International Journal of Information Technology and Knowledge Management
July-December 2011, Volume 4, No. 2, pp. 351-354

THE ENHANCING TECHNIQUE FOR DATA MINING TO WEB USAGE MINING

Sudhir B. Jagtap[1], Subhendu Kumar Pani[2] and G. N. Shinde[3*]

In the context of data mining the feature size is very large and it is believed that it needs a bigger population. Hence, this translates directly into higher computational load. With the huge amount of information available online, the web mining is a fertile area of research which applies the data mining techniques. It relates to several research communities such as Database, Information Retrieval and Visualization. We have categorized web data mining into threes areas; web content mining, web structure mining and web usage mining. In this research area that is receiving increasing attention from the data mining community.In this paper, we discuss some data mining techniques that could be used to enhance web-based learning environments.

Keywords: Data Mining, Web Mining, Web Data, Information Retrieval.

1. Introduction

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and resources. In addition, with the transformation of the Web into the primary tool for electronic commerce, it is imperative for organizations and companies, who have invested millions in Internet and intranet technologies, to track and analyze user access patterns. These factors give rise to the necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge both across the Internet and in particular Web localities. Many organizations and corporations provide information and services on the web such as automated customer support,on-line shopping, and a myriad of resources and applications.web based applications and environments for electronic commerce, distance education, on-line collaboration, news broadcasts etc., are becoming common practice and widespread. The WWW [2] is becoming ubiquitous and an ordinary tool for everyday activities of common people, from a child sharing music files with friends to a senior receiving photographs and messages from grandchildren across the world. It is typical to see web pages for courses in all fields taught at universities and colleges providing course and related resources even if these courses are delivered in traditional classrooms. It is not surprising that the web is the means of choice to architect modern advanced distance education systems.

[1]Swami Vivekanand Mahavidyalaya, Udgir Email: sudhir.
[2]RCMA, BPUT Bhubaneswar-751004, Orissa, India
[3]Indira Gandhi College, Nanded-431603,India
[*]Address for Communication:
Email:[1]jagtap 7@gmail.com, [2]Subhendu_pani@rediffmail.com, [3]shindegn @ yahoo. co.in

There are several important issues, unique to the Web paradigm that comes into play if sophisticated types of analyses are to be done on server side data collections. These include the necessity of integrating various data sources such as server access logs, user registration or profile information; resolving difficulties in the identification of users due to missing unique key attributes in collected data; and the importance of identifying user sessions or transactions from usage data, site topologies, and models of user behavior.

Feature selection is one of the very important preprocessing tasks of data mining and knowledge discovery in databases. It is obvious that the quality of knowledge discovered depends strongly on the quality of data being mined. No matter how intelligent a data mining algorithm is, it will fail to discover high quality knowledge if it is applied to low quality data. This has motivated the development of several feature selection algorithm. The main goal of the feature selection is to select a subset of relevant feature out of all available features of the data being mined.

In general feature selection can be visualized as the selection of a subset of features that will reduce the probability of misrecognition in the operational (classification) phase. A feature selection scheme based on the availability of a set of labeled samples from each of the predefined set of classes is referred to as feature selection in a supervised environment. But in practice one often comes across situations where the samples are unlabeled or at best imperfectly labeled. Again feature selection is very important for machine learning due to its potentiality of speeding up and reducing the costs of the followed stage of concept learning or instance classification, and improving the performance of the learned results. Therefore how to select the optimal feature subset to describe a learning

system is always regarded as a key technology in the domain of machine learning. Furthermore among the different categories of feature selection algorithms the genetic algorithm (GA) is a rather recent development. The GA-based feature selection is very essential because of the following reasons. Suppose there are 'm' numbers of features in the data being mined. Then the total number of candidate feature subsets is 2m that is the size of search space of the feature selection grows exponentially with the number of features.

The GA is biologically inspired and has many mechanisms mimicking natural evaluation. It has a great deal of potentiality in scientific and engineering optimization on search problems. The pioneering work by Siedlecki and Sklansky demonstrated evidence for the superiority of GA compared to representative classical algorithms. Subsequently many literatures that have shown advantages of GAs for feature selection were published. However, GAs are not guaranteed to find an optimal solution and their effectively is determined largely by the population size 'n'. As the population size increases, the GA has a better chance of finding the global solution, but at the same time the computation load also increases as a function of sizes of the population. With serial GAs, we have to choose between getting a good result with a high confidence and pay a high computational cost on loosen the confidence requirement and get (possibly poor) results fast. In contrast parallel GAs can keep the quality of the results high and find them fast because, using parallel machines larger populations can be processed in less time. This keeps the confidence factor high and the response time low, opening opportunities to apply genetic algorithms in time constrained applications. Additionally, parallel GAs may evolve several different independent solutions that may be recombined at later stages to form better solutions.

2. USEFUL DATA MINING TASKS

What is needed are summarization trends and patterns that can be interpreted by educators delivering their courses online. Due to the importance of e-commerce and the lucrative opportunities behind understanding on-line customer purchasing behaviors, there is tremendous research effort in developing data mining algorithms and system tailored for e-business related web usage data mining[10]. In addition to descriptive statistical analysis provided by most web access log analysis tools such as calculating hit frequency, average, median etc. length and duration of sessions and other limited low-level statistical measures, there have been some data mining approaches adapted specifically for web usages mining. The most used methods are association rules mining, clustering classification, sequential pattern analysis and dependency modeling [11], as well as predication. These techniques are primarily used for personalization, system improvement such as web caching and networking traffic improvement, site modification, and marketing intelligence .None of these applications, however, was tailored to distance learning, but the methods generals enough that e-learning systems could benefit from them. Association rules generation is the discovery of relationships between items in transaction. It is typically used from market basket analysis to discovered rules of the form x% of customers who buy item A and B also buy item C.'' Clustering is an unsupervised grouping of objects while classification is a supervised grouping. In web mining, the objects could be uses events, sessions, pages, etc. Sequential pattern analysis is similar to association rules but takes in to account the sequence of events. In other words, the fact that a page A is required before another page B is captured in the pattern discovered. All these techniques were designed for knowledge discovered from very large databases of numerical data [12] and were adapted for web mining and application in on-line business with success.

3. WEB DATA MINING

3.1 Overview

The web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services [5]. This area of research is so huge today partly due to the interest in e-commerce. This phenomenon partly creates confusion what constitutes Web mining and when comparing research in this area. Similar to [5], we suggest decomposing Web mining into these subtasks, namely

1. Resource finding: the task of retrieving intended Web documents.
2. Information selection and pre-processing: automatically selecting and pre-processing specific information from retrieved Web resources.
3. Generalization: automatically discovers general patterns at individual Web sites as well as across multiple sites.
4. Analysis: Validations and/or interpretation of the mined patterns

We should also note that humans play an important role in the information or knowledge discovery process on the web since the web is an interactive medium. This is especially important for validation and/or interpretation in step 4. So, interactive query-triggered knowledge discovery is as important as the more automatic data triggered knowledge discovery. However, we exclude the knowledge discovery done manually by humans. Thus, Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the web data. It implicitly covers the standard process of

knowledge discovery in databases (KDD) [2]. We could simply view web mining as an extension of KDD that is applied on the Web data. From the KDD point of view, the information and knowledge terms are interchangeable [3]. There is a close relationship between data mining, machine learning and advanced data analysis [4]. Web mining is often associated with IR or IE.However,web mining or information discovery on the web not the same as IR or IE [1].

3.2 Web Content Mining

Web content mining describes the automatic search of information resources available online [6], and involves mining web data contents. In the web mining domain, web content mining essentially is an analog of data mining techniques for relational databases, since it is possible to find similar types of knowledge from the unstructured data residing in web documents. The web document usually contains several types of data, such as text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured such as HTML documents or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. The unstructured characteristic of web data forces the web content mining towards a more complicated approach.

The web content mining is differentiated from two different points of view [7]: Information Retrieval View and Database View.R.kosla et al [8] summarized the research works done for unstructured data and semi-structured data from information retrieval view. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in training corpus as features. For the semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structures between the documents for document representation. As for the database view, in order to have the better information management and querying on the web, the mining always tries to infer the structure of the web site of to transform a web site to become a database. Multimedia data mining is part of the content mining, which is engaged to mine the high-level information and knowledge from large online multimedia sources.

3.3 Web Structure Mining

Most of the web information retrieval tools only use the textual information, while ignore the link information that could be very valuable. The goal of web Structure mining is to generate structural summary about the web site and web page. Technically, web content mining mainly focuses on the structure of inner-document, while web Structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different web sites. Web structure mining can also have another direction-discovering the structure of web document itself. This type of structure mining can be used to reveal the structure (schema) of web pages; this would be good for navigation purpose and make it possible to compare/ integrate web page schemes. This type of structure mining will facilitate introducing database techniques for accessing information in web pages by providing a reference schema. The detailed works on it can be referred to[9]

The structural information generated from Web structure mining includes the follows: the information measuring the frequency of the local links in the Web tuples in a web table containing links that are interior and the links that are within the same document: the information measuring the frequency of web tuples in a web table that contains links that are global and the links that span different web sites.web structure mining has a nature relation with the web content mining, since it is very likely that the Web documents contain links, and they both use the real or primary data on the web. Its quiet often to combine these two mining tasks in an application.

3.4 Web Usage Mining

Web usage mining tries to discovery the useful information from the secondary data derived from the interactions of the users while surfing on the web. It focuses on the techniques that could predict user's behavior while the user interacts with web. M. Spiliopoulou abstract the potential strategic aims in each domain in to mining goal as: predication of the user's behavior within the site, comparison between expected and actual web site usages, adjustment of the web site to the interests of its users. There are no definite distinctions between the web usage mining and other two categories. In the process if data presentation of web usage mining, the web site topology will as the information sources,Which interacts web usage mining with the web content mining and web structure mining moreover the clustering in the process of pattern discovery is a bridge to web content and structure mining from usage mining.There are lots of works have been done in the IR, Database, Intelligent Agents and topology, which provides a sound function for the web content, web structure mining. Web usages mining is a relative new research area, and gains more and more attentions in recent years. I will have a detailed introduction in the next section about mining, based on some up-to-date research works.

4. CONCLUSION

Web mining is a rapid growing research area. We survey the researches in the area of web mining. Three recognized types of web data mining are introduced generally Web content

mining is related but different from data mining and text mining. Web data are mainly semi-structured and/or unstructured. Web content mining requires creative applications of data mining and/or text mining techniques and also its own unique approaches. Due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still present many challenging research problems.

REFERENCES

[1] G.Salton and M.Mc Gill, "Introduction to Modern Information Retrival", McGraw Hill, 1983.

[2] U. Fayyad, G. Piatetsky – Shapiro, P. Smyth, "Data Mining to Knowledge Discovery: An Overview". In advances in Knowledge Discovery and Data Mining, pages 1-34.AAA Press, 1996.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Toward a Unifying Framework", In Proceeding of the Second int. Conference on Knowledge Discovery and Data mining", pages 82-88, 1996.

[4] M.A.Hearst.Untangling "Text Data Mining", In Proceedings of ACL'99", the 37th Annual Meeting of the Association For Computational Linguistics,

[5] O. Etzioni, "The World Wide Web: Quagmire or Gold Mine", Communications of the ACM, 39(11):65-68,1996.

[6] S.K.Madria, S.S.Bhowmick, W.K, Ng, and E.P.Lim, "Research Issues in Web Data Mining", In Proceedings of Data Ware Housing and Knowledge Discovery, First International Conference, DaWak'99, pages 303-312, 1999.

[7] R.Cooley, B.Mobasher, and J.Srivastava "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.

[8] R.kosala, H.Blockeel "Web Mining Research: A Survey".

[9] S.K.Madria, S.S Bhowmick, W.K.Ng, and E.P.Lim "Research Issues in Web Data Mining", In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWak'99, pages301-312, 1999

[10] M.N.Garofalaski, R.Rastogi, S.Seshadri, K.Shim, "Data Mining and the Web:past, Present and Future", Proceedings of WIDM99, Kansas City,U.S.A.,1999.

[11] J.Srivastava, R.Cooley, M.Deshpande, "P.tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations", 1, No.2,Jan.2000.

[12] J.han and M.kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publisher, 2001.