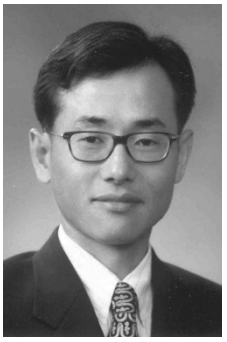
A new recommender system to combine content-based and collaborative filtering systems

Received: 22nd February, 2001



Byung-Do Kim

is Assistant Professor of Marketing at the School of Business Administration, Seoul National University, Korea. He was previously on the faculty of Carnegie Mellon University, Pittsburgh, USA. His current research interests include various econometric and statistical modelling issues on consumer choice behaviour, e-commerce, reward programmes and database marketing. His previous research has appeared in *Journal of Business & Economic Statistics*, *Journal of Interactive Marketing*, *Journal of Marketing Research*, *Journal of Retailing*, *Marketing Letters* and *Marketing Science*, among others.

Sun-Ok Kim

is a doctoral candidate at the School of Business Administration, Seoul National University, Korea. She received her BBA from Yonsei University, Korea and received her MBA from Seoul National University, Korea. Her current research interests include recommender systems, consumer choice modelling, database marketing and retailing.

Abstract The enormous number of choices often create confusion for consumers so they often like to get the opinion of other people in order to make better buying decisions. Many e-commerce sites are implementing recommender systems to help their customers find the most valuable products and services.

There are two fundamentally different approaches, the content-based and collaborative filtering techniques, to recommend products to customers based on their historical preferences. A new recommendation algorithm to combine these two systems is proposed in this paper. Applying the model to film rating data, the model is shown to perform better than the previous recommendation models in terms of predictive accuracy. How the model can be applied to personalise Internet shopping based on customer's transaction history is also discussed.

INTRODUCTION

Consumers use the evaluation or opinion of other people as an important information source.[1] People like to get recommendations when they perceive a risk in making a purchase decision or when they want to simplify their buying decision. For instance, when a consumer buys a camcorder, the consumer may ask their friends who have knowledge or experience of camcorders, or they may ask a salesperson to help them buy the best camcorder.

Recommendation becomes even more important in the Internet-based shopping environment where consumers do not make physical contact with products and face higher cognitive risk. In addition, e-commerce sites offer a very large number of alternatives since they do not have any physical constraint on inventory or shelf space. Hence, consumers may be confused by the number of choices. If the consumer is not familiar with the Internet, the problem becomes even more serious. In order to solve these

Byung-Do Kim
Seoul National University, School of Business Administration, 56-1 Shinlim-dong, Kwanak-ku, Seoul, 151-742, Korea.

Tel: 82-2-880-8258; Fax: 82-2-878-3154; e-mail: bxk@plaza.snu.ac.kr

problems, several e-commerce sites are employing recommender systems to help their customers make their purchase decisions more efficiently.[2]

A recommender system is an electronic agent that helps customers to find the most valuable products/services based on their historical preferences or tastes.[3,4] In fact, as the importance of e-commerce increases, the recommender system becomes an essential tool in implementing personalised marketing. The well-designed recommender system analyses the inferred or stated preference of each customer and automatically suggests a set of products/services.

This paper focuses on the recommender systems which suggest products/services based on customers' stated preferences or previous purchase histories even though there are several other types.[5] And in this class there are two fundamentally different approaches, the content-based and collaborative filtering techniques.

The content-based recommender system suggests products to consumers by analysing the content of items that they liked in the past.[6] Features and attributes of products can be contents of items. Its underlying assumption is that the content of an item is what determines the user's preference.[7] The content-based systems have been widely used with various applications. For example, search engines such as Yahoo and Alta Vista recommend relevant documents from user-supplied keywords.[8] Amazon.com recommends new books and/or albums based on customers' favourite authors or musicians.

The content-based approach is an effective recommendation tool, especially for new items. It has several limitations however.[9,10] First, it often provides bad recommendations since it only considers the pre-specified contents for products/services. If two items have the same contents, it will predict them to have identical ratings. Secondly, the content-based system tends to restrict the scope of the recommendation to items similar to those the consumer has already rated.[11] Finally, there is no way to provide recommendations for new customers because it knows nothing about their preferences.[12]

In contrast, the collaborative filtering technique recommends items that similar consumers have liked. Consumers in the collaborative filtering system share their evaluations and opinions regarding each product so that other consumers can better decide which items to choose,[13] it automates the process of word-of-mouth communication among consumers.

Collaborative filtering overcomes the limitations of the content-based systems by enabling consumers to share their opinions and experiences about products. It has been successfully applied to many e-commerce sites (eg books, music CDs, films, wines, etc.). It also has limitations though. First, collaborative filtering does not work very well when the number of evaluators/users is small relative to the volume of information in the system. That is, it is difficult to find similar users in predicting ratings for some unpopular products. Secondly, it has the early rater problem that occurs when a new product/item appears in the database. Collaborative filtering cannot provide predictive ratings for a new product until other consumers have evaluated it.

The main purpose of the paper is to develop a hybrid model that combines the content-based and collaborative filtering systems. Generalising from the previous models, the new model can be flexibly applied across various contexts and overcome the weakness of the content-based and collaborative filtering techniques. Applying the model to film rating data, the new model is shown to perform better than previous

recommendation models in terms of predictive accuracy.

The rest of the paper is organised as follows. In the next section the content-based and collaborative techniques are described more formally, and a hybrid model is developed to combine them. Why the new model is theoretically better than the existing models is also discussed. In the following section the new model is applied and shown to perform better than the existing recommender systems in terms of two statistical criteria. The marketing implications of the model and its extension to e-commerce sites are then explored. Finally, the limitations of the model are discussed along with future research directions and the authors' conclusions.

DEVELOPING A NEW RECOMMENDER SYSTEM

Recognising that the content-based and collaborative filtering system each has its advantages and disadvantages in recommending products, researchers have attempted to develop a hybrid model to combine the two approaches.[14–18] Claiming that their models take advantage of the collaborative filtering approach without losing the benefit of the content-based approach, they have shown that their models perform better than the individual approach.

Consistent with this research trend, the authors have developed a hybrid recommender system to combine the content-based and collaborative filtering systems. The point of departure of their model is extraction of the content component of products/items by employing a regression and then application of collaborative filtering to the consumer's preference unexplained by this (content-based) regression. Marketing researchers have traditionally used multiattribute approaches (eg preference regression) to explain consumer's preference for products by a set of their attributes. These models, however, often lead to poor predictions about customer preferences because of missing information such as undiscovered attributes or important attribute interactions, sensory or experiential attributes and word-of-mouth effects.[19] The collaborative filtering component of the new model can be used to capture this missing information.

Before describing the model in greater detail, it is helpful to look at the input data to understand the task more clearly. The typical input data for recommender system is represented in the form of (evaluation) ratings on each product/item. As shown in Table 1, it is an $n \times m$ user-item matrix with each cell representing a user/consumer's rating on a specific item/product. The main task is to predict the preference (or rating) for missing cells based on other observed evaluations. For example, Amy has rated Films 1, 2, 4 and *M*. Then what is Amy's predicted rating for Film 3? Similarly, the missing ratings for other customers are predicted. Once all the predicted film ratings have been obtained, film recommendations can be provided for each customer (eg suggest three highly-rated films for each customer).

The algorithm of the model consists of six major steps. First, a set of content components characterising all products/items needs to be determined. For example, consider a film recommendation site such as www.moviecritic.com. Here, site visitors can get film recommendations once they register and evaluate a minimum of 12 films. Key features (or contents) determining a visitor's preference for a film may be the genre of the film (eg comedy, drama, action), the director, the

Table 1: Input data for recommendation system

	Film 1	Film 2	Film 3	Film 4	...	Film *M*
Amy	5	2	.	4	...	1
Joseph	1	.	1	2	...	.
Michael	.	4	3	.	...	5
....	....	....	....	....	...	....
Jim	3	1	.	1	...	2
Laura	5	3	4	.	...	1

producer, the main actors/actresses and so on.

Secondly, the following regression model is applied for each customer once the key features have been identified:

$$R_{ij} = \beta_{0i} + \beta_{1i} X_{1ij} + \ldots + \beta_{Ki} X_{Kij} + \varepsilon_{ij} \quad (1)$$

where R_{ij} is the preference (or rating) of consumer i for product j and X_{1ij} is the value of the first feature for product j evaluated by consumer i. Note that in this regression K number of features for products are identified.

The parameters to be estimated, or βs in equation (1), measure how important each feature is in determining the preference of the consumer. Note that equation (1) is applied for each customer's observed ratings. Once the parameters have been estimated the consumer i's preference on products not yet evaluated can be predicted. For example, the regression is applied to Amy's observed film preferences in Table 1. Upon estimation, Amy's rating for Film 3 can be predicted with the estimated parameters and features of Film 3.

The procedure explained so far is no different from the content-based recommender system. That is, preferences of other consumers have not yet been used to predict consumer i's preference. As noted in the previous section, however, it is possible for a consumer to rate two films with identical features differently because there may be other factors influencing her preference. The rest of the algorithm is required to explain these discrepancies.

Thirdly, based on the estimated regressions, the fitted preferences/ratings ($\hat{R}_{ij}$) are computed for all consumers and all products/items. Note that here the predicted ratings for both observed and unobserved (or missing) products are computed.

The fourth step is to create a data matrix of prediction errors. The prediction errors are defined as the difference between the actual preference and the predicted preference. That is, $\varepsilon_{ij} = R_{ij} - \hat{R}_{ij}$. In the regression context, the errors are the residuals in regression model or the preferences unexplained by the regression model equation (1). Note that prediction errors cannot be calculated for products for which there are no actual ratings. Hence, consisting of a series of prediction errors with a set of missing values, the resulting data matrix of prediction errors looks similar to the input data matrix in Table 1.

Fifthly, the collaborative filtering technique is applied to the data matrix created in the previous step. The neighbourhood-based algorithm is employed among various collaborative filtering techniques.[20]

Here the goal is to calibrate the values for missing cells. In the neighbourhood-based method, it can be calculated as:

$$e_{t,j} = \bar{\varepsilon}_t + \tau \sum_{i=1}^{n} w_{t,i}(\varepsilon_{i,j} - \bar{\varepsilon}_i) \quad (2)$$

where $e_{t,j}$ is the predicted value/rating of

consumer t on product j and n is the number of consumers in the collaborative filtering database who have evaluated the product j. The weight $w_{t,i}$ is the similarity between consumer i and the (target) consumer t. τ is a normalising factor such that the absolute values of the weights sum to one.

Back to the (film) rating example given in Table 1, suppose that Amy's rating on Film 3 is predicted. In the neighbourhood-based method, it is given by the weighted average of Joseph, Michael, Laura and others' ratings on Film 3. In addition, the weights ($w_{t,i}$) are determined by how similar Amy is to other evalutors in terms of film ratings. There are many ways to specify this similarity measure including the Pearson correlation coefficient, the constrained Pearson correlation, the Spearman rank correlation coefficient and the vector similarity.[21] There are many other important issues in implementing collaborative filtering but they will not be described in this paper, interested readers should see Sarwar *et al.*[22]

The final step is to sum the output from the third step and the fifth steps. That is, the content-based approach in step 3 provides $\hat{R}_{ij}$ while the collaborative filtering in step 5 produces e_{ij}. The predicted preference of product j for consumer i is the sum of these two numbers. Now the algorithm can be summarised:

Step 1: determine a set of content components characterising all products/services
Step 2: fit the (contents) regression for each consumer
Step 3: calculate the fitted preferences for all consumers and all products
Step 4: create a data matrix of prediction errors
Step 5: apply the collaborative filtering technique into the data matrix
Step 6: sum the output from Step 3 and Step 5.

DATA AND ESTIMATION RESULTS

In this section the model is applied to actual film rating data — called EachMovie database — supplied by DEC systems. The database was collected for 18 months to September 1997. It includes 2,811,983 ratings for 1,628 different films from over 70,000 users. It also has some information on users (eg age, sex and zip-code) and films (eg name, genre, release date). Users were instructed to evaluate films on a six-point scale from 1 to 0 (1, 0.8, 0.6, 0.4, 0.2, 0). Higher value indicates stronger preference on the item.

Fifty users were randomly selected from the database, each with more than 120 film ratings, to validate the model. The 50 users selected have a total of 9,026 ratings on 1,103 film items. For each user, 5 per cent of the ratings were withheld as the validation sample. Sarwar *et al.*[23] adopted the same sampling method and this model is compared with their filter-bot hybrid model.

Four other competing models are applied to the film rating data. First, a baseline model is employed to benchmark the performance of other personalised recommender systems. It predicts the rating for each film by the mean rating across users.

Secondly, the content-based recommender system is fitted where the genres of the films are used as the contents of the film/item. A dummy variable is created for each of the ten genre variables including comedy, drama, action, art/foreign, classic, animation, family, romance, horror and thriller. A film can be simultaneously classified into more than one of these genres. The ten genre dummies are regressed on actual film ratings in the estimation sample for

each user. Based on the estimated regressions, the film ratings in the validation sample are predicted. The predicted ratings are evaluated against the actual ratings.

Thirdly, collaborative filtering is employed where the neighbourhood-based algorithm is implemented and the similarities between users are measured by Pearson correlation coefficients. In addition, 20 co-rated items are used as the cut-off for significance weighting, and the users with less than 0.01 correlations are not included as a set of neighbourhood.[24]

Fourthly, the hybrid recommender system suggested by Good *et al.* is fitted.[25] Their model attempted to overcome the sparsity and the early-rater problem of the collaborative filtering by using a few filter-bots. This model is easy to implement in the current collaborative filtering system because it can handle filter-bots as ordinary users. Ten genres are used as filter-bots in this model. That is, ten genre filter-bots with 50 common users are analysed through the collaborative filtering algorithm. These ten filter-bots act the same as the common users except that they rate every item. If an item belongs to a given genre, it rated the item as 0.8. Otherwise, the filter-bot rates the item as 0.2.

Finally, the model is applied following the six steps described above. Note that the algorithm of the model employs both the content-based and the collaborative filtering technique. The identical content-based and the collaborative filtering options used above are implemented.

Table 2 shows the validation results for each of the five models. The performance of each model is evaluated in terms of two evaluation criteria, the mean absolute error (MAE) and the Receiver Operating Characteristic (ROC) sensitivity measure.[26] Computed as $\Sigma_{i=1}^{n} |R_i - \hat{R}_i|/n$ *where* R_i is the actual rating and $\hat{R}_i$ is the predicted rating, the mean absolute error measures the statistical accuracy of the model. The lower the MAE, the more accurate the model is. On the other hand, the ROC measures the discriminating power of a filtering system. Operationally, it is the area under the ROC curve that plots the sensitivity and the specificity of the test.[27] Sensitivity refers to the probability of a randomly selected good item being accepted by the filter while specificity is the probability of a randomly selected bad item being rejected by the filter. The ROC sensitivity ranges from 0 to 1 where 1 is perfect and 0.5 is random.

As expected, the other four recommender systems incorporating some personalised components outperform the (aggregate) baseline model with respect to both MAE and ROC. Secondly, the performance of collaborative filtering turns out to be better than the content-based model. This result, however, should be tested in more cases in the future because the content-based model can be improved by incorporating the more important content variables.[28]

Finally, Table 2 also shows that the new model performs best in terms of both evaluation criteria. With respect to the ROC, the model improves the predictive performance of the content-based and the collaborative filtering by 6.8 per cent and 2.6 per cent respectively. In addition, the model is marginally better than a recent hybrid model (filter-bot) in terms of both evaluation criteria.

MARKETING IMPLICATIONS AND DISCUSSIONS

The recommender systems provide value to customers. First, a customer can reduce search costs by using

Table 2: Predictive accuracy of various recommender models

Type of model	MAE	ROC
Baseline model	0.2238	0.7398
Content-based model	0.2103	0.7640
Collaborative filtering	0.1955	0.8058
Filter-bot model	0.1982	0.8247
New model	**0.1832**	**0.8328**

recommender systems. Search costs include the cognitive effort and search time. Given that a consumer experiences cognitive difficulty in the Internet shopping environment and on-line buyers suffer from time starvation, the benefit from reducing the search effort and time is considerable.

Secondly, consumers can simplify their choices by using the recommender system. Since recommender systems replace one or more of the steps in a decision-making process, customers can buy products/services matching their needs with less effort.

Thirdly, consumers can improve their decision quality.[29] More specifically, consumers tend to have a consideration set without any dominating alternatives when a customer uses a recommender system. Moreover, consumers become more confident in their purchase decision making when they use a recommender system.

Finally, a recommender system can provide an enjoyable shopping experience. This is very important because it will enhance the experience of 'flow' in Internet shopping that influences repeat visits to websites.[30,31]

Recommender systems also provide many benefits to companies. Firms can increase their profits by increasing revenue and/or decreasing costs by employing the recommender systems.[32] There are various ways to increase revenue.[33] First, site browsers can be converted to buyers through recommender systems. Secondly, firms can increase cross-selling by recommending additional products related to items the customer has already purchased or shown interest in. Recommender systems can strategically provide complementary products to customers who buy related items. Thirdly, recommender systems improve customer loyalty by creating a value-added relationship between the site and the customer. The more a customer uses a recommender system, the more accurate the recommender system becomes. Recommender systems can build strong commitment from customers. Finally, recommender systems maximise the lifetime value of each customer by optimising each contact.[34]

CONCLUSIONS

Electronic commerce is growing explosively and the number of consumers who use the Internet for information search and on-line shopping is increasing dramatically. The unique characteristic of the Internet shopping environment, ie 'interactivity', is creating a new opportunity for personalised marketing. As the importance of e-commerce increases, recommender systems will be considered to be an essential part of personalised marketing. A recommender system is a sort of electronic agent suggesting the most valuable product to customers based on their preference. Many commerce websites are already using recommender systems to help their customers find products to purchase and many other companies have plans to adopt recommender systems in the near future.

This paper proposes a hybrid recommender system that combines the content-based and collaborative filtering systems. Generalising from previous competing models, the new model can

be flexibly applied across various contexts and overcome the weakness of the content-based and collaborative filtering techniques. Applying the model to film rating data, it was shown that the model performs better than previous recommendation models in terms of predictive accuracy.

The paper now concludes with a discussion about the model's limitations and future research directions. The model was applied to film rating data rather than Internet shopping data. The application of the results is, therefore, quite limited. The current model can only be applied when customers explicitly mention their preferences or ratings on products/services. Many e-commerce sites do not, however, have these customer evaluations. Instead, they know what kinds of products/services each of their customers has purchased. This purchase information can be treated as an indication of positive preferences. Similarly, information about customer returns can be treated as the indication of negative preference. The model should be modified to incorporate this implicit preference information when it is applied to Internet shopping data.

Finally, researchers have developed several other hybrid recommendation models. In this paper the new model was compared with the filter-bot, one of these hybrid models. In future research, each hybrid model should be evaluated more extensively in various contexts.

Acknowledgement

This research was supported by Research Development Fund from Seoul National University in Korea. We would also like to thank DEC systems research centre for providing the data.

References

1 Burnkrant, R. and Cousineau, A. (1975) 'Informational and normative social influence in buyer behavior', *Journal of Consumer Research*, Vol. 2, No. 4, pp. 206–215.
2 Schafer, B., Konstan, J. and Riedl, J. (1999) 'Recommender systems in e-commerce', Proceedings of ACM Electronic Commerce 1999 Conference.
3 *Ibid*.
4 Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2000) 'Analysis of recommender algorithms for e-commerce', Proceedings of ACM E-Commerce 2000 Conference.
5 See Hanson, W. (2000) 'Principles of internet marketing', South-Western College Publishing, Cincinnati, Ohio and Schafer, B., Konstan, J. and Riedl, J. (2001) 'E-commerce recommendation applications', *Journal of Data Mining and Knowledge Discovery*, forthcoming, for more discussion on these.
6 Balabanovic, M. and Shoham, Y. (1997) 'Fab: Content-based, collaborative recommendation', *Communication of the ACM*, Vol. 40, No. 3, pp. 66–72.
7 Balabanovic, M. (1997) 'An adaptive Web page recommendation service', First International Conference on Autonomous Agents, Marina del Rey, CA, February.
8 Ansari, A., Essegaier, S. and Kohli, R. (2000) 'Internet recommendation systems', *Journal of Marketing Research*, Vol. 37, No. 3, pp. 363–375.
9 Sarwar, B., Konstan, J., Borchers, A., Herlocker, J., Miller, B. and Riedl, J. (1998) 'Using filtering agents to improve prediction quality in the GroupLens Research Collaborative Filtering System', Proceedings of 1998 Conference on Computer Supported Collaborative Work.
10 Good, N., Schafer, J., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J. and Riedl, J. (1999) 'Combining collaborative filtering with personal agents for better recommendation', GroupLens Research Project, University of Minnesota.
11 Balabonovic and Shoham (1997) *op. cit*.
12 Maltz, D. and Ehrlich, K. (1995) 'Pointing the way: Active collaborative filtering', CHI '95 Proceedings Papers.
13 Herlocker, J., Konstan, J. and Riedl, J. (1999) 'An algorithmic framework for performing collaborative filtering', Proceedings of ACM SIGIR 1999, pp. 230–237.
14 Balabanovic (1997) *op. cit*.
15 Balabanovic and Shoham (1997) *op. cit*.
16 Basu, C., Hirsh, H. and Cohen, W. (1998) 'Recommendation as classification: Using social and content-based information in recommendation', Proceedings of the 1998 Workshop on Recommender Systems, pp. 43–52.
17 Sarwar *et al*. (1998) *op. cit*.
18 Herlocker, Konstan and Riedl (1999) *op. cit*.
19 Gershoff, A. and West, P. (1998) 'Using a community of knowledge to build intelligent agents', *Marketing Letters*, Vol. 9, No. 2, pp. 79–91.
20 More recently, model-based algorithms have been introduced (Ansari, Essegaier, and Kohli (2000) *op. cit*.). Compared to the neighbourhood-based algorithms, they are more practical for the environments in which user preference changes

slowly with respect to the time needed to build the model. However, they are not suitable for the environment in which the user preference model must be rapidly updated (Schafer, Konstan and Riedl (2000) *op. cit.*).

21 Breese, J., Heckerman, D. and Kadie, C. (1998) 'Empirical analysis of predictive algorithms for collaborative filtering', Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI.
22 Sarwar *et al.* (2000) *op. cit.*
23 Sarwar *et al.* (1998) *op. cit.*
24 See Herlocker *et al.* (1999) for greater detail on these specification issues.
25 Good *et al.* (1999) *op. cit.*
26 For various evaluation criteria, see Good *et al.* (1999) *op. cit.* and Herlocker *et al.* (1999) *op. cit.*
27 Swets, J. (1988) 'Measuring the accuracy of diagnostic systems', *Science*, Vol. 240, No. 6, pp. 1285–1289.
28 Gershoff and West (1998) *op. cit.*
29 Hauble, G. and Trifts, V. (2000) 'Consumer decision making in online shopping environments: The effects of interactive decision aids', *Marketing Science*, Vol. 19, No. 1, pp. 4–21.
30 Trevino, L. and Webster, J. (1992) 'Flow in computer-mediated communication: Electronic mail and voice mail evaluation and impacts', *Communication Research*, Vol. 19, No. 5, pp. 539–548.
31 Hoffman, D. and Novak, T. (1996) 'A new marketing paradigm for electronic commerce', Working Paper, Vanderbilt University.
32 Allen, C., Kania, D. and Yaeckel, B. (1998) 'Internet world guide to one-to-one Web marketing', Wiley Computer Publishing.
33 Schafer *et al.* (1999) *op. cit.*
34 Kania, D. (1999) 'Make database options pay off', *Advertising Age's Business Marketing*, Vol. 84, No. 3, pp. 31–32.