtain competitive advantage. For customers, the growth in the number of web retailers makes more products and suppliers available online. However it also makes customers struggle to search for and evaluate so many different products online and find the right one to satisfy their needs (Cho and Kim, 2004). Therefore

* This work was supported by Research Program 2010 of Kookmin University in Korea.

companies focus on enhancing their competitiveness by using one-to-one marketing, which differentiates their products to satisfy each individual customer, and/or Customer Relationship Management (CRM), which focuses on understanding customers from customers' perspectives and on building and maintaining the relationships with customers. One of the rising issues in CRM is recommender systems. Recommender system is the system which, by using automated information filtering technology, recommends products or services to the customers who are likely to be interested in. Those systems are widely used in many different Web retailers such as Amazon. com, Netfix. com, and CDNow.com. The most essential part of the recommender systems is to accurately analyze and predict customers' preferences in order to recommend the right products that customers want at the right time. Various recommender systems have been developed. Among them, Collaborative Filtering (CF) (Sarwar et al., 2000; Adomavicius and Tuzhilin, 2004; Cho and Kim, 2004) has been known as the most successful and commonly used approach.

The CF system recommends products to a target customer according to the following steps (Cho and Kim, 2004) : (1) By being provided with ratings from customers or by analyzing purchase transactions, the system builds the customer-product matrix which represents the preference scores of individual customer on each product. (2) Then, the system applies the statistical techniques or machine learning techniques to find a set of customers, known as neighbors, who have rated similarly or purchased similar set of products. Usually, a neighborhood is formed by the degree of similarity between the customers. (3) Once a neighborhood of similar customers is formed, the system generates a set of products that the target customer is most likely to purchase by analyzing the products the neighbors have purchased. These systems, also known as the nearest neighbor CF-based recommender systems, have been widely used in practice. However, as the number of customers and that of products managed in a Web retailer grows rapidly, CF suffers from two major problems that must be addressed (Adomavicius and Tuzhilin, 2004; Cho and Kim, 2004; Huang et al., 2004).

The first problem is related to sparsity. In a large Web retailer such as Amazon.com, there are millions of products and so customers may rate only a very small portion of those products. Most similarity measures used in CF do not work properly unless sufficient ratings are provided from customers. Such sparsity in ratings makes the formation of neighborhood inaccurate, thereby resulting in poor recommendation. Many approaches have been proposed to overcome the sparsity problem. These approaches can be classified into three categories : implicit rating, hybrid filtering and dimension reduction. The implicit rating approaches attempt to increase the number of ratings through observing customers' behavior. For this, many studies utilized analyzing click-streams in the Web retailer via data mining techniques (Cho and Kim, 2004; Lee et al., 2010). The hybrid filtering approaches combine content-based filtering and CF for augmenting sparse preference ratings (Aggarwal and Yu, 2000; Ziegler et al., 2000; Melville et al., 2002). These approaches learn to predict which products a given customer will like by matching properties associated with each product to those associated with products that he/she has liked in the past, and then use such a content-based prediction to convert a sparse customer profile into a dense one. The dimensionality reduction technique is used to project high dimensional (sparse) data into low dimensional (dense) one through concentrating most of information in a few dimensions. Singular Value Decomposition (SVD) is a well-known method for matrix factorization that factors the original rating space into three matrices and performs the dimensionality reduction by reducing the singular matrix (Billsus and Pazzani, 1998; Sarwar et al., 2000b; Kim and Cho, 2003).

The second problem is related to cold-start recommendations. The cold-start problem refers to the situation where a new customer or product has just entered the CF system (Schein et al., 2002). CF generates poor recommendations for the new customer because of the lack of previous ratings or purchases, and also cannot recommend the new product to many customers because very few customers have yet rated or purchased this product. The cold-start problem can be viewed as a special case of the sparsity problem in which most of cells in the customer-product matrix are empty (Huang et al., 2004). In order to solve the new customer recommendation problem, there have been three methods developed and commonly used : the best-seller based method, demographic information based method, and the explicit rating minimization method. The best-seller based method sorts the products based on their sales volume and recommends the high-ranked products to the new customers (Sarwar et al., 2000a). This method is used by many Web retailers because it is simple to apply and it does not need any additional information. However this method generates poor recommendations because personalized recommendations are not possible and the categories of recommended products are likely to be less heterogeneous. To recommend the products, the demographic information based method uses such information as gender, age, and occupation that customers provide to the Web retailer (Krulwich, 1997; Aggarwal et al., 1998). There are two different approaches for this method. One approach identifies customers whose demographic information is similar to that of the new customer and recommends the products which those customers have purchased (Krulwich, 1997). The other approach mines the association rules between demographic and purchase information and applies these rules to recommend (Aggarwal et al., 1998). This method can make personalized recommendations and utilize gathered demographic information for other demands. However, it is difficult to obtain sufficient amount of reliable demographic information. Moreover, it requires much processing time to make recommendations. The explicit rating minimization method learns new customers' preferences while reducing the burden of explicit ratings by decreasing the number of products customers have to rate explicitly (Schein et al., 2002; Yu et al., 2004; Park et al., 2006). It focuses on determining products to customers for rating based on the characteristics such as product popularity and product entropy. Nevertheless, it still requires the burdens for customers to actively rate, and generates poor recommendations when new customers provide inaccurate rating information.

Recently there are many studies paying attention to Social Network Analysis (SNA) as a method to analyze social relationships among people. SNA is a method to measure and visualize the linkage structure and status focusing on interaction among objects within communication group (Wasserman and Faust, 1994; Scott, 2000). The method has been widely used to search for relationships among social entities such as genetics network, traffic network, organization network, etc. (Kauffiman, 1993). CF analyzes the similarity among previous ratings or purchases of each customer, finds the relationships among the customers who have similarities, and then uses the relationships for recommendations. Thus CF can be modeled as a social network in which customers are nodes and purchase relationships between customers are links. (Ryu et al., 2006; Park et al., 2009; Cho and Kim, 2010). That is, CF system is an artificially structured social network system for effective and efficient recommendations (Kim et al., 2011).

Our previous work (Park et al., 2009) proposed a new method for the new customer recommendation using the centrality analysis (Freeman, 1979; Bonacich, 1987; Borgatti, 2005) that has been used to inspect the relative importance or influence of individual nodes within a social network. By using the degree centrality analysis, the method identifies the customers who are most likely to become neighbors of the new customer, and recommends the products which the neighbors have most purchased in the past. Even though the method significantly outperforms the best-seller-based method, it leaves much room for enhancing its performance because this method only uses a single centrality measure among numerous different measures. This study extends our previous method by using multiple measures of centrality analysis. To make centrality analysis more effective, we suggest a combined measure based on three well-used centrality measures including degree centrality, closeness centrality and betweenness centrality. To find which centrality measure or combination of them shows the best performance, we also conduct the experiments with several different weights. We propose the recommendation procedure modified to employ the combined measure.

This paper focuses on solving the problems of sparsity as well as the cold-start recommendation. Because the cold-start problem can be viewed as a case of the sparsity problem, we attempt to alleviate the sparsity problem with our proposed method. By regarding the customers with very few records as new customers, our method is applicable to solve the sparsity problem. This paper proposes a hybrid approach that applies our method to customers with very few purchase records while employing CF to other customer (with sufficient purchase records).

We begin by reviewing social network and centrality analysis which are prerequisite for our study in section 2. In section 3, our recommendation procedure is explained step by step, and experimental results and discussions are provided in section 4. Finally, section 5 concludes this paper with suggestions for future research.

2. Social Network and Centrality Analysis

Over the past 40 years, the studies in social network have attempted to understand the social interactions through the network structure represented by connection patterns. In other words, social network analysis has been used to explain the structures and behaviors of various social formations such as teams, organizations, and industries (Kukkonen et al., 2010). In general, the social network analysis uses data as a form of matrix. The matrix depicts the relations between rows as actors and columns as events, where the relations are represented as either binary or valued scores.

Even though there are no direct relations between customers who have purchased the products, the relations between customers or between products can be derived artificially as in the customer-product matrix, in which each cell has 1 for purchase or 0 for non-purchase. For example, a customer network can be configured in a way to connect the customers who have purchased one or more same products. In order to identify the relations between two customers, the number of purchases of same products, the cosine vector, the correlation coefficient, and Jaccard similarity are generally used.

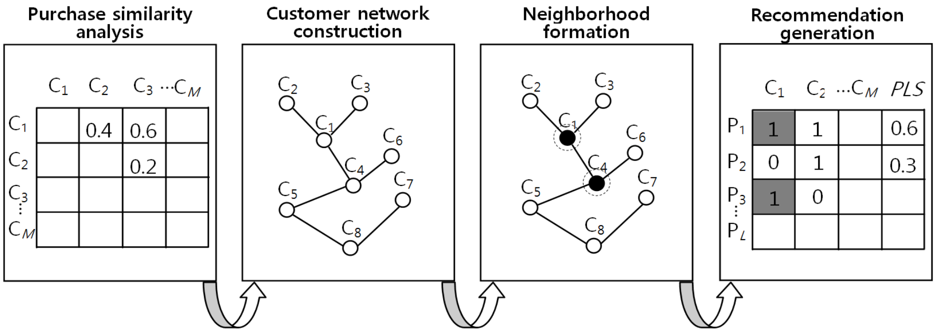
Social networks are categorized into the ego-centric network, the dyadic network, and the total network (Human and Provan, 2000; Weare et al., 2007). The ego-centric network is used to analyze the connections between one member and other members around the one while the dyadic network is used to analyze all the connections between certain two members. The total network is most often used to analyze all the connections among the entire members. Depending on the presence or absence of the direction of relations, social networks can be divided into directed networks and undirected networks (Wasserman and Faust, 1994). The directed networks have directions with starting and ending points between members while the undirected networks have the mutual relationship between the two members without any direction. The relationship among the customers based on their purchases is considered undirected. Therefore this study focuses on the undirected total network.

In social network analysis, identifying the most important or visible actors within the total network is one of the primary uses of this analysis (Zemljiˇc and Hlebec, 2005). Centrality in social networks is the concept which reflects different actors' varying importance for the structural properties of the network (Frank, 2002). Centrality has been used to investigate many different things such as the influence in inter-organizational networks, power in organizations, employment opportunities, and adoption of innovation (Borgatti and Everett, 2006). Measures of centrality describe actors' positions in a network relative to others and in relation to the total network (Costenbader and Valente, 2003).

Numerous measures have been developed, including degree centrality, closeness, betweenness, eigenvector centrality, information centrality, flow betweenness, the rush index, etc (Borgatti, 2005). Among them, many studies have used three well-known measures of centrality : degree, closeness and betweenness centrality (Freeman, 1979).

- Degree centrality : It is defined as the number of the nodes directly linked to a node. This measure considers only direct relations with other nodes, not indirect ones. Assume that A has only one relationship with B, and B only has a relationship with C. When degree centrality of A is calculated, C is not included because it only has an indirect relationship with A. Because degree centrality is measured by only relations within a regional boundary, it is viewed as a local centrality.
- Closeness centrality : It measures the centrality based on the distance between two nodes. Unlike degree centrality, it takes into account all the nodes which are connected both directly and indirectly in a network. Assume that a man attempts to meet everyone in a group that he belongs to. He cannot know every single person in the group by himself. However, He can get to know all the people through those people whom he knows and who know others that he doesn't know in the group. In this case, all the shortest paths through which he knows each person form closeness centrality. Closeness centrality is defined as the sum of the numbers of all the links counted in each shortest path to reach from a node to another. The higher closeness centrality, the more centered a node positions in the network.
- Betweenness centrality : Betweenness centrality measures the degree to which a node mediates or coordinates in a network. Assume that, among the three people, A, B, and C, A is able to reach to B only through C. C is potentially playing a role of 'broker', or 'gate keeper' which can control relationships among people. In this situation, C is viewed to have high betweenness centrality. The more one node places in the paths between other nodes, the higher betweenness centrality of the node becomes.

Recently, several studies attempted to apply the concept or analysis of social networks to solve some issues of CF. Cho and Bang (2009) suggested a new product recommendation method which applies centrality concept. In the method, the new products are recommended to the customers who are highly likely to buy the products, by analyzing the relationships among products with several centrality measures. Cho and Kim (2010) proposed an efficient approach to predict the performance of CF before building the CF system. They applied SNA to explore the topological properties of the network structure hidden in purchase transactions used for CF recommendations, and to build a prediction model. Kim et al. (2011) proposed a customer-driven recommender system which uses a local processing method in preference-based customer network, a form of social network for content recommendation. While the previous studies have focused on the structural concepts of social network, the focus of Kim et al. (2011) was on the active construction and operation of social networks for effective and efficient content recommendations.



<Figure 1> Overall Recommendation Process

3. Recommendation Procedures

3.1 Overview

<Figure 1> shows an overall procedure of new customer recommendation by using the social network analysis. The procedure consists of four phases : purchase similarity analysis, customer network construction, neighborhood formation, and recommendation generation. Purchase similarity analysis phase calculates the degree of purchase similarity among customers by using their purchase histories. A well- known similarity measure, Jaccard coefficient is used to calculate the similarity. Customer network construction phase represents individual customer as a node and builds a link between two nodes based on the similarity between two customers. For example, a link between two nodes can be connected when the two customers had purchased at least one common product previously. Neighborhood formation phase calculates the centrality of individual nodes in the customer network, and identifies the nodes whose centralities are high. Those nodes become new customers' neighbors whose roles are like those of neighbors in CF. Recommendation generation phase finds the products which neighbors of new customers have purchased and recommends among them those products which new customers are most likely to purchase.

3.2 Phase I : Purchase similarity analysis

This phase consists of two steps : representing affiliation relations and calculating purchase similarity.

First, from the purchase transaction database, which products each customer has purchased are analyzed and the results are represented as a $M \times N$ customer-product matrix $\mathrm{P} = (p_{ij})$:

$$p_{ij} = \begin{cases} 1 \ \text{if customer } i \text{ has purchasedproduct } j \\ 0 \ \text{ifcustomer } i \text{ has not purchasedproduct } j \end{cases} \quad (1)$$

where i = 1 to M, j = 1 to N, M is the total number of customers, and N is the total number of products.

Note that the values in the matrix are binary since multiple purchases of a product are not taken into account in Equation (1) while they are asymmetric since 1 (purchase) and 0 (non-purchase) are not equally important.

Second, based on the customer-product matrix, purchase similarity between each pair of customers is calculated. Treating binary variables as if they are interval-scaled can lead to misleading results. To compute similarity, therefore, we use Jaccard similarity coefficient (Choi et al., 2010), which is commonly used for assessing the similarity between asymmetric binary variables. Purchase similarity between customers a and b, *pur_sim*(a, b) is defined as

$$pur\_sim(a, b) = J(a, b) = \frac{m_{11}}{m_{11} + m_{10} + m_{01}} \quad (2)$$

where m_{11} is the number of columns that equal 1 for customers a and b, m_{10} is the number of columns that equal 1 for customer a but that are 0 for customer b, and m_{01} is the number of columns that equal 0 for customer a but equal 1 for customer b. The values for similarity range from 0 to 1. If the coefficient is 1, it means the two customers have purchased all the same products.

3.3 Phase II : Customer Network Construction

This phase constructs a customer network which connects the customers whose purchase patterns are similar. A customer network is an undirected graph $G(V, E)$, in which V is a set of nodes and E is a set of edges. Each node $V_a \in V$ represents a customer and each edge (V_a, V_b) where $V_a, V_b \in V$, represents the connectedness between customers a and b. The decision whether the two customers a and b are connected or not, follows Equation (3).

$$connectedness(a, b) = \begin{cases} 1, & pur\_sim(a, b) \geq \rho \\ 0, & pur\_sim(a, b) < \rho \end{cases} \quad (3)$$

where $o < \rho \leq 1$. That is, *connectedness*(a, b) equals 1 if the purchase similarity between customers a and b is more than an arbitrary threshold value ρ, and that means those two customers are connected in the network. The threshold value ρ is determined based on domain expertise or characteristics of purchase transactions. In this customer network, a customer with many links implies that the customer has similar purchase patterns to many other customers.

3.4 Phase III : Neighborhood Formation

This phase selects neighbors of new customers through centrality analysis for the customer network. Centrality for each customer in the network is calculated and top K customers with high centrality form a neighborhood (H) for new customers. The products that the neighbors have purchased are recommended to the new customers.

Three measures are used for the centrality analysis : degree centrality, closeness centrality and betweenness centrality. Degree centrality of customer a is calculated as follow.

$$d\_cen(a) = \frac{\sum_{j=1}^{M} connectedness(a, j)}{M-1} \quad (4)$$

where j $(j \neq a)$ is a customer and M is a total number of customers. Closeness centrality of customer a is calculated as follow.

$$c\_cen(a) = \frac{M-1}{\sum_{j=1}^{M} dist(a,j)} \tag{5}$$

where $j(j \neq a)$ is a customer, *M* is a total number of customers, and $dist(a, j)$ is a shortest path from customer *a* to *j*. Betweenness centrality of customer *a* is calculated as follow.

$$b\_cen(a) = \frac{2 \times \sum_{j<k} geod(j, k, a)/geod(j, k)}{M^2 - 3 \times M + 2} \tag{6}$$

where $j(j \neq a)$ is a customer, $geod(j, k)$ is the number of shortest pathways between customer *j* and *k*, and $geod(j, k, a)$ is the number of shortest pathways running through customer *a* among shortest pathways between customer *j* and *k*. Note that each value from the Equations (4), (5), and (6) ranges from 0 to 1 because they are all relative measures.

To calculate the centrality for each customer considering outputs of three measures together, we use Equation (7) :

$$cen(a) = w_d \times d\_cen(a) + w_c \times c\_cen(a) + w_b \times b\_cen(a) \tag{7}$$

where w_d, w_c, and w_b indicate weights for each measure, and $w_d + w_c + w_b = 1$. Thus, the value of Equation (7) ranges from zero to one.

To find which centrality measure or combination of them shows the best performance, we conduct the experiments with several different weights. Based on the Equation (7), we determine the neighborhood $H = \{h_1, h_2, \cdots, h_k\}$ such that $cen(h_1)$ is the highest centrality, $cen(h_2)$ is the next highest, and so on.

3.5 Phase IV : Recommendation Generation

This phase finds the products purchased by top *K* customers (neighborhood *H*) with high centrality identified in the previous phase, and calculates new customers' purchase likelihood of those products. Based on the purchase likelihood scores, top-*N* products are recommended. Given that *PLS* (*j*) denotes the purchase likelihood score of a new customer for product *j*, this phase generates a list of *n* products, $R = \{r_1, r_2, \cdots, r_n\}$ such that *PLS* (r_1) is the highest, $PLS(r_2)$ is the next highest, and so on. *PLS*(*j*) is computed as follows :

$$PLS(j) = \frac{\sum_{i \in H} p_{ij} \times cen(i)}{\sum_{i \in H} cen(i)} \tag{8}$$

4. Experimental Evaluation

4.1 Data and Evaluation Metric

For the experiments, we used the transaction data of a leading department store in Korea. The data set contains 198 products, 5,000 customers, and 128,720 transactions over one year period between May 2000 and April 2001. The transaction data are divided over the timeline into two sets-a training set and a test set. The initial customer network is formed from transaction data for the training period. The test set is used for the evaluation of the proposed method as well as for re-forming the network.

The *hit ratio* is used to measure the accuracy of recommendations in our experiments. It is defined as the fraction of successful recommendations (i.e., the probability that the target customer will purchase one of recommended products).

4.2 Experiment I : Solving New Customer Recommendation Problem

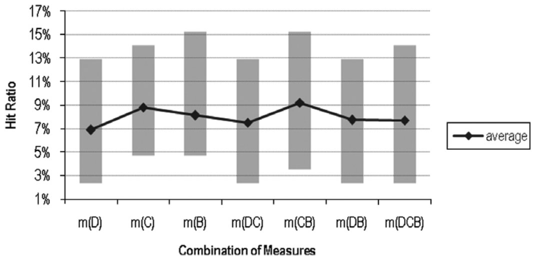
To evaluate the performance for new customer recommendations, the experiments were carried out with varying threshold values of purchase similarity (0.1 through 0.5) and varying numbers of neighbors (10 through 200). In order to find which centrality measure or combination of them, we also conducted the experiments with several different weights of the measures as shown in <Table 1>. In the <Table 1>, m(D) means that only the weight of degree centrality is 1 and all other weights are 0 in the equation (7) of section 3.4. It implies that we only use the degree centrality not other measures. Likewise, m(C) and m(B) denote the only use of closeness centrality, and betweenness centrality, respectively. m(DC), m(CB), and m(DB) denote the combinations of two measures while m(DCB) denotes the combination of three measures with equal weights.

<Table 1> Weights of Centrality Measures

combination	w_d	w_c	w_b
m(D)	1	0	0
m(C)	0	1	0
m(B)	0	0	1
m(DC)	1/2	1/2	0
m(CB)	0	1/2	1/2
m(DB)	1/2	0	1/2
m(DCB)	1/3	1/3	1/3

The experiments used the test set to observe whether the recommendations match with the new customer's real purchases in the test set. Here the new customer is defined as the one who makes a first purchase during the test period. The customer-product matrix was updated and the customer network was re-formed in accordance with the purchase transactions in the test set.

<Figure 2> shows the recommendation accuracy results with top-10 recommendations. In <Figure 2>, each bar depicts the range of hit ratio values which result from each combination of different weights of measures. As a result, it is shown that m(CB)'s average hit ratio is better than those of other combinations. Its maximum, average, and minimum values of hit ratio are 15.2%, 9.2%, and 3.5%, respectively. Even though the deviation of its values is relatively large, the average and the maximum values of m(CB) are highest. This result implies that the combination of closeness and betweenness centrality makes the best quality of recommendations.



<Figure 2> Accuracy Comparison between Combinations of Centrality Measures

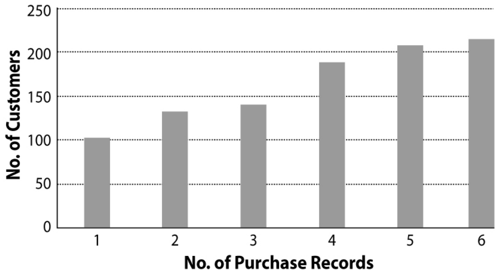
The experiment results also show that the recommendation accuracy of m(D) is the worst. Its maximum, average, and minimum value of hit ratio are 12.9%, 6.9%, and 2.4%, respectively. Our previous work (Park et al., 2009) reported that the new customer recommendation using only degree centrality analysis, m(D) in this study, is better than the best-seller-based recommendation. As <Figure 2> shows, the average hit ratio of m(CB) is about 33 percent better than m(D). These results indicate that the approach with the combination of closeness and betweenness centrality makes better recommendations to the new customer than not only the degree centrality analysis but also the best-seller-based recommendation.

In our experiment, any use of degree centrality in the combination decreases the performance. Therefore we need to focus on the combination of closeness and betweenness centrality. In this study we have used the same weights for the two measures, but it will be interesting to examine if the performance becomes better as we use different sets of weights. Furthermore we need to explain what the experiment results mean in the view of social network theory.

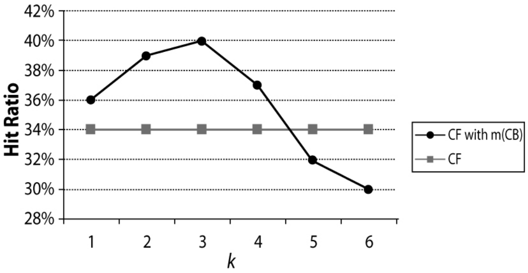
4.3 Experiment II : Alleviating Sparsity Problem

The second experiment focuses on alleviating the sparsity problem with our new customer recommendation method. The sparsity problem occurs when transactions are sparse and insufficient for identifying neighbors and it is a major issue because it reduces the quality of recommendations and limits the applicability of CF (Huang et al., 2004). Significant portion of customers with a few purchase records makes CF generate poor recommendations. By regarding the customers with few records as new customers, our method is applicable to solve the sparsity problem. That is, we propose a hybrid approach which applies CF to the customers with sufficient purchase records while employing the new customer recommendation method to the customers with a few purchase records.

The experiments used m(CB) for new customers (i.e., customers with few records) and the most referenced CF method (Sarwar, et al., 2000a) for other customers. For the purpose of simplicity of our experiment, we fixed the number of neighbors as 40 and the threshold value of purchase similarity as 0.1 in both m(CB) and CF. The experiments were carried out with varying numbers *k* (1 through 6) of purchase records. *k* is the determinant value where a new customer is defined as a customer with *k* or less purchase records. For example, customers with four or more purchases are regarded as existing customers while customers with three or less purchases are as new customers if *k* = 3. In this case, CF is applied to the customers with four or more purchases while our method is applied to those with three or less purchases. <Figure 3> depicts the distribution of customers with few purchase records. In the experiments, 990 customers among total 5,000 customers are considered as customers to whom our approach can be applied. We compared the performance of our hybrid approach with that of CF applied to all the customers.



<Figure 3> Distribution of Customers with Few Purchase Records



<Figure 4> Accuracy Comparison between Hybrid Approach and CF

<Figure 4> shows the recommendation accuracy results with top-10 recommendations. As <Figure 3> shows, the performance increases from k = 1 to 3 and decreases from k = 4 to 6. The accuracy of recommendation is the best (hit ratio of 40%) when k = 3. The hit ratio of CF is 34%. This result indicates that the performance gain of about 18 percent is produced when applying the hybrid approach to customers with three or less purchases instead of CF. The smaller the number of customers regarded as new customers, the less effect on performance. Because there are relatively fewer customers to whom our approach can be applied, as shown in <Figure 3>, the performance when k = 1 is less than when k = 3. The performance decreases from k = 4 to 6, because of increase in the portion of the customers, to whom CF makes better recommendations than the hybrid approach does. None the less, it is found that the performance of the hybrid approach is better than CF when k = 4. This means our approach can be applicable even when k = 4. However, k can vary depending on different datasets. Thus, it is needed to determine the best k for a given dataset depending on its characteristics such as mean and deviation of purchase frequencies. It remains for further study.

In sum, the experimental result validates that our method can be a viable solution to solve the sparsity problem in CF.

5. Concluding Remarks

This paper focused on solving the problems of the cold-start recommendation as well as sparsity. For the purpose, we proposed a method for the new customer recommendation by using multiple measures of centrality analysis. A combined measure based on three well-used centrality measures including degree centrality, closeness centrality and betweenness centrality was used to identify the customers who are most likely to become neighbors of the new customer. Among others, the combination of closeness and betweenness centrality was found to make better recommendations than not only a single measure but also the best-seller-based method. Next, we attempted to alleviate the sparsity problem with the proposed new customer recommendation method. We proposed a hybrid approach that applies the method to customers with very few purchase records while employing CF to the other customers. The experiments show that the performance is improved when applying the hybrid approach to customers with very few purchases instead of CF. Consequently, our study validates that our method can be a viable solution to solve both the sparsity and the cold-start problem in CF.

However, our method offers the same products to recommend to all the new customers because it uses the purchase information from the same neighbors. That means that the method is not able to recommend a personalized product to each new customer. In the experiments, we measured the recommendation accuracy with the past purchase data. However it is necessary to test in a future the proposed method with the live experiments because in reality, customers make purchase decisions in various and dynamic situations. This study proposed the recommendation method using only centrality measure. However, there are many other measures in SNA used to examine and analyze the characteristics of customer networks. To enhance the performance, we are currently extending our method by considering such measures as network density, inclusiveness, clustering coefficient, network centralization, and efficiency.

References

Adomavicius, G. and A. Tuzhilin, "Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions", *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.6 (2005), 734~749.

Aggarwal, C. C., Z. Sun, and P. S. Yu, "Online Algorithms for Finding Profile Association Rules", *Proceedings of the seventh international conference on Information and Knowledge Management*, (1998), 86~95.

Aggarwal, C. C. and P. S. Yu, "Data mining techniques for personalization", *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, Vol.23(2000), 4~9.

Billsus, D. and M. J. Pazzani, "Learning collaborative information filters", *Proceedings of the 15th International Conference on Machine Learning*, (1998), 46~54.

Bonacich, P., "Power and Centrality : A Family of Measures", *American Journal of Sociology*, Vol. 92(1987), 1170~1182.

Borgatti, S. P., "Centrality and network flow", *Social Networks*, Vol.27, No.1(2005), 55~71.

Borgatti, S. P. and M. G. Everett, "A Graph-theoretic perspective on centrality", *Social Networks*, Vol.28(2006), 466~484.

Cho, Y. H. and J. H. Bang, "Social Network Analysis for New Product Recommendation", *Asia Pacific Journal of Intelligent Technology and Management*, Vol.15, No.4(2009), 123~140.

Cho, Y. H. and I. H. Kim, "Predicting the Performance of Recommender Systems through Social Network Analysis and Artificial Neural Network", *Asia Pacific Journal of Intelligent Technology and Management*, Vol.16, No.4 (2010), 159~172.

Cho, Y. H. and J. K. Kim, "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce", *Expert Systems with Applications*, Vol.26, No.3(2004), 234~246.

Choi, S. S., S. H. Cha, and C. Tappert, "A Survey of Binary Similarity and Distance Measures", *Journal on Systemics*, *Cybernetics and Informatics*, Vol.8, No.1(2010), 43~48.

Costenbader, E. and T. W. Valente, "The stability of centrality measures when networks are sampled", *Social Networks*, Vol.25(2003), 283~307.

Frank, O., "Using centrality modeling in network surveys", *Social Networks*, Vol.24(2002), 385~394.

Freeman, L., "Centrality in Social Networks : Conceptual clarification", *Social Networks*, Vol.1(1979), 215~239.

Huang, Z., H. Chen, and D. Zeng, "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering", *ACM Transactions on Information Systems*, Vol.22, No.1(2004), 116~142.

Human, S. E. and K. G. Provan, "Legitimacy Building in the Evolution of Small-Firm Multilateral Networks : A Comparative Study of Success and Demise", *Administrative Science Quarterly*, Vol.45, No.2(2000), 327~365.

Kauffiman, S., *The Origins of Order*, Oxford University Press, 1993.

Kim, H. K., Y. U. Ryu, Y. H. Cho, and J. K. Kim, "Customer-Driven Content Recommendation over a Network of Customers", *IEEE Transactions on Systems, Man, and Cybernetics-Part A : Systems and Humans*, Forthcoming, 2011.

Kim, J. K. and Y. H. Cho, "Using Web Usage Mining and SVD to Improve E-commerce Recommendation Quality", *LNAI*, Vol.2891 (2003), 86~97.

Krulwich, B., "Lifestyle Finder : Intelligent User Profiling Using Large-Scale Demographic Data", *Artificial Intelligence Magazine*, Vol. 18, No.2(1997), 37~45.

Kukkonen, H. O., K. Lyytinen, and Y. J. Yoo, "Social Networks and Information System s : Ongoing and Future Research Streams", *Journal of the Association for Information Systems*, Vol.11(2010), 61~68.

Lee, S. K., Y. H. Cho, and S. H. Kim, "Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations", *Information Sciences*, Vol.180, No.11 (2010), 2142~2155.

Melville, P., R. J. Mooney, and R. Nagarajan, "Content-boosted Collaborative Filtering", *Proceeding SIGIR 2001 Workshop on Recommender Systems*, 2001.

Park, J. H., Y. H. Cho, and J. K. Kim, "Social Network : A Novel Approach to New Customer Recommendations", *Asia Pacific Journal of Intelligent Technology and Management*, Vol.15, No.1(2009), 123~140.

Park, S. T., Pennock, O. Madani, N. Good, and D. DeCoste, "Naive Filterbots for Robust Cold-Start Recommendations", *KDD*, 2006.

Ryu, Y. U., H. K. Kim, Y. H. Cho, and J. K. Kim, "Peer-oriented content recommendation in a social network", *Proceedings of the Sixteenth Workshop on Information Technologies and Systems*, (2006), 115~120.

Sarwar, B., G. Karypis, J. A. Konstan, and J. Riedl, "Analysis of recommendation algorithms for e-commerce", *Proceedings of ACM E-commerce 2000 conference*, (2000), 158~167.

Sarwar, B., G. Karypis, J. A. Konstan, and J. Riedl, "Application of Dimensionality Reduction in Recommender Systems-A Case Study", *Proceedings of ACM WebKDD Workshop*, 2000.

Schein, A. I., A. Popescul, D. M. Pennock, and L. H. Ungar, "Methods and Metrics for Cold-Start Recommendations", *SIGIR*, 2002.

Scott, J., *Social Network Analysis : A Handbook*. Thousand Oaks, CA : Sage, 2000.

Wasserman, S. and K. Faust, *Social network analysis : methods and applications*, Cambridge University Press, 1994.

Weare, C., W. E. Loges, and N. Oztas, "Email Effects on the Structure of Local Associations : A Social Network Analysis", *Social Science Quarterly*, Vol.88, No.1(2007), 222 ~243.

Yu, K., A. Schwaighofer, V. Tresp, X. Xu, and H. Kriegel, "Probabilistic Memory-based Collaborative Filtering", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No.1(2004), 56~69.

Zemljiˇc, B. and V. Hlebec, "Reliability of measures of centrality and prominence", *Social Networks*, Vol.27(2005), 73~88.

Abstract

협업필터링의 신규고객추천 및 희박성 문제 해결을 위한 중심성분석의 활용

조윤호[*] · 방정혜[**]

본 연구에서는 협업필터링의 두 가지 근본적인 문제인 신규고객 추천(cold-start recommendation)과 희박성(sparsity) 문제를 해결하고자 한다. 먼저, 사회 네트워크 분석에서 가장 많이 활용 되고 있는 세 가지 중심성 지표인 연결중심성(degree centrality), 근접중심성(closeness centrality), 매개중심성(betweenness centrality)을 결합한 다양한 중심성 지표들을 만든 후 이를 기반으로 신규고객의 잠재 이웃고객을 찾고 그 이웃고객들의 구매정보를 이용하여 신규고객에게 상품을 추천하는 새로운 방법을 제시한다. 다음으로 희박성 문제를 해결하기 위하여, 구매정보가 충분한 고객에게는 협업필터링을, 그렇지 않은 고객에게는 협업필터링 대신 제시한 신규고객 추천방법을 적용하는 하이브리드 추천 방법을 제안한다. 제시한 추천 방법의 효과성을 평가하기 위하여 국내 유명 백화점 중의 하나인 H백화점의 구매 트랜잭션 데이터를 사용하여 실험하였다. 실험결과로부터 근접중심성과 매개중심성을 결합한 지표를 신규고객 추천 시에 사용할 경우 추천 성능이 가장 우수한 것으로 판명되었으며, 제안한 하이브리드 추천 방법이 기존의 협업필터링의 성능을 상당히 개선함으로써 희박성 문제를 해결할 수 있는 새로운 대안임이 입증되었다.

Keywords : 중심성분석, 신규고객추천, 희박성, 사회연결망, 협업필터링

[*] 국민대학교 경영정보학부
[**] 국민대학교 경영학부

저 자 소 개

조윤호

현재 국민대학교 경영정보학부 전자상거래전공 부교수로 재직 중이다. 서울대학교 계산통계학과(전산학전공)를 졸업하고, KAIST 경영정보공학과에서 석사, KAIST 경영공학과에서 박사학위를 취득하였으며, LG전자(주)에서 6년간 주임연구원으로 재직하였다. 주 연구분야는 추천시스템, 모바일비즈니스, 고객관계관리, 데이터마이닝, 소셜네트워크 등이다.

방정혜

현재 국민대학교 경영학부 마케팅전공 조교수로 재직 중이다. 이화여대 경영학과와 대학원(MIS전공)을 졸업하고, University of Rhode Island에서 경영학박사학위(마케팅 전공)를 취득하였다. 한국 딜로이트경영컨설팅에서 컨설턴트로, 미국 Penn State University-Mont Alto에서 조교수로 재직하였다. 주 연구분야는 고객관계관리, 관계마케팅, 디자인경영 등이다.