International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

Mining Contenders from Huge Unstructured Datasets

S. Latha & S.A.MD. Noorulla Baig
Email: mlatha51995@gmail.com & Email: noorullabaig@gmail.com

[1]Student ,M.SC(CS), RIIMS Tirupati.

[2]M. tech Associate Professor, Dept. of Computer Science, RIIMS Tirupati.

Abstract:

Data mining is the popular area of the research which facilitates the business improvement process such as mining user preference, mining web information's to get opinion about the product or services and mining the competitors of a specific business. In the current competitive business scenario, there is a need to analyze the competitive features and factors of an item that most affect its competitiveness. The evaluation of competitiveness always uses the customer opinions in terms of reviews, ratings and abundant source of information's from the web and other sources. In this paper, a formal definition of the competitive mining is describes with its related works. Finally the paper provides the challenges and importance in the competitor mining tasks with optimal improvements.

Index Terms – Data mining, Web mining, Information Search and Retrieval, Competitor Mining, Firm analysis, Electronic commerce.

1. INTRODUCTION

The strategic importance of detecting and observing business competitors is an inevitable research, which motivated by several business challenges. Monitoring and identifying firm's competitors have studied in the earlier work. Data mining is the optimal way of handling such huge sua for mining competitors. Item reviews form online offer rich information about customers' opinions and interest to get a general idea regarding competitors. However, it is generally difficult to understand all reviews in different websites for competitive products and obtain insightful suggestions manually. In the earlier works in the literatures, many authors analyzed such big customer data intelligently and efficiently [1] [2] [3]. For example, a lot of studies about online reviews were stated to gather item opinion analysis from online reviews in different levels. However, most researchers in this field ignore how to make their findings be seamlessly utilized to the competitor mining process. Recently, a limited number of researches were noted to utilize the latest development in artificial intelligence (AI) and data mining in the e-commerce applications [4]. These studies help designers to understand a large amount of customer requirements in online reviews for product improvements. But, these discussions are far from sufficient and some potential problems. These have not been fully investigated such as, with product online reviews, how to conduct a thorough competitor analysis. Actually, in a typical scenario of a customer-driven new product design (NPD), the strengths and weakness are often analyzed exhaustively for probable opportunities to succeed in the fierce market competition.

The rest of this research is structured as follows. In Section 2, relevant studies are briefly reviewed. Section 3 outlines the problems in the existing work. In Section 4, the comparative study is given. In Section 5, concludes this survey.

2. LITERATURE REVIEW

This research provides the various methodologies implemented to mine competitors with reference to customer lifetime value, relationship, opinion and behavior using data mining techniques. The web growth has resulted in widespread usage of many applications like e-commerce and other service oriented applications. This varied usage of web applications has provided an enormous amount of data at one's disposal. Data is the input that exists in its raw form resulting in information for further processing. With huge amount of data, organizations faced the crucial challenge of extracting very useful information from them. This has led to the concept of data mining. Mining competitor's of a given item, the most influenced factor of the item which satisfies the customer need can be extracted from the data that is typically stored in the database. This section gives two types of literatures such as competitor mining and unstructured data management.

A. Unstructured data management:

The data collected from the web are sometimes semi-structured or unstructured. The semi-structured data's are in the format of XML, JSON etc., the unstructured data sources are in a different format, which is not fall under any predefined category. When managing thousands of customers, business will have difficulty sustaining the rising costs created by interactions among people. However, if all customer data is inserted into a database, the resulting records will provide a detailed profile of these customers and their interactions with one another, and will be an important resource for businesses that wish to probe customer data, customer needs, and customer satisfaction levels.

Data mining uses transaction data to gain a better understanding of customers and effectively discover hidden knowledge through the insertion of business intelligence into the process of competitor mining. In paper [5] authors argued that data mining is an approach to assist companies in developing more effective strategies to meet the competitions in the market. Data warehousing is useful and accurate for assembling a business' dispersed heterogeneous data and providing unified convenient information access technique. Data mining technology can be used to transform hidden knowledge into manifest knowledge. A competitor mining from web data system is extremely flexible. Therefore, one of the best competitive strategies is the successful utilization of web data for timely decision support.

Customer data for competitor mining is collected through several methods, which is usually unstructured; however, most data mining technologies can only handle structured data. Therefore, during competitor mining process, unstructured data is not taken into account and much valuable service information is lost. Structured systems are those where the data and the computing activity is predetermined and well-defined. Unstructured systems are those that have no predetermined

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

form or structure and are usually full of textual data. Typical unstructured systems include email, reports, letters, and other communications. The following figure 1.0 shows the unstructured and structured systems.

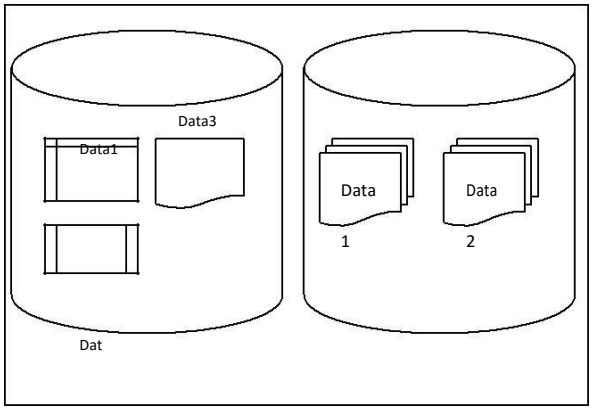

Fig 1.0 structured and un-structured systems

Information extraction from web pages is an active research area. Researchers have been developing various solutions from all kinds of perspectives to provide the comparative report. Many web information extraction systems rely on human users to provide marked samples so that the data extraction rules could be learned. Because of the supervised learning process, semi-automatic systems usually have higher accuracy than fully automatic systems that have no human intervention. Semi-automatic methods are not suitable for large-scale web applications [6] that need to extract data from thousands of web sites. Also web sites tend to change their web page formats frequently, which will make the previous generated extraction rules invalid, further limiting the usability of semi-automatic methods. That's why many more recent work [7], [8] focus on fully or nearly fully automatic solutions.

Web information extraction can be at the record level or data unit level. The former treat each data record as a single data unit while the latter go one step further to extract detailed data units within the data records. Record level extraction method generally involves identifying the data regions that contain all the records, and then partitioning the data regions into

individual records. Structured data extraction from Web pages has been studied extensively. Early works on manually constructed wrappers were found difficult to maintain and be applied to different Web sites, because they are very labor intensive.

Semi-automatic method known as wrapper induction [9] was proposed to tackle this problem. These methods need some labeled pages in the target domain as input to perform the induction. Thus, they still have limitation for large-scale applications. To overcome the above drawbacks, fully automatic methods have been developed. In paper [10] authors addressed the problem of unsupervised Web data extraction using a fully-automatic information extraction tool called ViPER. The tool is able to extract and separate data exhibiting recurring structures out of a single Web page with high accuracy by identifying tandem repeats and using visual context information. However, this technique lacks performance in few datasets.

B. Competitor Mining:

The earlier work on the competitor mining utilized the text data to collect comparative evidences between two items. But, the comparative evidences are based on the assumptions, which may not always exist. Competitor identification is referred to as a classification process through which competitors of a focal firm are identified based on "relevant similarities".

Authors in [11] developed an automatic system that discovers competing companies from public information sources. In this system data is crawled from text and it uses transformation oriented learning to obtain appropriate data normalization, combines structured and unstructured information sources, uses probabilistic modeling to represent models of linked data, and succeeds in autonomously discovering competitors. Bayesian network for competitor identification technique is used. The authors also introduced the iterative graph reconstruction process for inference in relational data, and shown that it leads to improvements in performance. To find the competitors, the authors used machine learning algorithms and probabilistic approaches. They also validate system results and deploy it on the web as a powerful analytic tool for individual and institutional investors. However, the technique has many problems like finding alliances and market demands using the machine learning approach.In the paper [12] [13], authors presented a formal definition of the competitiveness between two items. Authors used many domains and handled many shortcomings of previous works. In this paper, the author considered the position of the items in the multi-dimensional feature space, and the preferences and opinions of the users. However, the technique addressed many problems like finding the top-k competitors of a given item and handling structured data.

Authors in [14] proposed a new online metrics for competitor relationship predicting. This is based on the content, firm links and website log to measure the presence of online isomorphism, here the Competitive isomorphism, which is a phenomenon of competing firms becoming similar as they mimic each other under common market services. Through different analysis they find that predictive models for competitor identification based on online metrics are largely superior to those using offline data. The technique is combined the online and offline metrics to boost the predictive performance. The system also performed the ranking process with the considerations of likelihood.

Several works in the same strategy in literature have discussed the need for accurate identification of competitors and provided theoretical frameworks for that. Given the expected isomorphism between competing firms, the process of competitor identification through pair-wise analysis of similarities between focal and target firms is well founded. The unit of analysis is a pair of firms since competitor relationship s seen as a unique interaction between the pair. Authors in [15] have suggested frameworks for manual identification of competitors. The manual nature of these frameworks makes them very costly for competitor identification over a large number of focal and target firms, and over time.In the paper authors attempts to accomplish a novel task of mining competitive information with respect to an entity , the entity such as a company, product or person from the web. The authors proposed an algorithm called “CoMiner”, which first extracts a set of comparative candidates of the input entity and then ranks them according to the comparability, and finally extracts the competitive fields. But the CoMiner specifically developed to support for specific domain. However the effort for the further domains is still challenging.

The Authors in [17] have proposed ranking methods to give the competitor in a ranked way. They have used data from location-based social media. Authors proposed the use of Page-Rank model and it’s variant to obtain the Competitive Rank of firms. However mining competitors from the social media developed many privacy related issues.

C. Baseline Algorithm for Competitor Mining:

There are three base algorithms were used for the competitor mining such as Naïve base algorithm, GMiner, Cminer and CMiner++.

3. COMPARATIVE STUDY

The existing competitor mining algorithms such as Naïve base, GMiner, CMiner and Cminer ++ has been evaluated and compared with the time complexity. The fig 2.0 shows the computational time taken for the individual algorithm is plotted.

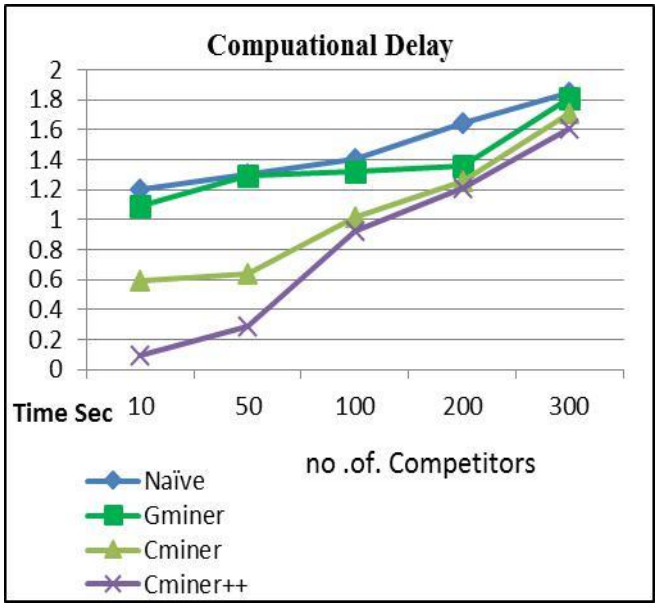

Fig 2.0 Computational efficiency analysis chart

4. CONCLUSION

Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Machine learning algorithms are widely used in various applications. Every business related application uses data mining techniques. To improve such business or providing appropriate competitors for the business to the user need the support of web mining techniques. The competitor mining is one such a way to analyze competitors for the selected items. In this paper, we gave a comprehensive analysis of the competitor mining algorithms with its advantages and drawbacks. Finally, the CMiner++ yielded least computation time when comparing others. The most important features and process are not considered in the all baseline algorithms. This can be improved in the further researches.

REFERENCES

[1] Ding, X., Liu, B., Yu, P.S., 2008. A holistic lexicon-based approach to opinion mining. In: Proceedings of the WSDM'08.

[2] Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. 26 (3), 12:1–12:34

[3] Chen, L., Qi, L., Wang, F., 2012. Comparison of feature-level learning methods for mining online consumer reviews. Expert Syst. Appl. 39 (10), 9588–9601.

[4] Zhan, J., Loh, H.T., Liu, Y., 2009. Gather customer concerns from online product

reviews – a text summarization approach. Expert Syst. Appl. 36 (2 Part 1), 2107–2115

[5] Jin, Jian, Ping Ji, and Rui Gu. "Identifying comparative customer requirements from product online reviews for competitor analysis." *Engineering Applications of Artificial Intelligence* 49 (2016): 61-73.

[6] Saxena, Prateek, David Molnar, and Benjamin Livshits. "SCRIPTGARD: automatic context-sensitive sanitization for large-scale legacy web applications." *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, 2011.

[7] Ghamisi, Pedram, Jon Atli Benediktsson, and Johannes R. Sveinsson. "Automatic spectral–spatial classification framework based on attribute profiles and supervised feature extraction." *IEEE Transactions on Geoscience and Remote Sensing* 52.9 (2014): 5771-5782.

[8] Petrucci, Giulio. "Information extraction for learning expressive ontologies." In *European Semantic Web Conference*, pp. 740-750. Springer, Cham, 2015.

[9] Gentile, Anna Lisa, Ziqi Zhang, Isabelle Augenstein, and Fabio Ciravegna. "Unsupervised wrapper induction using linked data." In *Proceedings of the seventh international conference on Knowledge capture*, pp. 41-48. ACM, 2013.

[10] K. Simon and G. Lausen, "ViPER: Augmenting Automatic Information Extraction with Visual Perceptions," Proc. 14th ACM Int'l Conf. Information and Knowledge Management, pp. 381-388, 2005