An Alternative Evaluation of the Item-based Collaborative Filtering Using Simulated Online Shopping

Hyung Jun Ahn*

Abstract

This paper presents a novel method for evaluating the usefulness of online product recommendation. Previous studies on evaluating recommendation systems have mostly relied on two methods : testing the accuracy of estimating user preferences by recommendation systems, or empirically testing the effectiveness with lab experiments involving human participants. The former does not measure the usefulness directly and hence can be misleading; the latter is expensive in that it requires a working online store system and test participants. In order to address the problems, the proposed approach uses simulation to imitate customer behavior and evaluate the usefulness of recommendation. Models for user behavior and an abstract Internet store are developed for simulation. Actual simulation experiments are performed to illustrate the use of the approach.

Keywords : Collaborative Filtering, Product Recommendation, Evaluation, Agent, Simulation

Received : 2009. 08. 21. Final Acceptance : 2009. 09. 16.

※ This work was supported by 2009 Hongik University Research Fund.

* Corresponding Author, Assistant Professor, College of Business Administration, Hongik University, hjahn@hongik.ac.kr

1. Introduction

As e-commerce matures, intelligent product recommendation has become a necessary tool for many Internet stores to help their customers shop more efficiently and boost their sales [Burke, 2002; Herlocker, et al., 2004; Komiak and Benbasat, 2006]. Accordingly, numerous studies have been conducted so far for developing such product recommendation systems. Many of the systems have been found to be useful and many Internet stores such as Amazon.com have been successfully using such systems for long time already [Linden, et al., 2003].

Because it usually involves much efforts and time to develop such recommendation systems, one of the most critical issues is to evaluate the performance of the systems before actual implementation and deployment. So far, two types of evaluation have been broadly used by most research. The first method of evaluation measures the accuracy of estimating users' preferences by the recommendation systems. This is because recommendations are usually based on the estimated preferences of users for candidate products. The second method is to conduct empirical experiments involving human participants and fully or partially working systems. This method is different from the former in that it measures the usefulness of the systems directly in the context of actual shopping.

Although both of the methods have respective advantages and have been used widely, they do have some limitations. First, the evaluation based on the accuracy of preference estimation may not always be able to measure how useful a system would be in the actual shopping context. This is because no matter how accurate the estimation is, the recommendation system may not be accepted or used by users if users find other functions or routes of shopping more efficient. Second, the empirical evaluation is usually expensive in that, first, it requires the implementation of the system to a significant degree, and second, it involves many human participants. The high costs of implementation and experiment is a serious issue when there are many factors to be tested over many experiments or frequent updates of the system is required.

There have been few studies so far, if any, that have attempted to address the above problems. Therefore, this research aims to present an alternative method of evaluating recommendation systems using simulation in an attempt to overcome the above limitations. Simulation approach can enable us to imitate the rational behavior of customers realistically at low cost without requiring working systems or human participants.

Among many different types of recommendation, this paper focuses its analysis on the item-based collaborative filtering method which recommends products that are similar to the one which has been just clicked by a customer. The analysis also tests the relative effectiveness of recommendation in comparison with another store function to test whether the quality of other functions affects the usefulness of recommendation.

This paper is organized as follows : Section 2 presents an overview of related literature.

Section 3 explains the models and methods used for the simulation-based evaluation. Then the results of a series of experiments are presented in Section 4 for illustrating the use of the presented approach. Section 5 concludes with discussion and further research issues.

2. Literature Review

Recommendation systems have been researched rigorously since the early days of e-commerce and many methods have been developed so far. Such recommendation methods are often classified into two types : collaborative filtering and content-based filtering [Burke, 2002; Herlocker, et al., 2004; Komiak and Benbasat, 2006]. Many collaborative filtering systems use virtual collaboration among users by recommending products to a given user based on the similarity between the user and other users. The similarity is often estimated using purchase history or ratings data. On the contrary, content-based filtering uses content information of products and users for recommendation. For example, titles and keywords for books can be used for recommending books to shoppers based on the content-based filtering.

Most studies that develop recommendation methods evaluate the performance of the methods based on the accuracy of estimating the user preferences for recommended products [Herlocker, et al., 2004]. There are several measures that are used widely for this purpose such as mean absolute error (MAE), precision, and recall. MAE measures the mean absolute difference between estimated preference and actual preference. Precision and recall are the measures that have been developed in the information retrieval (IR) field, where the former measures how many of the recommended products are actually preferred, and the latter how many of the target products were actually recommended. There are extensions and modifications of the above measures as well, including hybrids [Ahn, 2008].

Different from the above studies that design and develop methods of recommendation, there have also been many studies that empirically investigated the effects of using recommendation systems and also the factors that affect the effects. Many of the studies are very similar to behavioral studies in the management information systems (MIS) field in their research methods and theories. For example, some studies use Technology Acceptance Model (TAM) together with lab experiments to evaluate the effectiveness of recommendation systems [Jiang, et al., 2000; Klopping and McKinney, 2004]. A thorough review of this type of studies can be found in a recent article [Xiao and Benbasat, 2007].

Although the above approaches to evaluation of recommendation systems have been used widely so far, they have a couple of limitations. First, the evaluation based on estimation accuracy may not be able to predict whether the recommendation systems will actually be used in the context of online shopping. This is because there can be many different ways a user can use an online store depending on the available functions or structure of the store. Therefore, even if the estimation is very accurate,

users may choose to use a different route of shopping if it is perceived as more efficient. Second, empirical studies are usually expensive because many users have to be involved and a full or partially working system should be constructed for the lab experiments. Moreover, it is also difficult to motivate users to behave realistically in the experiments.

3. Research Models

3.1 Overview of the Research

The goal of this research is to present a novel method of evaluating recommendation systems to overcome the problems that were summarized in the literature review section. More specifically, there are two sub-objectives :

- The method should allow us to directly measure the performance improvement of online shoppers when using recommendation systems.
- The method should allow us to measure the performance improvement without human participants using simulation.

Since there are so many different styles of recommendation systems being used, this research constrains the scope of it to 'recommendation of similar items' used by many Internet stores such as Amazon.com. That is, the target recommendation systems of this research are those where a list of products is recommended to a user whenever a user clicks on a product. The products in the list are chosen according to the similarity with the current product calculated based on some similarity metrics.

There are several assumptions as well underlying this research :

- Different users may have different behavioral characteristics, and may find different ways of shopping more useful.
- Users collect information about a certain number of candidate products before making a purchasing decision [Kotler and Armstrong, 1991; Simon, 1959].
- Users are assumed to be rational. That is, users are assumed to pursue efficiency in shopping by trying to reduce the time taken to collect information about enough products.
- There are certain costs associated with each different action in shopping. Because users are assumed to be rational, they try to minimize the costs.

3.2 Description of the Research Models

In line with the assumptions above, there are several models needed for this research. First, an abstract model of an Internet store is needed so that simulation can be performed with the model. Second, a model for imitating users' shopping behavior is needed. Third, a model is needed for finding the rational behavior of users pursuing efficiency.

In order to develop the models, this research has modified and extended the models presented in a related work [Ahn, 2008]. First, the abstract model of an Internet store consists of

a main menu and many sub categories under the main menu. The sub categories cannot have further sub categories. Products in the store can belong to more than one category. Users can choose a product while browsing each category to see detailed information about the product. In this product detail page, the store presents a list of products as recommendation to the user.

Second, the user behavior model assumes that users will choose and browse a category that is most likely to yield the highest efficiency in shopping based on the estimation of the success probabilities of each category. That is, users have estimations about how successful each category would be in finding products that they are interested in. This estimation is updated as they browse products in each category. Based on this updated information, users can switch between categories whenever a better category is found. When presented a list of recommended products, users can choose to accept the recommendation and view the product or may just ignore the recommendation depending on the user's trust of the store's recommendation.

Third, this paper also adopts the meta-heuristic approach to find the rational behavior of users [Ahn, 2008]. The user behavior model explained in the above paragraphs has several parameters that need to be fixed for the rational, or optimal, behavior, and hence evolution strategy, a meta-heuristic, was applied. Readers are referred to [Beyer and Schwefel, 2002] for a detailed introduction to the heuristic. The optimization using the meta-heuristic finds the best values of the following parameters :

- The length of observation for updating the estimate of the success probability of a current category.
- The difference threshold that will trigger switching to a different category. That is, if a better category exists whose success probability is bigger than that of the current category plus the threshold, a user will make switching to the better one.
- Trust of recommendation which ranges between 0~1. This will determine the probability of a user's accepting recommendation. If trust is 1, recommendations will always be accepted. If it is 0, recommendations will never be accepted.
- Upper bound on the number of recommendations taken from a single product. That is, this sets the limit on the number of recommendations taken starting from a certain product in a certain category.

Along with the above, a simple cost model was developed to measure and compare the costs of user's browsing during the simulation. The cost model has four components :

- Component 1 : time taken to review a chosen product
- Component 2 : time taken to select a product among a list in a category
- Component 3 : time taken for every click
- Component 4 : time taken to return to a previous page, or pressing the 'Back' buttons of the browser

Using the above components, the meta-heuristic can measure how many time units it takes to accomplish a given shopping goal for each user with different behavioral parameters. By having a large population of experimental users and evolving the users through mutation and cross-over, the heuristic can find optimal or near optimal values of the parameters. The parameters are then assumed to represent the rational behavior of customers in the abstract Internet store.

4. Experiments

4.1 Goal of the Experiments

Several experiments were performed to illustrate how the models could be used as an alternative way of evaluating a recommendation method as explained in the previous section. In addition to this basic purpose, the experiments also aim to answer the following two interesting research questions.

Question 1 : Is recommendation useful for all users?

Recommendation systems have been generally known to be useful through many experiments and empirical studies. However, it may not be so for all users because different users have different characteristics and shopping behavior. Hence, a simulation experiment will attempt to find out whether all users can improve their shopping efficiency using a recommendation system.

Question 2 : Is the usefulness of recommendation affected by the usefulness of other functions?

Because there might be alternative paths or routes of browsing for desired products, the usefulness of a recommendation system might be affected by that of other functions that also help users improve shopping efficiency. Therefore, an experiment will be performed that varies the usefulness of some other functions and see the impacts on the relative usefulness of a recommendation system.

4.2 Overview of the Experiments

In order to perform the experiments, a virtual Internet store was constructed based on the models introduced earlier together with data on many movies, categories, and shoppers. More specifically, the dataset from the Netflix contest [Netflix, 2007] was reduced to a dense set of 1,000 most rated movies and 1,000 users with most ratings so that we can have as dense dataset as possible for the experiments. The reduction created a dataset with more than 80 percent of all possible ratings available, which means we know which product is preferred how much by each user for 80 percent of the all user-movie pairs. The 20 percent unknown ratings were regarded as non-preferred. This dataset was also combined with another dataset from Internet Movie Database (IMDB) [IMDB, 2008] to associate movies with 24 categories : {Action, Adult, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family,

Fantasy, Film-Noir, History, Horror, Music, Musical, Mystery, Romance, Sci-Fi, Short, Sport, Thriller, War, Western}. This combination of data was fed to the abstract Internet store model creating a virtual online DVD store where there are 24 movie categories and each movie may belong to multiple categories.

Among the 1,000 users, half of them were used for calculating association between movies and the other half were used as actual shoppers for the simulated shopping experiments. The association between any two products was produced using Pearson's correlation of the rating vectors of the two.

Experiments were performed with 50 randomly chosen users among the 500 shoppers in the combined dataset. For each user, evolution strategy was applied to optimize the behavioral parameters of the user that maximizes the shopping efficiency.

In all of the experiments, the following values in <Table 1> were used for the configuration of the experiments. It lists the four components for calculating shopping costs, the parameters for running the optimization, and the goal for shopping. Readers can note that each user should find information about 20 products that match their shopping goal to finish browsing.

Basically two experiments were performed to answer the research questions. The first experiment compared the effect of using recommendation in comparison with the effect of sorting each category according to the average rating of each product. The second experiment investigated whether the relative usefulness of the recommendation varies as the usefulness or

〈Table 1〉 Experiment Configuration

Category	Configuration Variable	Value
Cost components	Component 1	20
	Component 2	3
	Component 3	1
	Component 4	0.3
Evolution strategy parameters	Number of population	100
	Number of individuals selected at each iteration	20
	Number of maximum iteration	500
Shopping goal	Number of desired products to be browsed	20
	Target products	Products that belong to either of {Romance, Drama} with ratings higher than or equal to 4 by the given user
	Searched categories	Users were assumed to search for the targets in the related categories : {"Drama", "Romance", "Action", "Family", "Comedy", "Musical"}

〈Table 2〉 Summary of the experiments

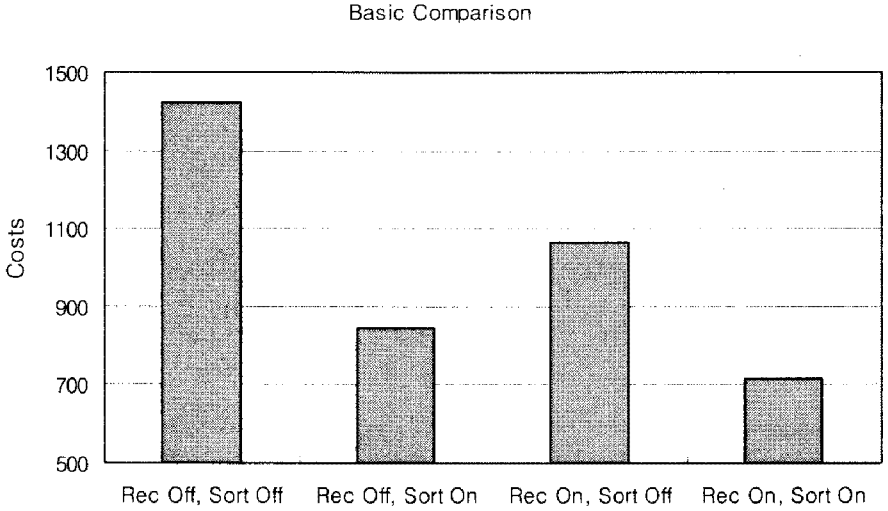
No	Experiment type	Independent variable	Dependent variable
1	Basic comparison (research question 1 and 2)	Whether recommendation was used or not Whether each category was sorted or not	Shopping performance
2	Effects of sorting integrity (research question 2)	Integrity (or noise) of sorting	The relative usefulness of recommendation

the integrity of the sorting changes. For this, each sorted category was given different levels of noise by swapping two random products in the sorted category for different number of times. <Table 2> summarizes the two experiments.

4.3 Basic Comparison

<Figure 1> shows how effective it is to use recommendation in comparison with sorting each category by average product ratings. It can be seen that recommendation is useful in improving the overall performance of the 50 tested users. However, it can also be noted that using sorting alone is more effective than using recommendation alone. The best result was observed when both were used together.

<Table 2> shows the results of the first experiment that is related with both research question 1 and 2. First, we can note that not all the 50 users accepted recommendations for efficient shopping. In the case of <Rec On, Sort Off>, we can see that only 47 users accepted at least 1 recommendation and 14 users did not always trust recommendations. In the case of <Rec On, Sort On> where sorting is also available, we can see that less users rely on recommendation because only 39 accepted one or more recommendations, and 25 users had trust values less than 1. All the above results indicate that recommendation may not always be useful equally to all users for improving their



〈Figure 1〉 Basic Comparison of the Effects of using Recommendation and Sorting Categories (Rec On/Off for using/not using recommendation; Sort On/Off for sorting or not sorting each category)

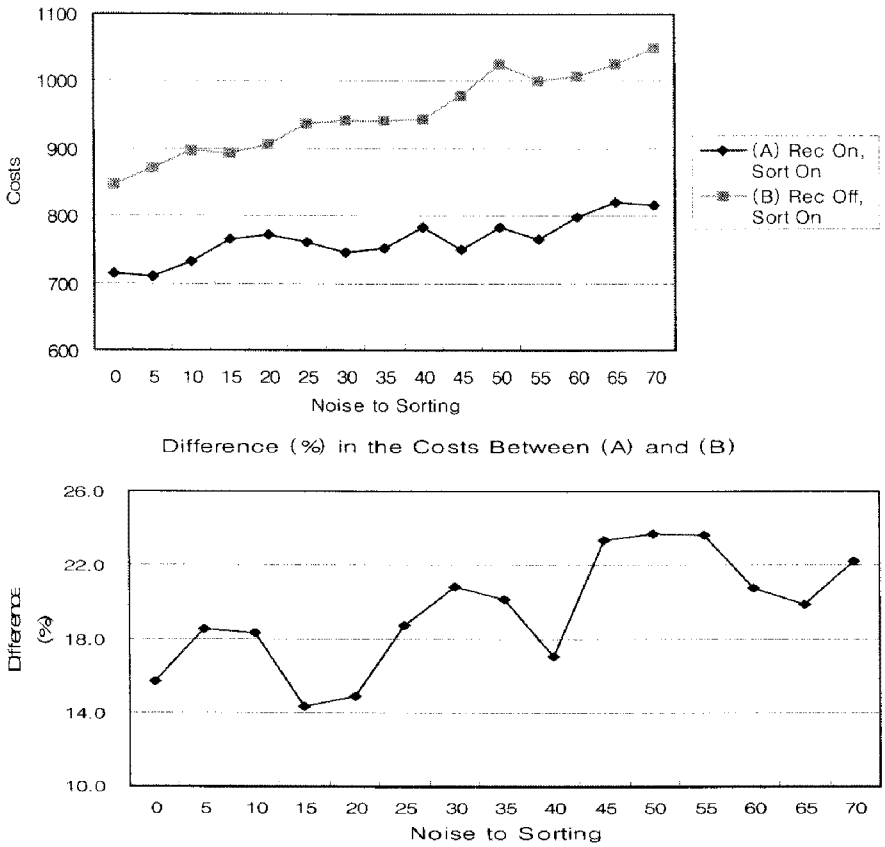
〈Table 3〉 Number of users who accepted recommendations and number of users with different trust values

	Rec On, Sort Off	Rec On, Sort On
Number of users who accepted at least 1 recommendation	47	39
Number of users with *Trust* = 1	36	25
Number of users with 0 < *Trust* < 1	14	25
Number of users with *Trust* = 0	0	0

shopping efficiency. Moreover, in relation to research question 2, it also shows that the availability of another function, sorting in this case, affects the acceptance of recommendation as well.

4.4 Effects of the integrity of sorting

The next experiment shows the effects of different integrity levels of sorting categories. That is, it was tested if the usefulness of the recommendation varies if the quality or integrity of sorting changes. In order to manipulate the levels of integrity, random noise was created to the order of products in each category after regular sorting. The noise was produced by swapping two random products in



[Figure 2] Upper chart for the effects of the integrity of sorting for the two options; lower chart for percentage difference calculated as [{Performance of (A)-Performance of (B)}/Performance of (A) × 100]

each category for a varying number of times, where more swapping implies less integrity of sorting. The results are shown in <Figure 2>.

First, the upper chart shows the changes in the shopping costs as the noise increases from 0 to 70, where 70 implies that each category was given 70 times of random swapping. The chart shows the costs in two cases : (A) when sorting is used but recommendation is not used, and (B) when both sorting and recommendation is used. It can be observed that the costs increase as more noise is fed to the categories. In order to address the research question 2, the lower chart shows the relative contribution of recommendation as the noise increases. It is observed that the relative difference also increases which implies that the recommendation can be relatively more useful when sorting has less integrity and is less useful.

5. Discussion and Conclusion

5.1 Discussion

The results and the findings of the experiments can be summarized as follows. First, the experiments showed how the research models can be combined together with users' ratings data to evaluate the effects of personalized product recommendation using simulation. Second, the experiments themselves yielded some interesting results as well. The first experiment showed that recommendation may not be equally accepted by all users, and that some users may even find the function not helpful. The second experiment showed that the relative usefulness of recommendation increases as the quality of sorting in the categories decreases.

There are meaningful practical implications from the research. First, the research presented a novel method for evaluating the usefulness of product recommendation directly at lower costs compared with the existing methods of evaluation. The presented approach allows us to evaluate recommendation systems without fully or partially implementing an Internet store, and without involving human participants. Second, the experiments performed to illustrate the use of the models showed that recommendation may not always be equally useful to all users. This suggests that different users might find different methods or styles of browsing more suitable for themselves. Internet stores might utilize this result by allowing the use of many different customer-aid functions, or by allowing the customization of shopping methods for users so that each user can find and adopt their own methods of shopping that can maximize their shopping efficiency.

There are also some limitations and further research issues remaining. First, the presented models are abstract ones that have simplified many structural components of Internet stores and customer behavior. The models need to be extended and refined to be used for analyzing real Internet stores with many different characteristics. Second, although the experiments yielded very interesting results, the findings need to be interpreted carefully. Because the main goal of this research is to present and illustrate a new method of evaluating recommendation systems, the experiments were performed with a limited set of configuration vari-

ables and their fixe values. Hence, adopting more variables or using different values of the experiment variables may produce different results. Third, the presented approach is for evaluating 'similar item recommendation' only. Hence, other styles of recommending products may again need modification of the models.

5.2 Conclusion

This paper presented a novel approach to evaluating personalized product recommendation systems at Internet stores, and illustrated the use of the approach through experiments using a virtual online DVD store. This research has several key contributions. First, this research introduced a new method of evaluating item-based product recommendation without the need of expensive experiments involving human participants or working Internet store systems. Second, the results of the experiments showed that recommendation may not always be helpful for all users equally. It suggests that more variety of functions or customizations of functions may be needed to address the different behavioral characteristics of customers. Although there are several further research issues remaining as discussed, the author believes that the careful application of the approach of this paper can help researchers and practitioners to better develop, analyze, and improve personalized product recommendation methods for Internet stores.

Reference

[1] Ahn, Hyung Jun, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem", Information Sciences, Vol. 178, No. 1, 2008, pp. 37-51.

[2] Ahn, Hyung Jun, "Analyzing Customer-aid Functions using Agents and Meta-Heuristic", in 2008 Fall Conference of Korean Operations Research and Management Science Society (KORMS), Seoul, Korea, 2008.

[3] Beyer, Hg, and Schwefel, Hp, "Evolution strategies : A comprehensive introduction", Natural Computing, Vol. 1, No. 1, 2002, pp. 3-52.

[4] Burke, and Robin, "Hybrid Recommender Systems : Survey and Experiments", User Modeling and User-Adapted Interaction, Vol. 12, No. 4, 2002, pp. 331-370.

[5] Herlocker, Jonathan L., Konstan, Joseph A., Terveen, and Loren G. et al., "Evaluating Collaborative Filtering Recommender Systems", ACM Transactions on Information Systems, Vol. 22, No. 1, January 2004, pp. 5-53.

[6] Imdb, "Internet Movie DataBase (IMDB)", September 2008; http://imdb.com.

[7] Jiang, Jj, Hsu, Mk, and Klein, G et al., "E-commerce user behavior model : an empirical study", Human Systems Management, Vol. 19, No. 4, 2000, pp. 265-276.

[8] Klopping, Inge M., and Mckinney, Earl, "Extending the Technology Acceptance Model and the Task-Technology Fit Model to Consumer E-Commerce", Information Technology, Learning and Performance Journal, 2004, pp. 35-48.

[9] Komiak, Sherrie Y. X., and Benbasat, Izak, "The Effects of Personalization And Fami-

liarity on Trust and Adoption of Recommendation Agents", MIS Quarterly, Vol. 30, No. 4, December, 2006, pp. 941-960.

[10] Kotler, Philip, and Armstrong, Gary, Principles of Marketing, 5 ed. : Prentice Hall International, 1991.

[11] Linden, Greg, Smith, Brent, and York, Jeremy, "Amazon.com recommendations : Item-to-Item Collaborative Filtering", IEEE Internet Computing, Vol. 7, No. 1, Jan/Feb 2003, pp. 76-80.

[12] Netflix, "Netflix movie dataset", November 2006; http://www.netflixprize.com/.

[13] Simon, Ha, "Theories of decision-making in economics and behavioral science", American Economic Review, Vol. 49, No. 3, 1959, pp. 253-283.

[14] Xiao, Bo, and Benbasat, Izak, "E-Commerce Product Recommendation Agents : Use, Characteristics, and Impact", MIS Quarterly, Vol. 31, No. 1, March, 2007, pp. 137-209.

Author Profile

Hyung Jun Ahn

Dr. Hyung Jun Ahn is an Assistant Professor at the College of Business Administration at Hongik University, Korea. He received his PhD in MIS from KAIST and served as a Senior Lecturer at Waikato University in New Zealand before joining Hongik University. His main research interests include intelligent systems, agent-based systems, and collaborative systems for business.

◈ 이 논문은 2009년 08월 21일 접수하여 1차 수정을 거쳐 2009년 09월 16일 게재확정되었습니다.