A MapReduce-Based User Identification Algorithm in Web Usage Mining

Mitali Srivastava, Department of Computer Science, Institute of Science, Banaras Hindu University, Varanasi, India
Rakhi Garg, Computer Science Section, Mahila Maha Vidyalaya, Banaras Hindu University, Varanasi, India
P.K. Mishra, Department of Computer Science, Institute of Science, Banaras Hindu University, Varanasi, India

ABSTRACT

This article contends that in the booming era of information, analysing users’ navigation behaviour is an important task. User identification is considered as one of the important and challenging tasks in the data preprocessing phase of the Web usage mining process. There are three important issues with the reactive strategies of User identification methods that need to be focused: the first is dealing of sharing IP address problem in a proxy server environment, the second is distinguishing users from Web robots, and the third is dealing with huge datasets efficiently. In this article, authors have developed a MapReduce-based User identification algorithm that deals with the above mentioned three issues related to user identification methods. Moreover, the experiment on the real web server log shows the effectiveness and efficiency of the developed algorithm.

KEYWORDS

Data Cleaning, Data Preprocessing, Hadoop, MapReduce, User Identification, Web Server Log, Web Usage Mining

1. INTRODUCTION

Apart from the content and structural information of the Website, server logs have also been considered as one of the valuable sources of information. This information can be used to analyse users’ navigation behaviour (Pabarskaite & Raudys, 2007). Web usage mining is a class of Web mining to mine server logs to find relevant patterns. These patterns are successfully applied in various applications like restructuring Websites, recommendation of pages and products, personalizing Web contents, and improving server activities like prefetching and caching (Facca & Lanzi, 2005; Kemmar, Lebbah, & Loudni, 2016). Web usage mining process can be divided into three important steps: Data preprocessing, Pattern extraction and Pattern evaluation (Liu, 2007). Due to the unstructured and huge nature of log data, Data preprocessing step has become the essential and time-consuming task in the Web usage mining process. It is a complex task and consumes more than 60% of whole Web usage mining process time (Tanasa & Trousse, 2004). Data preprocessing of server log incorporates several steps: Data fusion, Data cleaning, User identification, Session identification, Path completion, and Data transformation (Cooley, Mobasher, & Srivastava, 1999; Liu, 2007). Among them, User identification is one of the challenging tasks in Data Preprocessing.
due to the external/local proxy server, shared internet and cache systems (Pabarskaite & Raudys, 2007). This article focuses on User identification, a complex and challenging phase in the Web usage mining process. In User identification phase, users are identified and their activities are grouped and recorded into a user activity file. Several heuristics have been proposed for better identification of the user in last few years. Spiliopoulou et al. have classified user identification methods into two classes namely proactive methods and reactive methods. In proactive methods, users are identified by the previous or current interaction of the user with the Website. Proactive strategies incorporate methods such as user authentication, activation of cookies on the client-side, dynamic pages associated with the browser, etc. (Spiliopoulou, Mobasher, Berendt, & Nakagawa, 2003). However, these proactive approaches are most accurate and reliable methods for identifying users but they raise privacy concerns and purely dependent on users’ cooperation. In the absence of user authentication approach, the most popular proactive approach to distinguishing unique user is the use of client-side cookies information (Liu, 2007). Whenever a Web user navigates through a Website for the first time, the Web server sends a cookie i.e. a piece of information to the client browser. This information is stored on the client machine in the form of a text file (Facca & Lanzi, 2005). A cookie may contain various information including users’ unique id. Few researchers have applied the cookie based approach to identify users (Elo-Dean & Viveros, 1997; Ivanscy & Juhasz, 2007; Kamdar & Joshi, 2000). Although this approach is considered as one of the most accurate methods to identify users but cookies are not often recorded on client machine due to browser constraints or users’ non-cooperation e.g. Some browsers do not support cookies or disable cookies. Sometimes cookies are deleted by the user. On the other hand, in reactive methods, users are identified from existing log records after interaction with the Website. One of the basic approaches in reactive methods is identification by the IP address (Géry & Haddad, 2003). However, this approach is unable to deal with sharing IP address issue in the proxy server. According to Cooley et al., two heuristics can be used to solve this issue: the first heuristic assumes that two log entries having same IP addresses but different User agents may belong to two different users. In the second heuristic, some additional information like Web site topology and referrer log are used to identify users. This heuristic assumes that a user is considered as a new user if requested page is not accessible through hyperlink of previously requested pages of the same IP address (Cooley et al., 1999). Tanasa et al. have used IP address and User agent information to identify users if authentication of the user is not available (Tanasa & Trousse, 2004). Castellano et al. and Suneetha et al. also, have used IP address and user agent information to identifying users (Castellano, Fanelli, & Torsello, 2007; Suneetha & Krishnamoorthy, 2009). Further, researchers have applied the combined approach to identify users. According to their approach, if IP address is same and User agent is different then consider a new user. Further, if both are same and requested resource is not accessible through previously accessed pages then consider a new user (Reddy, Reddy, & Sitaramulu, 2013).

However, all above-discussed methods are successfully applied in various applications but they are not suitable for large datasets. In the last few years, MapReduce programming framework has become a popular framework for distributed computation of big data that is executed on a cluster of nodes and Hadoop is an open source implementation of MapReduce framework (Bhandarkar, 2010; Dean & Ghemawat, 2008). Few researchers have focused on scalability issues of Data Preprocessing methods in the Web usage mining process. They have identified Web users by using IP address information in MapReduce framework (Savitha & Vijaya, 2014; Zhang & Zhang, 2013). However, their methods are appropriate for large datasets but are unable to deal with proxy server problem. Huang et al. have given an improved referrer based algorithm for user session identification using MapReduce programming framework. For user identification, they have considered a specific user is under same Asymmetric Digital Subscriber Line (ADSL) and same User agent (Huang, Chen, & Le, 2013). This method is suitable for large datasets however it is not able to distinguish users from Web robots at User identification phase.
Prototyping in Web Development
www.igi-global.com/chapter/prototyping-web-development/29584?camid=4v1a

A Novel Architecture for Deep Web Crawler
www.igi-global.com/article/novel-architecture-deep-web-crawler/52804?camid=4v1a