WDS'13 Proceedings of Contributed Papers, Part I, 59–62, 2013. ISBN 978-80-7378-250-4 © MATFYZPRESS

A Short Survey of Web Data Mining

J. Just

Charles University in Prague, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. Usage of Web Services and Web-based application are rapidly growing at exponential rate resulting in a huge amount of data which can be used for mining. The web mining requires different methods used than in traditional data mining. It is due to the nature of the data used. The data on the web are in different form such as web serve logs, news pages and hyperlinks. Based on the type of the data web mining can be categorized into: Web structure mining, Web content mining and Web usage mining. In this paper, we give a short description of each web mining category. Furthermore, we describe subcategories with examples of the methods used to mine each subcategory. Reader will find preliminary focus of my future research in web usage mining section.

Introduction

The World Wide Web is daily used by millions of people. The data are added, edited and read on the web. It is the reason why the World Wide Web can be viewed as biggest database in the world. This dynamically changing database is good subject for data mining research. The data mining is basically discovering unknown patterns in large amount of data. If data mining techniques are used on web data, we are calling it web data mining or web mining. Two different approaches were originally taken to define Web mining. The first one was a 'process-centric view', which defined Web mining as a sequence of tasks [*Etzioni*, 1996]. The second one was a 'data-centric view', which defined Web mining in terms of the types of Web data that was being used in the mining process [*Cooley, Srivastava, and Mobasher,* 1997]. In recent days, the second approach is more acceptable in research community and will be used in this paper. Based on this approach, [*Kosala, Blockeel and Neven*, 2002] defined web data mining as the whole of data mining and related techniques that are used to automatically discover and extract information from web documents and services. The next section of this paper will give taxonomy of the web data mining and describe each web mining type in more detail with examples of the methods used. In web usage mining section, reader will find preliminary focus of my future research which will be targeting improvement of pre-processing algorithms.

Taxonomy of Web Mining

In this section I will present taxonomy of Web mining. This taxonomy is shown in Figure 1 which gives an overview of web mining categories. The web data mining can be divided based on the data which are used in mining process to three types: Web structure mining, Web content mining and Web usage mining.

Web structure mining uses hyperlinks or web page tree-like structure as source to discover knowledge. Web content mining extracts information from web pages content, this can also include pictures, audio files and videos. Web content mining can be categorized in two sections based on the point of the view: Information Retrieval and Database views. Web usage mining is the process where usage data such as web server logs, application server data for example application server logs tracking a various kinds of events and logs of events specially define in an application are used to find user activity patterns. Web content and Web structure can be combined in order to mine hyperlinks together with their content.

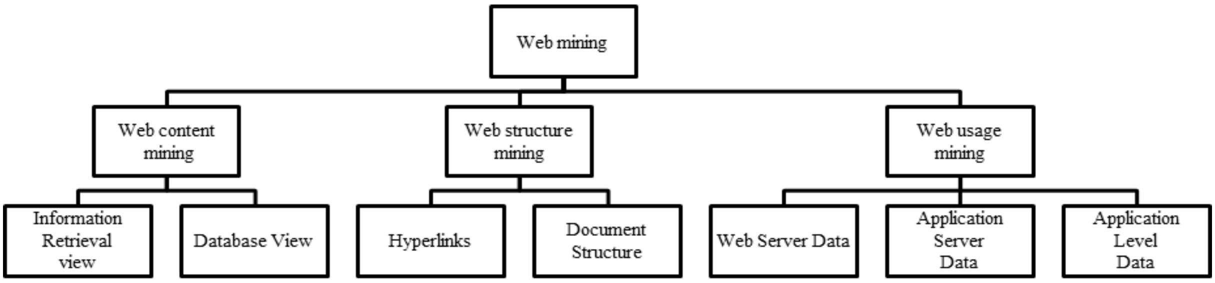

Figure 1. Web mining Taxonomy.

Web structure mining

Web structure mining discovers useful information based on hyperlinks structure of the Web Site. It explores topology of hyperlinks of the Web site. The structure of the World Wide Web can be viewed as graph where Web pages are nodes, and hyperlinks are edges connecting related pages. Then, Web structure mining is the process of discovering structure information from the Web. Web structure mining can be divided to two main parts based on the kind of structure information used: Hyperlinks and Document structure.

A Hyperlink is a structural unit which connects a location in Web page to either different place on the same Web page or to different Web page. The Intra -Document Hyperlink is the hyperlink pointing to place within the same Web page. On other hand, Inter-Document Hyperlink connects two different Web pages. There are a number of algorithms proposed based on the link analysis. Three important algorithms are PageRank [*Brin and Page,* 1998], Weighted PageRank (WPR) [*Xing and Ghorbani*, 2004] and Hypertext Induced Topic Search (HITS) [*Kleinberg,* 1999].

Document structure is a Web page organized in a tree-structured format, based on the various HTML and XML tags within the page. Here main effort is focused on automatically extracting document object model (DOM) structures out of documents. One of the algorithms is DTD-Miner [*Moh, Lim, and Ng,* 2000].

Web content mining

Web content mining is the process where useful information is extracted from the contents of Web documents. Content data correspond to the collection of facts a Web page was designed to pass on to the users. Data on the Web page can be in the form of text, video, pictures and audio. The text mining of Web content has been most researched. The multimedia mining is working with all forms of data such as video, pictures, audio and text. One of the examples of multimedia mining is a Unified Learning Framework for Auto Face Annotation by Mining Web Facial Images [*Wang, Chu Hong Hoi, and He*, 2012]

The Web content data are in the form of unstructured data such as free texts, semi-structured data such as HTML and XML documents, and a more structured data such as data in the tables or database generated HTML pages. Much of the data on the Web are in unstructured form. Web content mining can be divided in two sections based on the point of the view: Information Retrieval and Database views.

Information Retrieval view deals with unstructured and semi-structured documents. The unstructured documents are free texts such as news stories. There are mainly three main types of the unstructured documents pre-processing: the bag of words or vector representation [*Salton and McGill,* 1983], Latent Semantic Indexing (LSI) [*Deerwester, Dumais, Furnas, Landauer, and Harshman,* 1990] and using information about word position in the document [*Cohen,* 1995]. The bag of words or vector representation uses single words find in training corpus as feature which can be Boolean (occurs or not) or frequency based (number of occurrences in document). LSI is an indexing and retrieval method that uses a mathematical technique called singular value decomposition to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts [*Deerwester,* 1988].The information about word position in the document is using n-grams representation (word sequences of length up to n) [*Kargupta, Hamzaogiu, and Stafford,* 1997]. The semi-structured documents have as addition structure (HTML and hyperlink) when comparing to unstructured documents. The most of the researchers are focused on HTML structure inside the document. Methods used including Rule learning [*Fiirnkranz,* 1999] and Neural networks with reinforcement learning [*Shavlik and Eliassi-Rad,* 1999].

Database view on Web content mining is focused on techniques for organizing the semi-structured data on the Web into more structured collections of resources and using standard database querying mechanisms and data mining techniques to analyze it. There are two different approaches: Multilevel Databases and Web Query Systems. Multilevel Databases approach uses idea that the lowest level of the database contains semi-structured information such as hypertext documents stored in various Web repositories. The Meta data or generalizations are extracted from lower levels to the higher level(s) and organized in structured collections, i.e., relational or object-oriented databases. Example of Multilevel Database approach was the ARANEUS system [*Merialdo, Atzeni and Mecca,*1997] which extracted relevant information from hypertext documents and integrated them into higher-level derived Web Hypertexts which were generalizations of the notion of database views. The ARANEUS project was closed around year 2000. Web Query systems approach is using fact that many Web-based query systems and languages use standard database query languages such as SQL, structural information about Web documents, and even natural language processing for the queries that are used in World Wide Web searches [*Cooley et al.,* 1997]. One of the methods is Unstructured Query Language; it queries heterogeneous and semi-structured information on the Web using a labeled graph data model.

Web usage mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications [*Srivastava, Cooley, Deshpande,* 2000]. Web servers collect large amounts of data from the Web sites' usage. These data are stored in Web access log files. Together with the Web access log files, other data can be used in Web Usage Mining like the Web structure information, user profiles, refers logs (contain information about the referring pages for each page reference), etc. This data analysis can be used by organization or e-commerce for cross-marketing strategies across the products, effectiveness of promotions and other things. Following the standard data mining process, the overall Web usage mining process can be divided into three inter-dependent stages: data collection and pre-processing, pattern discovery, and pattern analysis [*Fayyad, Piatetsky-Shapiro, and Smyth*, 1996].

Data collection and pre-processing is the gathering of data and their transformation to format to which mining algorithms can be applied. It is the most important stage of usage mining because data are usually collected from multiple resources and across different channels. Pre-processing of collected data is challenging because of time consuming and intensive use of computation power. Usage data preparation presents a number of unique challenges which are leading to a variety of algorithms and heuristic techniques for pre-processing tasks such as data cleaning, user and session identification, pageview identification [Cooley, Srivastava, and Mobasher, 1999]. Data cleaning involves removal of uninteresting data and references of crawler navigations. User identification deals with identification of individual users with help of client-side cookies, a combination of IP addresses or other information such as user agents and referrers. Session identification is the process of separating the user activity record of each user into sessions, each representing a single visit to the site. The pageview is a collection of Web objects or resources representing a specific user event such as clicking on a link or viewing a product page. Identification of pageviews is dependent on the intra-page structure of the site, page contents and the underlying site domain knowledge. Due to describe complexity of pre-processing of web usage data, I will focus in my future research on improvements of the algorithms by implementing them in distributed environment using MPI.

Pattern Discovery mines knowledge from the datasets which are result of pre-processed raw logs. The data mining techniques to accomplish this are mainly association rule mining, sequential pattern mining and clustering. The association rule mining is based on identification of strong rules discovered in databases using different measures of interestingness. Apriori [*Agrawal and Srikant*, 1994] is the first and still the most used algorithm for this task. Sequential pattern mining is similar to association rule mining with addition of time element (order of events, i.e. clicks). GPS [*Srikant and Agrawal*, 1996] is one of the sequential pattern mining algorithms which is incorporated in IBM data mining products. Clustering is the division of data into groups of similar objects [*Gordon*, 1999]. CHAMELEON [*Karypis, Han and Kumar*, 1999] is clustering algorithm.

Pattern Analysis is used to understand, visualize and interpret the patterns which are results of patterns discover. WebVis [*Pitkow and Bharat*, 1994] is example of pattern analysis tool. WebViz allows the analyst to selectively analyze the portion of the Web that is of interest by filtering out the irrelevant portions. OLAP techniques are used to simplify the analysis of usage statistics from server access logs using data cubes [*Gray, Bosworth, Layman, and Pirahesh*, 1996].

Web usage mining can be also divided into three main categories based on origin of the data: Web Server data, Application Server Data and Application Level data. Web Server data are data collected from Web Server logs such as IP addresses, page reference and access time. Application Server Data are data coming from commercial application servers and are used to track various kinds of business events. Application Level data are data which are resulting from events specially define in an application. It is important to point out that many end application requires a combination of one or more of the techniques applied in the above categories [*Srivastava et al.*, 2000].

Conclusion

The World Wide Web is rapidly growing in the size every day and it is becoming important part of every day life. Therefore, web data mining is more and more important research topic. In this paper, we gave a short survey of web data mining. The main three categories of web data mining listed are agreed by most of the paper authors which we used in this paper. However, further division of the categories is not standardized. This variety of the web mining categories division should be unified in the future to avoid confusion of the students and researchers.

I will be focusing in my future research on web usage mining mainly on data collection and pre-processing stage. The goal of the research will be improving performance and implementation of pre-processing algorithms in distributed environment using MPI.

References

R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules, In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proceedings of the Twentieth International Conference on Very Large Data Bases*, VLDB, pp. 487-499. Morgan Kaufmann, 1994.

S. Brin and L. Page, The anatomy of a large scale hypertextual web search engine, Comput. Network ISDN Syst., 30, pp. 107-117, 1998.

W. Cohen, Learning to classify english text with ilp methods, In Advances in Inductive Logic Programming (Ed. L. De Raedt), IOS Press, 1995.

R. Cooley, J. Srivastava, and B. Mobasher, Web mining: Information and pattern discovery on the world wide web, In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997.

R. Cooley, J. Srivastava, and B. Mobasher, Data preparation for mining world wide web browsing patterns, Knowledge and Information systems, 1(1), pp. 5-32, 1999.

S. Deerwester, Improving Information Retrieval with Latent Semantic Indexing, Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, pp. 36–40, 1988.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, 41(6), pp. 391-407, 1990.

O. Etzioni, The world wide web: Quagmire or goldmine, Communications of the ACM, 39(i1), pp. 65-68,1996.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, From data mining to knowledge discovery: An overview, *In Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, pp. 1-34,1996.

J. Fiirnkranz, Exploiting structural information for text classification on the www, *In Advances in Intelligent Data Analysis*, Third International Symposium, IDA-99, pp. 487-498, 1999.

D. Florescu, A. Y. Levy, and A. O. Mendelzon, Database techniques for the world-wide web: A survey. SIGMOD Record, 27(3):59-74, 1998.

A. Gordon, Classification. Chapman and Hall, 1999.

J. Gray, A. Bosworth, A. Layman, and H. Pirahesh, Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. In *IEEE 12th International Conference on Data Engineering,*pp. 152-159,1996.

H. Kargupta, I. Hamzaogiu, and B. Stafford, Distributed data mining using an agent based architecture, In *Proceedings of Knowledge Discovery And Data Mining*, pp. 211-214. AAAI Press, 1997.

G. Karypis, E. Han, V. Kumar, CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling, *IEEE Computer* 32, pp. 68-75, August 1999.

J. Kleinberg, Authoritative sources in a hyperlinked environment, *J. ACM*, 46, pp. 604-632, 1999.

R. Kosala and H. Blockeel, Web mining research: A survey, *SIGKDD Explorations 2*, no. 1, pp. 1-15, 2000.

R. Kosala, H. Blockeel, en F. Neven, An overview of web mining, Dealing with the data flood. Mining data, text and multimedia, (Meij, Jeroen, ed.), *STT/Beweton*, The Hague, pp. 480-497, 2002.

P. Merialdo, P. Atzeni and G. Mecca, Semistructured and structured data in the web: Going back and forth. In *Proceedings of the Workshop on the Management of Semistructured Data (in conjunction with ACM SIG-MOD),* 1997.

C.H. Moh, E.P. Lim, and W.K. Ng. DTD-Miner, A Tool for Mining DTD from XML Documents, *WECWIS*, 2000.

J. Pitkow and Krishna K. Bharat, Webviz: A tool for world-wide web access log analysis. In *First International WWW Conference*, 1994.

G. Salton and M. McGill, Introduction to Modern Information Retrieval. McGraw Hill, 1983.

J. Shavlik, and T. Eliassi-Rad, Intelligent agents for web-based tasks: An advice-taking approach, In *Working Notes of the AAAI/ICML-98 Workshop on Learning for Text Categorization*, Madison, WI, pp. 588-589, 1999.

R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements", *EDBT*, 1996.

J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *SIGKDD Explorations*, 1(2), pp. 12–23, 2000.

D. Wang, S. Chu Hong Hoi, and Y. He, A unified learning framework for auto face annotation by mining web facial images. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (CIKM '12). ACM, New York, NY, USA, pp. 1392-1401,2012.

W. Xing and A. Ghorbani, Weighted PageRank algorithm, *Proceeding of the 2nd Annual Conference on Communication Networks and Services Research, May* 19-21, IEEE Computer Society, Washington DC., USA., pp. 305-314, 2004.