INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

A Survey on Web Research for Data Mining

Gaurav Saini[1]
gauravhpror@gmail.com[1]

Abstract— Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. The process of extracting useful information from the contents of web document is data mining. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. The large and dynamic information source that is structurally complex and ever growing, the World Wide Web is fertile ground for data mining principles, or Web mining. Here, it defines the information retrieval and information extraction from web and making research for data mining.

I. INTRODUCTION

Data mining is the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses, or other information repositories. It is also popularly referred to as knowledge discovery in databases (KDD). Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, etc [1]. Data mining is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too much time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.
Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line [3].

The Foundations of Data Mining
Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery [7]. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

The Scope of Data Mining
Data mining derives its name from the similarities between searching for valuable business information in a large database. For example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviors**. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.
- **Automated discovery of previously unknown patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions [5, 11].

Databases can be larger in both depth and breadth:

- **More columns**. Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.
- **More rows**. Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

The most commonly used techniques in data mining are:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- **Genetic algorithms**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms. The appendix to this white paper provides a glossary of data mining terms [8].

II. ARCHITECTURE FOR DATA MINING

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on [9].

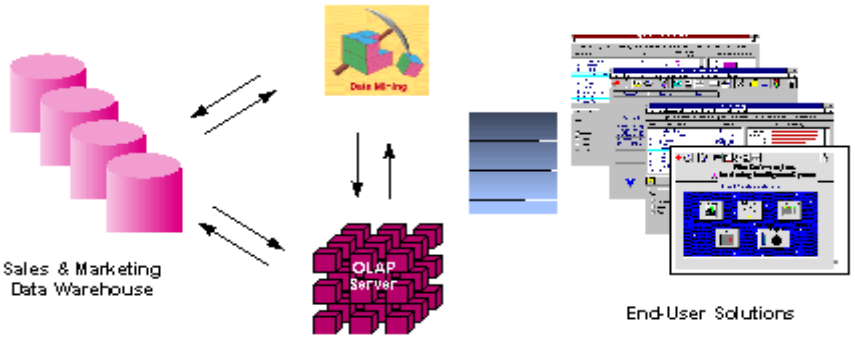

Figure 1 - Integrated Data Mining Architecture

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions [5].

This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans.

III. WEB DATA MINING

The term Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest.

Data Mining is done through various types of data mining software. These can be simple data mining software or highly specific for detailed and extensive tasks that will be sifting through more information to pick out finer bits of information. For example, if a company is looking for information on doctors including their emails, fax, telephone, location, etc., this information can be mined through one of these data mining software programs. This information collection through data mining has allowed companies to make thousands and thousands of dollars in revenues by being able to better use the internet to gain business intelligence that helps companies make vital business decisions.

Before this data mining software came into being, different businesses used to collect information from recorded data sources. But the bulk of this information is too much too daunting and time consuming to gather by going through all the records, therefore the approach of computer based data mining came into being and has gained huge popularity to now become a necessity for the survival of most businesses [2]. This collected information is used to gain more knowledge and based on the findings and analysis of the information make predictions as to what would be the best choice and the right approach to move toward on a particular issue. Web data mining is not only focused to gain business information but is also used by various organizational departments to make the right predictions and decisions for things like business development, work flow, production processes and more by going through the business models derived from the data mining.

A strategic analysis department can undermine their client archives with data mining software to determine what offers they need to send to what clients for maximum conversions rates. For example, a company is thinking about launching cotton shirts as their new product. Through their client database, they can clearly determine as to how many clients have placed orders for cotton shirts over the last year and how much revenue such orders have brought to the company [4].

After having a hold on such analysis, the company can make their decisions about which offers to send both to those clients who had placed orders on the cotton shirts and those who had not. This makes sure that the organization heads in

the right direction in their marketing and not goes through a trial and error phase to learn the hard facts by spending money needlessly. These analytical facts also shed light as to what the percentage of customers is who can move from your company to your competitor. The data mining also empowers companies to keep a record of fraudulent payments which can all be researched and studied through data mining. This information can help develop more advanced and protective methods that can be undertaken to prevent such events from happening. Buying trends shown through web data mining can help you to make forecast on your inventories as well. This is a direct analysis, which will empower the organization to fill in their stocks appropriately for each month depending on the predictions they have laid out through this analysis of buying trends. The data mining technology is going through a huge evolution and new and better techniques are made available all the time to gather whatever information is required. Web data mining technology is opening avenues on not just gathering data but it is also raising a lot of concerns related to data security [6, 4].

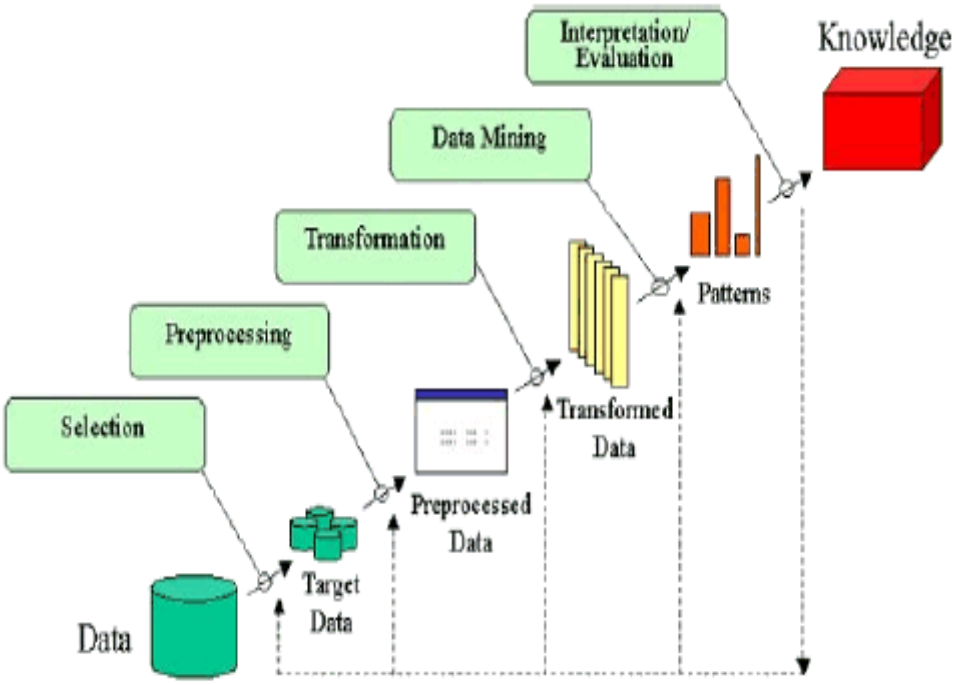

Figure 2: Data mining as part of KDD process

IV. RELATED WORK

Recent research has mostly focused on Web usage analysis, partly because of its applicability in e-business. We expect privacy issues, distributed Web mining, and Semantic Web mining to attract equal, if not more, interest from the research community. Increased use of Web mining techniques will require that privacy issues be addressed, however. Similarly, aggregating data in a central site and then mining it is rarely scalable, hence the need for distributed mining techniques. Finally, researchers will need to leverage the semantic information the Semantic Web provides. Exposing content semantics and the link explicitly can help in many tasks; including mining the hidden Web is defines or means that is, data stored in databases and not accessible through search engines. As new data is published every day, the Web's utility as an information source will continue to grow.

V. CONCLUSION

The topic of this paper defining the web mining and data researches can be found, particularly in the commercial area as common data mining applications. However, there is space for research in the public scientific field and especially the functionality of Clear Research, which is literally determined for researchers, might be a source of inspiration.

ACKNOWLEDGMENT

Thanks to my Guide and family member who always support, help and guide me during my dissertation. Special thanks to my father who always support my innovative ideas.

REFERENCES

[1] Yong Shi, Yuqing Song and Aidong Zhang. A shrinking-based approach for multidimensional data analysis. In the 29th VLDB conference, September 2003.
[2] M Eirinaki and M Vazirgiannis, "Web Mining for Web Personalization," ACM Trans. Internet Technology, vol. 3, no.1,2003.
[3] Diligenti M., Gori M., Maggini M.: "A Unified Probabilistic Framework for Web Page Scoring Systems" , IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 1, pp. 4 – 16, 2004.
[4] J.B. Schafer, J. Konstan, and J. Riedl, "Electronic Commerce Recommender Applications," J. Data Mining and Knowledge Discovery, vol. 5, nos. 1/2, 2000.
[5] M. Spiliopoulou, "Web Usage Mining for Site Evaluation," Comm. ACM, , vol. 43, no. 8, 2000.
[6] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.
[7] Kosala R., Blockeel H.:, "Web Mining Research: A Survey", ACM SIGKDD Explorations Newsletter, Vol. 2, No. 1, pp. 1 – 15, 2000.
[8] Xiaoqing Zheng,Yiling Gu,Yinsheng Li,"Data Extraction from Web Pages Based on Structural Semantic Entropy", International World Wide Web conference Committee (IW3C2),April 2012.
[9] Ananthi.J, "A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014.
[10] Monika Yadav, Mr. Pradeep Mittal, "Web Mining: An Introduction",IJARCSSE, March – 2013.
[11] Govind Murari Upadhyay, Kanika Dhingra, "Web Content Mining: Its Techniques and Uses", IJARCSSE, November, 2013.

