INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(5), 661-676, October 2017.

Mining Users' Preference Similarities in E-commerce Systems Based on Webpage Navigation Logs

P. Li, C.X. Wu, S.Z. Zhang, X.W. Yu, H.D. Zhong

Ping Li
College of Biological and Environmental Sciences,
Zhejiang Wanli University
No. 8 South Qianhu Rd., Ningbo, Zhejiang, 315100,
P. R. China, liping_kaixin@163.com

Chunxue Wu
School of Optical-Electrical and Computer Engineering,
University of Shanghai for Science and Technology
No. 516 Jun Gong Road, Shanghai 200093,
P. R. China, tyfond@126.com

Shaozhong Zhang
School of Electronic and Computer Science,
Zhejiang Wanli University
No. 8 South Qianhu Rd., Ningbo, Zhejiang, 315100,
P. R. China, dlut_z88@163.com

Xinwu Yu
The Information Center,
Zhejiang Wanli University
No. 8 South Qianhu Rd., Ningbo, Zhejiang, 315100,
P. R. China, herrison@163.com

Haidong Zhong*
Logistics and E-commerce School,
Zhejiang Wanli University
No. 8 South Qianhu Rd., Ningbo, Zhejiang, 315100,
P. R. China
*Corresponding author: zhonghaidong@zwu.edu.cn

Abstract: Mining users' preference patterns in e-commerce systems is a fertile area for a great many application directions, such as shopping intention analysis, prediction and personalized recommendation. The web page navigation logs contain much potentially useful information, and provide opportunities for understanding the correlation between users' browsing patterns and what they want to buy. In this article, we propose a web browsing history mining based user preference discovery method for e-commerce systems. First of all, a user-browsing-history-hierarchical-presentation-graph to established to model the web browsing histories of an individual in common e-commerce systems, and secondly an interested web page detection algorithm is designed to extract users' preference. Finally, a new method called UPSAWBH (User Preference Similarity Calculation Algorithm Based on Web Browsing History), which measure the level of users' preference similarity on the basis of their web page click patterns, is put forward. In the proposed UPSAWBH, we take two factors into account: 1) the number of shared web page click sequence, and 2) the property of the clicked web page that reflects users' shopping preference in e-commerce systems. We conduct experiments on real dataset, which is extracted from the server of our self-developed e-commerce system. The results indicate a good effectiveness of the proposed approach.
Keywords: web browsing history mining, e-commerce, preference, recommendation.

1 Introduction

E-commerce (or electronic commerce) usually refers to products or services exchange oriented activities based on the Internet technology, which covers online payment, information security, logistics distribution, etc [25, 29].

Pushed by the widespread availability of the Internet, more and more consumers prefer to shift from traditional face-to-face transactions to web-based commercial activities [16]. Meanwhile, with the advances of network technology and fast-growing e-commerce systems, such as Amazon, Stimulated by network technology advances, the rapid growth of networking systems and the boom in netizens, an ever-increasing number of traders and entrepreneurs have participated in e-commerce [19].

Nowadays, supply chain management, online transaction processing and many other e-commerce relevant industries have attracted thousands of workers and companies to provide a great many products or services for the online transactions. With online information growing exponentially, it will eventually result in "Information Overload" and "Information Loss" in e-commerce systems [11], which seriously hinder the development of e-business or e-commerce industries.

Since the born of collaborative filtering approach in the 1990s, recommendation systems have become an intensively investigated independent discipline and deemed as an effective means to ease the "Information Overload" problem. Generally, the existing recommendation approach can be divided into three categories [2, 26, 28]: (1) collaborative filtering recommendation, which is based on the idea of similarity of users to make predictions (filtering) about the interests of an individual by calculating preferences from other users (collaborating) who have brought the same items as the target user; (2) content-based recommendation, which is the technology of choosing items as recommended for a target user according to other similar users' preferred items; (3) hybrid recommendation, which is proposed to combine different recommendation technologies according to different mixed strategies (e.g., weighted, switch, mixed characteristics, combination, series, meta level hybrid, etc.). Although, the specific steps of these recommendation methods vary, the fundamental principles of them are similar: finding users with similar preferences with target consumer and using them to make recommendations.

However, there are still many challenges, such as data scarcity and cold-start [20, 21], in utilizing these recommendation approaches. Searching for candidate users more precisely is deemed as an important solution to improve these challenges. In the majority of the existing researches, insufficient effort has been made to solve the problem that users' preference is time-varying and can be measured in different granularity.

From the view point of Srivastava, web usage logs, which is an important part of web data, contain abundant information and can be exploited in many Web personalization applications [24]. They provide demographic data (for instance name, age, country, marital status, education, interests, etc.) of each website user, also implicit knowledge about users' behavior patterns and other preferences. To discover the information explicitly from a novel perspective, a web page navigation logs mining based method is proposed to extract users' preferences dynamically.

The paper directs toward excavating users' browsing histories in typical e-commerce systems and tries to establish a hierarchical presentation model for the data, which will have wide utilization potentiality in e-commerce system intelligence, monitoring users' preferences and analyzing on different levels. Based on the proposed hierarchical presentation model and interested web page detection algorithm, a user's shopping preference measure algorithm UPSAWBH is put forward.

2 Related work

2.1 Web page navigation logs mining

Web page navigation logs generally refers to the detailed information that can be gathered from the Internet browsing of users, which are lists of links network clients clicked and the elapsed time between them [22,30]. They can be represented as a quintuple *WebLog={time, remote host, method, page, request status}.*

The format of the web browsing log may vary slightly in different application servers, but the elements listed in the quintuple are essential. The meaning of each element in the quintuple is explained as follows [32]:

"*time*" denotes the time the server responds to the user's request and returns the requested resources. In the article, time between two URLs is the time interval between the request of a URL and the followed URL page, which contains of web page load time and page time as shown in Fig.1 [1].

"*remote host*" denotes the logic name or IP address of the Network server that a user visits. A proxy server may exists between the user and the Web server, so "*remote host*" may represent the final proxy server that the user has visited.

"*method*" denotes the request method, which include GET, POST, HEADER, OPTIONS, PUT, and so on of the user. Among these methods, GET and POST are the most usually adopted.

"*page*" denotes the requested web page. The "page" can be functionally divided into two types: navigation pages and content pages. Navigation pages act as "guiding people" in the Internet, while the content page is the place where people usually spend most of their time.

"*request status*" denotes the status code that request the user to return to the server. The status code consists of three digits, which represent the status response of the server to the browser's request.

Web page navigation log data contains a lot of valuable information, such as the records of links that a user has visited and the elapsed time between them, the number of clicks on each web page, the complete visit path and the time spent on the web pages. This kind of statistical information can be explored by a variety of methods and used for many scenarios, such as users' preference prediction and the prefetching of pages to improve users' browsing experience.

A great many scholars have paid the much attention on website access pattern investigation of users by their browsing logs with the help of statistical analysis methods to reduce server-side response time and improve access efficiency of web pages [27]. Magdalini Eirinaki and Michalis Vazirgiannisrely on the application of statistical analysis and intelligent data mining methods (for instance, clustering, association rule mining, sequential pattern discovery and classification) to the Web log data, resulting in a set of valuable patterns that imply individuals' access patterns, and the knowledge is then employed to personalize pages for users according to their navigational behavior and profile [9]. Based on the theory of probability, Borges and Levene put forward a data mining method that captures users' web page access patterns: individuals' navigation sessions are treated as hypertext probabilistic grammar whose higher probability strings correspond to the interested tails of an individual and the last N visited web pages affect the affect the probability of the following page to be navigated [4]. Ezeife and Lu proposed a Web access pattern tree (WAP-tree) approach to explore frequent visit sequences for users, which can response dynamically without numerous re-constructions of WAP-tree during knowledge mining [10]. To overcome the weakness of ineffective content management of websites and the incapability of providing personalized web page services for the users in traditional web usage mining approaches, Yao-Te Wanga and Anthony J. T. Lee introduced the concept of throughout-surfing patterns and present an advanced access pattern mining model [27]. Also, they put forward a compact graph

model, termed a web page navigation path traversal graph, to store knowledge about the web page access paths of the website users.

2.2 User similarity measurement

In website access scenario, a user profile contains static part, which changes seldom (such as demographic information) and dynamic part that changes frequently. The ability to find users with similar preference or distinguish between different individuals is a matter of cardinal significance in various information system applications, especially in e-commerce systems. The process of finding similar users is usually conducted based on users' profile mining, which focus on knowledge about web page access preferences and characteristics of the users. Due to the convenience of collecting users' web page navigation and other potential valuable information from server-side, users' profile exploration has attracted much attention of scholars all over the world in recent years. Personalized recommendation in e-commerce systems is deemed as one of the most popular applications that based on users' profile and preference extraction.

User similarity measurement is one of the research aspects of users' preference mining. Although, much effort has been paid on this topic, it is still a problem-rich area. The existing researches focus mainly on how to measure user similarity in different circumstance. Some investigations follow the perspective of geography to probe users' similarity. For example, Li, Zheng, Xie, et al. found it important to discover valuable knowledge from large scale spatio-temporal data, and proposed hierarchical-graph-based similarity measurement (HGSM) framework to model an individual's trajectories. The model considered both the sequence property of people's movement behaviors and hierarchy attribute of geographic feature, which proved to be an effective way of measuring similarities among users [17]. Guy, Jacovi, Perer, et al. studied nine kind of sources (friending, communities, blogs, forums, et al.) that can be used for users' similarity measurement in social media applications [13]. Their research shows that the aggregation of sources may be valuable to measure the similarity between people. All these approaches are based on the hypothesis that the more two users share the same geographical overlap areas, the more likely they can be similar to each other. However, it may be challenge to evaluate the similarity of two users who are living very close. Take two persons, an old man and a young one, living in the same community as an example: they may share the same geographical overlap area (both stay at home) in a period of leisure time (week end or holiday), but it is hard to say that they are similar. While still other works pay much attention to semantic analysis to exploit users' similarity. Ying, Lu, Lee, et al. argued that geographically close uses' trajectories may not have to be similar, because the activities implied by nearby landmarks that they passed through may vary and so they put forward a MSTPS (Maximal Semantic Trajectory Pattern Similarity) method, which measures the similarity between users based on the calculation of semantic similarity of their trajectories [31]. Lee and Chung proposed a method to calculate user similarity according to the semantics of frequently visited locations and the user's potential preference [15]. Still others studies focus on mining users' purchase history or web page access record to measure their similarity. For example, Eckhardt A. put forward a collaborative filtering based user preference model to explore users' similarity [8]; Wei, Shijun, Yunlu, et al. held the opinion that users sharing similar purchase in history are very likely to have similar preference in the future, and constructed a user similarity network to get rid of the negative affect of popular objects or items for personalized recommendation [12].

In this work we focus on web usage logs mining to find users' preferences, based on the opinion that web page navigation patterns resemble users may have similar interests. We combine the approach of hierarchical-graph-based similarity measurement with Web usage mining techniques on web page access records.

2.3 E-commerce recommendation

With the boom in e-commerce in recent years, the structure of the e-commerce system has become much more complicated than ever before as it provides a great many customized services for both customers and enterprises. Meanwhile, a wide variety of goods is provided by sellers in e-commerce virtual shops, which makes it impossible for a user to view all the products when he/she has something to buy. Under this circumstance, the demand for understanding users' preferences in e-commerce systems and find useful knowledge to make recommendations has greatly increased. By providing precise and useful suggestions to a potential consumer, recommendations in e-commerce systems have made good profits for many popular e-commerce enterprise, such as Tabao.com and Amazon.com [18], who recommend new products related to the items purchased previously by similar users of the target user.

Since the collaborative filtering approach first proposed in the mid-1990s, scholars all over the world have devoted their effort on recommender system, and make it a high-profile research area [33]. The interaction between a recommendation system and the user can be divided into explicit approach, using users provided registration information to get a general picture of their static profiles, as well as implicit methods, by exploring the web page access records to infer users' preferences [14]. The former make recommendations based on the users' input conditions, while the latter would automatically collect or observe users' behavior to detect their profile. In e-commerce systems, recommendation seems extremely import because it aim sat customizing/personalizing a given product according to interests of the consumers and help them make decisions.

Although, recommendation methods (such as Collaborative filtering, Content-based filtering and rule-based filtering) may different from one another, the process of e-commerce recommendation generally involves three main steps [3, 5–7]:

(1) the users' buying records are collected, processed and analyzed to find their preferences;

(2) based on the conclusion of the above first step, commence recommendation for a target user and

(3) provide a recommended goods list for the target users to buy. In addition to mining customers' properties and filtering unnecessary information, an e-commerce recommendation system focuses, as far as possible, on the matter of the ability to suggest items of interest to the user.

Over the last decade many new algorithms and methods have been put forward to improve recommendation accuracy and efficiency in both practical application and theoretical research. However, it still faces many challenges, such as cold start data sparse. At present, recommender system related research is still a popular issue because it constitutes problem-rich research areas concerning not only about finding accurate recommendation algorithms, but also a great many crucial factors, such as diversity, recommender persistence, robustness, serendipity, privacy etc.

In many e-commerce systems, web page browsing pattern mining plays a very important role in generating accurate recommendations. When a user accesses to a web page or a browser-server based e-commerce system, the URLs of the pages visited will be stored in the server access log. It plays a crucial role in conveying knowledge of customers' activities and preferences, which are very useful for personalized recommendation. Furthermore, it dynamically reflects their interests in a sense and the similar click sequence of two different users may imply that they have similar preferences at that time, which is very import for real-time recommendations. Preference similarity is one of the most useful pieces of knowledge that can be extracted by many kind of data mining models [13]. This knowledge will help in finding groups of visitors with similar preferences and making effective recommendations. In recent years, many scholars have turned

their attention to web browsing history mining based e-commerce recommendation, for example, Qinbao Song and Martin Shepperd put forward a vector analysis and fuzzy set theory based model to explore similar users, frequently visited web pages and navigation paths and designed a web browsing mining based recommendation model for e-commerce systems [23].

3 Users' preference similarities exploration

In this section, we conduct a detailed introduction to the processes involved in user preferences extraction, which includes web browsing trajectory definition, user browsing history hierarchical presentation and users' preference similarity measure algorithm.

3.1 Preliminary

Definition 1. Web Browsing Trajectory. A web browsing trajectory ($WBTraj$) is a clicked URL sequence of pages viewed by a user across the entire web visit process: Each $WBTraj$ contains a page URL ($l_i.URL$) request time ($l_i.RTime$) and leave page time ($l_i.LTime$). Thus, a web browsing trajectory can be represented as $WBTraj = l_1 \longrightarrow l_2 \longrightarrow ...l_{n-1} \longrightarrow l_n$, where $l_i.RTime < l_i.LTime$ and $l_i.LTime < l_{i+1}.RTime$.

Definition 2. Interested Web Page. Generally, an interested web page ($IURL$) can be represented by a URL where a user stays on longer than a certain time interval. Therefore, the main factor involved in extraction of the interested web page depends on the time threshold (θ_t), which implies the time a user stays on a certain web page. Formally, a set of interested web pages can be defined as $IURL = \{l_i \in WBTraj,\ \ |l_i.\ LTime - l_i.RTime| >= \theta_t\ \}$.

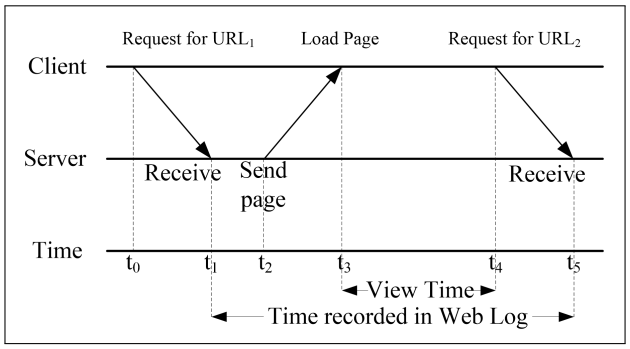

Figure 1: Time of request between two URLs

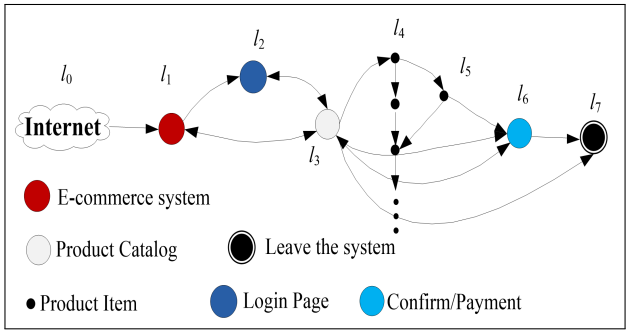

Figure 2: Web browsing trajectory of a general e-commerce system

As demonstrated in Figure 2, $l_1 \longrightarrow l_2 \longrightarrow l_3 \longrightarrow l_4 \longrightarrow l_5 \longrightarrow l_6$ formulates a general web browsing trajectory in a common e-commerce system, and the interested web page can be extracted according to the web page browsing time threshold.

Typically, interested web pages may occur in the following conditions: (1) a person enters an e-commerce system, opens a web page and then diverts his (or her) attention away, and (2) a user opens more than one product item web page to compare their characteristics and decide which one to buy, exceeds a time limit at a certain web page. It is impossible to grasp customers' interested produce web page navigation logs under the former circumstance. We focus on the second case, and put forward an Interested-Webpage-Detection algorithm. Detailed process of the algorithm can be described below.

Algorithm 1 Interested Webpage Detection

Require: Web browsing trajectory WBTraj and a time threshold θ_t
Ensure: A set of URLs (IURL=$\{L\}$) of the interested web pages
1: $i = 0, \quad iUrlCount = |WBTraj| \quad //The\ \ number\ of\ WebUrl\ in\ WBTraj$
2: **while** $i < iUrlCount$ **do**
3: **if** $|l_i.\ LTime - l_i.RTime| >= \theta_t$ **then** $IURL.insert(l_i)$
4: **end if**
5: $i++$
6: **end while**
7: **return** $IURL$

To capture interested web pages in a server-browser based e-commerce system as accurate as possible, and we need to find a suitable time threshold to detect every stay on a certain web page. A too small time threshold can lead to too many navigated web pages over-detected as IURLs. A small time threshold value, for example 1 second, might be more capable of identifying much IURLs for an e-commerce system; however, this could cause too many IURLs detected, making us get lost and don't know which is the real web page or produce that users interested. In addition, time interval between the server response and the requested web page shown in the user's screen may be greater than 1 second (depending on network situation). This is obviously not in accordance with people's intuitiveness, as the web page has not shown in that short time interval. Meanwhile, too large time threshold (θ_t) value is not appropriate either. It could result in many interested web pages, which indicate users' real preference, cannot be detected efficiently.

3.2 User browsing history hierarchical presentation

From the above section, we can ascertain that the more interested web pages two people share in an e-commerce system, the more likely it is that they may have the same preferences and the similar product purchase inclination. However, it is subjective to measure the similarity of two customers' preferences directly based on the web pages of interest that they have in common. Moreover, it doesn't make sense to judge users' preference similarities just by yes or no. Therefore, we aim to measure the degree of similarity of two users' interest quantitatively, and then rank a group of people according to the preference similarities among them. To solve the key point of the issue, we put forward a hierarchical graph to present users' browsing histories in an e-commerce system, as shown in Figure 3. Three procedures need to be preformed before building such a graph for an e-commerce system browse path.

(1) Formulate a set of user click logs according to the time sequence in which he (or she) visits an e-commerce system and form the web browsing trajectory;

(2) Filter out the common web URLs that everyone will click in the e-commerce system, such

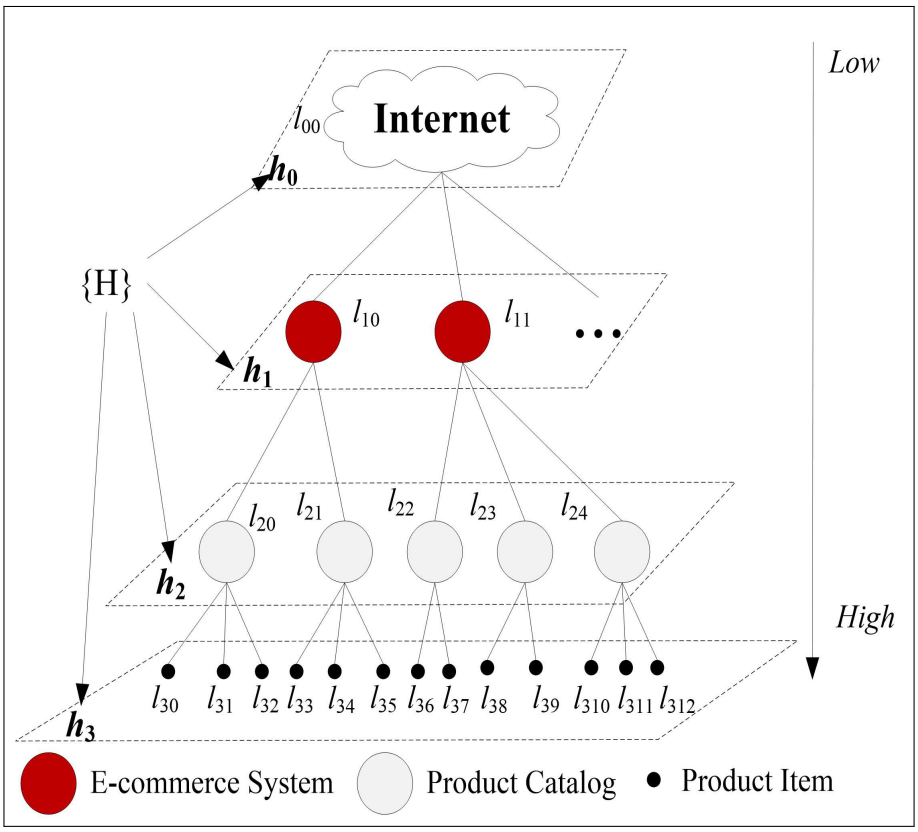

Figure 3: Hierarchical structure of the user browsing trajectory

as login, logout, product payment or other related URLs;

(3) Construct the hierarchical graph according to the web page properties of the clicked URL. We are not concerned with the web page clicked time sequence while creating the graph (in other words, a web page clicked more than one time can be repetitively treated as a multiple node in the graph). Also, a web page clicked many times can be represented as multiple nodes in the hierarchical graph.

Definition 3. Hierarchical Graph (HG). HG is a collection of clicked URLs in an e-commerce system, with a hierarchical structure $HG = \{H, L\}$, where $H = \{h_1, h_2, ..., h_{n-1}, h_n\}$ represents the collection of layers of the hierarchy graph. $L = \{l_{ij} | 0 \leq i \leq |H| \ , \ 0 \leq j \leq |L_i|\}$, where l_{ij} denotes the jth nodes on the layer h_i ($h_i \in H$), and L_i is the set of nodes on layer l_i

3.3 Users' preference similarity measurement

Concepts of similar web click sequences

Definition 4. Similar Web Click Sequence (SWCS). A similar web click sequence represents for two users (u_p and u_q) who have viewed the same sequence of URLs within the same period. Formally, a pair of web-click-sequences, $webclickseq_i^p$ and $webclickseq_i^q$, for two users, u_p and u_q,

$$webclickseq_i^p = \left[url_1^p(t_1^p) \xrightarrow{\Delta t_1^p} url_2^p(t_2^p) \xrightarrow{\Delta t_2^p} url_3^p(t_3^p) ... \xrightarrow{\Delta t_{n-1}^p} url_n^p(t_n^p) \right]$$

$$webclickseq_i^q = \left[url_1^q(t_1^q) \xrightarrow{\Delta t_1^q} url_2^q(t_2^q) \xrightarrow{\Delta t_2^q} url_3^q(t_3^q) ... \xrightarrow{\Delta t_{n-1}^q} url_n^q(t_n^q) \right]$$

where $url_j(1 \leq j \leq n)$ are the graph vertices that u_p and u_p share on the layer l_i, $t_j(1 \leq j \leq n)$ stands for the times the visitor successively stay on the web page url_j and Δt_j $(1 \leq j \leq n)$ represents the times interval that the user transfer from url_j to url_{j+1}. $webclickseq_i^p$ and $webclickseq_i^q$ are similar web click sequences only if they meet the following conditions:

(1) The two users, u_p and u_q, share the same vertex in one layer of their HGs. Formally, $\forall 1 \leq j \leq n,\ url_j^p = url_j^q$.

(2) The two users, u_p and u_q, have assemble transition times between the orderly accessed web pages. Formally, $\forall 1 \leq j \leq n,\ \ 0 \leq \frac{|\Delta t_j^p - \Delta t_j^q|}{\max(\Delta t_j^p, \Delta t_j^q)} \leq T_{threshold}$, where $T_{threshold}$ is a predefined time threshold.

If the above two conditions hold, a similar web click sequence, $simwebclickseq_i^{q,p}$, included in $webclickseq_i^q$ and $webclickseq_i^q$ can be extracted as:
$simwebclickseq_i^{p,q} = \langle url_1^{p,q}(\min(\Delta t_1^p, \Delta t_1^q) \rightarrow url_2^{p,q}(\min(\Delta t_2^p, \Delta t_2^q) \rightarrow ... \rightarrow url_n^{p,q}(\min(\Delta t_n^p, \Delta t_n^q))\rangle$, where $\min(\Delta t_1^p, \Delta t_1^q)$ stands for the minimum of the time intervals Δt_1^p and Δt_1^q.

Definition 5. n-Length Similar Web Click Sequence. If there is n nodes in the similar web click sequence $simwebclickseq_i^{q,p}$ for two users, u_p and u_q, we call the sequence *n-length* similar web click sequence.

Similar web click sequences extracting

It can be found in Figure 3 that the bottom nodes in the hierarchy graph reveal more specific preferences than those at the top. Also, from the top to the bottom, the customers' shopping intentions increase gradually. Therefore, the hierarchical characteristic of this graph is efficient in depicting individuals' preference similarity. Customers who share the same web browsing trajectory on a lower layer are more likely to have similar shopping preferences than those who have web browsing trajectories in common on a higher layer.

According to the user browsing history hierarchical presentation model, we propose the following users' preference similarity (UPS) measure approach and define preference similarity between user p and q as:

$$UPS(p,q) = \frac{SumScore_{p,q}}{|HG_p| + |HG_q|} \qquad (1)$$

where HG_p and HG_q denote the hierarchical graphs of the user p and q,$SumScore_{q,q}$ denotes the sum of the score for the same node in each layer, $|HG_p|$ is the number of nodes in HG_p, $|HG_q|$ is the number of nodes in HG_q. The sum score ($SumScore_{q,q}$) value of two users (p and q) need to be calculated according to the URLs they share in each layer. The detailed process of acquiring $SumScore_{q,q}$ can be described in the following algorithm.

Obviously, the weighted factor α, in the algorithm, can impact the result of users' preference similarity to a certain degree. We set $f(i) = \frac{1}{|H|-i+1}$($|H|$ is the total layers of the hierarchical graph) to normalize the preference similarity value to [0,1]. Additionally, the greater $UPS(p,q)$ value means more preference similarities between users p and q. The process of calculating the total length of the shared trajectory for users p and q in each layer is illustrated in the following algorithm (**GetSimilarClickSeqLength**), and the demonstration of similar web page click sequence matching is presented in Figure 4.

4 Experiment evaluation

To examine the effectiveness of the proposed web page navigation logs mining based user preference similarity measure approach, we conduct the experiments on dataset collected from

Algorithm 2 UPSAWBH

Require: Hierarchically structured web browsing trajectory HG_p, HG_q of users p and q
Ensure: $UPS(p,q)$ which shows the preference similarities of users p and q

$SumScore_{q,q} = 0$, $UPS(p,q) = 0$
while $h_i \in H$ **do**
  $score_h = 0$// preference similarity on a layer
  $\alpha = f(i)$ //α is an i-dependent factor
  while $l_i \in L$ **do**
    $len = GetSimilarSeqLength(HG_p^i, HG_q^i)$// get similar web page length
    $score_h = 2len * \alpha$//get the total length of a similar web page
  end while
  $SumScore_{q,q} = SumScore_{q,q} + score_h$
end while
$UPS(p,q) = \frac{SumScore_{p,q}}{|HG_p|+|HG_q|}$ // calculate UPS according to formula (1)
return $UPS(p,q)$

Algorithm 3 GetSimilarClickSeqLength

Require: Web browsing list set L_p, L_q
Ensure: the total length ($TotalLength$) of the similar web page click sequence that user p and q share in L_p and L_q

Sort L_p, L_q according to the web click time order, and form the web browsing trajectories $Traj_p$ and $Traj_q$
$indictor = 0$, $TotalLength = 0$//variable initialization
while url_p **in** $Traj_p$ **do**
  while $indictor < |Traj_q|$ **do**
    if $url_p == Traj_q[indictor]$ **then** $TotalLength++$
    end if
    $indictor++$
  end while
end while
return $TotalLength$

the server of our self-developed e-commerce system. The dataset covers all the system click logs during a whole month. Table 1 depicts the profile of the dataset of the dataset that we used in the experiment, and Figure 5 details the click trajectories we extracted. It suggests that most of the system users has trajectory length of 1, which means many users just access a web page of the system and exit within a very short time, because the search engine www.Baidu.com is chosen to promote our system, and many link-clicks from the search engine. Meanwhile, the numbers of users decrease sharply with the increase of the click trajectory length in our system.

All the proposed algorithms and models are implemented and run in a computer with Intel (R) Core™ i3-2310 CPU @2.10 GHz (4 CPUs), 6GB RAM, Windows 7 Ultimate 64-bit using visual Programming Language C# (in Microsoft Visual Studio 2010 Professional). To differentiate the significance of similar web page sequences with various length values on different layers, we set $\alpha = \frac{1}{|H|-i+1}$. Here α increases in accordance with the layer i in the hierarchical graph, since we intuitively observe that the likelihood of two individuals' preference similarity rises sharply. There are thousands of combinations of time threshold (θ_t) with factor (α) in algorithm UPSAWBH. Honestly, it is a great challenge to determine what time interval is proper to detect interested

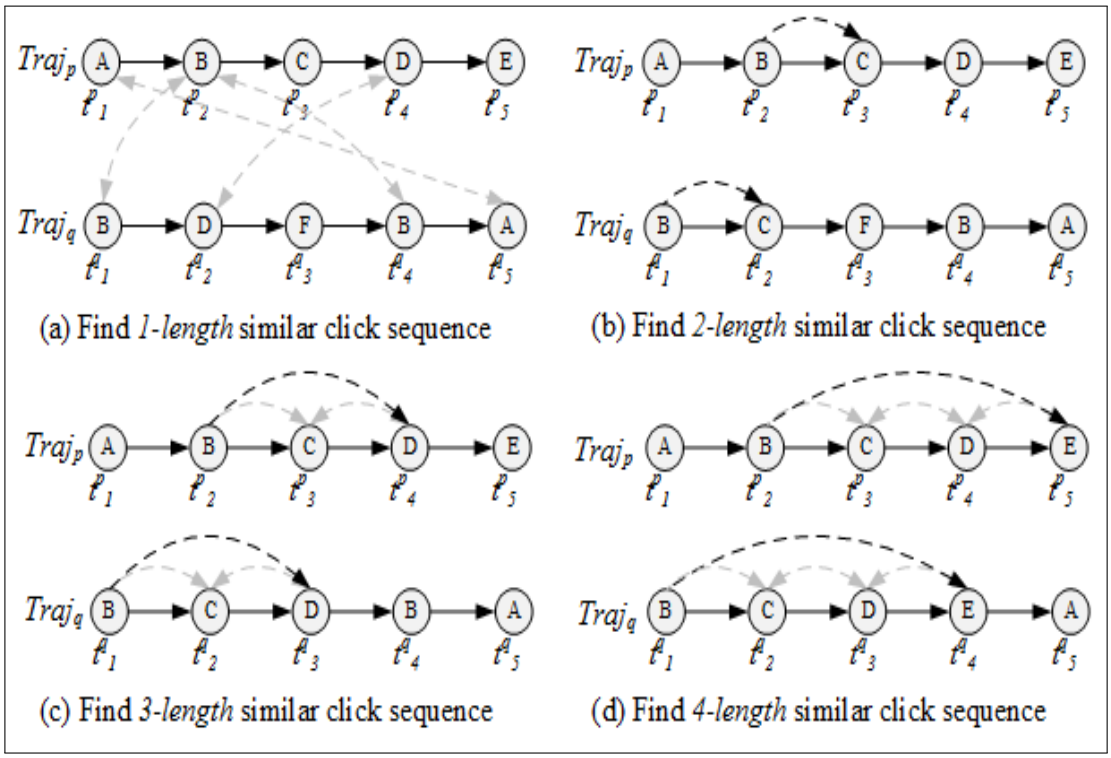

Figure 4: Schematic of similar web page click sequence matching

Table 1: Statistical information of the web browsing log data

Item	Count
Total request(Number of entries)	1685
Aver request/Hour	86.46
Average bytes transferred(MB)	4198.56
Average bytes transferred per hour	164.61
No of request after pre-processing	5417
Number of users	305
Number of user sessions	317

web pages of an individual. Therefore, we set the values of θ_t with reference to commonsense knowledge in real-world web system design. In our experiment, θ_t is set to 5 seconds, 10 seconds, 15 seconds, 20 seconds, 25 seconds, 30 seconds, 35 seconds, 40 seconds and 45 seconds. The result, as shown in Figure 6, reveals that that the number of interested web pages decrease rapidly as the time threshold θ_t increases. Take $\theta_t = 5$ as an example, nearly 290 interested web pages are detected, but the value drops to less than 80 when θ_t is set to 10 seconds.

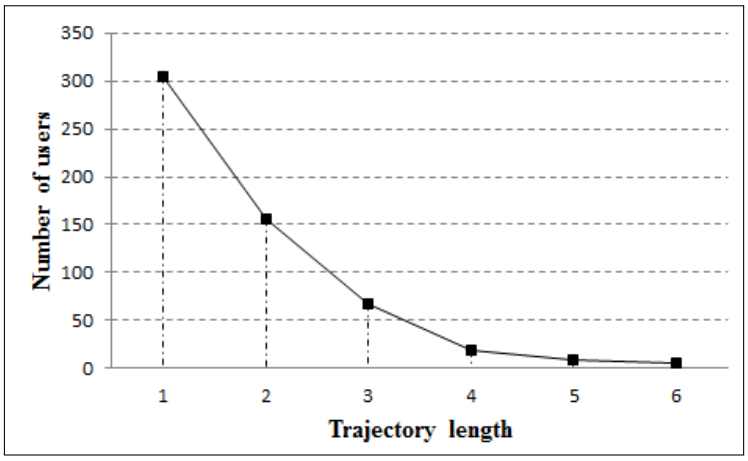

Figure 5: Relationship between trajectory length and number of users

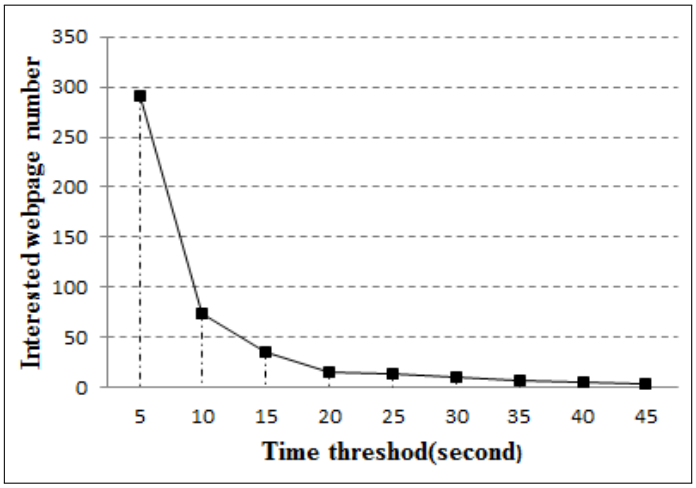

Figure 6: Number of interested web page changing over time threshold

We mainly focus on verifying the feasibility and effectiveness of hierarchical graph, interested-webpage-detection algorithm, web click sequence, in measuring users' similarity in e-commerce systems. To further explore the accuracy of proposed UPSAWBH algorithm, six candidate total layer values in the hierarchical graph are tested in the experiment. As we can see in figure 7, with the increase in the total layer in the hierarchical structure of user the browsing trajectory, the accuracy of UPSAWBH nearly increases linearly, and the calculated preference similarity value (UPS) rises simultaneously. Intuitively, the more different layer levels in hierarchical graph considered the more similar web click sequences can be checked, that is, both preference similarity and average accuracy are better while 6 total layers are considered. All these prove that our approach has good effectiveness in mining users' preference similarity knowledge using webpage navigation logs of users in e-commerce systems.

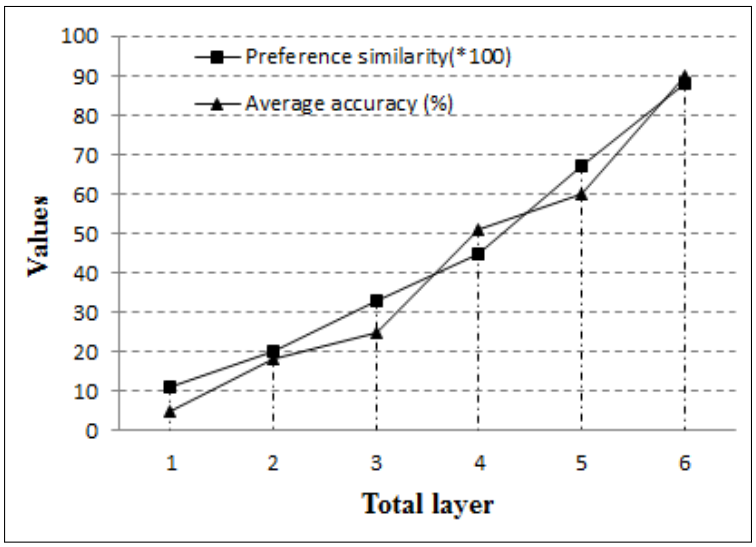

Figure 7: The influence of total layer parameter on the accuracy of the UPSAWBH algorithm

5 Conclusion and future work

Users' preference similarities mining is an important research area in many information systems, especially in e-commerce application. To some extent alleviate challenges, such as cold start and data sparsity, in e-commerce environment, the article presents a method to explore users 'preference similarities based on web browsing history data mining. In this research, we introduce a web page-navigation-log data extraction based users' similarity calculation approach which (1) starts from a perspective of geography, treating individuals' visits of web pages in e-commerce systems as trajectories, (2) devotes to explore the users' preference based on web page click sequence discovering and (3) tries to check users' preferences similarity in terms of their web click pattern resemblance in distinct levels. In order to construct a complete a research framework of our approach, a hierarchical graph model is proposed to present the structure of web navigation history and an interested web page detection algorithm is put forward to explore users' preference patterns. Meanwhile, an algorithm, UPSAWBH, is put forward to figure out the preference similarities among users, and the experimental results on dataset from real-world e-commerce system sever prove good effectiveness of our approach. Two traits, the sequence peculiarity of user click and hierarchy feature of webpage browsing levels, have been considered in this similarity measure.

One limitation of our method is that it has been tested using only dataset from our self-developed e-commerce system and we just set the value of weighted factor (α), by intuition. In the future, we will pay more attention to collect web page navigation dataset from other B/S based e-commerce systems to compare the result and improve our approach. We plan to focus on the following two areas: (1) further explore the impact of weighted control factor (α) on the result of UPSAWBH quantitatively, and (2) investigate how to improve the veracity of the proposed user preference similarity computation method to meet the requirement of personalization recommendation. We have just put forward a rough research paradigm to mine users' preference similarities using web page navigation logs in e-commerce system from another perspective, and still have a long way to go.

Acknowledgments

This work is partly supported by the Ministry of Education, Humanities and Social Sciences Research Project (Grant No. 14YJC630210), the Zhejiang Public Technology Research and Application Project (Grant No. 2015C33065), the Zhijiang Youth Action Project: study on mobile e-commerce recommendation (Grant No. G306), the Ningbo Education Science Planning Key Project (Grant No. 2017YZD010), the Zhejiang Business Economics Association Project (Grant No. 2016SJYB01), the China National Natural Science Foundation Project (Grant Nos. 61202376, 71071145 and 41201550), the Natural Science Foundation of Zhejiang (Grant No. LY16G020012), the Major Research Projects of Humanities and Social Sciences in Colleges and Universities of Zhejiang (Grant No. 2014GH015), the Ningbo Huimin Project of Science and Technology (Grant No. 2016C51040), the China Press and Publication Administration Key Laboratory Project and Shanghai Key Lab of Modern Optical System, the Modern Port Service Industry and Culture Research Center of the Key Research Base of Philosophy and Social Sciences of Zhejiang Province and the China Scholarship Council.

Bibliography

[1] Abraham S., Lai P.S. (2011); Spatio-temporal Similarity of Web User Session Trajectories and Applications in Dark Web Research, *Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics*, Beijing, China, 2011. doi:10.1007/978-3-642-22039-5_1.

[2] Adomavicius G., Tuzhilin A. (2005); Towards the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, *IEEE Transaction on Knowledge and Data Engineering (TKDE)*, 17(6), 734-749, 2005. doi:10.1109/TKDE.2005.99.

[3] Becchetti L. et al. (2014); A lightweight privacy preserving SMS - based recommendation system for mobile users, *Knowledge and Information Systems*, 40(1), 49-77, 2014.

[4] Borges J., Levene M. (2000); *Web usage analysis and user profiling, Chapter: Data mining of user navigation patterns (92-112)*, San Diego, CA, USA; Springer Berlin Heidelberg, 2000.

[5] Chen D.-N. et al. (2010); A Web-based personalized recommendation system for mobile phone selection: Design, implementation, and evaluation, *Expert Systems with Applications*, 37(12), 8201-8210, 2010. doi:10.1016/j.eswa.2010.05.066.

[6] Cheng A.-J. et al. (2011); Personalized travel recommendation by mining people attributes from community-contributed photos, *Proceedings of the 19th ACM international conference on Multimedia*, Scottsdale, Arizona, USA; ACM, 83-92, 2011.

[7] Dao T.H., Jeong S.R., Ahn H. (2012); A novel recommendation model of location-based advertising: Context-Aware Collaborative Filtering using GA approach, *Expert Systems with Applications*, 39(3), 3731-3739, 2012. doi:10.1016/j.eswa.2011.09.070.

[8] Eckhardt A. (2012); Similarity of users' (content-based) preference models for Collaborative filtering in few ratings scenario, *Expert Systems with Applications*, 39(14), 11511-11516, 2012. doi:10.1016/j.eswa.2012.01.177.

[9] Eirinaki B.M., M Vazirgiannis M. (2003); Web mining for web personalization, *ACM Transactions on Internet Technology*, 3(1), 1-27, 2003.

[10] Ezeife C.I., Lu Y. (2005); Mining web log sequential patterns with position coded pre-order linked WAP-tree, *Data Mining and Knowledge Discovery*, 10(1), 5-38, 2005. doi:10.1007/s10618-005-0248-3.

[11] He S., Fang M. (2008); Personalized recommendation based on ontology inference in E-commerce, *Proceedings of the International Conference on Management of e-Commerce and e-Government*, 192-195, 2008. doi:10.1109/icmecg.2008.24.

[12] Gan M., Jiang R. (2013); Constructing a user similarity network to remove adverse influence of popular objects for personalized recommendation, *Expert Systems with Applications*, 40(10), 4044-4053, 2013. doi:10.1016/j.eswa.2013.01.004.

[13] Guy I. et al. (2010); Same places, same things, same people?: mining user similarity on social media, *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, Savannah, Georgia, USA, 41-50, 2010. doi:10.1145/1718918.1718928.

[14] Leavitt N. (2006); Recommendation technology: Will it boost E-Commerce, *Computer*, 39(5), 13-16, 2006. doi:10.1109/MC.2006.176.

[15] Lee M.-J., Chung C.-W. (2011); A user similarity calculation based on the location for social network services, *Proceedings of the Database Systems for Advanced Applications*, Hong Kong, China, Springer, 38-52, 2011. doi:10.1007/978-3-642-20149-3_5.

[16] Li P. et al. (2007); Preference update for e-commerce applications: Model, language, and processing, *Electronic Commerce Research*, 7(1), 17-44, 2007. doi:10.1007/s10660-006-0061-0.

[17] Li Q. et al. (2008); Mining user similarity based on location history, *Proceedings of the ACM SIGSPATIAL GIS*, Irvine, CA, USA, 2008. doi:10.1145/1463434.1463477.

[18] Linden G., Smith B., York J.(2003); Amazon.com recommendations: item-to-item collaborative filtering, *IEEE Internet Computing*, 7(1), 76-80, 2003. doi:10.1109/MIC.2003.1167344.

[19] Papazoglou M.P. (2001); Agent-oriented technology in support of e-business - Enabling the development of "intelligent" business agents for adaptive, reusable software, *Communications of the ACM*, 44(4), 71-77, 2001. doi:10.1145/367211.367268.

[20] Park S.T., Pennock D., Madani O. (2006); Collaborative filtering for robust cold-start recommendations, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, 2006.

[21] Sarwar B.M. (2001); *Sparsity, scalability and distribution in recommender systems*, Minneapolis, USA, University of Minnesota, 2001.

[22] Shahabi C. et al. (1997); Knowledge discovery from users web-page navigation, *Proceedings of the Seventh International Workshop on Research Issues in Data Engineering*, Birmingham, England, 1997. doi:10.1109/RIDE.1997.583692.

[23] Song Q., Shepperd M. (2005); Mining web browsing patterns for E-commerce, *Computers in Industry*, 57(7), 622-630, 2006. doi:10.1016/j.compind.2005.11.006.

[24] Srivastava J. et al. (2000); Web usage mining: Discovery and applications of usage patterns from web data, *ACM SIGKDD Explorations Newsletter*, 1(2), 12-23, 2000.

[25] Turban E. et al. (2009); *Electronic Commerce*, NJ, USA: Prentice Hall Press, 2009.

[26] Waga K., Tabarcea A., Franti P. (2011); Context aware recommendation of location-based data, *Proceedings of the 15th International Conference on System Theory, Control, and Computing (ICSTCC)*, Sinaia, Romania, 1-6, 2011.

[27] Wang Y.-T., Lee A.J.T. (2011); Mining web navigation patterns with a path traversal graph, *Expert Systems with Applications*, 38(6), 7112-7122, 2011. doi:10.1016/j.eswa.2010.12.058.

[28] Woerndl W., Brocco M., Eigner R. (2009); Context-aware recommender systems in mobile scenarios, *International Journal of Information Technology and Web Engineering*, 4(1), 67-85, 2009. doi:10.4018/jitwe.2009010105.

[29] Wu C., Hou F. (2011); Design and Optimization of Redundant ControlNet Networking Control System, *Process Automation Instrumentation*, 32(3), 50-56, 2011.

[30] Wu C., Yu Z. (2005); Data transmission with data package dropout and control method on NCS, *Control & Automation*, 10, 39-41, 2005.

[31] Ying J. J.-C. et al. (2010); Mining user similarity from semantic trajectories, *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, San Jose, California, 19-26, 2010. doi:10.1145/1867699.1867703.

[32] YU X.B., GUO S.S., HUANG X.R. (2010); Intelligent e-commerce based on Web usage mining and its application (in Chinese), *Computer Integrated Manufacturing Systems*, 16(2), 439-448, 2010.

[33] Zhong H. et al. (2014); Study on Directed Trust Graph Based Recommendation for E-commerce System, *International Journal of Computers Communications & Control*, 9(4): 510-523, 2014.