Electronic Commerce Research and Applications 14 (2015) 1–13

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 22 October 2014

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/

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