International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3
ISSN: 2454-4248
363 – 366

A Survey on Mining Top-k Competitors from Large Unstructured Dataset Using k_means Clustering Algorithm and Sentiment Analysis Approach

Miss Ankita A. Kushwah
Department of Information Technology,
Bharati Vidyapeeth Deemed University College of Engineering, Pune, India
ankita.kushwah92@gmail.com

Prof.Y.C. Kulkarni
Department of Information Technology,
Bharati Vidyapeeth Deemed University College of Engineering, Pune, India
yckulkarni@bvucoep.edu.in

***Abstract*—** Along line of research has shown the vital significance of recognizing and observing company's contestants. In the framework of this activity various questions are emerge like: In what way we formalize and measure the competitiveness between two items? Who are the most important competitors of a specified item? What are the various features of an item that act on competitiveness? Inspired by this issue, the advertising and administration group have concentrated on observational strategies for competitor distinguishing proof and in addition on techniques for examining known contenders. Surviving examination on the previous has concentrated on mining near articulations (e.g.one product is superior then other product) from the web or other documentary sources. Despite the fact that such articulations can without a doubt be indications of strength, they are truant in numerous spaces. By surveying the various papers, we found the conclusion of basic significance of the competitiveness between two items on the basis of market segments.

Keywords— *Data mining, Web mining, Information Search and Retrieval, Electronic commerce*

*****

I. INTRODUCTION

Competitive intelligence initially classifies the potential risk and chances by collecting the information about the context to handle the manager in making tactical decisions for an organization. Many organization recognizes the significance of competitive intelligence in enterprise risk management and decision support system. They also invest a great amount of money in competitive intelligence. The fundamental significance of customer choices, e.g., in correlation with new product expansion procedures. These procedures are broadly affirmed in marketing research. Usually customer choices are evaluated through conjoint analysis using online or paper-pencil survey. Though, this type of choices can highly price with reference to time and money[1][2].

Distinguishing potential risks, it is essential for companies to evaluate the information about their competitors' products and tactics. Depending on this information, a company can examine the comparative weaknesses and strengths of its own products and can then plan new pointed products and promote to countervail those of its competitors. Conventionally, the information about contestants come from press releases, like analyst reports and trade journals and also from competitors' websites and news sites. Tactlessly, this information is typically created by the company that fabricates the product. As a result, the amount of accessible information is partial and its objectivity is doubtful. The unavailability of adequate information sources about competitors highly confines the competence of competitive intelligence[1][2].

By examining the environment of the company or group of companies denotes the quality of business. To validated information about the competitor relations, people utilizes various options, like enquiring business associates, analyzing news articles, searching the web, take a part in business conventions, etc. While the company summarizing resources have truncated search efforts and made some business relationship information accessible, due to their restricted resources and variances in criteria, they can endure a scalability issue and deliver partial information [3]. Existing research based on mining comparative articulations (e.g. "product A is superior than product B") from the web or other documentary sources [3], [4], [5]. However, this articulation can certainly be sign of competitiveness and they are missing in numerous domains. For example, while competing brand names at the company level (e.g. Google vs Yahoo or Sony vs Panasonic). While comparing these patterns, it can be found by simply questioning on the web. But, it is easy to classify mainstream domains where such facts are tremendously uncommon, such as jewelry, hotels, restaurants and furniture. Inspired by these limitations, we present a new description of the competitiveness between two items on the basis of market sectors.

II. LITETATURE SURVEY.

This survey provides the basic significance of competitiveness between two items. It also provides the numerous methodology to mine competitors in respect of customer lifespan value, relationship, opinion and activities using data mining techniques. The web development has

 ISSN: 2454-4248
363 – 366

resulted in huge usage of many applications like e-commerce and other service-oriented applications.

i. Z. Ma, G. Pant, and O. R. L. Sheng, "Mining competitor relationships from online news: A network-based approach," Electronic Commerce Research and Applications, 2011.

In this paper, authors propose an approach to classifying competitors which is significant for businesses. Author present an approach of graph theoretic measures and machine learning methodologies to conclude competitor correlations on the basis of structure of an intercompany system derived from company quotations in online news artefacts. Author proposed a neutral language method in that it does not uses natural language processing methods on news. Author's methodology involves a given collection of news stories that is controlled by company and classify the company quotations in the news stories. Author have developed a directed and weighted intercompany network on the basis of company's quotations. Author have constructed the intercompany network which distinguishes network structure by classifying four types of attributes from network structure. For identify company's pairs it labels some arbitrarily selected pairs with respect to Hoover's and Mergent. Author have identified attributes and competitors on the basis of Hoover's and Mergent to instruct the classifiers to conclude the competitor relationship between a pair of companies that are joined in the network and compute competitor performance from various metrics (e.g. precision, recall, false positive rate, F1, etc.) with four dissimilar classifiers.

In this paper, Author concluded a method that utilized company quotations i.e. co-occurrences in online news to form an intercompany network that structural attributes are used to conclude competitor correlation between the companies. In this paper, author discover the citation-based network that conveys latent information and structural properties can be used to conclude competitor relationship. The main drawback of this paper is that this method is not examine with new stories written in another language and the intercompany network has not exploring the future prediction of competitor relationships.

ii. R. Decker and M. Trusov, "Estimating aggregate consumer preferences from online product reviews," International Journal of Research in Marketing, vol. 27, no. 4, pp. 293–307, 2010.

In this paper, author discussed on customer reviews that are widely accessible on the internet for huge number of product classifications. The experts and cons articulated in this manner discover individually observed the strengths and weaknesses of the particular products, while the typically allocated product rankings represent their complete valuation. The important query from this point is by what means to turn the accessible plentitude of individual customer opinions into aggregate customer first choice which can be utilize in product expansion or improvement procedures. To overcome these shortcomings author presents an econometric framework that can be utilized to the cited type of data and used natural language processing methodologies. The suggested procedure simplifies the evaluation of parameters allow implications on the comparative outcome of product features and brand names on the complete evaluation of the products.

In this paper, author proposes the concept of data pre-processing and attribute extraction steps. This method includes review-wise partitioning of the experts and cons into individual words and phrases. After that it removes those words and phrases that neither explicit nor implicit product review. Next, it combines redundant words and phrases. Later, it modifies the implicit candidate attributes into explicit attribute. Afterward, it merges the synonyms. Then it removes those candidate attributes that are unfrequented. Finally, the binary coding of the experts/con summarizes using the accessible set of attributes.

In this paper, using product data, author present a methodology of econometric preference analysis. This methodology determines the effects of attributes which have overall evaluation of the products. For example, in the opinion of future product development. The basic significance of this approach is that each product review can be denoted by a combination of positively ("the experts") and negatively ("the cons") evaluated characteristics which is accomplished by overall valuation of the product. This paper includes three formulations of the models i.e. homogeneous preference model, heterogeneous model with discrete distribution of preferences, heterogeneous model with a continuous distribution of preferences.

In this paper, author concluded an econometric structure that can be useful to turn the plentitude of individual customer views made accessible by online product reviews into collective customer preference data. The experts/con summaries that naturally accompany full-text reviews with the help of natural language processing methods. The considered methods, and the negative binomial regression (NBR) model enables the estimation of meaningful parameters. These parameters permit for the implications on the comparative effect of useful attributes and brand names on product evaluations and purchase decisions. The future should be devoted to the development of powerful filters for detecting fake reviews and to further automation of time consuming data, pre-processing and attribute extraction.

iii. C. W.-K. Leung, S. C.-F. Chan, F.-L. Chung, and G. Ngai, "A probabilistic rating inference framework for mining

 ISSN: 2454-4248
363 – 366

user preferences from reviews," World Wide Web, vol. 14, no. 2, pp. 187–215, 2011.

This paper proposes a novel Probabilistic Rating Inference Framework, well known as PREF, for mining user choices from reviews and then map out such choices onto numerical rating scales. PREF utilizes existing linguistic processing methods to extract opinion words and product attributes from reviews. It then estimates the sentimental orientations (SO) and strength of the opinion words by means of our proposed relative-frequency-based technique. In this paper, author presents a Pref technique i.e. a probabilistic rating inference framework. Pref is a probabilistic rating inference framework model which helps to developed and to support the integration of sentiment analysis and CF. It includes four steps: 1) Data preparation: This method processes the user opinions for the subsequent analysis. Different preprocessing methods may be needed to depend on the data sources. For example, if user preferences are downloaded as HTML pages then the HTML tags and non-textual contents then, they contain are removed in this step. 2) Feature extraction: This method extracts the attributes of opinion words and product attributes from the reviews. Author presented an NLP technique of POS tagging and two heuristics i.e. negation tagging and feature generalization for the identification of interesting features. 3) Opinion dictionary construction: This method consists of opinion words and their estimated SO and the strength of their SO. 4) Rating inference: This method is used to defining the overall SO of a review on the basis of SO of the opinion words it contains. It has been viewed as a multi-category classification task in that the class labels are scalar ratings, such as 1 to 5 "stars". Though, dissimilar from standard topic-based classification for the reason that class labels in the rating inference task are ordered, and there exist different degrees of similarity between the class labels.

In this paper, author concluded a probabilistic rating inference framework that also recognized as Pref. It includes the key tasks and design problems in each step. Author also showed an extensive experimental study for confirming the effectiveness of Pref which shows higher performance to various interrelated algorithms. Additionally, Author's results indicate that Pref does not depend on a large training corpus to function which can be an important concern when applying sentiment analysis to new domains where labeled (rated) reviews are limited. A feasible solution to model text-based CF as an information retrieval issue having reviews written by a target user as the query and those written by other similar users as the relevant documents from which recommendations for the target user can be generated.

iv. "E. Marrese-Taylor, J. D. Vel´asquez, F. Bravo-Marquez, and Y. Matsuo,"Identifying customer preferences about tourism products using an aspect-based opinion mining approach," Procedia Computer Science, vol. 22, pp. 182–191, 2013"

In this paper, author present and extend an approach of Bing Liu's aspect-based opinion mining methodology to utilize it to the tourism domain. Author also suggested a method for considering a novel alternative to uncover customer opinions regarding tourism products, specifically hotels and restaurants using opinions accessible on the web as reviews. To estimate this suggestion, author also conducted an experiment using hotel and restaurant reviews found from Trip Advisor. This outcome displayed that tourism product reviews available on web sites comprise valuable information about customer opinions that can be extracted using an aspect-based opinion mining method. In this paper, author proposes an extension of Liu's aspect-based opinion mining methodology in accordance to the tourism domain. This extension shows that the user signifies differently to the dissimilar types of products when writing reviews on the web. Think on a generic product, that shows to the conceptual commodity formed by an industry. This product shows the extensive variability of real forms which each having the same functionality. In the literature, author including Kotler frequently identify these generic products using two types: 1) Physical goods and 2) intangible services. Though, these products are not in separate types but rather in range or time with pure service on one terminal point and pure objective on the other. Existing work in product review mining involving Liu's, is presents on physical product reviews. In these types of reviews, users usually go in sequence to the point and talk directly about product features and after that they liked it or did not liked it. Additionally, some people will care about problems like who has designed or manufactured the product. Though, for the different types of products other kind of phenomena occur. For example, when a user writes a review, he possibly comments not only on movie elements, but also on movie correlated people. Consider other terminal point, tourism products like restaurants, provides physical goods i.e. the foods; but also, services in procedure of ambience and settings.

In this paper, author proposed the methodology to describe and extract the opinions from web documents present in easy and effective way of transforming the unstructured data about opinions that are presented on the web. Though, the procedure for aspect expressions extraction is based on frequent nouns and NPs occurring in reviews, accomplished a poor performance in the tourism domain. The application of NLP rules for defining semantic orientation showed to be very effective for extracted aspect expressions and accomplishing an average precision and recall of 90%. The main limitations of the proposed methodology are that they are not domain sensitive. Particular sentences about context or domain

 ISSN: 2454-4248
363 – 366

dependent topics need to be specifically treated. In the tourism domain, this could denote a major problem since a lot of opinions could indicate a positive or negative sentiment depending on the product of opinion is given on. Considering that in tourism product reviews a important number of sentences do not contain opinions which directed to poor precision in the task of subjectivity classification.

v. "K. Xu, S. S. Liao, J. Li, and Y: "Mining comparative opinions from customer reviews for Competitive Intelligence" Decision Support System, 2011"

In this paper, author proposed a graphical method to extract and predict comparative interactions between products from customer reviews through the interdependencies among interactions consider helping enterprises to discover potential risks and further design innovative products and marketing tactics. In this paper, authors remark on a corpus of Amazon customer reviews demonstrate that our suggested method can extract comparative relations more precisely than the benchmark procedures. In this paper an author proposes a graphical model to model the complex issues in a natural way. This methodology is used to identify semantic relations in bio-science texts. The Graphical model can be classified into directed graphical model (Bayesian networks) and undirected graphical model. The newly emerging Conditional Random Fields (CRF) is an undirected graphical model which is compared with the Bayesian networks and CRF directly models the conditional probability distribution of the output is given to the input, consequently it can utilize the rich and global features of the inputs without describing the dependencies in the inputs. CRF models needs to evaluate the less parameters than the Bayesian networks, consequently it has outstanding performance when the training sample is small. The comparative relation extraction includes multiple entities and long-range dependencies which needs to detect rich features from the inputs, so CRF is a supreme tool to use for it.

In this paper, author proposed a methodology to extract comparative relations from customer opinion data to form the comparative relation maps for supporting enterprise managers in identifying the potential operation risks and supporting strategy decisions. Author also proposed a CRF model with unfixed interdependencies can better extract the comparative relations, through utilizing the complicated dependencies between relations, entities and words, and the unfixed interdependencies among relations. In this paper, author strategy to extend the model to mutually identify the comparative relations and entities to reduce the errors collected in the pipeline process.

III. CONCLUSION.

Data mining has significance with respect to finding the examples, estimating, disclosure of learning and so forth., in various business areas. Machine learning methodologies are broadly utilized as a part of different applications. Each business-related application utilizes information mining systems. To enhance such business or giving proper competitor to the business to the client require the help of web mining systems. The competitor mining is one such an approach to investigate competitors for the preferred items. In this framework, we denote a far-reaching examination of the comprehensive mining calculations with its points of advantages and disadvantages. At last, the K_means and Cminer++ yielded slightest calculation time when competing at others. For finding competitiveness in the item K_means, and sentimental analysis of user review method is used. Subsequently, this development also works with more accuracy for efficient results. This strategy is also used to apply in different workspace. It can also be developed into an android application.

REFERENCES

[1] K. Xu, S. S. Liao, J. Li, and Y. Song, "Mining comparative opinions from customer reviews for competitive intelligence," Decis.Support Syst., 2011

[2] R. Decker and M. Trusov, "Estimating aggregate consumer preferences from online product reviews," International Journal of Research in Marketing, vol. 27, no. 4, pp. 293–307, 2010.

[3] Z. Ma, G. Pant, and O. R. L. Sheng, "Mining competitor relationships from online news: A network-based approach," Electronic Commerce Research and Applications, 2011.

[4] R. Li, S. Bao, J. Wang, Y. Yu, and Y. Cao, "Cominer: An effective algorithm for mining competitors from the web," in ICDM, 2006.

[5] Z. Ma, G. Pant, and O. R. L. Sheng, "Mining competitor relationships from online news: A network-based approach," Electronic Commerce Research and Applications, 2011.

[6] R. Li, S. Bao, J. Wang, Y. Liu, and Y. Yu, "Web scale competitor discovery using mutual information," in ADMA, 2006

[7] C. W.-K. Leung, S. C.-F. Chan, F.-L. Chung, and G. Ngai, "A probabilistic rating inference framework for mining user preferences from reviews," World Wide Web, vol. 14, no. 2, pp. 187–215, 2011.

[8] E. Marrese-Taylor, J. D. Vel´asquez, F. Bravo-Marquez, and Y. Matsuo, "Identifying customer preferences about tourism products using an aspect-based opinion mining approach," Procedia Computer Science, vol. 22, pp. 182–191, 2013.