A RECOMMENDER SYSTEM BASED ON CONTEXTUAL INFORMATION OF CLICK AND PURCHASE DATA TO ITEMS FOR E-COMMERCE

Duo Lin1, 2*, Su Jingtao2

1 Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, P.R. China
2 School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing, 100083, P.R. China
*e-mail: duolin2003@126.com

Abstract

A recommender system with good performance for an e-commerce web site is important for both customers and merchants. In most of the existing recommender systems, only the purchase information is utilized data and the navigational and behavioral data are seldom concerned. In this paper, we design a novel recommender system for comprehensive online shopping sites. In the proposed recommender system, the contextual information data, such as access, click, read, and purchase information of a customer, are utilized to calculate the preference degree to each item; then items with larger preference degrees are recommended to the customer. In addition, nonexpendable items are distinguished from expendable ones and handled by a different way. Lastly, we structure an example to show the performance of the proposed recommender system. The results show that the proposed method is well-performed.

Keywords: Recommender System, Online shopping, E-commerce, Preference Degree

1 Introduction

With the development of internet and smart mobile devices, online shopping has been more and more convenient and frequent in our lives. Generally, the types of goods are very large in a comprehensive online shopping site like Alibaba, therefore there is a trouble for customers to find out an appropriate item from all the items. That would affect the customers' enthusiasm for buying and then decrease the sales of businesses. For that reason, a suitable recommender system seems to be very important for both customers and businesses of a comprehensive online shopping site.

Dozens of recommender techniques have been designed for different kinds of web sites and generally they can be divided into two categories: content-based recommendations and collaborative filtering recommendations [1]. The former chooses recommended items based on the purchasing and rating history of a customer himself, while the latter selects recommended items according to the history of the similar customers. Both content-based recommendations and collaborative filtering ones have disadvantages. Content-based methods work badly for new items or new customers; it cannot distinguish two different items with the same keywords either. The collaborative filtering methods have several shortcomings: it cannot find the neighbors for new users and then cannot recommender any items to him; a new item does not have any ratings and its utilities to customers cannot be measured; the ratings are too sparsity and which leads that the recommender is inaccurate. To overcoming that, hybrid approaches are further proposed. Hybrid approaches can be obtained by the following four ways: a) implement two methods separately and combine their recommended items by a specially designed way [2]; b) adding the characters of content-based methods into collaborative filtering models[3,4]; c) adding the collaborative filtering ideas to content-based ways[5]; d) constructing a unifying recommendation model which includes the characters of both content-based and collaborative filtering ways[6].

The above-mentioned methods are based on the purchasing and rating information and the navigational and behavioral data of customers are not utilized. In 2001, Kelly and Belkin [7] and Lee et al.[8] separately consider the impact of reading time and click data to the ratings of items. In 2005, Kim et al.[9] consider more comprehensive navigational and behavioral data into their recommender system for e-commerce sites. In 2011, Kim and Yum[10] proposed an improved recommender system based on the method of Kim et al. The improved system calculates the confidence levels between clicked items, between the items placed in the basket, and between purchased items, respectively, and calculates the preference degree of an item through a linear combination of the above three confidence levels. The utilization of navigational and behavioral information can remedy the lack of rating information and improve the performance of recommender system.

In this paper, we propose a recommender system for comprehensive online shopping sites according to the contextual information of navigational and behavioral data of customers. The innovations of the proposed method are threefold:

(1) All the navigational and behavioral information of a customer to an item are converted to a uniform measure.
A ripple-like algorithm is used to measure the similarity between items. Nonexpendable items are distinguished from expendable ones.

The remainder of this paper is organized as follows. In Section 2, the details of the proposed recommender method are introduced gradually; in Section 3, an example is carefully structured, which exhibits the performance of our recommender method. Lastly, a brief conclusion is drawn in Section 4.

2 The Proposed Method

2.1 Selected Navigational and Behavioral Information

In the method of Kim et al., the navigational and behavioral patterns of customers are considered. The navigational information include browsing, searching, product click, basket placement, and actual purchase, while behavioral information consist of the click ratio of a certain type of product, length of reading time spent on a specific product, number of visits to a specific product, printing and bookmarking[9,10]. The consideration of data is comprehensive; however we think that some of the data are overlapping with each other for a comprehensive online shopping site.

In a nowadays online shopping site, the navigate system of products (or items) are very well and the products are organized into a tree structure. A customer can directly reach the most relative class of his interested product. Hence, the significance of some traditional navigational and behavioral information would decline. The analyses are as follows:

- **Browsing and searching information can be reduced to click information.** No matter browsing or searching for a product, the direct goal is to click the product and reading its parameters and descriptions.
- **Basket placement information can be classified to click or purchase information.** When a customer place a product into his shopping basket, he has two possible purposes: purchase it right now or easily click it at next access.
- **The information of click ratio and number of visits to a product can be calculated by the click information.** As long as the click information of all products are captured, the above two information can be easily calculated.
- **Printing and book-marking information are meaningless for comprehensive online shopping sites.** The behaviours may often occur in music, movie, or book sites, but they are unusual behaviours for comprehensive online shopping sites.

For this reason, we would calculate the equivalent attention and then the preference degree of a customer to each product, by only considering the click, read and purchase information.

2.2 Equivalent Attentions and Corresponding Preference Degrees

2.2.1 Basic Formula for Preference Degree

If a customer wants to purchase a product, he would click it and view its detailed descriptions first. Hence, there is relevance between the number of clicks and preference degree. The relation ought to have the following characters:

- **The preference degree should grow fast first.** If a product is clicked by a customer only one time, that maybe occurs by chance; yet, if a product is repeatedly clicked, it means that the customer has a great preference to the product.
- **The growth rate of preference degree ought to slow down later.** For example, one customer clicks a product 50 times while another customer clicks the same product 100 times; it does not mean that the latter has a double preference to the product, but is possible that the latter have more sparse time to surf the internet.
- **The preference degree ought to have an upper bound.**

According to the above analyses, we should find a first fast increasing, later slow increasing, and bounded function. After screening among the usual functions, we find that arc tangent function is a suitable candidate. Hence, we use arc tangent function to express the relation between preference degree and number of clicks.

Suppose that the number of clicks to a product $p$ is $N_p$ and the corresponding preference degree is $D_p(p)$, then

$$D_p(p) = \frac{2}{\pi} \arctan(C_0 N_p)$$

(1)

In which, $C_0$ is a parameter to adjust the variation speed of preference degree. Besides, in order to coincide to the traditional recommender system, a coefficient of $\frac{2}{\pi}$ is used to normalize the preference degree into interval $[0,1]$.

![Figure 1. The changing rules of the basic preference degree $D_p(p)$ under different values of $C_0$.](image)

Figure 1 shows the curves of $D_p$ under different values of parameter $C_0$. We can see that under each value of $C_0$, the image increases fast first and slow later. For a bigger $C_0$, the value of $D_p$ would be rapidly close to 1 while for a smaller one, the value would approach slower to 1. To explicitly distinguish the preference degrees for different numbers of clicks, a value no more than 0.5 would be reasonable.

2.2.2 Impact of Reading Time Added
After a product is clicked, its detailed description would be read by the customer. Generally, the longer the Reading time is, the more favorite the customer is to the product.

In our method, we convert the impact of both click and read into a uniform quantity named equivalent attention. The impact of one time of click is always one. To measure the impact of reading time, we give a reference length of time $T_{ref}$. The reference length can be selected as the average reading time of all products or other carefully determined value by an optimization method. Record the reading time of a product as $t_p$ and which can be divided into several sections by the reference length. Then the equivalent attention $A_i(p)$ of click and read can be calculated by the following formula:

$$A_i(p) = \text{Impact}_{\text{click}} + \text{Impact}_{\text{read}}$$

$$= \frac{1}{\pi} \arctan(C_i A_i(p))$$

$$= \frac{1}{\pi} \arctan(C_i \sum_{p \neq q} b^{d_{p,q}} \sqrt{\sum_{k=0}^{N-1} \left(\frac{1}{2}\right)^k})$$

In which, “\(\lfloor \cdot \rfloor\)" is the lower integer notation and it would obtain the maximum integer no more than a given real number.

2.2.3 Impacts of Relevant Products Added

Products may have relevances with each other. When a customer clicks and reads a product, it implies that he is interested in not only the clicked product, but also the relevant products.

In a comprehensive online shopping site, the products are always well classified and placed in a tree structure. The relevance between two products can be estimated according to their distance on the tree.

In this paper, we design a ripple-like algorithm to measure the relevances among products. The name is coming from the following reason: for a water wave, the maximum amplitude occurs at the center, and the farther a point is to the center, the smaller the amplitude is; we would use this character to define the relevances among products. For two products $p$ and $q$, record their distance on the tree as $d_{p,q}$, then the relevance level $R_{p,q}$ between them is defined as

$$R_{p,q} = b^{d_{p,q}}$$

In which, the base number $b$ is a parameter chosen from interval $(0,1)$ and whose value can adjust the relevances among products. Specifically, if $p = q$ , $d_{p,q} = 0$ and then $R_{p,q} = b^0 = 1$.

Supposing that the set of all products is $P$, according to the relevance level between products, we can calculate the total equivalent attention of the click and read information of all the products to a specific product $p$ by the following formula:

$$A_i(p) = \sum_{p \neq q} R_{p,q} A_i(q)$$

$$= \sum_{p \neq q} b^{d_{p,q}} \sum_{k=0}^{N-1} \left(\frac{1}{2}\right)^k$$

Substituting $N_p$ with $A_i(p)$ in Formula 1, we can get the preference degree of a customer to a product $p$ by considering all the click and read information of all the products as follows:

$$D_i(p) = \frac{2}{\pi} \arctan(C_i A_i(p))$$

$$= \frac{2}{\pi} \arctan(C_i \sum_{p \neq q} b^{d_{p,q}} \sqrt{\sum_{k=0}^{N-1} \left(\frac{1}{2}\right)^k})$$

2.2.4 Impact of Purchase Information Added

In the scheme of Kim et al.[9], once a customer purchased a product, his preference degree to the product increases to 1 (the maximum value). They do not consider the difference of the durability of products. In fact, some products are durable and some are nondurable; the buying cycles of products with different durability would be different either.

Depending on the durability, the items in a comprehensive online shopping site can be classified into expendable products and nonexpendable products:

- An expendable product is easily exhausted and may be repeatedly purchased in a short time.
- A nonexpendable product can be used for a rather long time and would not be repurchased recently.

We use the symbol $\sigma(p)$ to express the durability of a product $p$, and whose value can be express as follows:

$$\sigma(p) = \begin{cases} +1, & p \text{ is expendable;} \\ -1, & p \text{ is nonexpendable.} \end{cases}$$

Purchasing of an expendable or a nonexpendable product would generate different impact to the preference degree: purchasing of an expendable product would make a positive impact to the preference degrees of all the relevant products, and buying of a nonexpendable item would make a negative impact to the similar products while a positive impact of other relevant items.
To reality the effect in a uniform way, we give sign function to roughly express the purchase impact of a product to product $q$:

$$\delta_q(p) = \begin{cases} \sigma(p), & d_{p,q} \leq d_q; \\ +1, & d_{p,q} > d_q. \end{cases}$$  

In which, the threshold $d_q$ varies with the levels of classification of the online shopping site. A rough classified site ought to use a small threshold while an elaborate site should use a large one. Specifically, if $q = p$, $\delta_q(p) = \delta_p(p) = \sigma(p)$.

In our method, We regard the purchase behavior as a time of super click. We appoint that a product is clicked $N_p$ times and purchased $M_p$ times, then the equivalent attention of click, read, and purchase information of itself is

$$A(p) = \delta_q(p)C_pM_p + A(q)$$

$$= \delta_q(p)C_pM_p + \sum_{i=1}^{n} \sum_{k=1}^{\gamma_i} \left( \frac{1}{2} \right)^{i}$$  

Furthermore, the total equivalent attention of click, read, and purchase information of all the product to a specific product $p$ can be calculated by the following formula:

$$A(p) = \sum_{q \neq p} R_{q,p}A(q)$$

$$= \sum_{q \neq p} b^{q,p} \left[ \delta_q(p)C_pM_p + \sum_{i=1}^{n} \sum_{k=1}^{\gamma_i} \left( \frac{1}{2} \right)^{i} \right]$$

Take $A(p)$ into Formula 1 we can get the preference degree of a customer to a product $p$ by comprehensively considering the click, read, and purchase information of all the products as:

$$D_p(p) = \frac{2}{\pi} \arctan(C_pA(p))$$

$$= \frac{2}{\pi} \arctan(C_p\sum_{q \neq p} b^{q,p} \left[ \delta_q(p)C_pM_p + \sum_{i=1}^{n} \sum_{k=1}^{\gamma_i} \left( \frac{1}{2} \right)^{i} \right])$$

### 3.1 Accessing Data and Selected Parameters

Figure 2 is a part of the familiar product-tree of comprehensive online shopping sites. The tree shows the classification of computer and related products. The products are divided into three classes at the first level: complete machine, computer parts, and peripherals. At the second and third levels, the products can be divided into dozens of classes, but only a few of them are listed in this example. Specific products are represented by leaves in the tree.

![Product Tree](image)

### 3.2 An Example to Exhibit The Operations of The Proposed Recommender Method

<table>
<thead>
<tr>
<th>Access No.</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>22</td>
<td>18</td>
<td>12</td>
<td>15</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>12</td>
<td>15</td>
<td>9</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>28</td>
<td>22</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>0</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>15</td>
<td>20</td>
<td>10</td>
</tr>
</tbody>
</table>
In the proposed method, there are several parameters to be determined. In this example, we directly choose the values of the parameters as Table 2.

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Selected Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_r$</td>
<td>0.4</td>
</tr>
<tr>
<td>$C_p$</td>
<td>2</td>
</tr>
<tr>
<td>$T_{of}$</td>
<td>10 seconds</td>
</tr>
<tr>
<td>$b$</td>
<td>$\sqrt{2}$</td>
</tr>
<tr>
<td>$d_i$</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 2: The Values of The Parameters Chosen in The Example

3.2 The Recommended Products

According to the method in section 2, we can calculate the preference degrees of all the ten products. Then we can recommend the items with higher preference degrees to the customer. The calculating results and recommended items are listed in Tables 3 and 4.

Table 3 shows the preference degrees of all the products after each time of access. After the preference degrees are obtained, we recommend the top four items to the customer. In the table, the underlined numbers mean that the corresponding items are recommended to the customer at his next access.

In Table 4, We illustrate the performance of the proposed method. The indices include the precision, recall and $F_1$, which are the most familiar indices used in recommender systems. Supposing that $R(p)$ is the set of recommended items to a customer $p$ and $C(p)$ is the set of clicked items by the customer. Then the definitions of the indices are as follows[11,12]:

$$\text{Precision} = \frac{|R(p) \cap C(p)|}{|R(p)|}$$

$$\text{Recall} = \frac{|R(p) \cap C(p)|}{|C(p)|}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4 Conclusions
In this paper, a novel recommender method for e-commerce based on the contextual information of click, read and purchase is proposed. The remarkable point is that, which distinguishes nonexpendable items from expendable ones. The innovations are hoped to improve the performance of the recommender system.

At the last of the paper, a simple experiment are implemented and the performance of the method are concerned. The results show that the proposed method is well-performed.

One of our further work is to find suitable data set to investigate the performance the proposed method and optimize the parameters. We also expect that some online shopping sites can test our method on their sites, for the parameters determined by that way would be more convicive.

Acknowledgment

This work is supported by the Yunnan Provincial Department of Education Foundation Project (No. 2014Y087). Fundamental Research Funds for the Central Universities No. FRF-TP-15-027A3.

References
