ISSN 2250-3153

Extraction of user specified web knowledge using Spatial Data Mining

Priyanka Tiwari

Shri Ram institute Of Technology Jabalpur
Jabalpur (M.P.), India

Abstract- Nowadays the World Wide Web has becoming one of the most comprehensive information resources. It probably, if not always, covers the information need for any user. Those differences make it challenging to fully use Web information in an effective and efficient manner. Web mining is the application of data mining techniques to extract knowledge from web data including web documents, hyperlinks, log usage of website etc. In this paper we extract data from web using spatial data mining. Spatial data mining is the process of trying to find patterns in geographic data. Spatial data mining is the application of data mining techniques. Spatial data mining follows along the same functions in data mining, with the end objective to find patterns in geography. In this paper we provide an introduction of spatial data mining as well as web. Then we focus on how data is extracted from web using some preprocessing techniques or some steps. It describes a method to extract useful information from a web page using spatial data mining. We are extracting hyperlinks and email from single and multiple websites that's why it is using spatial data mining because in spatial mining data is extracted from different locations. Different websites will have different web servers means different locations. This method includes some preprocessing steps to extract information. That extracted information will be knowledge.

Index Terms- spatial data mining, web mining, preprocessing techniques, hyperlinks, email extraction

I. INTRODUCTION

Data mining, a branch of computer science and artificial intelligence, is the process of extracting patterns from data. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. The related terms data dredging, data fishing and data snooping refer to the use of data mining techniques to sample portions of the larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These techniques can, however, be used in the creation of new hypotheses to test against the larger data populations. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

A. *Spatial data mining*

Spatial data mining is considered a more complicated challenge than traditional mining because of the difficulties associated with analyzing objects with concrete existences in space and time. As with standard data mining, spatial data mining is used primarily in the world of marketing and retail. Spatial data mining is the process of trying to find patterns in geographic data. Most commonly used in retail, it has grown out of the field of data mining, which initially focused on finding patterns in textual and numerical electronic information. It is a technique for making decisions about where to open what kind of store. It can help inform these decisions by processing pre-existing data about what factors motivate consumers to go to one place and not another. Spatial data mining is the application of data mining techniques to spatial data. Spatial data mining follows along the same functions in data mining, with the end objective to find patterns in geography. It has grown out of the field of data mining, which initially focused on finding patterns in textual and numerical electronic information. Spatial data mining is considered a more complicated challenge than traditional mining because of the difficulties associated with analyzing objects with concrete existences in space and time. Spatial data mining is the application of data mining techniques to spatial data. Spatial data mining follows along the same functions in data mining, with the end objective to find patterns in geography. data mining and Geographic Information Systems have existed as two separate technologies, each with its own methods, traditions and approaches to visualization and data analysis. As with standard data mining, spatial data mining is used primarily in the world of marketing and retail. It is a technique for making decisions about where to open what kind of store. It can help inform these decisions by processing pre-existing data about what factors motivate consumers to go to one place and not another.

B. Spatial Data

Also known as geospatial data *or* geographic information it is the data or information that identifies the geographic location of features and boundaries on Earth, such as natural or constructed features, oceans, and more. Spatial data is usually stored as coordinates and topology, and is data that can be mapped. Spatial data is often accessed, manipulated or analyzed through Geographic Information Systems . Spatial data is about information that has several dimensions. It is sometimes referred

to as spatial data. It includes both geospatial data and structo-spatial data.

C. Process of Spatial Mining

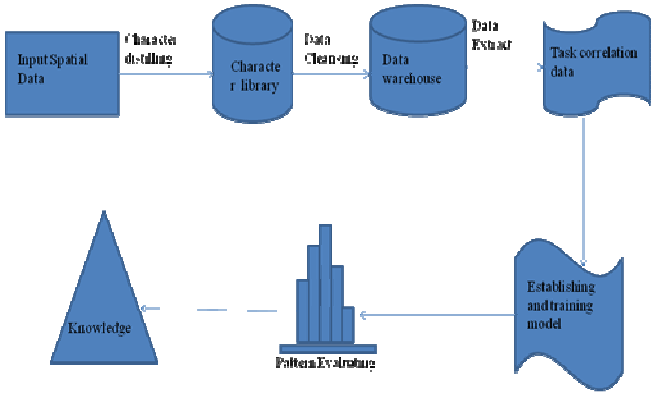

Figure 1: Process of Spatial Data Mining

1. *Input Spatial Data:* It is the input to the spatial mining process.
2. *Character Library:* Feature extraction and Feature database establishment. Spatial datum is the object of data mining. Because the existing databases, data warehouses, OLAP and data-mining technologies can not be used to process spatial datum, it is necessary to convert non-structuring spatial datum into structuring relational table or extensional table firstly, in order to take full advantage of such technologies. Store the features extracted out and the original pixel value in the feature data base, which is one source of data mining case set.
3. *Data Warehouse:* This definition of the data warehouse focuses on data storage. The main source of the data is cleaned, transformed, catalogued and made available for use by managers and other business professionals for data mining, online analytical processing, market research and decision support. However, the means to retrieve and analyze data, to extract, transform and load data, and to manage the data dictionary are also considered essential components of a data warehousing system. Many references to data warehousing use this broader context. Thus, an expanded definition for data warehousing includes business intelligence tools, tools to extract, transform and load data into the repository, and tools to manage and retrieve metadata.
4. *Data Extraction:* Data extraction is the act or process of retrieving (binary) data out of (usually unstructured or poorly structured) data sources for further data processing or data storage (data migration). The import into the intermediate extracting system is thus usually followed by data transformation and possibly the addition of metadata prior to export to another stage in the data workflow.
5. *Establishing and Training Model:* Data mining model is an abstract data structure, created, filled and provided query by data mining algorithm. Through the input and output strings of specified data mining model and the use of mining algorithm, data mining algorithm can set up an empty data mining model structure; and through inserting the training set into data mining model, the training of the model is implemented.
6. *Pattern Evaluating:* A data mining system can discover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user. The challenge is to develop techniques able to assess the interestingness of discovered patterns. Evaluating the mined-out model, discovering the hidden knowledge. Usually large numbers of models are to be found in a training set, and among these models some cases are low in support and reliability. So, it is necessary to evaluate the model discovered. Through evaluation, eliminate unreasonable models and meanwhile, store the considerably reliable models into the data mining model as “knowledge”, to prepare for further analysis and forecast by using knowledge.
7. *Knowledge:* All the useful information at the end will be knowledge.

D. Web mining

Nowadays, the World Wide Web has becoming one of the most comprehensive information resources. It probably, if not always, covers the information need for any user. However, the Web demonstrates many radical differences to traditional information containers such as databases, in schema, volume, topic-coherence. Those differences make it challenging to fully use Web information in an effective and efficient manner. Web mining is right for this need In fact, Web mining can be considered as the applications of the general data mining techniques to the Web. However, the intrinsic properties of the Web make us have to tailor and extend the traditional methodologies considerably. Firstly, even though Web contains huge volume of data, it is distributed on the internet. Before mining, we need to gather the Web document together. Secondly, Web pages are semi-structured, in order for easy processing, documents should be extracted and represented into some format. Thirdly, Web information tends to be of diversity in meaning, training or testing data set should be large enough. Even though the difficulties above, the Web also provides other ways to support mining, for example, the links among Web pages are important resource to be used. In general web mining tasks can be classified into three categories: Web content mining, web structure mining, web usage mining. All of the three categories focus on the process of knowledge discovery of useful information from the web. Each of them focuses on different mining object from the web.

II. PROBLEM STATEMENT

Nowadays, the World Wide Web has becoming one of the most comprehensive information resources. It probably, if not always, covers the information need for any user. Before mining, we need to gather the Web document together. Secondly, Web pages are semi-structured, in order for easy processing; documents should be extracted and represented into some format. In this paper the method has to extract all hyperlinks and email from a web page as well as multiple web pages.

III. PROPOSED METHOD

This paper addresses how to filter knowledge from a web page as well as multiple web pages. We propose a method to extract some predefine knowledge by some preprocessing techniques. Instead of removing noise or redundant data from a web page this method prefers to extract useful information based on our method. Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns already present in the data, the target dataset must be large enough to contain these patterns while remaining concise enough to be mined in an acceptable timeframe. Pre-process is essential to analyze the multivariate datasets before clustering or data mining. The target set is then cleaned. Cleaning removes the observations with noise and missing data.

Data preprocessing techniques can improve the quality of data, there by helping to improve the accuracy and efficiency of the mining process. Data preprocessing is a very important step in knowledge discovery process. This method of extraction of data includes some steps like giving input, read input, pattern evaluation, cleaning and then data extraction.

A. Read Input

In this step reading of input is done provided by the user in a web page format.

B. Pattern Evaluation:

A data mining system can discover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user. The challenge is to develop techniques able to assess the interestingness of discovered patterns. Usually large numbers of models are to be found in a training set, and among these models some cases are low in support and reliability. So, it is necessary to evaluate the model discovered. Through evaluation, eliminate unreasonable models and meanwhile, store the considerably reliable models into the data mining model as knowledge to prepare for further analysis and forecast by using knowledge.

C. Cleaning:

Data cleaning is one of the most important preprocessing techniques. It is used to clean redundant data. Data cleaning removes entries which are unused to data analyzing and mining. Data cleansing or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant etc. parts of the data and then replacing, modifying or deleting this dirty data.

D. Data extraction:

Data extraction is the act or process of retrieving (binary) data out of (usually unstructured or poorly structured) data sources for further data processing or data storage (data migration). The import into the intermediate extracting system is thus usually followed by data transformation and possibly the addition of metadata prior to export to another stage in the data workflow. Usually, the term data extraction is applied when (experimental) data is first imported into a computer from primary sources, like measuring or recording devices. Today's electronic devices will usually present a electrical connector through which raw data' can be streamed into a personal computer.

IV. PERFOFMANCE EVALUATION

To evaluate the performance of method we use a proper sequence of preprocessing steps. First user enters a url then the whole data of the web page is extracted and stored in html format. Then data is readied in the html format. We are extracting all the url's and email id present in the web page. For this we match the patterns of url's and email id's in the web page data. After pattern evaluation the data other than url's and email id's is cleaned. Because that data is redundant means unused. Then at the end data extraction is done means the data we want all the url's and email id's is extracted from the web page. This extracted data is knowledge.

A. Experiment Result:

In above experiment we use an example of a web site. When user enters a website url then the data of the website is stored in html format. Then in the pattern evaluation phase the pattern of the url's and the pattern of the email's is evaluated. In this example there is website of a collage and all url's and email id's is extracted from a web page. These url and email are the knowledge.

Figure 2: Output of the extraction of url's from single web page.

Figure 3: Output of the extraction of email's from single web page.

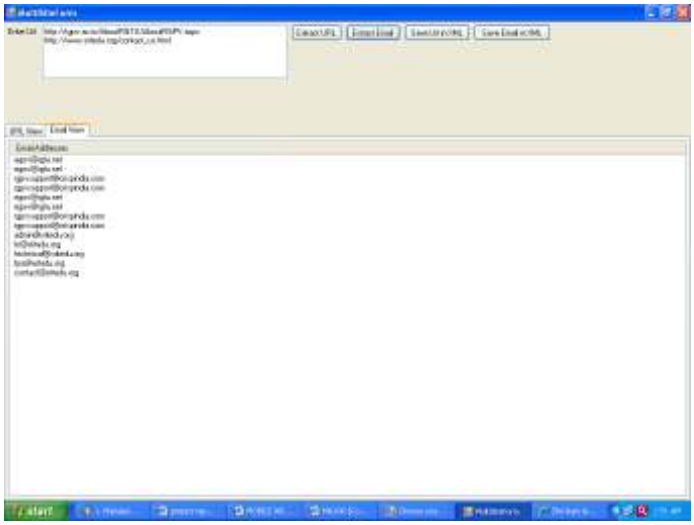

Figure 4: Output of the extraction of email's from multiple web pages

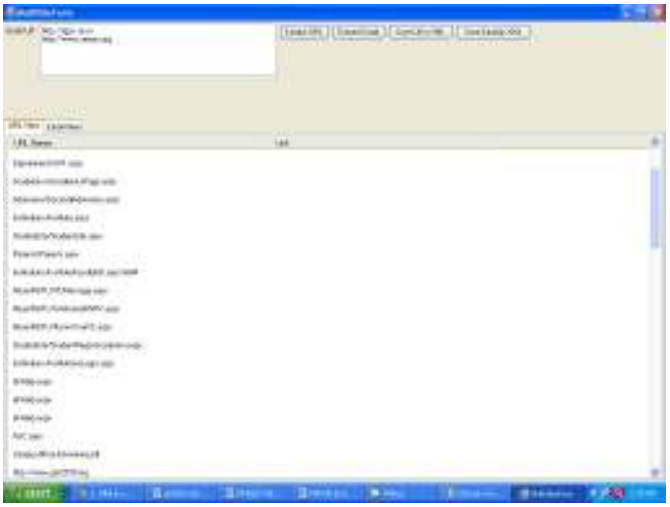

Figure 5: Output of the extraction of url's from multiple web pages

V. CONCLUSION

The research area of spatial mining has focused on the category of Web mining. Later in the paper when I had discussed spatial mining, and web mining. In this paper we provide an efficient method to extract useful information from a web page an multiple web pages. We explore the extraction of some predefine knowledge from a website. Our approach is based on some preprocessing steps in data mining. The method identifies all url's and email's presented at the web page. First the method is applied to single web page after that to multiple web pages because in spatial data mining data is extracted from different locations. Multiple websites have different web servers. The result obtains from a real world website shows that the result is correct.

REFERENCES

[1] He yue-shun ding qiu technique research on data mining based on semi structured data source. Journal of hurbin institute of technology.

[2] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, ACM SIGKDD Explorations Newsletter, June 2000, Volume 2 Issue 1.

[3] Jidong Wang, Zheng Chen, Li Tao, Wei-Ying Ma, Liu Wenyin, Ranking User's Relevance to a Topic through Link Analysis on Web Logs, WIDM' 02, November 2002.

[4] Han, J., Kamber, M. Kamber. Data mining: concepts and techniques. Morgan Kaufmann Publishers, 2000.

[5] Wang Jicheng, Huang Yuan, Wu Gangshan, Zhang Fuyan. Web mining: knowledge discovery on the Web. Systems, Man, and Cybernetics, 1999.

[6] Cooley, R.; Mobasher, B.; Srivastava, J.; Web mining: information and pattern discovery on the World Wide Web. Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference. Page(s):558 – 567 - 3-8 Nov. 1997.

[7] L. Getoor, Link Mining: A New Data Mining Challenge. SIGKDD Explorations, vol. 4, issue 2, 2003.

[8] O. Etzioni. The world wide web: Quagmire or gold mine. Communications of the ACM, 39(11):65-68, 2007.

[9] www.google.com

[10] www.wikipedia.com

First Author – Priyanka Tiwari, M.E- C.S.E. IVth Semester, Shri Ram institute of Technology Jabalpur (M.P.), India.
Email- Priyanka8684@gmail.com