OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

FINDING COMPETITIVENESS IN LARGE UNSTRUCTURED DATASETS

Jyoti Pandurang Mohite[1] Prof. Dr. B. K. Sarkar[2]
Department of Computer Engineering Padmbhushan Vasantdada Patil Institute of Technology, Bavdhan Pune, Maharashtra
jyoti.mohite91@gmail.com,dr.bksarkar2003@gmail.com

Abstract: Identifying competitors is important for businesses. In the current competitive business scenario, it is necessary to identify the competitive characteristics and factors of an article that most affect its competitiveness. This study shows the importance of recognizing and observing contestants of the company. This activity in the framework, many questions arise, such as: Who are the key competitors of a particular item? What are the different features of the item that affect its competitiveness? Motivated from this issue, management and advertising groups concentrate on observing strategies for competitors that distinguish evidence. From this previous inspection, concentrate on mining the nearby products, e. g one product is better than the other from other documentary sources. To find the top competitors the system proposes a KNN algorithm. KNN algorithm can be used for classification. Using K- means algorithm unstructured data is structured. After that this structured data is clustered into the appropriate domain. In previously system Apriori algorithm is used for pattern matching, but the proposed system uses the Eclat algorithm for pattern matching. Eclat algorithm identifies each frequent item. It finds the frequent patterns in dataset, for example if user buys milk he also buys bread. Finally, the system shows that the proposed system is more accurate than the existing system.

Keywords: Competitive business, Competitiveness assessment, KNN algorithm, Apriori algorithm, Eclat algorithm. Data mining,

---∴∴∴---

I INTRODUCTION

The strategic importance of detecting and observing business competitors is the inevitable research motivated by several business challenges. Monitoring and identifying competitors in the company studied in earlier jobs. Data mining handles such a huge amount of information in a way that is optimal for mining competitors. Product reviews are a form of online and rich information that gives your customers the opportunity to share their opinions and profits with your competitors. However, under such circumstances, it is an insightful suggestion to get a competitive product that all understand from different websites.

In the previous paper, many authors analyzed such large customer data intelligently and efficiently. For instance, there are a lot of different levels of online reviews from the analysis of the opinions of the items collected by the online reviewers. However, most researchers in this area ignore the method of seamlessly utilizing the findings to competitors mining processes.

Recently, there has been limited research to take advantage of the latest developments in e-commerce applications in artificial intelligence (AI) and data mining. These researches help designers to understand a large number of customer requests for an online review of products for improvement. But these arguments are not enough and some potential problems. These have not been fully investigated, including how to conduct an online review of the product and a thorough competitive analysis. In fact, in the typical scenario of customer-driven new product design (NPD), we need to exhaustively analyze your strengths and weaknesses for opportunities that could be successful in intense market competition. Consider the example shown in Figure 1.

Mainly, the paradigm of competitiveness is based on the following observation the competitiveness between two items, they are the same group of customers (i.e, the same market segments) for example, the two

restaurants present in different countries, and there is no overlap between the target groups, so they are obviously not competitive.

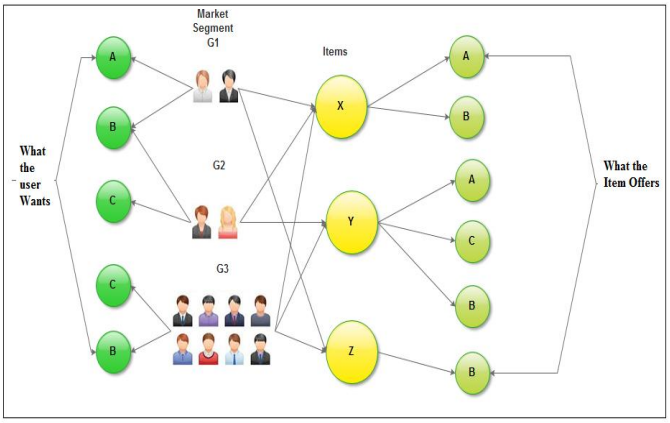


Figure 1: The example of our competitiveness paradigm

Contribution:

- To find the top competitors the system proposes a KNN algorithm.
- The system unstructured data is structured and structured data is clustered into the appropriate domain.
- Use the Eclat algorithm for pattern matching. Eclat algorithm identifies each frequent item. It finds the frequent patterns in the dataset.

II LITERATURE REVIEW

The data mining is a way of handling a huge amount of information for mining competitors. In this paper author present an efficient method for competitiveness in large review datasets. Here to describe the finding the top-k competitor's problem [1].

In this paper, they propose an improved method for comparative sentence recognition. These mixed rules are based on the registration of trade names or names in the dictionary. Compared with previous studies, the proposed method can achieve better recognition accuracy with less man-made work and supervision. As the experimental results show, this method is better than many others. By categorizing sentences directly into "Equative", "Non-Equal" and "Non-Comparison", the recognition result is good, and then mining operations can be simplified. Finally, based on this method, a visual restaurant competitiveness analysis which proves the practical value of this paper was carried out. To realize real business intelligence, comparative sentence recognition technology can be combined with product function mining, sentiment analysis, and comparison network structure etc., as a result, it can extract the potential useful business value and help decision makers to make the right decision or judgment [2].

Using data from location-based social media, the first assumption here is that the stores being ranked are of the same type, by turning check-in into a competition between the restaurant and its neighbors. The second assumption is that there is competition between stores that are close to each other. They evaluate the performance of the actual data set from Foursquare, and the probability options PNAR and PNUR work as well. They are also qualitatively analyzing the results through case studies and verifying the correctness of this model through "truth" [3].

This paper [4] proposes a way to incorporate competitive intelligence into BI systems using disaggregated aggregated data. The key competitive measure that we already infer the development and verification approaches is detailed cross-company data without the need for customer activity. Instead, this method derives these competitive metrics from simple, generally available, "site-centric" data to online companies, and then uses the syndicated data provider.

The author proposes and evaluates an approach to constructing a business-to-Business Network by using the quotations of companies in online news. As mentioned earlier, a company quoted on the news does not necessarily represent a competitor relationship. They found that such citation-based networks carry potential information, and structural properties can be used to infer the relationship of competitors. This is currently evaluated quickly and observed. First, the intercompany network gets a signal about the relationship of its competitors. Second, if the structural attributes combined with various types of the classification model, to guess the relationship of competitors. For disproportionate portions of the data, more advanced modeling techniques (e.g., data segmentation, DTA) are needed to achieve reasonable performance. Third, the two commercial data sources quantify the extent to which they are imperfect in their competitor's coverage and ensure that the approach is consistent with it while maintaining adequate performance [5].

The paper [6] Based on NPs, an algorithm for extracting aspect representations with frequent nouns appearing in the review, however, though the unstructured data on opinions available on the web, it is possible to easily extract the views from the web document, then propose a model for defining and

extracting the views from the web document. In the tourism sector, poor performance is achieved not only by providing results but also by showing that it is actually used in multiple representations of the same attributes or components of tourism products. Therefore, the most frequent words that you must take into account when the extraction surface in order to express the recovery of such a thing.

In this paper, they propose an index of on-line isomorphism based on web site content and linkage structure. Then, the existence of on-line isomorphism is utilized for the competitive identification problem. While the identification of competitors is highlighted as an important and challenging step in competitive analysis and strategy, there is limited literature on automated identification of competitors. They use online metrics as input for predictive models that classifies a company pair as a competitor or non-competitor. The resulting predictive model yields high accuracy, F measurements, and AUC. The model also provides a distinct advantage compared to the use of individual controls, using the various web metrics proposed by us this advantage is the proportion of pairs of competing companies and non-competing companies in different data sets companies. They Benchmark a predictive model that uses online metrics [7].

This article, proposes a new Framework for inference of Classification Probability, well-known as PREF, for the extraction of the choices of the users of the reviews and then mapping those options on the scale of a numeric rating. PREF uses existing language processing methods to extract opinion words and product attributes from comments. He then estimates the sentimental orientations and the strength of the words of opinion through the proposed technique based on relative frequency [8].

In this work, the author proposed the method to extract and predict interactions comparison. The comparison is done between the products of the customer feedback through the interdependencies between interactions to consider to help companies discover potential risks and to design more innovative products and marketing tactics. In this article, the authors comment on a corpus of comments from Amazon clients show that the suggested method can extract comparative relationships more accurately than the reference procedures. The paper proposes a graphical model for modeling complex problems in a natural way. The methodology is used to identify semantic relations in the text of Biological Science [9].

The classifier is based on information retrieval techniques for feature extraction and scoring, and the results of various metrics and heuristics vary depending on the test situation. Operating in individual phrases collected from the web searches, a limited performance due to noise and ambiguity. But in the context of a completely web-based tool and aided by a simple method of grouping phrases into attributes, the results are qualitatively very useful [10].

III PROPOSED APPROACH

Problem Statement

Identifying competitors is important for businesses. In the current competitive business scenario, it is necessary to identify the competitive characteristics and factors of an article that most affect its competitiveness.

Proposed System Overview

Figure 2 shows, the detailed description of the proposed system.

1) In the system, the first unstructured data takes as an input.

2) The system performs clustering on data by using K-means algorithm into five domains like shopping, E-commerce, Helth, Finance, Restaurants or hotels, and others.

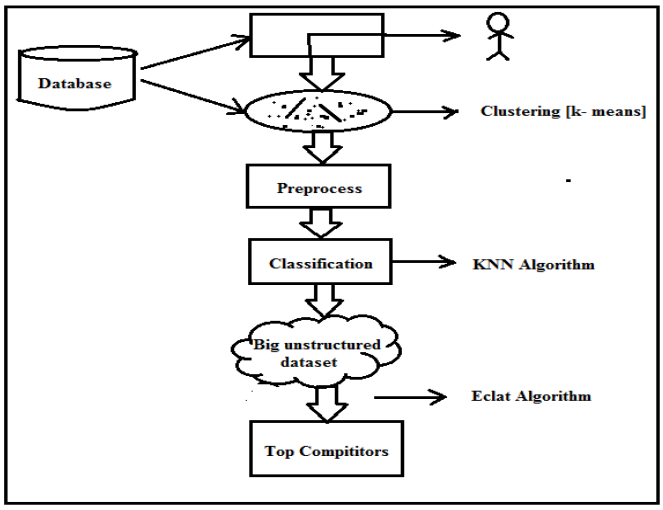


Figure 2 Proposed System Architecture

3) Preprocessing is performed on data for eliminating stop words and stemming data. For example of stop words are a, about, above, after, again, against all, am, the system removes this types of stop words and stemming data

words like "increasing", the system removes the "ing" and return increase.

4) After that, the system performs the classification, they classify the domain data according to particular features, for example Restaurants services, user ratings etc. For classification and better accuracy, the system uses the KNN algorithm.

5) Suppose user want to find particular item sets for pattern matching by using association rule mining algorithm. The system uses Eclat algorithm for generating a rule for mining.

6) From all reviews and ratings of users, the system gives top competitors or top products and its services.

Algorithms

- **Algorithm 1: K-means Clustering Algorithm**
  1. input the initial set of *k* cluster Centers *C*
  2. set the threshold TH_{min}
  3. while k is not stable
  4. generate a new set of cluster centers C_Θ by k-means
  5. for every cluster centers $C_{\Theta\lambda}$
  6. get the minimum relevance score: $min\ (S_i)$
  7. if the $min\ (S_i) < TH_{min}$
  8. add a new cluster center: $k = k + 1$
  9. go to while
  10. until k is steady
- **Algorithm 2: Naive Bayes**

Step 1: Convert the data set into a frequency table

Step 2: Create a Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.

Step 3: Now, use a Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

$$P(x\backslash c) = \frac{P(x\backslash c)P(c)}{P(x)}$$

- **Algorithm 3: Apriori Algorithm**
- join step: C_k is generated by joining L_{k-1} with itself.
- prune step: any (k-1) itemset that is not frequent cannot be a subset of a frequent k-item set
- Pseudo code: C_k: Candidate itemset of size k
- L1={Frequent items };

  for(k=1; L_k!=∅; $k++$)do begin

  C_{k+1}=Candidate generated from L_k

  For each transaction tin database do

  Increment the count of all candidates in C_{k+1}

  That is contained in t

  L_{k-1} = Candidate in C_{k+1} with min_support
- end
- return $\cup_k L_k$;
- **Algorithm 4: KNN**

1) Classify (X, Y, x)// X is training data \, Y is a Class labels of X, x is an unknown sample.

2) for i=1 to m do

3) Compute distance d(X_i,x)

4) end for

5) Compute set I containing indices for the k smallest distance dX_i,x

6) return majority label for {Y_i,where i ∈I}

- **Algorithm 5: Eclat Algorithm**

1) Identify each frequent item and sort the tid-list for atoms in descending order according to support count: After a transformation from horizontal to vertical on-the-fly, we scan the tid-list of items successively and incrementing the items support for each entry. After the computation of support and comparison with the minimum support, wegetthe1-frequent item sets sorted in descending order of support.

2) Construct the equivalence class lattice and generate candidate itemsets: Begin with the set of atoms of class sorted in ascending order according to their supports, we construct the equivalence class lattice. Firstly, an item set is generated as a union of subsets of the sets of atoms, and secondly, its support is computed by the intersection of tid-lists of atoms. In fact, by intersecting every two subsets with the length of k-1 in tid-list, we get the support so fallk –item sets in each class.

3) Prune the candidate itemsets and mine frequent itemsets. Then our Eclat algorithm mines all frequent itemsets.

Mathematical Model

Let S be a system such that,

{S = {I, P, O, Sc, Fc}

where,

I = Input of the system

O = Output of the system

P = Processes in the system

Sc = Success Case of the output of the system

Fc = Failure Case of the output of the system
I= {Trip Advisor Dataset, Hotel Dataset, yelp Dataset};

Process:
1) P1={I}; //Read dataset

2)P2= {P1}; //Clustering into domain category

i] K-means Algorithm

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_1^{(i)} - C_j \|^2 \quad (1)$$

Where,
j= Objective function
k=Number of clusters
n= Number of cases
i= Case i

Cj= Centroid for cluster j

3) P3= {P2}; // Preprocessing

4) P4= {P3}; //create train and test files

5) P5= {P4}; //Classification

i] Naive Bayes Algorithm

$$P(A\backslash B) = \frac{P(B\backslash A)P(A)}{P(B)} \quad (2)$$

Where;
P= category name
A\B =Class Attributes
B=Attribute
P (A) = Class prior
P (B) =Predictor P

ii] KNN Algorithm

$$C_n^{inn}(x) = Y_{(1)} \quad (3)$$

6) P6= {P5};
Getting relevant data from Big data set.

7) P7= {P6};
To find the best competitor by using Apriori Algorithm

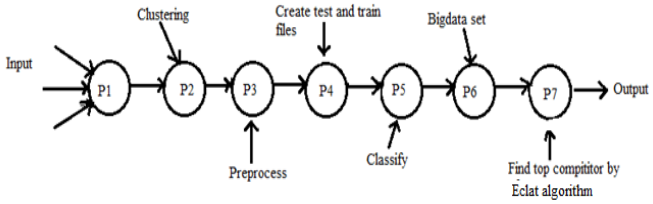
$$Support(x) = \frac{Number\ of\ occurances}{Total\ frames\ action} \quad (4)$$

$$Confidance\ \{x \rightarrow y\} = \frac{Support\{x,y\}}{Support\{x\}} \quad (5)$$

8) P8= {P7}

To find the best patterns matching we use Eclat Algorithm

Mathematical Diagram:



IV. RESULTS AND DISCUSSION

A. *Experimental Setup*

The system is built using the Java framework on the Windows platform. The Net bean IDE is used as a development tool. The system doesn't require any specific hardware to run; any standard machine is capable of running the application.

B. *Dataset*

The system uses Yelp Training dataset.

C. *Experimental Result*

In this section discussed the experimental result of the proposed system.
Table I shows, the accuracy comparison between the existing and proposed system algorithm. Figure 3 shows, accuracy comparison between the Naïve Bayes and KNN algorithms. From the graphs, it is concluded that the proposed KNN algorithm is more accurate than Naïve Bayes algorithm.

TABLE I: ACCURACY COMPARISON

Algorithms	Accuracy in (%)
Naive Bayes	78
KNN	84

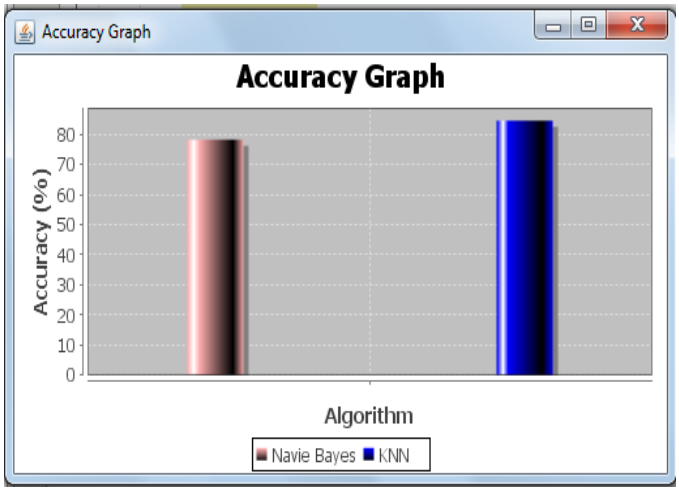


Figure 3: Accuracy Graph

V.CONCLUSION

Data mining is important in terms of pattern discovery, prediction, knowledge discovery, and so on in different business domains. Machine learning algorithms are widely used in a variety of applications. All business relationships are made using data mining techniques. The improvement of such companies provides appropriate competitors in the business, which the user needs to support web mining technology. Competitor mining is one such way to analyze the competitors of the selected item. So propose a KNN algorithm for better accuracy. Results show that the proposed system is more accurate and efficient.

REFERENCES

[1] George Valkanas,Theodoros Lappas, and Dimitrios Gunopulos, "Mining Competitors from Large Unstructured Datasets", [2016].

[2] Song Gao, H. Wang, Y. Song and Ting Lu., "Mining Comparative Opinions From Customer Reviews for Competitive Intelligence", Decis. Support Syst., AIS Electronic Library [2011].

[3] DOAN T. N., F. C. T. Chua and E.-P. Lim "Mining Business Competitiveness From User Visitation Data", [2015].

[4] Z. Zheng, P. Fader, and B. Padmanabhan, "From Business Intelligence to Competitive Intelligence: Inferring Competitive Measures using Augmented site-centric data," [2012].

[5] ZhongmingMaa, Gautam Pant, Olivia R.L. Sheng, "Mining competitor relationships from online news: A network-based approach" [2011].

[6] Edison Marrese-Taylora, Juan D. Velsqueza, Felipe Bravo-Marquezb, Yutaka Matsuoc, "Identifying Customer Preferences about Tourism Products using an aspect-based opinion mining approach", Procedia Computer Science, vol. [2013].

[7] Gautam Pant and Olivia R. L. Sheng, "Web Footprints of firms: Using Online Isomorphism for Competitor Identification", [2015].

[8] wang, f, l, chen, qi, l, " comparison of feature-level learning methods for mining online consumer reviews", [2012]

[9] ruigu. jin, jian, and ping ji, "identifying comparative customer requirements from product online reviews for competitor analysis", [2016]

[10] S. Lawrence, K. Dave, and D. Pennock "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews" WWW, [2003]