Web Usage Mining

MapReduce-based Emerging Pattern Mining in Hypergraph Learning

Xiuming Yu, Meijing Li and Keun Ho Ryu
Database/Bioinformatics Lab, Chungbuk National University, Cheongju, South Korea

1 RESEARCH PROBLEM

Web usage mining is a popular research area in data mining. As the rapid growth of internet, more and more log information is collected by the web servers around the world. It becomes difficult to extract useful information from these huge web log data.

Classic techniques of web usage mining are performed with low efficency in large number of web log data, because a lot of system resource is needed to deal with large computation.

Most techniques of web mining are performed based on assosiation rule mining or frequent pattern mining, which aim to find relationships among web pages or predicting the behavoir of web users. That's difficult to find some favourite web pages in different web users.

2 OUTLINE OF OBJECTIVES

In this research, we aim to find some favourite web pages in different web users in large web log data. We propose an efficient approach based on hypergraph learning by considering the programming model of MapReduce and the techniques of emerging pattern mining.

Comparison with traditional pagerank method in web mining, our proposed method which is based on hypergraph learning can save a lot of cost in computation by considering fewer nodes in generated graphs of page links.

The programming model of MapReduce is used to improve our proposed approach by counting visiting time which accessed by web users in one web page. It can greatly improve the efficiency of computation.

Emerging patterns in web log data can be represented as favourite web pages of different web users, we can obtain useful information by performing the techniques of emerging pattern mining in web log data.

3 STATE OF THE ART

Web Usage Mining (WUM), also known as web access, the web access pattern tracking can be defined as web page history, the mining task is a process of extracting interesting patterns in web access logs. There are so many techniques of web using mining have been proposed (Yu, 2011, Yu, 2012, Li, 2014). It is still a hot topic in the research area of data mining.

A hypergraph is a specific graph whose edges can connect more than two vertices or nodes. Because of this characteristic, hypergraph is suitable to solve the problem of high-order relation. In this research, we consider that it can reduce the number of vertices or nodes in generated graphs of page links, to save the cost of computation in the process of mining task. Hypergraph has been widely used in the area of data mining, and there are many articles (Xie, 2014, Pliakos, 2014, Chen, 2014, Yu, 2014) about hypergraph learning have been published.

Hadoop-MapReduce (Narayanan, 2014, Ghit, 2014, Doulkeridis, 2014) is a programming model and an associated implementation for processing and generating large datasets, which processes large datasets in parallel fashion. It allows users to not worry about the nuisance details of coordinating parallel sub-tasks and maintaining distributed file storage. It greatly increases user productivity while users process a large amount of parallel data. A MapReduce program consists of two parts: a map function, which processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function, which merges all of the intermediate values associated with the same intermediate key.

The discovery of class comparison or discrimination information is an important problem in the field of data mining. Recently, a novel kind of knowledge pattern, called EPs (emerging patterns), is introduced in (Dong, 1999). Emerging patterns, defined as multivariate features whose supports change significantly from one class to another, are

Yu X., Li M. and Ryu K..
Web Usage Mining - MapReduce-based Emerging Pattern Mining in Hypergraph Learning.

very useful as a mean of discovering distinctions between different classes of data. By aggregating the differentiating power of EPs, the constructed classifiers are usually more accurate than other existing state-of-the-art classifiers (Dong, 1999), (Li, 2015). EP-based classifiers also use this concept in analysis of high-dimensional biomedical data for rule discovery, diagnosis, prognosis and for better understanding of mechanics of the disease. Emerging patterns are patterns whose supports change significantly from one dataset to another. By using the techniques of emerging pattern mining, we can find emerging web pages in web log data. This technique has been applied in many articles (Sherhod, 2014, Yu, 2014) and it is still a hot topic in computer science.

4 METHODOLOGY

In this research, we get large pages as vertices in hypergraph from processed web log data, transform original data set into hypergraphs grouping by web users, and find emerging patterns in these hypergraphs based on the programming model of mapreduce. Our work flow is shown in Figure 1.

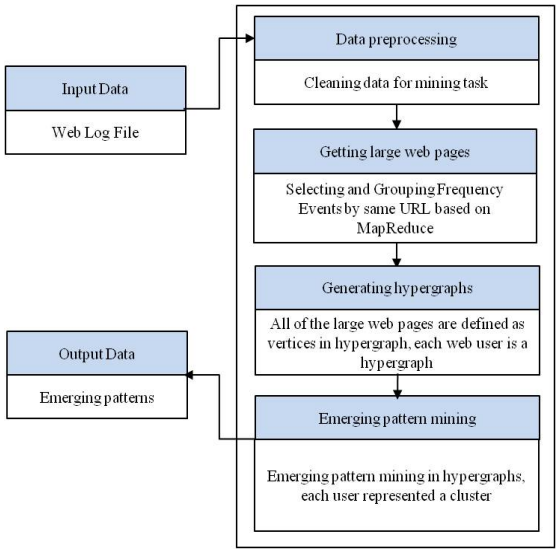

Figure 1: Work flow of our proposed approach.

4.1 Data Pre-Processing

Web log data is automatically recorded in web log files on web servers when web users access the web server through their browsers. Not all of the records sorted into the web log files are legal or necessary for the mining task, so before analysis of the web log data, the data cleaning phase needs to be implemented.

4.1.1 Removing Records with Missing Value Data

Some of the records sorted in web log file are not complete because some of the parameters of the records are lost. For example, if a click-through to a web page was executed while the web server was shut down, then in the log file only the IP address, user ID and access time are recorded; the method, URL, referrer and agent will be lost. This kind of record is illegal for our mining task, so these records must be removed.

4.1.2 Removing Illegal Records with Exception Status Numbers

Some illegal records are caused by errors in the requests or by the server. Although the records are intact, the activity did not execute normally. For example, records with the status numbers of 400 or 404 are caused by HTTP client errors, bad requests or a request not found. Records with status numbers 500 or 505, are caused by HTTP server errors, when the internal server cannot connect, or when the HTTP version is not supported. These kinds of data are illegal for our task, so the records must be removed.

4.1.3 Removing Irrelevant Records with No Significant URLs

Some URLs in the records consist of .txt, .jpg, .gif or .js extensions, which are automatically generated while a web page is requested. These records are irrelevant to our mining task, so they must be removed.

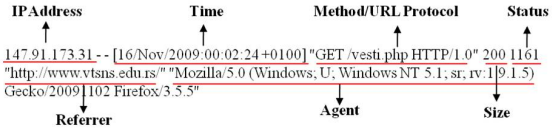

Figure 2: Common log format for web log data.

4.1.4 Selecting the Essential Attributes

As shown in the common log format of web log data in Figure 2, there are many attributes in one record, but for web-usage mining, not all the attributes are necessary. In this paper, the attributes for IP address, time, URL and referrer are essential attributes to our task, so they should remain but the rest of the attributes should be discarded.

4.2 Generate Large Page

A large page set is a set of frequent web pages. We define frequent web pages as whose support thresholds are greater than or equal to user specified minimum support threshold.

In this paper, a web log file denotes a data set, and one web page is defined as an event and LP denotes the set of web pages that are accessed by web users with enough frequency over a period of time. An important definition for generating LP is user session. Here the user session time is defined as one hour for simplicity. Then, the example data is grouped by one hour for each web user. According to the experimental data, candidate event types are extracted and their supports are calculated. To calculate the support count for each candidate, we need to count the visit times that are accessed by different web users. Finally, a user specified Minimum Support threshold for Large Page (MSLP) must be defined. MSLP denotes a kind of abstract level that is a degree of generalization. The support value will be determined by the proportion of web users accessing times of web pages. Selecting MSLP is very important; if it is low, then we can get a detailed event. If it is high, then we can get general events. In this example, MSLP is defined as 75%. In other words, if a web page is accessed by greater than 75% web users, then this web page can be denoted as a large page.

4.3 Generate Hypergraph

After generating large web pages, all of the large web pages are defined as vertices in hypergraph. In web log data, each web user is represented as a class which may contain hundreds or thousands of accessed records. For example, for one web user U1, he or she visited 20 web pages {p1, p2... p20} in a user session, we can get generalized graph of page links shown in Figure 3. We assume that large page set in Figure 3 is {p6, p12, p16, p19}, then for web user U1, his or her hypergraph of accessed pages can be described in Figure 4.

4.4 Emerging Pattern Ming in Generated Hypergraphs

After generating hypergraphs for all of web users, we aim to find emerging patterns in these hypergraphs. An example of hypergraphs of some web users is shown in Figure 5.

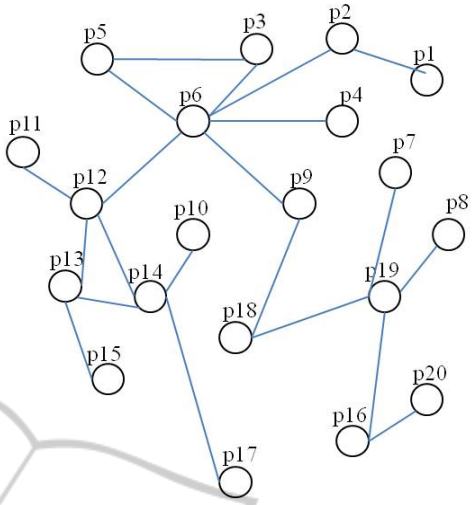

Figure 3: Generalized graph of page links of U1.

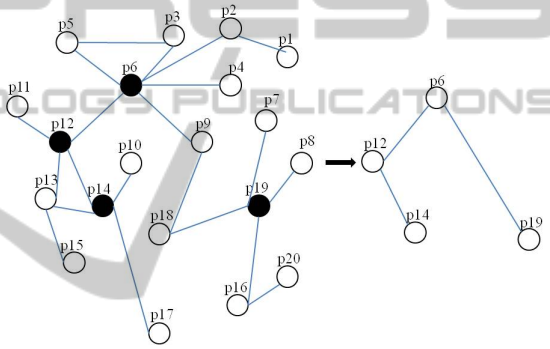

Figure 4: Hypergraph of user U1.

In the process of emerging pattern mining, we will use the idea of ρ -EP and JEP (Jumping Emerging Pattern). An example of emerging pattern mining is presented using the example data shown in Figure 5. We make the hypergraph in web user U1

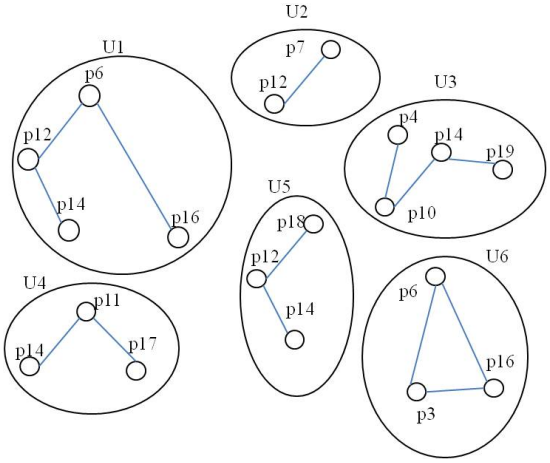

Figure 5: Example hypergraphs of some web users.

as class 1 and other hypergraphs in other web users as class 2. Table 1 shows that the items in these two classes. Table 2 lists out all possible EPs, their support and the growth rates of all EPs. If we set the minimum growth rate threshold ρ is =1.5, there are six EPs: Class 1 has four EPs ({p6}, {p12}, {p14}, {p16}) and five JEPs ({p6, p12}, {p6, p16}, {p12, p14}, {p6, p12, p14} and {p12, p6, p16}).

Table 1: A sample dataset with two classes.

Class 1	Class 2
p6	p3
p12	p4
p14	p6
p16	p7
p6, p12	p10
p6, p16	p11
p12, p14	p12
p6, p12, p14	p14
p12, p6, p16	p16
p6, p12, p14, p16	p17
	p18
	p19
	p7, p12
	…

Table 2: Discovering Eps for $\rho = 1.5$.

Items	Support		Growth Rate of EPs	
	Class 1	Class 2	Class 1	Class 2
p6	0.56	0.08	7	-
p12	0.56	0.17	3.3	-
p14	0.33	0.33	-	-
p16	0.33	0.08	4.1	-
p6, p12	0.33	0	∞	-
p6, p16	0.22	0.08	2.75	-
p12, p14	0.22	0.08	2.75	-
p6, p12, p14	0.11	0	∞	-
p12, p6, p16	0.11	0	∞	-
p6, p12, p14, p16	0	0	-	-

However, JEPs ({p6, p12}, {p6, p12, p14} and {p12, p6, p16}) are JEPs of class 1 with support of not zero values in class 1 and zero in class 2. It can be seen that these JEPs, appearing one in class 1 but zero time in class 2, are not useful for classification, especially when there are much noise present in the data.

5 EXPECTED OUTCOME

In this section, we will perform our proposed approach to web log data and do some experiments to proof the efficiency of our proposed approach.

5.1 Experiment Data

In the experiments, we used weblog data come from a NASA website (http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html), which is cited in [18], because that web log is large-scale temporal data with time points. The event file can be generated from a web log file including the URL requested, HTTP method requested, the IP address from which the request originated, and a timestamp.

Two web log files, NASA_access_log_Jul95 and NASA_access_log_Aug95, were used as our two experiment datasets. The first dataset was collected from 00:00:00 July 1, 1995, through 23:59:59 July 31, 1995 (a total of 31 days). The second dataset was collected from 00:00:00 August 1, 1995, through 23:59:59 August 31, 1995. The uncompressed content of the first dataset is 205.2 MB, and it contains 1,891,715 records. The uncompressed content of the second dataset is 167.8 MB, and it contains 1,569,898 records.

5.2 Analysis of Our Proposed Approach

In this section, we want to evaluate our proposed approach by three ways.

- Algorithm based on mapreduce programming model vs. traditional algorithm;
- Prediction based on hypergraph learning vs. generalized graph mining;
- Accuracy of emerging pattern mining;

For the way of Algorithm based on mapreduce programming model vs. traditional algorithm, we want to get expected outcome which our proposed approach performed with less time than traditional algorithm.

For the way of Prediction based on hypergraph learning vs. generalized graph mining, we want to see our proposed approach can save a lot of time in computation.

6 STAGE OF THE RESEARCH

In this research, we propose a process of getting large pages from processed web log data, define these large web pages as vertices in hypergraph, then transform original data set into hypergraphs grouping by web users, and find emerging patterns

in these hypergraphs based on the programming model of mapreduce. The main stage of the research is that how to apply the techniques of web usage mining, hypergraph learning and emerging pattern mining to web usage mining. And in the stage of experiment, we try to use right way to evaluate our proposed approach, and do sufficient experiments to prove our research points.

ACKNOWLEDGEMENTS

This research was supported by Export Promotion Technology Development Program, Ministry of Agriculture, Food and Rural Affairs (No.114083-3), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (No.2013R1A2A2A01068923) and (No. 2008-0062611).

REFERENCES

Xie, Y., Yu, H. M., and Hu, R. 2014. Probabilistic hypergraph based hash codes for social image search. *Journal of Zhejiang University SCIENCE* C, 15(7), 537-550.

Pliakos, K., and Kotropoulos, C. 2014. Simultaneous image tagging and geo-location prediction within hypergraph ranking framework. *In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on (pp. 6894-6898)*. IEEE.

Chen, X., Peng, Q., Han, L., Zhong, T., and Xu, T. 2014. An effective haplotype assembly algorithm based on hypergraph partitioning. *Journal of Theoretical Biology*.

Yu, J., Hong, C., Tao, D., and Wang, M. 2014. Semantic embedding for indoor scene recognition by weighted hypergraph learning. *Signal Processing*.

Narayanan, A. H., Krishnakumar, U., and Judy, M. V. 2014, An Enhanced MapReduce Framework for Solving Protein Folding Problem Using a Parallel Genetic Algorithm. *In ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol I (pp. 241-250)*. Springer International Publishing.

Ghit, B., Yigitbasi, N., Iosup, A., and Epema, D. 2014. Balanced Resource Allocations Across Multiple Dynamic MapReduce Clusters. *In ACM SIGMETRICS*.

Doulkeridis, C., and Nørvåg, K. 2014. A survey of large-scale analytical query processing in MapReduce. *The VLDB Journal, 23(3), 355-380.*

Dong, G., Zhang, X., Wong, L., and Li, J. 1999. CAEP: Classification by aggregating emerging patterns. *In Discovery Science (pp. 30-42)*. Springer Berlin Heidelberg.

Li, G., Law, R., Vu, H. Q., Rong, J., and Zhao, X. R. 2015. Identifying emerging hotel preferences using Emerging Pattern Mining technique. *Tourism Management, 46, 311-321.*

Sherhod, R., Judson, P., Hanser, T., Vessey, J., Webb, S. J., and Gillet, V. 2014. Emerging Pattern Mining to Aid Toxicological Knowledge Discovery. *Journal of chemical information and modeling.*

Yu, Y., Yan, K., Zhu, X., Wang, G., Luo, D., and Sood, S. 2014. Mining Emerging Patterns of PIU from Computer-Mediated Interaction Events. *In Agents and Data Mining Interaction (pp. 66-78)*. Springer Berlin Heidelberg.

Yu, X., Li, M., Kim, H., Lee, D. G., Park, J. S., and Ryu, K. H. 2011. A novel approach to mining access patterns. *In Awareness Science and Technology (iCAST), 2011 3rd International Conference on (pp. 350-355)*. IEEE.

Li, M., Yu, X., and Ryu, K. H. 2014. MapReduce-based web mining for prediction of web-user navigation. *Journal of Information Science, 0165551514544096.*

Yu, X., Li, M., Lee, D. G., Kim, K. D., and Ryu, K. H. 2012. Application of closed gap-constrained sequential pattern mining in web log data. *In Advances in Control and Communication (pp. 649-656)*. Springer Berlin Heidelberg.