Electronic Commerce Research and Applications xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

journal homepage: www.elsevier.com/locate/ecra

A method for discovering clusters of e-commerce interest patterns using click-stream data

Qiang Su, Lu Chen [*]

School of Economics & Management, Tongji University, Shanghai 200092, China

ARTICLE INFO

Article history:
Received 3 January 2014
Received in revised form 7 October 2014
Accepted 14 October 2014
Available online xxxx

Keywords:
Click-stream data
User interest
Behavior analysis
Leader clustering algorithm
Rough set theory

ABSTRACT

Having a good understanding of users' interests has become increasingly important for online retailers hoping to create a personalized service for a target market. Generally speaking, user's browsing behaviors (when looking at websites) represent a comprehensive reflection of their interests. Users with various interests will visit multiple categories and research various items. Their browsing paths, the frequency of page visits and the time spent on each category all vary widely. Based on these considerations, a novel approach to discovering consumers' interests is proposed and is systematically studied in this paper. The browsing behavior of a number of consumers – including their visiting sequence, frequency and time spent on each category – are mined via the click-stream data recorded on an e-commerce website. Given this behavioral data, we construct an improved leader clustering algorithm and leverage it with a rough set theory in order to generate users' interest patterns. Furthermore, a case study is conducted based on nearly three million click-stream data, which was collected from one of the largest Chinese e-commerce websites. Using this data, the parameters of the algorithm are tested and optimized to make the algorithm more effective in terms of large data analysis and to make it more suitable for discovering users' multiple interests. Using this algorithm, three typical user interest patterns are derived based on a real click-stream dataset. More importantly, further calculations based on different click-stream datasets verify that these three interest patterns are consistent and stable. This study demonstrates that the proposed algorithm and the derived interest patterns can provide significant assistances on webpage optimization and personalized recommendation.

1. Introduction

To attract more customers, e-commerce companies are continuously diversifying their products and increasing their category range. Large e-commerce organizations frequently see more than a million customers per day log on to their websites. Those potential customers view hundreds of thousands of catalog items every day. As a result, a specific challenge arises for these e-commerce companies; namely how to discover the website users' interests and promote sales by effectively managing an ever-increasing number of categories and products.

Most of the existing techniques used to measure consumer interest mainly rely on customer ratings. Whether or not a user rates an item indicates, at least to some extent whether they are interested in it. The rating values themselves represent how much the users like the target items (Zhao et al. 2013, Cleger-Tamayo et al. 2012). However, ratings information is too limited to describe users' website navigation processes. Besides, a product rating is a final comprehensive evaluation which incorporates users' perceptions of price, service and logistics. The rating is provided by and relates more closely to the e-business company than the products themselves. In addition, ratings from new customers are insufficient for reference purposes, while experienced customers may not be willing to give ratings every time they use a website. These factors make it more difficult to discover the users' true interests based on ratings alone.

Some scholars studied the topic of users' interests in social network media (Zeng et al. 2008, Li et al. 2012). They found that users' interests are frequently reflected by the posts they visit and those posts to which they reply. This idea can be similarly applied to an e-commerce website. Users will look at the items that interest them and attract their attention (Xing et al. 2007, He et al. 2012). Users with a variety of interests will visit different categories and multiple items. For different types of users, their browsing paths, the frequency with which they visit web pages and the time spent on each category will all vary. Compared with user ratings, this more detailed information can be used to describe users' interests far more precisely.

[*] Corresponding author.
E-mail address: lucia1119@gmail.com (L. Chen).

http://dx.doi.org/10.1016/j.elerap.2014.10.002
1567-4223/© 2014 Elsevier B.V. All rights reserved.

Thanks to the development of information technology, the internet allows for the real-time, low cost and unobtrusive collection of detailed information regarding individuals' activities. The record of an internet user's actions online has come to be known as click-stream data (Bucklin and Sismeiro 2009). Click-stream data captures a wide variety of information in a complete, timely and accurate manner. This data covers user activities such as browsing paths, purchased products and clicked banner ads. Click-stream data is becoming one of the most useful resources for researchers and practitioners attempting to understand individuals' behaviors in terms of choice. Up to now, many researchers have explored the click-stream data from websites that sell a single type of product, such as automotive products (Sismeiro and Bucklin 2004), books (Moe and Fader 2004), digital music (Aguiar and Martens 2013), wine (Van den Poel and Buckinx 2005) and nutrition products (Moe 2003). Unlike the click-stream data taken from these single-category product websites, the click-stream data mined from a comprehensive e-business website will be far more complex. The e-commerce website data will usually encapsulate considerably more details of an individual's behavioral history. This excessive detail makes the dataset itself large and cumbersome, and consequently leads to difficulties in data mining.

In this paper, with the aim of finding users' interests based on their click-stream data, a novel approach to discovering user interests is developed and studied systematically. Meanwhile, large amounts of real click-stream data are collected and utilized to validate the effectiveness of the newly-devised approach. The remainder of the paper is structured as follows: In Section 2, the related literature regarding click-stream data mining and clustering algorithms is thoroughly reviewed. Section 3 describes the website topology structure of various product catogeries and defines the indicators for measuring user interest. In Section 4, a rough leader clustering algorithm is developed and analyzed in detail. In Section 5, a case study is conducted to test the effectiveness of the method. In addition, the pre-processing of the click-stream data is elaborated upon, and the parameters of the algorithm are optimized. In Section 6, the performance of the proposed algorithm is discussed, based on the comparative study of a number of other algorithms. The stability of the interest patterns are also verified through different click-stream datasets. Moreover, some managerial suggestions are proposed. Finally, Section 7 presents the paper's conclusions and makes suggestions for the direction of future research.

2. Related works

Along with the development of largescale data analytics, the e-business sector has witnessed a boom in the application of web data mining aimed at researching customers' preferences and interests. Several studies have applied user purchasing patterns, web page visit numbers and web browsing paths to construct a model designed to predict customer preferences (Chiang et al. 2013). To measure any user's interest, several characteristics of that user's behavior are examined, e.g., product ratings (Zhao et al. 2013, Cleger-Tamayo et al. 2012), purchasing records (Li et al. 2005, Park and Chang 2009), page discussed sequence (Hong and Hu 2012, Li and Tan 2011), page detention time (Zheng et al. 2010, Kim and Yum 2011), and page browsing frequency (Rathipriya and Thangavel 2010, Liu et al. 2012). More specifically, Rathipriya and Thangavel (2010) proposed a fuzzy co-clustering algorithm to identify a subset of users with similar navigation behavior over a specific set of web pages. However, this work simply defined user interest by considering how many times a user visited each product's webpage, while the study neglected other important factors, such as the amount of time spent on each page. Conversely, Zheng et al. (2010) calculated user interest rates based on user browsing time, but without considering factors such as visit frequency and sequence. Kim and Yum 2011 proposed a more comprehensive evaluation method and described a user's interest as being based on that user's purchasing decisions and the time spent on each webpage.

In recent years, click-stream data mining has become more and more important in the area of web data analysis. Bucklin and Sismeiro (2009) defined click-stream data as the electronic record of a user's activity on the internet. The data is the natural by-product of a user accessing web pages, and click-stream data refers to the sequence of pages visited and the number of times these pages were viewed. Based on click-stream data mining, Bucklin and Sismeiro 2003 proposed a model to predict whether a visitor decides to continue browsing or to exit the site, as well as how long the visitor would spend browsing a web page. Moreover, they developed a task-completion approach to estimate the user's online shopping behavior (Sismeiro and Bucklin 2004). Meanwhile, many of the studies used click-stream data to explore users' behavioral characteristics, including users' browsing behavior (Moe and Fader 2004, Montgomery et al. 2004), users' responses to website design (Danaher et al. 2006, Lam et al. 2007), and how users move across different websites (Park and Fader 2004, Goldfarb 2006). Another main objective of click-stream data mining is to model people's online shopping behavior and to determine how to predict shoppers' online purchasing behavior (Moe 2006, Aguiar and Martens 2013). In addition, click-stream data makes it possible to track users' exposure to internet advertising, as well as their subsequent actions (Rutz and Bucklin 2012, Nottorf 2014).

The understanding of customer segmentation is critical for retailers who wish to build customer relationships, facilitate customer support and build a more effective interactive online shopping environment. Numerous studies of online customers have come to a similar conclusion that there is heterogeneity but consistency between customer groups. Many of the researchers divided online consumers by using their various navigation methods. Moe (2003) cluster analyzed store visits by considering a variety of factors, such as how much time the shopper spent on each page and how many brands they visited. Moe detected five different categories of online consumers' shopping strategies. Chen et al. (2009) divided customers into different groups by considering how frequently they made purchases and how much money they spent. In addition, consumers appear to be universally driven by two motivations to engage in online shopping. Some consumers may seek utilitarian benefits, while others prefer hedonic benefits (Bridges and Florsheim 2008, López and Ruiz 2011). Ganesh et al. (2010) also used shopping motivation measures and e-store attribute measures separately, in order to develop three unique online shopper subgroups. Wu and Chou (2011) developed a soft-clustering approach and used multi-category data to segment customers, including customers' satisfaction with service, the level of customers' internet usage, their shopping behavior and their demographics.

As one of the most critical techniques in mining web data, clustering has been widely applied to various purposes, such as webpage design (Carmona et al. 2012), web usage analysis (Zhai et al. 2011, Kou and Lou 2012) and user segmentation (Hussain et al. 2010, Wei et al. 2012). A K-means algorithm is one of the most popular clustering approaches, which is well known for its high efficiency. Nevertheless, Voges et al. (2002) stated that K-means is not suitable for web data clustering analysis, because K-means' prerequisites (that variables are normally distributed, and all groups have an equal variance–covariance matrix) are not fulfilled in most realistic web datasets. In addition, a K-means clustering algorithm is sensitive to the outliers, even though it is quite efficient in terms of computational time. Given these considerations, a K-medoids clustering algorithm was proposed (Park and

Jun 2009). K-medoids is based on the most centrally located object in a cluster; it is not as sensitive to abnormal data as is K-means (Xu and Wunsch 2005). However, due to its time complexity, the computation efficiency of a K-medoids clustering algorithm is low (Velmurugan and Santhanam 2010). Moreover, as stated by Jain (2010), predetermining the number of clusters K is one of the most difficult problems with K-means-type algorithms.

Under these circumstances, a centroid-based model entitled a "leader clustering algorithm" is put forth. The leader clustering algorithm can find a set of leaders as centroids of the clusters, through only one time of data set screening calculation (Yu and Luo 2011). In a leader clustering algorithm, the cluster numbers do not need to be given in advance, as they would in K-means-type algorithms. The partition of the data sets is automatically controlled by a user-specified threshold. The algorithm randomly chooses one of the objects as the starting leader. Then, for each object, the algorithm may either assign the object to the most similar cluster, or it may regard the object as a new leader if its similarity with the existing leaders does not satisfy the threshold requirement (Asharaf and Murty 2003). Compared to the K-medoids clustering algorithm, the low time expense of a leader clustering algorithm makes the latter an efficient method that has been widely applied in web data mining (Suresh Babu and Viswanath 2009, Yu and Luo 2011).

Generally speaking, conventional clustering techniques mandate that every object in a data set should be assigned to one and only one cluster. However, such a requirement is not suitable for social and behavioral web data mining applications, in which an object may possess characteristics similar to two or more clusters. As a result, the clusters should be able to overlap, at least to some extent. To cope with this concern, Lingras and West (2004), do Prado et al. (2002) and Voges et al. (2003) proposed rough cluster algorithms referring to a rough sets theory (Pawlak 2002). In a rough clustering algorithm, each cluster has two approximations, namely a lower and an upper approximation. The members of the lower approximation certainly belong to the cluster, and thus, they cannot belong to any other cluster. The members of the upper approximation, however, may belong to one or several other clusters. In addition, since their membership is uncertain, these members must belong to at least one other cluster (Peters 2006).

This paper aims at developing an effective algorithm to extract users' interests based on a huge amount of click-stream data. Therefore, an efficient clustering algorithm is needed, in order to derive a reasonable solution in a short period of time. Meanwhile, considering the multiple interests of users, the clustering algorithm should be able to assign a single user into several different clusters. Therefore, a novel algorithm is developed, based on the integration of a leader clustering algorithm and a rough set theory.

3. Catogery structure and user interest measurement

Before developing the algorithm, some concepts and measurement methods for evaluating user behavior and interest in an e-business website should firstly be defined. Therefore, the topologic structure of the product categories is described, and three indicators for measuring a user's interest are accordingly defined.

3.1. Catogery structure and assumptions relating to a user's interest

In order to attract more customers, e-business companies have been expanding to provide more and more products on their websites. Normally, an e-commerce website, such as Amazon.com, presents its product information using a hierarchical tree structure, as shown in Fig. 1. A high level category (father node) can be split into several lower level categories (child nodes), until it gets to the lowest level corresponding to the product items. For example, as shown in Fig. 1, $category_2$ is a father node of $category_{21}$, and $item_n^2$ is a child node of $category_{21}$. The depth from the highest category to a specific item may vary between products, depending on the product taxonomy of the company.

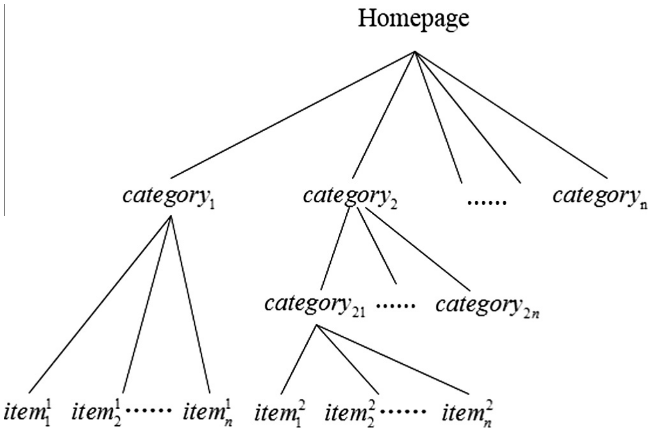

Fig. 1. E-commerce website topology.

Regarding the measurement of user interest, the following assumptions should first be given:

- Every user has a preference when he visits a website. Users only visit products that interest them. Users with similar interests show similar browsing behaviors.
- The frequency of visits to the product pages is positive correlate to users' interests (Rathipriya and Thangavel 2010, Liu et al. 2012).
- The duration for which users stay on a product webpage is positive correlate to their interests (Gong and Cheng 2008, Kim and Yum 2011).
- The sequence of visits to a product webpage is positive correlate to users' interests (Hong and Hu 2012, Li and Tan 2011). Users will firstly choose a product they most prefer. Users' browsing sequence can also be influenced by the design and layout of a website, as well as the promotions provided. However, these factors are not taken into account in this paper.

Given the above assumptions, a user's interest can be defined as a set of product categories that the user has visited. In order to unify measurement units and simplify calculations, we only evaluate user's interest in terms of the third level categories in this paper.

3.2. Indicators of a user's interest

According to the above assumptions, the visiting sequence, frequency of visits and time spent on a product page are utilized as the three main indicators to measure a user's interest. These indicators are defined as follows:

3.2.1. Indicator 1: category visiting path

Session: In this paper, we study users' browsing behaviors in units of session. A session is defined as a period of sustained web browsing or a sequence of page viewings. The website can automatically create a session ID every time a user visits a page, after handling the first page request of any given session. Also, the session will be terminated if the user has not requested any pages for a given period of time.

Browsing path: The browsing path $P_i\{url_1, url_2, url_3, \ldots, url_n\}$ of $user_i$ is a sequence of web pages browsed during a session. The length of P_i is expressed as $len(P_i) = n$. A user's browsing path can be mapped into the website topology as shown in Fig. 1. For

Please cite this article in press as: Su, Q., Chen, L. A method for discovering clusters of e-commerce interest patterns using click-stream data. Electron. Comm. Res. Appl. (2014), http://dx.doi.org/10.1016/j.elerap.2014.10.002

example, the browsing path $P_1\{ctg_1, item_1^1, ctg_1, item_2^1, item_1^2\}$ represents that $user_1$ first visits category ctg_1, then visits $item_1^1$ which belongs to ctg_1, and after that returns to ctg_1 and visits $item_2^1$, before finally visiting $item_1^2$ which belongs to category ctg_2.

Category visiting path: The category visiting path $CtgP_i\{ctg_1, ctg_2, ctg_3, \ldots, ctg_m\}$ is a sequence of categories visited by $user_i$ during any given session. In this paper, only the third level categories are listed in the category visiting path. The length of $CtgP_i$ is m, where $m \leqslant n$. For example, the category visiting path of $user_1$ is $CtgP_1\{ctg_1, ctg_2\}$ according to his browsing path $P_1\{ctg_1, item_1^1, ctg_1, item_2^1, item_1^2\}$.

3.2.2. Indicator 2: visiting frequency

Hits: For a website, $hits_i^j$ is utilized to count the number of visits (or hits) to category ctg_j by $user_i$ in a session. As shown in formula 1, hits consist of two parts: the number of visits to ctg_j and the length of the visits to product items which belong to ctg_j.

$$hit_i^j = count(user_i, ctg_j) + \sum_{k=1}^{l} count(user_i, item_k^j) \tag{1}$$

where $count(u, v)$ denotes the number of visits of a page v visited by user u during a session, $item_k^j$ belongs to ctg_j, and $l \leqslant len(P_i)$.

For example, according to the browsing path of $user_1$, $P_1\{ctg_1, item_1^1, ctg_1, item_2^1, item_1^2\}$, we can get $hits_1^1 = 4$ and $hits_1^2 = 1$.

Visiting frequency: Visiting frequency is defined as the ratio of category hits to the length of the user's browsing path. We use $freq_i^j$ to denote the visiting frequency to category ctg_j by $user_i$ during a session:

$$freq_i^j = \frac{hits_i^j}{len(P_i)} \quad (0 \leqslant freq_i^j \leqslant 1) \tag{2}$$

For example, according to the browsing path of $user_1$, $P_1\{ctg_1, item_1^1, ctg_1, item_2^1, item_1^2\}$, we can derive the visiting frequency of the two categories as $freq_1^1 = 0.8$, and $freq_1^2 = 0.2$. Other than hits, visiting frequency can be used to normalize the measurements.

Category-user associated matrix with frequency characteristics: A category-user associated matrix with frequency characteristic $\mathbf{R}_{t\times s}^{freq}$ is used to describe the relationship between categories and users in terms of visiting frequency. Let t be the total number of users and s be the total number of categories. Accordingly, each row of the matrix represents a user, and each column represents a category. Each element in the matrix denotes the visiting frequency of a specific user to a specific category.

3.2.3. Indicator 3: relative duration

Duration: In this paper, we use $duration_i^j$ to denote the total time that $user_i$ spends browsing category ctg_j in a session. The time spent browsing each product item will be accumulated to its father node, the corresponding third level category. A user may repeatedly browse a category or item page. Therefore, the time spent on each visit should be added up (see formula (3)).

$$duration_i^j = \sum_{s=1}^{count(user_i, ctg_j)} time(user_i, ctg_j, s) + \sum_{k=1}^{l} \sum_{t=1}^{count(user_i, item_k^j)} time(user_i, item_k^j, t) \tag{3}$$

In which, $time(u, v, s)$ denotes the time spent on page v by user u for the sth time in a session, $item_k^j$ belongs to ctg_j and $l \leqslant len(P_i)$.

In reality, a user may load a webpage on the browser, but that user may not be viewing this page for the whole period of time the page is open. Because of this concern, a time threshold is applied in the following experimental study.

Relative duration: Relative duration $redu_i^j$ is defined as the ratio of time $user_i$ spends on category ctg_j to the time of the entire session.

$$redu_i^j = \frac{duration_i^j}{time(P_i)} \tag{4}$$

where $0 \leqslant redu_i^j \leqslant 1$.

In this paper, relative duration rather than absolute duration is utilized as one of the main indicators to reflect a user's interest.

Category-User associated matrix with time characteristics: A category-user associated matrix with time characteristic $\mathbf{R}_{t\times s}^{time}$ is used to describe the relationships between categories and users in terms of time. The elements in the matrix represent a user's relative duration in the corresponding category.

4. Design of the rough leader clustering algorithm

A huge amount of click-stream data is required for an efficient clustering algorithm. In addition, users' multiple-interest characteristics require the algorithm to allow one user to be assigned to multiple clusters. In these circumstances, a rough leader clustering algorithm is developed using a leader clustering algorithm leveraged by a rough set theory. The leader clustering calculation is based on users' similarities in terms of browsing behavior. At the same time, rough set analysis makes it possible for a user to be assigned to more than one cluster.

4.1. Measurements for similarity

According to the assumptions in Section 3, users with similar interests display similar browsing behaviors. Given the indicators of browsing behaviors, i.e., category visiting path $CtgP_i$, visiting frequency $freq_i^j$ and relative duration $redu_i^j$, the similarity of behaviors can be evaluated quantitatively. Thereby, users can accordingly be assigned to different clusters.

Various functions have been developed to estimate the degree of similarity between two users, such as the Pearson correlation coefficient (Albadvi and Shahbazi 2009, Lee and Kwon 2008), cosine similarity (Jeong et al. 2009, Symeonidis et al. 2008) or distance measures (Kim et al. 2009, Park and Chang 2009). The choice of similarity function should be made properly, based on the data set at hand (Choi et al. 2012). In our case, the data set is high dimensional, with three different indicators: sequence, frequency and duration. First, the cosine function is utilized to estimate the similarity of two users from the perspectives of frequency and duration. Then, a path similarity is defined for estimating the two users in terms of sequence. Finally, the three similarities are integrated as the total similarity of a pair of users.

Frequency similarity: As shown in equation (5), the frequency similarity between two users $user_p$ and $user_q$ is defined as a cosine similarity measure of two vectors $\vec{p}, \vec{q}$, which are row vectors extracted from matrix $\mathbf{R}_{t\times s}^{freq}$.

$$sim_{pq}(freq) = \cos(R^{freq}[p, \cdot], R^{freq}[q, \cdot]) \tag{5}$$

In which, $0 \leqslant sim_{pq}(freq) \leqslant 1$.

Duration similarity: Similarly, the duration similarity between two users $user_p$ and $user_q$ is defined as follows:

$$sim_{pq}(time) = \cos(R^{time}[p, \cdot], R^{time}[q, \cdot]) \tag{6}$$

In which, $0 \leqslant sim_{pq}(time) \leqslant 1$.

Path similarity: As shown in equation (7), the path similarity $sim_{pq}(path)$ between two users $user_p$ and $user_q$ is defined as the common path length divided by the maximal path length. A common path is defined as the common segment in the two category paths of $CtgP_p$ and $CtgP_q$. In this common path, two users visit the same categories and in the same order. If there is more than one common path between two users, the longest one is used in the calculation of path similarity.

$$sim_{pq}(path) = \max\frac{common(CtgP_p, CtgP_q)}{[len(CtgP_p), len(CtgP_q)]} \tag{7}$$

In which, $0 \leqslant sim_{pq}(path) \leqslant 1$.

Given the above three similarities, the total similarity between two users $user_p$ and $user_q$ is defined as follows:

$$sim_{pq} = \alpha \times sim_{pq}(seq) + \beta \times sim_{pq}(freq) + \gamma \times sim_{pq}(time) \tag{8}$$

In which, α, β, γ are used to adjust the weight of the three dimensions of sequence, frequency, and duration (time). Also, $\alpha + \beta + \gamma = 1, 0 \leqslant sim_{pq} \leqslant 1$.

4.2. Cluster algorithm

The leader clustering algorithm can discover a set of leaders (the cluster representatives) by making only a single pass through the data set (Yu and Luo 2011). The algorithm starts with a randomly selected object as the initial leader. For each object in the data set, we calculate the similarity between the object and the leader. If the similarity meets the predefined threshold, then the object is assigned to the cluster represented by the leader; otherwise, the object will be regarded as a new leader. These procedures will be repeated until all objects are either assigned to a cluster or regarded as leaders. Considering the rough characteristics of a user's interest, a rough set theory is utilized and integrated to the leader cluster algorithm to make it more practical, and the following rough set properties should be satisfied (Lingras et al. 2014).

Property 1. *An object can be a member of at most one lower bound. This is consistent with the definition of lower approximation, where members belong to only one cluster.*

Property 2. *The lower bound is contained in the upper bound. An object that is a member of the lower approximation of a cluster is also a member of the upper approximation of the same cluster.*

Property 3. *An object cannot belong to only a single upper bound region. An object that does not belong to any lower approximation must be a member of at least two upper approximations.*

Given the above considerations, the rough leader clustering algorithm is defined as follows:

Let:

U	dataset
τ	general threshold
ζ	rough threshold
sim_{ij}	similarity between $user_i$ and l_j
$L\{l_1, l_2 \ldots l_j \ldots l_m\}$	a set of all current leaders
$C\{l_k\}$	a cluster represented by leader l_k
$\underline{C_k}$	lower approximation of $C\{l_k\}$
$\overline{C_k}$	upper approximation of $C\{l_k\}$
U/P	classification scheme that partitions the dataset U based on an equivalence relation P

At the very beginning of the algorithm, a user will be randomly chosen from the dataset as a leader of the first cluster. Then, the following two criteria are utilized to determine whether or not a user should be assigned to an existing cluster, and how the users would be assigned to the approximations.

General threshold τ: General threshold τ is utilized to control whether or not a user should be assigned to an existing cluster.

$$\begin{aligned}
&\forall \overline{user_i} \in U: \\
&\exists l_k \in L\{l_1, l_2 \ldots l_m\} \\
s.t\ & \\
&sim_{ik} \succ \tau, sim_{ik} = \max_{1 \leqslant j \leqslant m} sim_{ij}, user_i \in C\{l_k\} \\
&\forall C\{l_k\}, 1 \leqslant k \leqslant m : C\{l_k\} \in U/P
\end{aligned}$$

Rough threshold ζ: The overlap among clusters is controlled by rough threshold ζ. According to property 1, if $user_i$ belongs to a lower bound of $C\{l_k\}$, that user only belongs to $C\{l_k\}$. In addition, according to property 2, if $user_i$ is a member of the lower approximation of $C\{l_k\}$, that user is also a member of the upper approximation of the same cluster. According to property 3, $user_i$ cannot belong to just a single upper boundary region. A criterion is designed to determine whether a user belongs to the upper or lower bound of a cluster.

$$\begin{aligned}
&\forall \overline{user_i} \in U: \\
&\exists l_k, l_s \in L\{l_1, l_2 \ldots l_m\}, l_s \in L^*, L - L^* = \{l_k\} \\
&sim_{ik} : \iff \max_{1 \leqslant j \leqslant m} sim_{ij}, sim_{is} : \iff \max_{1 \leqslant j^* \leqslant m} sim_{ij^*} \\
s.t\ & \\
&\forall user_i \subset \underline{C_k}: \\
&sim_{ik} \succ \tau, sim_{is} \succ \tau, sim_{is}/sim_{ik} \prec \zeta, user_i \in \underline{C_k}, user_i \in \overline{C_k} \\
&\forall C\{l_k\}, 1 \leqslant k \leqslant m : C\{l_k\} \in U/P, \underline{C_k} \subset \overline{C_k} \\
&\forall user_i \subset \overline{C_k}: \\
&\tau \prec sim_{ik} \prec sim_{is}/\zeta, user_i \in \overline{C_k} \cap \overline{C_s} \\
&\forall C\{l_k\}, 1 \leqslant k \leqslant m : C\{l_k\} \in U/P, \overline{C_k} \cap \overline{C_s} \neq \emptyset
\end{aligned}$$

Based on the above analysis, the procedure of a rough leader clustering algorithm can be depicted as follows:

```
The procedure of rough leader clustering algorithm
INPUT
U: dataset
$\tau$: general threshold
$\zeta$: rough threshold
OUTPUT
leaders: array of all leaders
clusters: array of all clusters
STEP:
Randomly choose a user from dataset as a leader of the first
  cluster
FOR EACH $user_i$ IN dataset DO
  FOR EACH $l_j$ IN leaders DO
    calculate the similarity $sim_{ij}$ between $l_j$ and $user_i$
   according to Eqs. ((5)–(8))
  sort leaders by corresponding similarities
  max_sim = largest value of similarity
  IF max_sim $\leqslant \tau$:
    create a new cluster led by $user_i$
  ELSE:
    FOR EACH $l_j$ IN sorted leaders DO
      IF $(sim_{ij}/max\_sim) \geqslant \zeta$:
        add $user_i$ into the cluster led by $l_j$
RETURN clusters, leaders
```

Another critical issue of the algorithm is that the two thresholds should be carefully studied in advance, so as to control the cut borders of clusters and the degree of overlap between clusters.

Table 1
Descriptive dataset.

Variables	Original clickstream data	Filtered dataset
Number of records	3,000,000	198,325
Number of sessions	188,619	30,452
Average visiting records	7	20
Average browsing duration	7.78 min	19.58 min
Number of categories	2370	1823

Unfortunately, there are no commonly accepted standard values assigned to these thresholds. To make the algorithm more practical in our case study, the values of the two thresholds are carefully explored in Section 5.2.

5. A case study

Given the algorithm, a case study is systematically conducted, based on the click-stream data collected from one of the largest e-commerce websites in China. Similar to Amazon.com, the website used in our study sells a broad assortment of items, such as household necessities, electronic devices, apparel and cosmetics. The website users occupy approximately eight percent of online consumers in China's online shopping market. An average of three million users visit the website every day.

5.1. Data collection and its preprocess

Only click-stream data from PC users was collected. The requested URLs, as well as timestamp requests, are recorded in the click-stream data. The detailed content of each URL page, however, is not included in the dataset. The website does not trace how the user navigates to or between pages (e.g., by selecting a hyperlink, opening a bookmark, or directly entering the URL in the address bar). The records, which belong to the same user over a certain period, are organized into sessions, and each session is assigned a unique session ID. The original dataset is comprised of approximately three million records, which were randomly selected from server logs which cover the complete day of May 6, 2013.

Potentially, the click-stream is a very rich data source, because the HTML file of each user's requests can be retrieved through the URLs contained in the record. However, without a number of particular structures in place, it is difficult to analyze this free-formatted and textual data. Therefore, considerable effort is put into preprocessing and transforming the row data to a high quality dataset. These preprocess tasks include user identification, category identification, time estimation and data cleaning.

Users are distinguished via session IDs, rather than the user IDs in the dataset. A user ID is only accessible after the user logs in. However, many users prefer not to log in when they navigate the website in question, because it is not mandatory to do so. Everyone who visits the website, whether identifiable or not, is treated as a customer, and all visitors' activities are recorded over time. Therefore, adopting the use of session IDs gives rise to a huge numbers of unlogged users being brought into the analysis, resulting in a much richer user profile. The risk of distinguishing users via session IDs stems from the fact that a user may trigger multiple sessions simultaneously if they run multiple web browser sessions covering different areas of a website. Nevertheless, in Montgomery's study (2004), most users had only one active session producing page requests during a given period of time. Therefore, it's reasonable to identify users through session IDs.

A product category is identified according to the category ID in a URL. In addition, all product item records are backtracked to their own corresponding categories. Because we concentrate on categories and items, records are filtered by URLs in order to eliminate irrelevant records, such as ad clicks. Additionally, in order to improve data quality, we eliminate a number of unvalued categories and users. We discard categories and irrelevant records which are very rarely visited by users. We also omit certain invalid customers who visit very few categories.

With respect to the length of time spent on a webpage, an original dataset records the request time of each URL. However, the records do not trace the total time a user spends on a page or the time when a user leaves the page. Therefore, the time spent on a webpage is estimated as the time interval between two adjacent page requests. In this case, an extraordinarily long period of time will be considered an outlier. Since outliers may exert undue

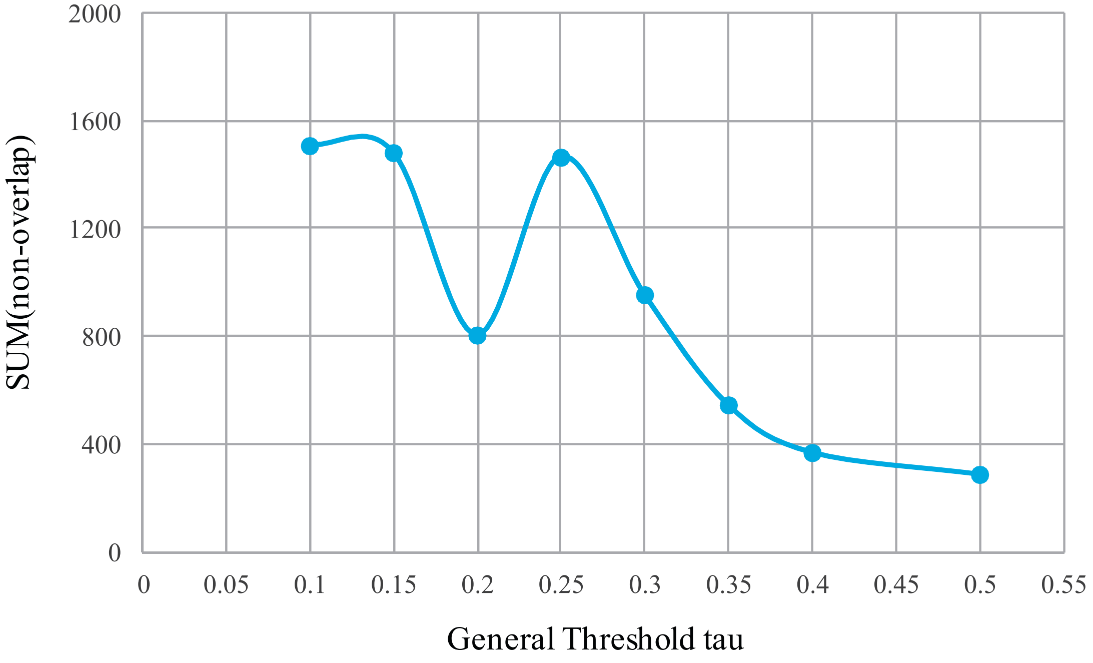

Fig. 2. Impact of the general threshold on the sum of lower approximations ($\zeta = 0.8$).

influence in determining the clusters and skewing the cluster leaders, the length of time spent on a webpage is limited to no longer than 180 s.

The detailed descriptions of the click-stream data before and after pre-processing are listed in Table 1. One can see that the number of records and sessions are decreased tremendously, compared with the original dataset. However, the number of categories does not reduce by much. This implies that, after preprocessing, the information of interest is retained to the greatest extent by a relatively small dataset. At the same time, the increases in average visiting records and average browsing duration in each session suggests the data quality is significantly improved.

To simplify the calculation, the dataset is evenly divided by a fixed interval, which ensures the data still covers the full day and thereby ensuring the results will not be affected by time. As a result, the experimental dataset contains 10,000 users and 1810 categories. Although the user number in the experimental dataset is only about one-third of that in the whole dataset, the number of categories is decreased only very slightly, from 1823 to 1810.

5.2. Parameter analysis

Before the application of the algorithm, the parameters of the algorithm should be carefully calibrated, in order to make the algorithm more practical for large data analysis and more suitable for finding users' multiple-interest characteristics. Here, two parameters, i.e., general threshold and rough threshold, are experimentally studied using real data. The general threshold controls the similarity of users in each group, while the rough threshold makes it possible for a user with multiple interests to be involved in multiple clusters. A general threshold which is set too low will not provide a useful interpretation of the partitioning of the data, while an overly high general threshold makes it too difficult to form groups. A rough threshold which is too low will lead to an unreasonable overlap between groups, while an extremely high rough threshold makes it too difficult for a user to be involved in multiple clusters. Therefore, in order to achieve a good clustering result, an optimization trade-off between these situations should be carefully conducted.

In order to keep the separate characteristics and interpretations of each cluster, the degree of overlap between the clusters must be controlled. Accordingly, Voges et al. (2003) suggested maximizing the sum of the lower approximations of the clusters when rough clustering. Hence, we utilize the sum of the lower approximations $\sum|\underline{C_k}|$ to determine the best value of general threshold τ (with a fixed rough threshold of $\zeta = 0.8$).

As shown in Fig. 2, one can find that the sum of the lower approximations $\sum|\underline{C_k}|$ increases rapidly with the decrease of general threshold τ from 0.5 to 0.3, and the sum arrives at its peak value when τ is 0.25. The sum then decreases again when τ is 0.2. However, when τ is lower than 0.15, $\sum|\underline{C_k}|$ is abnormally high. This is because when the general threshold is too low, a cluster may cover objects which are actually quite far away from the leader, resulting in a decreased number of clusters. Consequently, those objects on the fringes will belong to a single cluster and increase the lower approximations. Considering that a general threshold which is too low will lead to meaningless clusters, we determine here that the best value of a general threshold is $\tau = 0.25$.

To find out the optimal value of the rough threshold ζ, we applied the method proposed by Lingras et al. (2008), in which the percentage of a boundary region is defined as the ratio of cardinality of the union of all boundary regions, divided by the total number of objects.

$$BoundarySize = \frac{\|U_{C\{l_k\}\in U/P}(\overline{C_k} - \underline{C_k})\|}{\|U\|} \times 100\% \tag{9}$$

The optimal value of a rough threshold can be recognized by the fact that further increases in a threshold do not lead to any significant change in *BoundarySize*. To analyze the impact of rough threshold ζ on *BoundarySize*, we fix the general threshold τ as 0.25. As shown in Fig. 3, the *BoundarySize* goes down relatively slowly, until the ζ reaches the value of 0.8. Therefore, it is reasonable to consider rough threshold $\zeta = 0.8$ as an appropriate value in terms of the variance in *BoundarySize*.

In the following computational experiments, we hereby set the general threshold τ as 0.25 and the rough threshold ζ as 0.8.

5.3. Mining the interest patterns

By employing the above parameter values, the rough leader clustering algorithm is developed and applied in the prepared dataset. As a result, the original 10,000 users in the test dataset are classified into 734 separate clusters. Among these clusters, the top three clusters in terms of size contain more than 300 members. In addition, the involved categories in each cluster are illustrated by a multi-circle colored graph. The graph represents the interest patterns of the users in the corresponding cluster, as shown in Figs. 4–6. Each colored circle represents a category that was visited by group members. In addition, the length of the colored arc displays what percentage of the group's members visit the category. Circles are arranged in descending order, based on the number of visitors.

Fig. 4 shows the interest patterns in the largest user population of 402 members. All the categories in this pattern relate to women's dresses. The most popular category is dresses, followed by T-shirts, blouses, and so on. Since the data was collected in early summer, this category becomes the most popular cluster. The interest pattern implies that women care about dress collocation when they browse female clothing pages. Fig. 5 shows an interest pattern as represented by a group of 352 users. Compared with Interest Pattern 1, the categories in Interest Pattern 2 are highly diversified and include drinks, food and household necessities. These categories are characterized by high frequency use and relatively low prices. A combination of some categories can satisfy a user's eating requirements; for example, milk and cereal.

Fig. 6 shows an interest pattern as represented by a group of 300 users. This pattern demonstrates the users' interest in electronics and computers. Some categories in that pattern, such as desktop PCs and laptops, offer mostly overlapping functionalities. Some other categories, on the other hand, share similar characteristics, such as cell phones and tablet PCs, which all have mobile computing capability. According to the calculation results, for these more expensive products, customers tend to investigate their options more carefully. They crosscheck different categories to make trade-off decisions between different brands and different features. Therefore, if e-commerce websites provide online comparison tools on their websites, more potential customers may be attracted and their customers' overall experience will be improved. Since a Bluetooth headset is an accessory for both cell phones and computers, online retailers can recommend such headsets on their own, or the retailers can sell the headsets as part of a bundle with the relevant products.

6. Discussions

To validate the effectiveness of the proposed approach, the characteristics of the proposed leader clustering method were studied in comparison with that of the K-medoids clustering method. Moreover, the stability of the three interest patterns elaborated above is verified using different datasets. Thereafter, some managerial insights are discussed in detail.

Please cite this article in press as: Su, Q., Chen, L. A method for discovering clusters of e-commerce interest patterns using click-stream data. Electron. Comm. Res. Appl. (2014), http://dx.doi.org/10.1016/j.elerap.2014.10.002

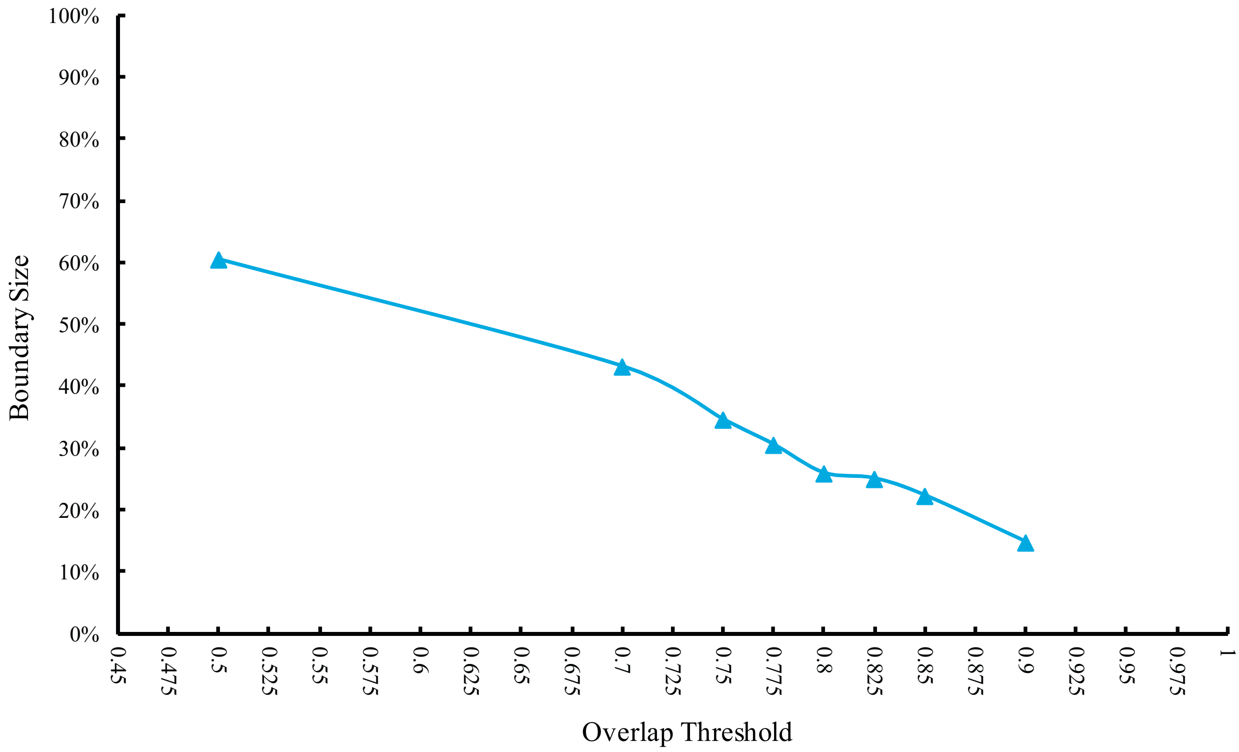

Fig. 3. Change in *BoundarySize* with rough threshold ($\tau = 0.25$).

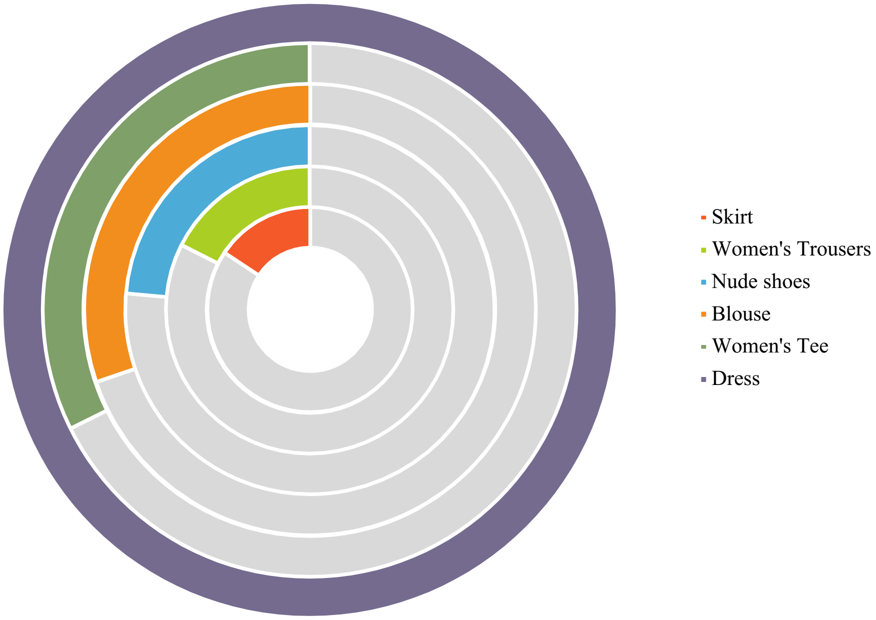

Fig. 4. Interest Pattern 1.

6.1. Comparison with other algorithms

As is already well accepted, a good clustering algorithm should be able to search out clusters with the shortest distances between members of a single cluster, and with the longest distances between members of different clusters. From this perspective, Dunn's index (Brun et al. 2007) and Davies–Bouldin's index (Bolshakova and Azuaje 2003) are usually employed to compare clustering results.

The Dunn's index is defined as the ratio of the minimum distance between two clusters and the largest intra-distance. For any partition $U \leftrightarrow C\{l\}:C\{l_1\} \cup C\{l_2\} \cdots \cup C\{l_k\}$, where $C\{l_k\}$ represents the kth cluster of partition and l_k is the leader of $C\{l_k\}$, the Dunn's index $D(U)$ is defined as:

$$D(U) = \frac{\max_{1 \leqslant i,j \leqslant k, i \neq j} sim(l_i, l_j)}{\min_{1 \leqslant i \leqslant k} \Delta C\{l_i\}}, \tag{10}$$

where l_i is the leader of $C\{l_i\}$ in the proposed algorithm, while in the K-medoids algorithm, l_i is the medoids, and $sim(l_i, l_j)$ is the similarity between two clusters representing the inter-cluster distances of

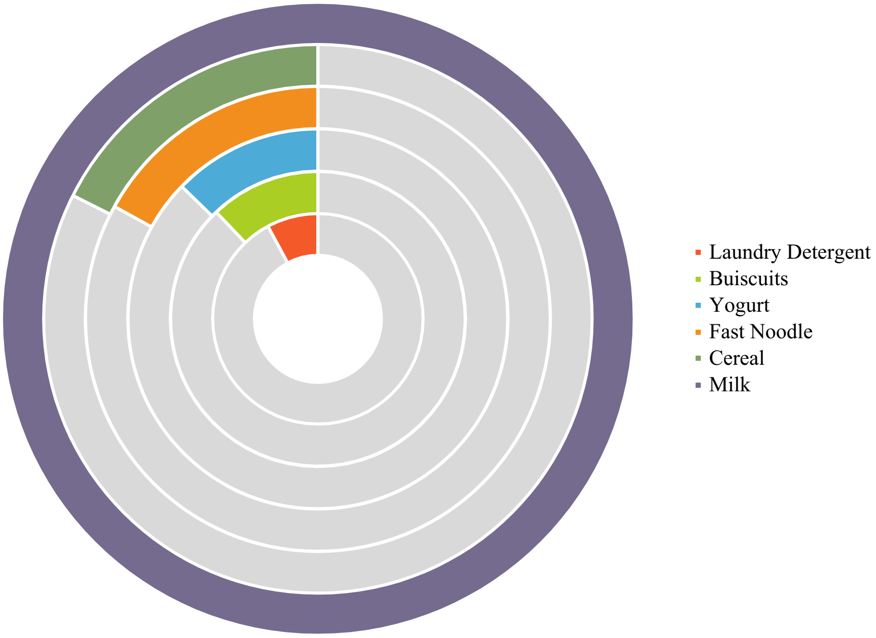

Fig. 5. Interest Pattern 2.

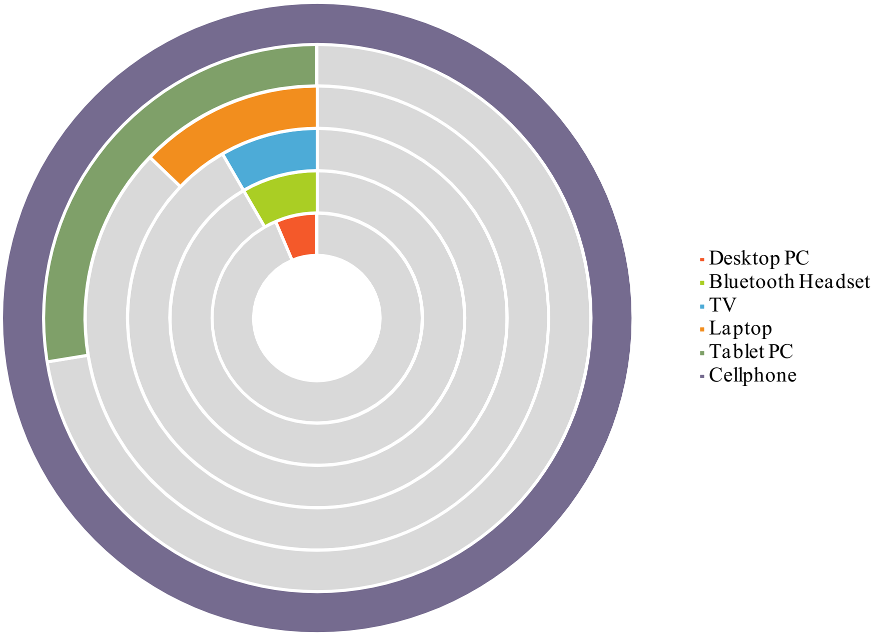

Fig. 6. Interest Pattern 3.

two clusters. $\Delta C\{l_i\}$ are the intra-cluster distances of $C\{l_i\}$ and $\Delta C\{l_i\} = \frac{1}{|C\{l_i\}|}\sum_{user_j \in C\{l_i\}} sim(user_j, l_i)$.

The Davies–Bouldin index aims at identifying sets of clusters that are compact and well separated, but which take into consideration the number of clusters. The Davies–Bouldin index, $DB(U)$, is defined as:

$$DB(U) = \frac{1}{k}\sum_{i=1}^{k} \max_{1 \leqslant i,j \leqslant k, i \neq j} \left\{ \frac{\Delta C\{l_i\} + \Delta C\{l_j\}}{sim(l_i, l_j)} \right\}, \qquad (11)$$

where k is the number of clusters.

A smaller $D(U)$ value and a larger $DB(U)$ value mean that derived clusters are compact, and their centers are far away from each other.

In this study, based on the same dataset, we comparatively studied three algorithms, i.e., the proposed rough leader clustering, the traditional leader clustering and the K-medoids clustering. As shown in Table 2, the proposed rough leader algorithm partitions the dataset into 79 groups. In addition, considering the impact of the pre-defined k value, K-medoids experiments are conducted five times, with different k values of 50, 60, 70, 80, 90. Compared with a traditional leader clustering algorithm and the K-medoids clustering method, the proposed rough leader clustering algorithm is the

Table 2
Internal validation index by various clustering algorithms.

	k	Time (s)	*D*(*U*)	*DB*(*U*)
Rough leader	79	14	0.196164	3.11073e+15
Traditional leader	77	12	0.323233	1.14371e+15
K-medoids	90	25	14.7486	2.16869e+14
K-medoids	80	22	5.54882	1.48146e+14
K-medoids	70	21	15.2974	1.14647e+14
K-medoids	60	19	42.1225	2.62941e+14
K-medoids	50	18	44.7526	1.10177e+14

best in terms of achieving the smallest $D(U)$ and the largest $DB(U)$. Moreover, a rough leader clustering algorithm performs much better than a K-medoids clustering method in terms of efficiency. These experiment results imply that the proposed algorithm is more suitable for large dataset clustering with multiple dimensions.

6.2. Stability of the interest patterns

To make use of the three interest patterns derived in Section 5.3, the stability of those interest patterns should be verified. In other words, the three interest patterns could be derived from different click-stream datasets collected from the same website. To achieve this aim, we conduct two more comparative experiments with different click-stream datasets. As shown in Fig. 7, the experiment results demonstrate that the same three interest patterns can be discovered in Dataset 2 and Dataset 3. What's more, the percentage of users in each category is very similar among the patterns derived from the three datasets respectively. Therefore, the three interest patterns are regarded as stable and independent interest patterns existing on the case website.

6.3. Managerial insights

Understanding users' browsing behaviors and discovering users' interest patterns are very important tasks for e-business companies who wish to personalize their services to the customers' unique requirements and preferences. Gaining knowledge of users' behavior and interests can provide valuable managerial insights for e-business companies looking to facilitate their website reengineering, optimize their recommendation systems and enhance their cross-selling ability.

With this knowledge, the categorization of products, as well as the navigation structure of various categories, can be reengineered and aimed specifically at target consumers. Taking the case website as an example, Interest Pattern 3 tells us that users who are interested in cell phones tend to also look for information on tablets. Nevertheless, on the studied company's current website, cell phones and tablet PCs are regarded as two unrelated categories. The navigation labels of the two categories are separated by four different navigation labels. Moreover, there's no link on the cell phone web pages that will guide users directly to the tablet web page. If users want to visit the tablet web page after browsing cell phones, they must either go back to the site's homepage and reenter the tablet label, or they have to search for tablet pages using the search bar. Obviously, this website design is not user-friendly and should be redesigned.

In fact, some commercial websites have already started to change from their conventional navigation designs. They are now restructuring categories and items according to their users' interest patterns. For example, Taobao.com, the largest commercial website in China, carried out a radical change of its homepage layout this year. In addition to the traditional category navigation bars, several groupings are constructed to specifically target different user groups, such as women, men and children. In each grouping, the

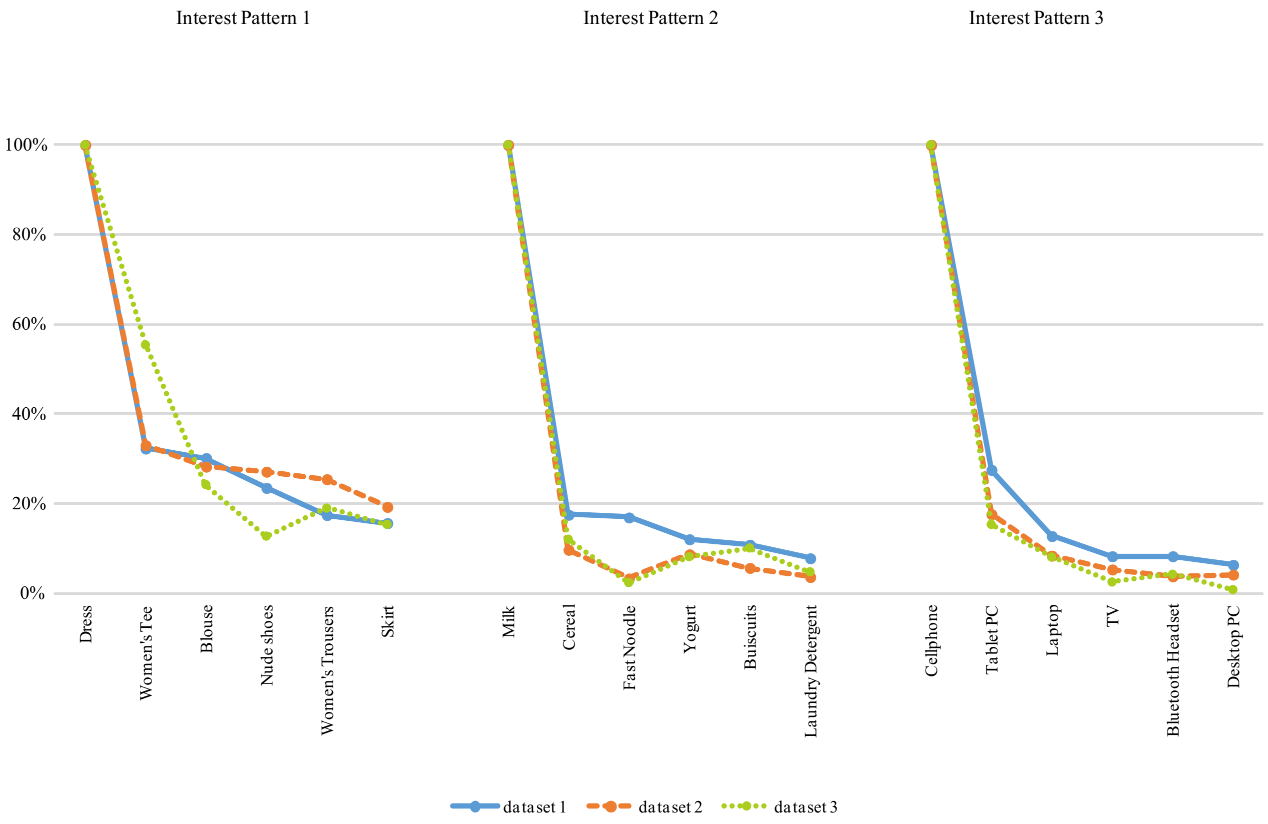

Fig. 7. Stability analysis of the three interest patterns.

Please cite this article in press as: Su, Q., Chen, L. A method for discovering clusters of e-commerce interest patterns using click-stream data. Electron. Comm. Res. Appl. (2014), http://dx.doi.org/10.1016/j.elerap.2014.10.002

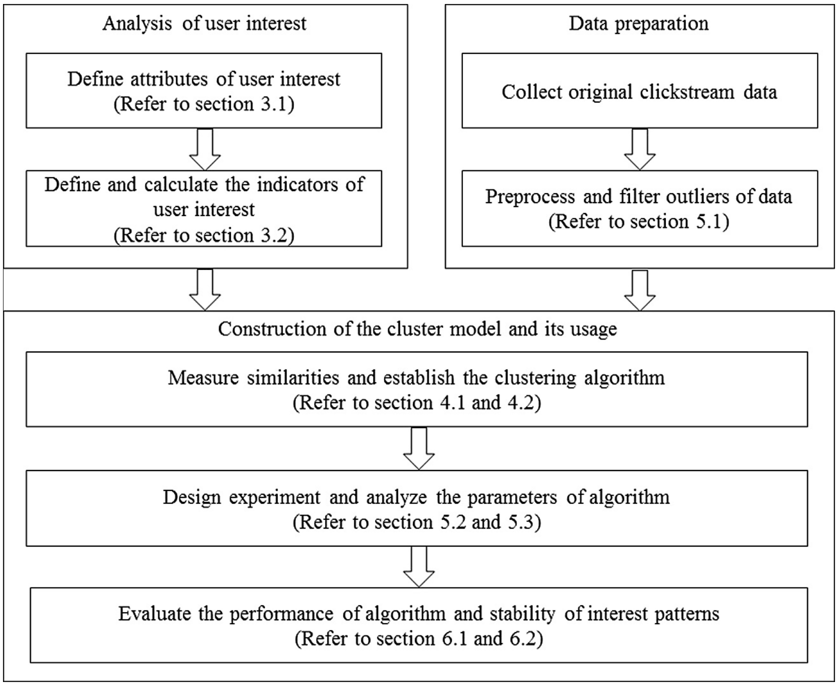
Analysis of user interest
Define attributes of user interest (Refer to section 3.1)
Define and calculate the indicators of user interest (Refer to section 3.2)
Data preparation
Collect original clickstream data
Preprocess and filter outliers of data (Refer to section 5.1)
Construction of the cluster model and its usage
Measure similarities and establish the clustering algorithm (Refer to section 4.1 and 4.2)
Design experiment and analyze the parameters of algorithm (Refer to section 5.2 and 5.3)
Evaluate the performance of algorithm and stability of interest patterns (Refer to section 6.1 and 6.2)

Fig. 8. General research route for clustering interest patterns.

most popular categories are displayed and updated frequently. However, the segmentation of user groups on Taobao.com is simply based on the users' most basic attributes, such as gender and age. By referring to the mining results in our study, user groups can be more precisely segmented in terms of various interest patterns, such as "fashion lady", "3C fans", and "housewife favorites." This segmentation of a website can help ensure the navigation structure more accurate in terms of reaching the targeted customers.

In recent years, the competition in e-commerce has been increasingly fierce, not just in China but all over the world. In order to improve users' experience, many online retailers have begun to personalize their services and provide more friendly recommendations. Nevertheless, most online retailers only take advantage of very limited information before speculating on users' preferences. They rely on very basic information, such as user ratings or browsed products. Generally speaking, websites simply recommend similar products from the same category in which users are looking. For example, if a user is browsing a type of milk, all milks of different brands and flavors are recommended to that user. This type of recommendation method, which has been widely used in the sales of online movies and books, is based on the assumption that the products are similar to each other, and users are normally interested in a specific type of product. However, for a large-scale e-commerce website with diverse categories, the situation is completely different. As discovered in our experiment's results, users tend to view multiple categories during a session. Therefore, recommending items of the same category brings very little added value to customers. On the contrary, online retailers should recommend items from different categories, based on the interest patterns discovered through data mining. For instance, according to Interest Pattern 2, biscuits should be recommended on the web pages where milk brands are listed, and vice versa.

7. Conclusion

Nowadays, e-commerce websites can generate tremendous amounts of click-stream data on a daily basis. With such enormous amounts of data becoming available, users' browsing behaviors and their corresponding interest patterns can be discovered by employing data mining technology. This paper is dedicated to finding users' interest patterns based on their click-stream data analysis. A novel rough leader clustering is proposed and systematically tested. This algorithm can be used to group users with similar browsing behaviors and determine their interest patterns. Based on datasets collected from a real case website, the parameters of the algorithm are optimized, and the effectiveness of the algorithm is validated.

The experiment's findings demonstrate that the newly-proposed approach can outperform the traditional leader clustering algorithm and the K-medoids algorithm, both in terms of effectiveness and efficiency. Using this approach, three main interest patterns are discovered, and the stability of the three patterns is tested. These interest patterns can provide significant assistance to online retailers who wish to personalize their website's layout and navigation structure. More importantly, these interest patterns can improve the site's recommendation strategies and make the site more effective.

Although significant progress has been made in this area, many issues are still open to investigation in terms of future study. For example, how to use our approach to improve the recommendation algorithm, and how much better the improved recommendation is compared with other recommendation methods, such as the collaborative filtering (CF) method could be examined. The CF recommendation is a well-known method based on ratings information. The method predicts users' ratings on each item and recommends top-rated items to the user. Given the novel approach proposed in this paper, the CF recommendation could be refined by basing it on users' browsing behaviors and users' preferences. Firstly, the preference of user u on item i can be formulized according to the user's browsing behaviors, including browsing sequence, frequency and duration. Then, using the proposed rough leader clustering algorithm, the top k similar users will be selected. Finally, the preference information of the top k similar users can be used to predict the preferences of user u, and some common items of interest will be recommended to user u. Meanwhile, the accuracy of the recommendation should be evaluated and compared with that of the traditional rating-based CF method. To this end, we can employ the two indices of recall and precision, which are widely used to measure the quality of recommendations (Albadvi and Shahbazi 2009, Huang and Huang 2009, Choi et al. 2012). In addition, competition between online retailers is now reaching as far as mobile terminals. Considering that a user's shopping experience on a mobile phone is different from that on a PC in terms of the user interface, response speed and the price of data traffic, researchers must explore how users' behaviors and interests may differ because they are shopping on different terminals. The research approaches and the corresponding results could provide a reference framework for future studies of users' behaviors and users' interests, particularly in terms of the consistency of customer groups. According to the authors' experience, the general research route can be summarized as in Fig. 8.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Project Nos. 71090404 and 71072026).

References

Aguiar, L., Martens, B., 2013. Digital Music Consumption on the Internet. Institute of Prospective Technological Studies, Joint Research Centre.

Albadvi, A., Shahbazi, M., 2009. A hybrid recommendation technique based on product category attributes. Expert Systems with Applications 36 (9), 11480–11488.

Asharaf, S., Murty, M.N., 2003. An adaptive rough fuzzy single pass algorithm for clustering large data sets. Pattern Recognition 36 (12), 3015–3018.

Bolshakova, N., Azuaje, F., 2003. Cluster validation techniques for genome expression data. Signal Processing 83 (4), 825–833.

Bridges, E., Florsheim, R., 2008. Hedonic and utilitarian shopping goals: the online experience. Journal of Business Research 61 (4), 309–314.

Brun, M., Sima, C., Hua, J., et al., 2007. Model-based evaluation of clustering validation measures. Pattern Recognition 40 (3), 807–824.

Bucklin, R.E., Sismeiro, C., 2003. A model of web site browsing behavior estimated on clickstream data. Journal of Marketing Research, 249–267.

Bucklin, R.E., Sismeiro, C., 2009. Click here for Internet insight: advances in clickstream data analysis in marketing. Journal of Interactive Marketing 23 (1), 35–48.

Carmona, C.J., Ramírez Gallego, S., Torres, F., et al., 2012. Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. Expert Systems with Applications 39 (12), 11243–11249.

Chen, Y., Kuo, M., Wu, S., Tang, K., 2009. Discovering recency, frequency, and monetary (RFM) sequential patterns from customer's purchasing data. Electronic Commerce Research and Applications 8 (5), 241–251.

Chiang, R.D., Wang, Y.H., Chu, H.C., 2013. Prediction of members' return visit rates using a time factor. Electronic Commerce Research and Applications 12 (5), 362–371.

Choi, K., Yoo, D., Kim, G., et al., 2012. A hybrid online-product recommendation system: combining implicit rating-based collaborative filtering and sequential pattern analysis. Electronic Commerce Research and Applications 11 (4), 309–317.

Cleger-Tamayo, S., Fernández-Luna, J.M., Huete, J.F., 2012. Top-N news recommendations in digital newspapers. Knowledge-Based Systems 27, 180–189.

Danaher, P.J., Mullarkey, G.W., Essegaier, S., 2006. Factors affecting web site visit duration: a cross-domain analysis. Journal of Marketing Research 43 (2), 182–194.

do Prado, H.A., Engel, P.M., Chaib Filho, H., 2002. Rough clustering: an alternative to find meaningful clusters by using the reducts from a dataset. In: Rough Sets and Current Trends in Computing. Springer, Berlin, Heidelberg, pp. 234–238.

Ganesh, J., Reynolds, K.E., Luckett, M., Pomirleanu, N., 2010. Online shopper motivations, and e-store attributes: an examination of online patronage behavior and shopper typologies. Journal of Retailing 86, 106–115.

Goldfarb, A., 2006. The medium-term effects of unavailability. Quantitative Marketing and Economics 4 (2), 143–171.

Gong, S.J., Cheng, G.H., 2008. Mining user interest change for improving collaborative filtering. Second International Symposium on Intelligent Information Technology Application, 2008, IITA'08, Vol. 3. IEEE, pp. 24–27.

He, Y., Ma, L.X., Teng, G.E., 2012. Web log mining based on user's accessing interest. Systems Engineering Theory & Practice 32 (6), 1353–1361.

Hong, Y.U., Hu, L.U.O., 2012. Possibilistic fuzzy clustering algorithm based on web user access paths. Journal of Chinese Computer Systems 1, 024.

Huang, C.L., Huang, W.L., 2009. Handling sequential pattern decay: developing a two-stage collaborative recommender system. Electronic Commerce Research and Applications 8 (3), 117–129.

Hussain, T., Asghar, S., Fong, S., 2010. A hierarchical cluster based preprocessing methodology for web usage mining. In: 2010 6th International Conference on Advanced Information Management and Service (IMS). IEEE, pp. 472–477.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31 (8), 651–666.

Jeong, B., Lee, J., Cho, H., 2009. An iterative semi-explicit rating method for building collaborative recommender systems. Expert Systems with Applications 36 (3), 6181–6186.

Kim, Y.S., Yum, B.J., 2011. Recommender system based on click stream data using association rule mining. Expert Systems with Applications 38 (10), 13320–13327.

Kim, H.K., Kim, J.K., Ryu, Y.U., 2009. Personalized recommendation over a customer network for ubiquitous shopping. IEEE Transactions on Services Computing 2 (2), 140–151.

Kou, G., Lou, C., 2012. Multiple factor hierarchical clustering algorithm for large scale web page and search engine clickstream data. Annals of Operations Research 197 (1), 123–134.

Lam, S.Y., Chau, A.W.L., Wong, T.J., 2007. Thumbnails as online product displays: how consumers process them. Journal of Interactive Marketing 21 (1), 36–59.

Lee, K.C., Kwon, S., 2008. Online shopping recommendation mechanism and its influence on consumer decisions and behavior: a causal map approach. Expert Systems with Applications 35 (4), 1567–1574.

Li, Y., Tan, B.H., 2011. Clustering algorithm of web click stream frequency pattern. Journal of Tianjin University of Science & Technology 3, 018.

Li, Y., Lu, L., Xuefeng, L., 2005. A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-commerce. Expert Systems with Applications 28 (1), 67–77.

Li, D., Lv, Q., Xie, X., et al., 2012. Interest-based real-time content recommendation in online social communities. Knowledge-Based Systems 28, 1–12.

Lingras, P., West, C., 2004. Interval set clustering of web users with rough k-means. Journal of Intelligent Information Systems 23 (1), 5–16.

Lingras, P., Chen, M., Miao, D., 2008. Precision of rough set clustering. In: Rough Sets and Current Trends in Computing. Springer, Berlin, Heidelberg, pp. 369–378.

Lingras, P., Chen, M., Miao, D., 2014. Qualitative and quantitative combinations of crisp and rough clustering schemes using dominance relations. International Journal of Approximate Reasoning 55 (1), 238–258.

Liu, H., Xing, H., Zhang, F., 2012. Web personalized recommendation algorithm incorporated with user interest change. Journal of Computational Information Systems 8 (4), 1383–1390.

López, I., Ruiz, S., 2011. Explaining website effectiveness: the hedonic–utilitarian dual mediation hypothesis. Electronic Commerce Research and Applications 10 (1), 49–58.

Moe, W.W., 2003. Buying, searching, or browsing: differentiating between online shoppers using in-store navigational clickstream. Journal of Consumer Psychology 13 (1), 29–39.

Moe, W.W., 2006. An empirical two-stage choice model with varying decision rules applied to internet clickstream data. Journal of Marketing Research, 680–692.

Moe, W.W., Fader, P.S., 2004. Dynamic conversion behavior at e-commerce sites. Management Science 50 (3), 326–335.

Montgomery, A.L., Li, S., Srinivasan, K., et al., 2004. Modeling online browsing and path analysis using clickstream data. Marketing Science 23 (4), 579–595.

Nottorf, F., 2014. Modeling the clickstream across multiple online advertising channels using a binary logit with Bayesian mixture of normals. Electronic Commerce Research and Applications 13, 45–55.

Park, Y.J., Chang, K.N., 2009. Individual and group behavior-based customer profile model for personalized product recommendation. Expert Systems with Applications 36 (2), 1932–1939.

Park, Y.H., Fader, P.S., 2004. Modeling browsing behavior at multiple websites. Marketing Science 23 (3), 280–303.

Park, H.S., Jun, C.H., 2009. A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications 36 (2), 3336–3341.

Pawlak, Z., 2002. Rough sets and intelligent data analysis. Information Sciences 147 (1), 1–12.

Peters, G., 2006. Some refinements of rough k-means clustering. Pattern Recognition 39 (8), 1481–1491.

Rathipriya, R., Thangavel, D.K., 2010. A fuzzy co-clustering approach for clickstream data pattern. Global Journal of Computer Science and Technology 10 (6).

Rutz, O.J., Bucklin, R.E., 2012. Does banner advertising affect browsing for brands? Clickstream choice model says yes, for some. Quantitative Marketing and Economics 10 (2), 231–257.

Sismeiro, C., Bucklin, R.E., 2004. Modeling purchase behavior at an e-commerce web site: a task-completion approach. Journal of Marketing Research, 306–323.

Suresh Babu, V., Viswanath, P., 2009. Rough-fuzzy weighted k-nearest leader classifier for large data sets. Pattern Recognition 42 (9), 1719–1731.

Symeonidis, P., Nanopoulos, A., Manolopoulos, Y., 2008. Providing justifications in recommender systems. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 38 (6), 1262–1272.

Van den Poel, D., Buckinx, W., 2005. Predicting online-purchasing behaviour. European Journal of Operational Research 166 (2), 557–575.

Velmurugan, T., Santhanam, T., 2010. Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points. Journal of Computer Science 6 (3).

Voges, K.E., Pope, N., Brown, M.R., 2002. Cluster analysis of marketing data examining on-line shopping orientation: a comparison of k-means and rough clustering approaches. Heuristics and Optimization for Knowledge Discovery, 207–224.

Voges, K. E., Pope, N. K., and Brown, M. R. A rough cluster analysis of shopping orientation data. In Proceedings Australian and New Zealand Marketing Academy Conference, Adelaide, 2003, 1625–1631.

Wei, J., Shen, Z., Sundaresan, N., et al., 2012. Visual cluster exploration of web clickstream data. In: 2012 IEEE Conference on Visual Analytics Science and Technology (VAST). IEEE, pp. 3–12.

Wu, R.S., Chou, P.H., 2011. Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. Electronic Commerce Research and Applications 10 (3), 331–341.

Xing, C.X., Gao, F.R., Zhan, S.N., 2007. A collaborative filtering recommendation algorithm incorporated with user interest change. Journal of Computer Research and Development 44 (2), 296–301.

Xu, R., Wunsch, D., 2005. Survey of clustering algorithms. IEEE Transactions on Neural Networks 16 (3), 645–678.

Yu, H., Luo, H., 2011. A novel possibilistic fuzzy leader clustering algorithm. International Journal of Hybrid Intelligent Systems 8 (1), 31–40.

Zeng, J., Zhang, S., Wu, C., 2008. A framework for WWW user activity analysis based on user interest. Knowledge-Based Systems 21 (8), 905–910.

Zhai, Z., Liu, B., Xu, H., et al., 2011. Clustering product features for opinion mining. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, pp. 347–354.

Zhao, X., Niu, Z., Chen, W., 2013. Interest before liking: two-step recommendation approaches. Knowledge-Based Systems.

Zheng, L., Cui, S., Yue, D., et al., 2010. User interest modeling based on browsing behavior. 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), Vol. 5. IEEE, V5-455–V5-458.

Please cite this article in press as: Su, Q., Chen, L. A method for discovering clusters of e-commerce interest patterns using click-stream data. Electron. Comm. Res. Appl. (2014), http://dx.doi.org/10.1016/j.elerap.2014.10.002