International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey of Novel Clustering and Knowledge Extraction from Log

Vasim Dilawar Mujawar[1], Prof. Pratima Bhati[2]

[1]Research Scholar, Department of Computer Engineering, Dhole Patil College of Engg, Wagholi, Pune, Maharashtra, India
[2]Department of Computer Engineering, Dhole Patil College of Engg, Wagholi, Pune, Maharashtra, India

Abstract: ***Now, a days the explosive growth of Internet is rapidly increasing and knowledge extraction become the objective of most business like advertisement. Due to the biggest source of information the information available to the user is not relevant .to overcome this problem we have to understand the user need and preferences. Web mining is nothing but mining the information related to web users. This is crucial task to get web users interest and need. In web mining there are several task that has to be perform such as preprocessing ,clustering and knowledge extraction .we use clustering technique in this paper for grouping web users having same interest. There are different clustering systems that need to be survey. The knowledge extraction deals with extracting user's interest through their preferences .here in this paper we are using log files as a source for extracting information. In the existing system k-means clustering, priori algorithm are used. A preprocessing identifies each user through his transaction logged in file. World Wide Web is intermediate of web pages and links. In web usage mining is also the application of data mining techniques to determine usage pattern from web data for understanding the needs.***

Keywords: Web mining, clustering, knowledge extraction, web usage mining.

1. Introduction

Now a day's internet is the important part of human day to day life. Internet is the biggest source of information available in this era. Providing information to web user according to their need is important challenge. The information available to user may be irrelevant or not as per user interest due to diverse of information available today on web repository. To get the required information user has to browse more and more so that the navigational path also increases this result into wasting the time of users in browsing only. For providing relevant information to the user we have to know the user interest, need and preferences. For this web mining technique is used. Web mining is nothing but is part of data mining which is used for mining the user's task from user's log.

Web log is nothing but a file which is stored at server side and captures all the transactions performed by user on web. We collect the information from logs which are needed for mining the task through web mining. This step is called preprocessing. Web log contains several different fields which are used for further knowledge extraction. A web log also contains other fields which are not required for web mining. To get required and relevant fields from web log we need to perform data cleaning. This process is called data cleaning in preprocessing step. The web log contains required fields for web mining such as IP address, user id, time and page to which users visit.

Here in this paper we have to cluster users according to their preferences this is called as clustering process of web mining. Web mining is application of data mining that contain web data and identify user visiting information and extract their interest using the patterns. Web mining is important in computer and information science. By knowledge extraction we can come to know the users preferences. This information can be further utilize for personalize the web site, search engine optimization, for advertisement used by commercial business application. Web servers collect data of users by their request and responses. Log files stored client information such as IP address, URL request etc.

2. Literature Survey

Web users clustering is a crucial task for mining information related to users' needs and preferences. Up to now, popular clustering approaches build clusters based on usage patterns derived from users' page preferences. This paper emphasizes the need to discover similarities in users' accessing behavior with respect to the time locality of their navigational acts. In this context, we present two time aware clustering approaches for tuning and binding the page and time visiting criteria. The two tracks of the proposed algorithms define clusters with users that show similar visiting behavior at the same time period, by varying the priority given to page or time visiting. The proposed algorithms are evaluated using both synthetic and real datasets and the experimentation has shown that the new clustering schemes result in enriched clusters compared to those created by the conventional non-time aware users clustering approaches. These clusters contain users exhibiting similar access behavior not only in terms of their page preferences but also of their access time.

.

1) A New Clustering and Preprocessing for Web Log Mining, B. Uma Maheswari, Dr. P. Sumathi

World Wide Web is a massive repository of web pages and links. It provides information about vast area for the Internet users. There is tremendous growth and development in internet. Users' accesses are documented in web logs. Web

usage mining is application of mining techniques in logs. Since due to tremendous usage, the log files are growing at a faster rate and the size is becoming huge. Preprocessing plays a vital role in efficient mining process as Log data is normally noisy and indistinct. Reconstruction of sessions and paths are completed by appending missing pages in preprocessing. Additionally, the transactions which illustrate the behavior of users are constructed exactly in preprocessing by calculating the Reference Lengths of user access by means of byte rate. Using Web clustering several types of objects can be clustered into different groups for various purposes. In this internet era, web sites on the internet are useful source of information in everyday life. Therefore there is an enormous development of World Wide Web in its volume of traffic and the size and complexity of web sites. As per August 2010 Web Server survey by Net craft there are 213,458,815 active sites. Web mining is the application of data mining, chart technology, artificial intelligence and so on to the web data and identifies user's visiting behaviors and extracts their interests using patterns. Due to its usual application in Web analytics, e-learning, e-commerce, information retrieval etc., web mining has become one of the important areas in computer and information science.

2) Mining User Tasks from Print Logs, Lei Zhang1, Ping Luo

With lots of applications emerging in World Wide Web, many interaction data from users are collected and exploited to discover user behavior or interest patterns. In this paper, we attempt to exploit a new interaction data, namely print logs, where each record is printing URLs selected by a user using a popular web printing tool. Users usually print web contents based on an intention (subtask or task). Apparently, mining common print tasks from print logs is able to capture users' intentions, which undoubtedly benefits many web applications, such as task oriented recommendation and behavior targeting. However, it is not an easy job to perform this due to the difficulty of URL topic representation and task formulation. To this end, we propose a general framework. Additionally, print tasks can also help online advertising. Since lots of print tasks, such as travel planning or conference organization, contain commercial interests, they may help advertisers to identify potential customers and provide more relevant ads to such users. This behavioral targeting service is very helpful for all businesses targeting different types of customers named UPT (Users Print Tasks mining framework), for mining print tasks from print logs. Specifically, we attempt to leverage delicious (a social book marking web service) as an external thesaurus to expand the expression of each URL by selecting tags associated with the domain of each URL.

3) Model based clustering and visualization of navigation patterns on a web site, Cadez , D. Heckerman, C. Meek, P. Smyth, and S. White

This system present a new methodology for exploring and analyzing navigation patterns on a website. The patterns that can be analyzed consist of sequences of URL categories traversed by users. In this approach, first partition site users into clusters such that users with similar navigation paths through the site are placed into the same cluster. Then, for each cluster, system displays these paths for users within that cluster. The clustering approach employ is model-based (as opposed to distance-based) and partitions users according to the order in which they request web pages. In particular, System clusters users by learning a mixture of first-order Markov models using the Expectation-Maximization algorithm. The runtime of algorithm scales linearly with the number of clusters and with the size of the data; and implementation easily handles hundreds of thousands of user sessions in memory

4) Model-based cluster analysis for web users sessions, G. Pallis, L. Angelis, and A. Vakali

One of the main issues in Web usage mining is the discovery of patterns in the navigational behavior of Web users. Standard approaches, such as clustering of users' sessions and discovering association rules or frequent navigational paths, do not generally allow characterizing or quantifying the unobservable factors that lead to common navigational patterns. Therefore, it is necessary to develop techniques that can discover hidden and useful relationships among users as well as between users and Web objects. Correspondence Analysis (CO-AN) is particularly useful in this context, since it can uncover meaningful associations among users and pages. This system considers the duration of users accessing on page.

5) Clustering of web users using session-based similarity measures, J. Xiao and Y. Zhang

This is one of the important research topic in web usage in the clustering of web user based on their common properties. This type of system presents an approach for measuring similarity of interests among web users from their past access.

6) Click stream clustering using weighted longest common subsequences, Banerjee , J. Ghosh

Categorizing visitors based on their interactions with a website is a key problem in web usage mining. The click streams generated by various users often follow distinct patterns, the knowledge of which may help in providing customized content. In this paper, propose a novel and effective algorithm for clustering web users based on a function of the longest common subsequence of their click streams that takes into account both the trajectory taken through a website and the time spent at each page.

This system proposed an algorithm for clustering web users based on the function of the longest common subsequences of their click stream.

7) Knowledge Discovery from Users Web-Page Navigation, Cyrus Shahabi, Amir M. Zarkesh, Jafar Adibi

The orthodox view on the web-pages assumes a unidirectional flow of information from the server to the client. In this view, the flow of information in the other direction is possible only if the client chooses to respond through email or web-page forms. The previous studies on the design of the web-pages are also concentrated on how to

make this uni-lateral flow more efficient to the client. In this paper, we discuss a broader view in which the servers also could continuously receive useful information from the clients. Capturing the characteristics of the users of a business web site is an important task for their marketing department.

8) Time Aware Web Users Clustering, Sophia G. Petridou, Vassiliki A. Koutsonikola

DESPITE Web's remarkable adoption, Web users often experience problems of low precision or irrelevance in their searching, low accessing speeds (due to information overload) and outdated or out of their interests personalized information. Thus, from the Web information providers' perspective, it is important to organize their data and to address their users according to their preferences and needs. Web users clustering so far has proposed techniques to organize users into clusters based on their navigational behavior, i.e. visiting patterns are identified and compared in order to assign users to the same or different clusters. The task of Web users' clustering is crucial and has been studied in various application frameworks, since for example based on users clusters, Web-based companies may provide dynamic content (advertisements, offers, customized guides) and decide their market strategies and administrators may restructure or redesign their sites and improve their performance (by user tailored caching and pre-fetching policies) etc. This paper emphasizes the need to develop clustering system which considers both time and page preferences using tuning and binding algorithms.

9) Divergence Measures Based on the Shannon Entropy, Jianhua Lin

New class of information-theoretic divergence measures based on the Shannon entropy is introduced. Unlike the well-known Kull back divergences, the new measures do not require the condition of absolute continuity to be satisfied by the probability distributions involved. More importantly, their close relationship with the variational distance and the probability of misclassification error are established in terms of bounds. These bounds are crucial in many applications of divergence measures. The new measures are also well characterized by the properties of nonnegativity, finiteness, semiboundedness, and boundedness.

10) Ordering Points to Identify the Clustering Structure, Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, J&g Sander

Cluster analysis is a primary method for database mining. It is either used as a stand-alone tool to get insight into the distribution of a data set, e.g. to focus further analysis and data processing, or as a preprocessing step for other algorithms operating on the detected clusters. Almost all of the well-known clustering algorithms require input parameters which are hard to determine but have a significant influence on the clustering result. Furthermore, for many real-data sets there does not even exist a global parameter setting for which the result of the clustering algorithm describes the intrinsic clustering structure accurately. We introduce a new algorithm for the purpose of cluster analysis which does not produce a clustering of a data set explicitly; but instead creates an augmented ordering of the database representing its density-based clustering structure. This cluster-ordering contains information which is equivalent to the density-based clustering corresponding to a broad range of parameter settings.

3. Conclusion

We have to develop a clustering system which will overcome the limitations of existing clustering system. We have to extract the user interest through knowledge extraction. This information will be utilized by commercial business application for business or web advertisement purpose.

References

[1] A New Clustering and Preprocessing for Web Log Mining, B.Uma Maheswari, Dr. P.Sumathi.
[2] Data Mining for Web Personalization, Bamshad Mobasher
[3] New Path Filling Method on Data Preprocessing in Web Mining, Chungsheng Zhang & Liyan Zhuang, School of Mathematics and Computer Science, Inner Mongolia University for the Nationalities
[4] Mining User Tasks from Print Logs, Xin Li1, Lei Zhang1, Ping Luo2, Enhong Chen1, Guandong Xu3, Yu Zong4
[5] A Large-Scale Empirical Study on Software Reuse in Mobile. Israel J. Mojica, McAfee
[6] Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "Modelbased clustering and visualization of navigation patterns on a web site,"
[7] G. Pallis, L. Angelis, and A. Vakali, "Model-based cluster analysis for web users sessions,"
[8] L. Columbus, "Roundup of Mobile Apps & App Store Forecasts, 2013," Forbes, 9 June 2013; www.forbes.com/sites/louiscolumbus/ 2013/06/09/roundup-of-mobile-apps-app-store-forecasts-2013
[9] T. Li and L. Chen, "Internet of Things: Principles, Frameworks and Applications," Proc. Future Wireless Networks and Information Systems, pp. 477-842, 2008.
[10] G. Xu, Y. Zhang, J. Ma, and X. Zhou, "Discovering user access pattern based on probabilistic latent factor model,.
[11] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2001.
[12] Y. Zhao and G. Karypis, "Criterion functions for document clustering: experiments and analysis," Department of Computer Science, University of Minnesota," Technical Report, 2001.
[13] Bianco, G. Mardente, M. Mellia, M. Munafo, and L. Muscariello, "Web user session characterization via clustering techniques," in GLOBECOM'05, IEEE, Dec. 2005.
[14] D. Hand, H. Mannila, and P. Smyth, Principles of Data Mining. Cambridge, MA, USA: MIT Press, 2001