Evaluating Recommender Systems

Wen Wu¹, Liang He², Jing Yang²
Department of Computer Science and Technology
East China Normal University
Shanghai, China
¹5111201013@ecnu.cn
²{lhe, jyang}@cs.ecnu.edu.cn

Abstract—Recommender systems now tend to gain popularity and significance. The proliferation of many recommender systems leads to the difficulty of locating a good recommender system. The algorithms contained in the recommender system determine the efficiency of the recommender systems. The question now is to find the most appropriate algorithms to meet users' needs. So far, the research carried out has focused on improving the accuracy of recommender systems. In this paper, we propose that the recommender system should move beyond the conventional accuracy criteria and take some other criteria into account, such as coverage, diversity, serendipity, scalability, adaptability, risk, novelty and so on. Experimental results with data from VELO indicate that people with different interest degrees tend to prefer different algorithms; thus the use of various evaluation criteria to judge the performance of algorithm is meaningful.

Keywords—Recommender systems, collaborating filtering, evaluation of algorithms

I. INTRODUCTION

As E-Commerce has proliferated in terms of variety and quantity, the end-users spend considerable time to select the products and services. Recommender systems can now be found in many modern applications that expose the user to a huge collection of items. Such systems [1] typically provide the user with a list of recommended items they prefer, or supply guesses of how much the user prefers each item.

The recommender systems are supported by well founded and incremental algorithms. These algorithms differ considerably with respect to their strengths and weaknesses. Thus, the users encounter with choices for the selection of the most effective.

Today most algorithms focus on improving the accuracy of the recommender system, however providing accuracy alone is inadequate. An example is as follows:

Imagine you are using a video recommender system. Suppose all of the recommendations you got are for videos you have already watched. Even if the system was very good at ranking all of the videos you have watched in order of preference, this still would be a poor recommender system. Would you like to use such a system? Absolutely not, on the accuracy of the example has good performance, but it is not difficult to see the shortcomings: repeated redundant and nothing new at all. Unfortunately, this is exactly how we currently test our recommender systems. In the standard methodology, the video recommender would be penalized for recommending new films instead of videos the users have already watched! Current accuracy metrics, such as MAE measure recommender algorithm performance by comparing the algorithm’s prediction against a user’s rating of an item, which leads to the result of redundancy.

By focusing on the accuracy of testing recommenders, are we really helping users find the items, they are interested in? We claim that there are a lot of different facets to the recommendation process which current accuracy metric systems do not measure. It is essential to evaluate recommender system from different facets in order to make the system more diverse, specific and comprehensive. In this paper, we will address such facets: accuracy, coverage, diversity, serendipity, scalability, adaptability, risk, novelty, etc. We hope that after experiments, we could draw some meaningful conclusions.

The main contributions in this paper are:
- We design five simple recommendation algorithms based on the single item for coupon recommendation.
- Some special and novel evaluation metric systems are chosen and redefined to evaluate our recommender system.
- We conduct a user study aiming to know the preference of people with different interest degree.

The paper is structured as follows: we begin with some necessary background on recommender system and proposed purpose and significance of this study (Section 1). Some evaluation metrics for recommender systems is proposed in Section 2. Then we choose and redefined some of them to evaluate some algorithms with VELO datasets, compared with manual scoring (Section 3). Finally, draw some meaningful conclusions and give future work in Section 4.

II. EVALUATION METRICS FOR RECOMMENDER SYSTEM

This paper will focus on research questions of evaluating recommender system from different facets. Being accurate is not enough, the quality of recommendation is more important. The Recommended quality [2] is defined to meet or exceed customers' expectations, and obviously, the user will not only be satisfied with some of the tedious, nothing new recommendation. So some new evaluation metrics should be considered for recommender system.
Research at home and abroad related to evaluation metrics are as follows:

We divided this metrics into two parts [3]: evaluation criteria based on the recommender algorithm and evaluation criteria independent on the recommender (from systems’ angle and users' angle). Figure 1 shows the classification table:

![Evaluation metrics for Recommender](image)

The brief introduction of the Figure 1’s evaluation metrics is as follows:

A. Evaluation criteria based on recommender algorithms

- Accuracy

It has been largely acknowledged that accuracy can be divided among three categories [5]: the accuracy of rating’s predictions, the accuracy of usage predictions and the accuracy of rankings of items.

For the different categories, we need to use different metrics or formula to express, for example, if we use scoring prediction, we often use Root Mean Squared Error (RMSE), while ordering items according to the user's preferences, we can try to determine the correct order on a set of items for each user and measure how close a system comes to this correct order.

- Coverage

In many cases, accuracy and coverage always come in a pair. An example is usage predictions [1]which mentioned above.

In paper [2], the coverage also has two definitions: (1) the percentage of the items for which the system can able to generate a recommendation, and (2) the percentage of the available items which effectively are ever recommended to a user. Though different authors differ with respect to terminology, here we adopt the definition from [4]and refer to (1) as prediction coverage, and to (2) as catalogue coverage.

- Novelty

Novelty recommendations are recommendations for items that the user did not know about. In applications that require novel recommendation, an obvious and easy to implement an approach is to filter out items that the user already rated or used. Otherwise, it will reduce trust in the system, resulting in the loss of the user.

- Serendipity

Serendipity is a measure of how surprising the successful recommendations are. Imaging in a video recommendation system, user A is Spielberg’s fans. The system makes a list based on user preferences of Spielberg’s other films, although the user hasn't watched before, this can be only considered as a new recommendation rather than a surprise one. Serendipity has two characteristics: surprising and attractive [4]. That means a highly serendipitous recommendation would help a user to find a surprising and interesting item.

- Diversity

Diversity is generally defined as the opposite of similarity [5]. In some cases suggesting a set of similar items may not be as useful to the user, because it may take longer to explore the range of items. Consider, for example, for a coupon, presenting a list with 5 recommendations, all for the same company, varying only on the choice of items, may not be as useful as suggesting five different items from different companies.

B. Evaluation criteria depend on recommender algorithms

1) From system's angle

- Confidence

Confidence in the recommendation [6] can be defined as the system’s trust in its recommendations or predictions. If the content has a better explain, the system's trust will go up, and having a good degree of trust systems tend to have more quality.

- Scalability

As recommender systems are designed to help users navigate in large collections of items, one of the goals of the designers of such systems is to scale up to real data sets. Algorithm needs to be further optimized, such as the use of in-memory calculation method to resolve the massive data problems.

- Adaptivity

In paper [5], real recommendation systems may operate in a setting where the item collection changes rapidly, or where trends in interest over items may shift. Perhaps the most obvious example of such systems is the recommendation of news items or related stories in online newspapers.

2) From users' angle

- User Preference

When we wish to improve a system, it is important to know why people favor one system over the other. Typically, it is easier to understand that when comparing specific properties. So, while user satisfaction is essential to measure, breaking satisfaction into smaller components is helpful to measure the system and improve it.
• Trust

While confidence is the system trust in its ratings, in trust we refer here to the user’s trust in the system recommendation [7]. In the recommender system, sometimes the user will find something he/she viewed or purchased in the recommendation list, though the user might think that novelty is not enough, but at least that will increase the trust in the system, because to some extent the user would believe that the system can able to predict the tastes accurately.

• Privacy

Nowadays, due to increasing hacking, users need to take security into account. In a collaborative filtering system, a user willingly discloses his/her preferences over items to the system in the hope of receiving useful recommendations. However, it is important for most users whom their preferences stay private [8].

Next, some of the evaluation metrics will be chosen and redefined to evaluate several algorithms with VELO datasets.

III. EXPERIMENT

A. Experimental purposes

Evaluating recommender systems from different facets is essential, since it not only makes the system more diverse, specific and comprehensive, but also meets users’ needs. Therefore, we choose and redefined some of metrics to evaluate some algorithms with VELO dataset, such as three collaborative filtering recommendation algorithms and two popular algorithms (Hot Recommendation and Fresh Recommendation), then a user study will be proposed to evaluate the recommender system from other way. Finally, we will draw a conclusion.

B. Experimental data sets and environment

1) Data sets

In this paper, we select VELO dataset as our experimental data sets. VELO is a company providing a discount coupon service in China. For our purpose presented above, we make use of only part of the data and part of the variables collected from a commercially operational service. The dataset analyzed in this paper was collected from time period 1.8.2009-22.2.2010. In the filtered dataset 500 customers were selected from the overall customer database. The selection was based on the need to pick those customers with a lot of background information such as age, occupation, job, salary, interest and marriage records.

According to the “Log Date,” we set data before January 1, 2010 as a training set, while data after that time as the test set. The data of training set only to be used to make a prediction while the recommendation of test set is the key of evaluating.

2) Data environment

In this paper, hardware platform is the PC configured with Intel Core 4 CPU and 2G RAM. All the algorithms are achieved by Microsoft Visual Studio 2008, and we use SQL Server 2008 to support our database service.

C. Brief introduction of experimental algorithms

In this paper, we design five recommendation algorithms based on a single item. Three of them are collaborative filtering algorithms; others are non-personalized algorithms. How to choose these algorithms depends on whether we are able to get the history of the behavior of the users.

The following three collaborative filtering algorithms are used when it is possible to get users' history behavior.

• Collaborative filtering algorithm based on coupons' similarity (CSCF)

According to the coupon which a user had printed often in a certain period, we find the most similar and unexpired coupons to make a recommendation by comparing keyword of each coupon.

• Collaborative filtering algorithm based on users' similarity (USCF)

According to the coupon which a user had printed often in a certain period, we find the users who print such as coupon most as the neighbor, then recommend the unexpired coupons which neighbors print most of the users.

• Back to higher Recommendation (BR)

According to the coupon which a user had printed most in a certain period, we find the coupon's company, and then recommend the unexpired coupons which most printed and belong to these companies to the user.

If a customer is a newcomer or has none history behavior, it is difficult to use the above algorithms, so we address two non-personalized algorithms.

• Hot Recommendation (HR)

According to the coupons sorted by the print count descent, we find the hottest and unexpired coupons to the user. Each user has same recommendation.

• Fresh Recommendation (FR)

According to the coupons sorted by the edit time from now to the previous, we find the newest and unexpired coupons to the users. Each user has same recommendation.

In summary, the simple process of the algorithm contains four steps:

Step1: Check the history of the behavior of the users.
Step2: If records exit, then go to Step3; otherwise goto Step4
Step3: Recommend items by USCF, CSCF and BR.
Step4: Recommend items by HR and FR.

D. Experimental evaluation metrics

In this paper, according to practical application of VELO, we propose several evaluation metrics and redefined them.

• Accuracy = Nh / Nt (1)
• Coverage = Nh / Nu (2)
• Diversity = Nc / Ntrc (3)
• Serendipity = (Nuhc - Ntpc((uhc)<1000)) / Ntrc (4)
• Scalability: time complexity and space complexity

Explain:
• Nhc means the number of hit coupons.
• Ntrc means the number of total coupons system recommended.
• Ntpc means the number of coupons which user actually printed.
• Ncc means the number of different companies of the recommended coupons.
• Nuhc means the number of coupons except hit ones.
• Ntpc((uhc)<1000) means the number of non-hit coupons which total print count <1000

It is not hard to see that the more we focus on increasing accuracy, the less catalog coverage is, so just consider one-side is not very reasonable. We should make a balance between accuracy and coverage. Therefore, we design comprehensive evaluation criteria to balance accuracy and coverage. In summary,

- Comprehensive evaluation criteria = accuracy * 0.65 + coverage * 0.35 (5)

E. Experimental proposal

This paper provides emphasis on evaluating recommender system from different facets. So we make a proposal as follows:

- Clarify a hypothesis: users with different degree of interest may tend to prefer different kinds of algorithms.
- For 500 users we choose, we use three collaborative filtering algorithms and two non-personalized algorithms to make five recommendation lists respectively. We evaluate and compare these five algorithms by using five metrics present above, such as accuracy, coverage, diversity, serendipity, scalability.
- Choose some representative recommendation list, then adopting manual scoring in order to judge the performance of the algorithms from the user's perspectives.
- Verify whether the hypothesis is reasonable and analyze the importance of evaluating recommender system by using different metrics.

F. Experimental results and analysis

1) According to the above parts, we were interested in clarifying the following hypothesis:

Hypothesis 1: users with different degree of interest may tend to choose different kinds of algorithms.

Notice: the definition of the degree of interest will be mentioned in F(3).

2) Objective Evaluation

- Comprehensive evaluation criteria & Diversity & Serendipity

According to the formula (3) * (4) * (5), it is not difficult to get the value of objective evaluation, Figure 2 shows the result:

From the table above, we can see:

If we use "comprehensive criteria" to evaluate five algorithms, the rank of order is: USCF>CSCF>HR>BR>FR.

That means USCF and CSCF performs better than the other three recommender algorithms in our conventional metrics, such as accuracy and recall since users' history behavior are considered into recommendation, items which have high similarity or preferred by the neighbors are much easier to be recommended.

If we use the criteria "diversity" to evaluate five algorithms, the rank of order is: BR>USUF>FR>HR>CSCF.

The reason that BR's diversity value is much higher than the others is that the all items we recommended by BR algorithm are from different companies in order to increase the diversity.

If we use the criteria "serendipity" to evaluate five algorithms, the rank of order is: BR>HR>USCF>FR>CSCF.

It is not difficult to find that CSCF and USCF which have good performance in accuracy and recall does not stand out in this evaluation metric. They put more emphasis on the similarity but just ignore the importance of novel and serendipity while BR and HR attach more importance to them.

- Time complexity (time consumption)

Their time complexities are both O(n), but the specific time spent there is still a big difference. According the core sql statement in five algorithms, we set time spent on “order by” statement as t1 while time spent on “group by” statement as t2. The results can be seen in Table 1:

<table>
<thead>
<tr>
<th>TABLE I. TIME COMPLEXITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time complexity</td>
</tr>
<tr>
<td>HR</td>
</tr>
<tr>
<td>FR</td>
</tr>
<tr>
<td>CSCF</td>
</tr>
<tr>
<td>USCF</td>
</tr>
<tr>
<td>BR</td>
</tr>
</tbody>
</table>
It is clear to see that algorithms' order by time complexity is: BR>CSCF>USCF>HR>FR

- Space complexity (memory consumption)

In our five algorithms, CSCF has more memory consumption, because it should occupy a large number of memories when they calculate similarity each other.

So we can see that algorithms' order by space complexity is: CSCF>USCF>HR>FR=BR

3) User Study

In this paper, we adopt three collaborative filtering algorithms and two non-personalized algorithms to make five top-10 items recommendation lists and invite users to rate them. Rating guidelines is 1 star to 5 stars, from low to high satisfaction.

50 people were invited to conduct this user study. We will distinguish between the 50 people to be the users with broad interests or the users with narrow interests by analyzing their selecting interests in various topics, such as food, shopping, cinema, gaming, hairdressing, cosmetics, reading, dancing, body building, club, photo, fashion, outdoors sport, pet, KTV, others. If he chooses more than five kinds of interests, then he will be considered as the user with broad interests. Otherwise, he may be the people with narrow interests. According to statistics, 22 people belong to the users with broad interests and 28 people belong to the users with narrow interests.

The manual score results are as follows:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Narrow Interests</th>
<th>Board Interests</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>3.25</td>
<td>3.25</td>
</tr>
<tr>
<td>FR</td>
<td>2.24</td>
<td>2.64</td>
</tr>
<tr>
<td>CSCF</td>
<td>4.22</td>
<td>3.31</td>
</tr>
<tr>
<td>USCF</td>
<td>4.46</td>
<td>4.26</td>
</tr>
<tr>
<td>BR</td>
<td>3.42</td>
<td>3.80</td>
</tr>
</tbody>
</table>

Analysis of possible reasons:

Combined with Figure 2 and Table 2, we can see that:

- From the result, we can find that USCF always ranks the first. The reason is that USCF based on users' similarity, the recommender items will not only be to users' taste, but also be very fresh and rich.

- FR is different from USCF, the algorithm always ranks the last. The reason we think is that FR doesn't recommend item according to different people, although it seems to be fresh, but the accuracy and coverage are too poor.

- The scoring of CSCF, BR and HR ranked by different kind of people is not same. If we evaluate them only by the accuracy, the result is very clear: CSCF performs better than BR and HR. But from user' perspectives, the fact is not just as easy as we thought.

- We can see that from the scoring of users with broad interests, BR ranks better than CSCF and HR. The reason is that CSCF recommends items based on coupons' similarity, that will cause the items they recommend be boring and lack of surprise. On the contrary, although BR has lower accuracy than CSCF, but it is more diversity and interesting. Actually it is good news for these people with broad interests.

- On the other hand, if our customer is a person with narrow interests, that is to say he may like just one or two thing or his taste never changes, CSCF is a better recommender algorithm. The items recommended by CSCF always have high accuracy and coverage. In this case, BR doesn't have many advantages.

- As a non-personalized algorithm, HR has a not bad accuracy, diversity and serendipity. So if the user is a newcomer or has narrow history behavior, HR is just OK.

- From the users' perspective, they only need to care about the quality of the coupons, but to the designer, we should also take the algorithms' performance into account. In this case, CSCF may spend more time and space than other algorithms. Efficiency is a question we can't ignore.

All in all, this hypothesis holds good.

IV. CONCLUSION AND FUTURE WORK

In this paper we discussed how we should evaluate recommendation algorithms so as to select the best algorithm from a set of candidates. We considered that being accuracy is not enough. Recent works suggest that beyond accuracy, a variety of other metrics should be taken into account when evaluating a recommender system. These metrics put emphasis on the quality of the recommender system.

Through the Experiments, we compared five algorithms from different facets and proved that the preference of different algorithms vary from person to person. So to different recommender systems and different users, we should take the essence and discard the dregs.

In the future, we are interested in studying the impact of personality on users’ needs for recommendation quality, such as diversity and novelty.

ACKNOWLEDGMENT

This work is supported by a grant from the National High Technology Research and Development Program of China (863 Program) (No.2011AA01A107) and the ShangHai Science and Technology commission Foundation (No.10dz1500103 and 111511502203).

REFERENCES


