iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** VI **Month of publication:** June 2019

DOI: http://doi.org/10.22214/ijraset.2019.6361

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

Efficient Data Mining Method for Finding Competitors from Large Unstructured E-Commerce Data

Komal Ghadage[1], Dr. Sunil Rathod[2]
[1, 2]*Department of Computer Engineering, Dr. D. Y. Patil School of Engineering Lohgoan, Savitribai Phule, Pune University*

Abstract: ***Data mining is the popular area of research that facilitates the improvement process of the business, such as the preference of the user of the mining. Mining information from the web used to obtain the opinion on the products or services. In the current competitive business scenario, it is necessary to identify the competitive characteristics and factors of an item that most affect its competitiveness.***
The competitiveness assessment always uses the opinions of customers in terms of rating reviews and an abundant source of information from the web and other sources. The problem is to find the top competitors in other domain by considering the features of a particular domain.
The system proposes a C 4.5 algorithm for better accuracy to find the top competitors. The unstructured data is structured by using the K- means algorithm. Then this structured data is clustered into an appropriate domain by using our proposed C 4.5 algorithm.
For pattern matching previously used Apriori algorithm has some disadvantages so the system use FPgrowth algorithm. The rules provided by FPgrowth algorithm are more accurate than the Apriori algorithm. Finally the result shows that the proposed system is require less time to find top competitors and it more accurate than existing systems.
Keywords: ***Data mining, Competitive business, Competitiveness assessment, C 4.5 algorithm, Apriori algorithm, FPgrowth algorithm.***

I. INTRODUCTION

The identification of competitors acts as an important fact in various spheres. In the economics of industrial organizations, this involves the task of defining the market, which is important for regulation and antitrust policy.
Marketing supports the analysis of pricing policy, product design, development and positioning, communication strategies and distribution channels.
All companies have competition, and potential entrepreneurs ignore competitors at their own risk. If the enterprise does not have an absolute monopoly on important products, then there are competitors offering replacement products and services. In any business plan, the competitive status of the "market space" organization is stated, because the analysis of competitors is an important requirement, the partners and the business plan readers are required to comply with it.
The main goal of the analysis and execution of competitors in the process of collecting relevant information, interpreting it, identifying information needs and major competitors. Management and the marketing community have focused on empirical methods for analyzing competitors.
Extensive research has focused on comparative expression discovery cases: "article A is better than Article B" from different websites and other text sources.
The paradigm of competitiveness is primarily based on the following observations: the competitiveness among two different factors, they compete for business with the attention of the same customer groups.
For example, 2 restaurants that exist in different countries are obviously not competitive because there is no overlap between Target groups. Consider the example shown in Figure 1. Each item is mapped to a set of features that can be provided to the customer. In this example, three functions are taken into account: A, B, C. The actual formalization covers a wider range, such as binary functions, categorical functions, and numeric functions. Users are grouped by preference in terms of features. For example, G2 customers are only interested in feature B and C. Y conflicts with both X(for G1 and G2 groups)and Z(for G3).

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

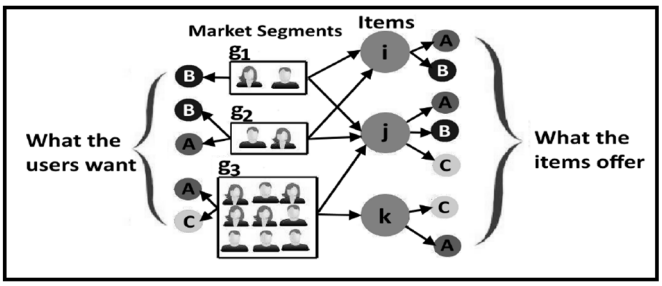


Fig 1: The example of our competitiveness paradigm

In this paper study about the related work is done, in section II, the proposed approach modules description, mathematical modeling, algorithm and experimental setup in section III is discussed and at final we provide a conclusion in section IV.

II. LITERATURE REVIEW

In paper [1], the author proposes an approach to classify the competition which is important for enterprises. Based on the structure of the intercompany system obtained from the estimation of the company in the online news fact, the approach of graph theory scale and methodology of machine learning for the conclusion of the competitor correlation was presented. The author proposed a neutral method in the point of not using the natural language processing method in the news. The author's methodology includes a specific collection of news articles controlled by the company and classifies the company quotes by news articles.

The author in [2], discussed the customer reviews that are widely accessible on the internet due to the huge number of product classifications. They propose the concept of data pre-processing and attribute extraction steps. This paper review social partitioning experts and consolidated results in words and phrases. Then remove any product review words or phrases that are neither explicit nor implicit.

This article [3], 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.

In this article [4], the author proposes an extension of Liu's methodology based on aspects of opinion mining according to the domain of Tourism. This extension shows that the user means differently to different product types when writing comments on the web. Generic products show the conceptual goods formed by an industry. This product shows the wide variability of real forms that each has the same functionality.

In this work, the author [5], 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.

In this paper [6], propose a technique for mining rivals automatically from the web. Here CoMiner algorithm is proposed for all the information about mining, the strength of competitors and the competitive sphere. This algorithm is a method that does not depend on the scale mining domain to the web.

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 [7].

The author proposes to use some language rules to address the problem, Along with a new opinion aggregation function. Extensive experiments show that these rules and function are highly effective. In this paper, Observer opinion system is built. They determine the semantic orientation of each opinion expressed on every given product feature [8].

In this paper, the author focus on specific domain Movie Reviews. Multi-knowledge. They propose a base approach that integrates WordNet, statistical analysis, and film knowledge. The results show that the proposed method is effective in movie review mining and summarization [9].

In this paper, propose a system to automatically extract advantages and disadvantages from online reviews. While many approaches have been developed for extracting opinions from the text, the focus here is to take advantage of an online review site with pros and cons that may themselves be in any form of facts or opinions generated by the author. Review text strengths and weaknesses provide a system that matches the review text the largest entropy model is trained in the results set to extract the strengths and weaknesses that are not explicitly offered from online review sites, then the experimental results show that results in the 76% of accuracy [10].
From this research paper conclude that data mining is a very important research topic and that facilitates the improvement process of the business, such as the preference of the user of the mining. Many approaches have been developed for extracting opinions from the text, news and customer reviews that are widely accessible on the internet. Study the various methods and graphical model for modeling complex problems in a natural way.

III. PROPOSED APPROACH

A. Problem Statement

In the current competitive business scenario, it is necessary to identify the competitive characteristics and factors of an item that most affect its competitiveness. The competitiveness assessment always uses the opinions of customers in terms of rating reviews and an abundant source of information from the web and other sources. The problem is to find the top competitors in other domain by considering the features of a particular domain.

B. Proposed System Overview

Figure 1 shows, the detailed description of the proposed system.

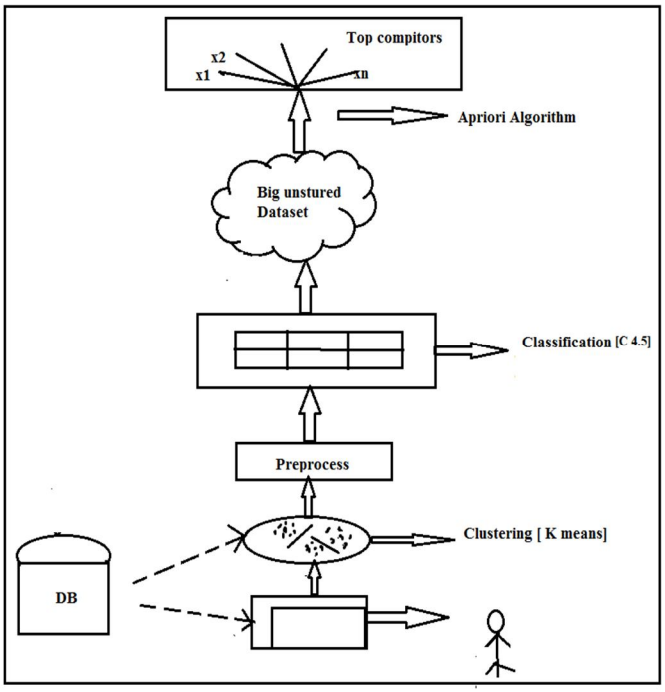


Fig. 1.Proposed System Architecture

1) In the system, first unstructured data take as an input.
2) Preprocessing is performed on data for eliminating stop words and stemming data.
3) Perform clustering on data by using K- means algorithm into five domains like shopping, E-commerce, Helth, Finance, Restaurants or hotels, and others.
4) After that, we classify the domain data according to particular features, for examples Restaurants services, user ratings etc. For classification and better accuracy, we use C 4.5 algorithm.
5) Suppose user want to find particular item sets for pattern matching by using association rule mining algorithm. We use FPgrowth algorithm for generating a rule for mining.
6) From all reviews and ratings of users, the system gives a top competitors or top products and its services.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

C. Algorithms

1) Algorithm 1: K-means Clustering Algorithm

a) input the initial set of *k* cluster Centers *C*
b) set the threshold TH_{min}
c) while k is not stable
d) generate a new set of cluster centers C_{Θ} by k-means
e) for every cluster centers $C_{\Theta\lambda}$
f) get the minimum relevance score: *min* (S_i)
g) if the *min* (S_i) $<TH_{min}$
h) add a new cluster center: $k = k + 1$
i) go to while
j) until k is steady

2) Algorithm 2: Naive Bayes

a) Step 1: Convert the data set into a frequency table
b) Step 2: Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.
c) 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.

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

3) 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 $\cup_k L_k$;

4) Algorithm 4: FPgrowth Algorithm

a) Procedure: FPGrowth DB, ξ

i) Define and clear F-List : F[];
ii) For each transcation Ti in DB do
iii) Foreach item aj in Ti do
iv) F[ai] ++;
v) End;
vi) Sort F [];
vii) define and clear the root of FP tree : r;
viii) for each transaction Ti in DB do;
ix) make Ti ordered according to F;
x) call ConstructTree(Ti,r);
xi) end
xii) foreach item ai, in I do
xiii) call growth (r, a, ξ);
xiv) end ;

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

5) Algorithm 5: C 4.5 Algorithm

a) From training databuildsC4.5 decision tree classification.

b) Training data are set A = a1, a2, a3 … of already classified samples.

c) Every sample Si contains a p dimensional vector (X 1,i , X 2,i , X p,i ,) andXj represent attribute values or features after sample, as well as a class in which Si falls.

D. Mathematical Model

P (q) is the percentage of users represented by query q and let $V_q^{i,j}$. The two items i and j provided are pairwise coverage to the space defined by the q feature. Next, we define the competitiveness between i and j in the market with the feature subset F as follows:

$$C^F(i,j) = \sum_{q \in 2^F} p(q) \times V_{i,j}^q \qquad (1)$$

There is a clear probabilistic interpretation of this definition.

Competitiveness CF (i, j) represents the probability that two items are included in the random user consideration set.

The Pair Wise coverage $V_{i,j}^F$, which is characterized by two items i, j as a percentage of all possible values of f, which can be covered by both I and j. Formally, we were given a set of all possible values V^F for f defined :

$$V^{fi,j} = \left|\{u \in V^f : u < f[i] \wedge u < f[j]\}\right| -_{|values(f)|} \qquad (2)$$

The pairwise coverage of a numeric feature f by two items i, j can be computed as follows:

$$V_{i,j}^F = \min(f[i], f[j]) \qquad (3)$$

The PairWise query q coverage for two items I and j is based on the Pair Wise coverage provided by two items for each feature$f \in q$.

$$V_{i,j}^q = \prod_{f \in q} V_{i,j}^f \qquad (4)$$

The probability that a random user will be interested in exactly the set of features includes in q. formally:

$$p(q) = freq(q,R) \sum_{q \in 2} freq(q',R) \qquad (5)$$

IV. RESULTS AND DISCUSSION

A. Experimental Setup

The system is built using the Java framework on 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. Yelp is a basic local business directory and review site with social networking features. Allows users to give ratings and review business.

C. Expected 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 2 shows, accuracy comparison between the Naïve Bayes, CMiner and C 4.5 algorithms. From the graphs, it is concluded that the proposed C 4.5 algorithm is more accurate than Naïve Bayes and CMiner algorithm.

Table II: Accuracy Comparison

Algorithms	Accuracy in (%)
Naive Bayes	87
CMiner	89
C 4.5	92

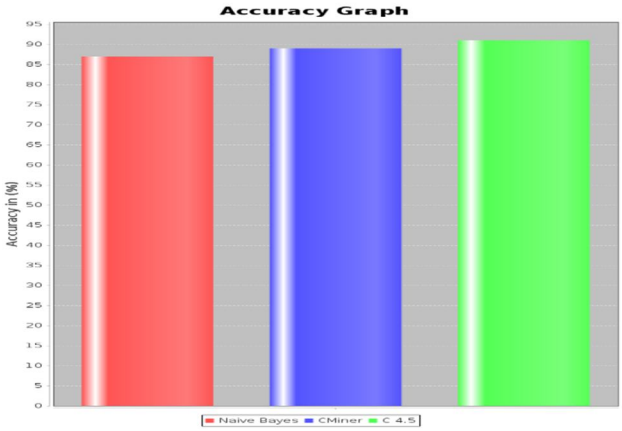


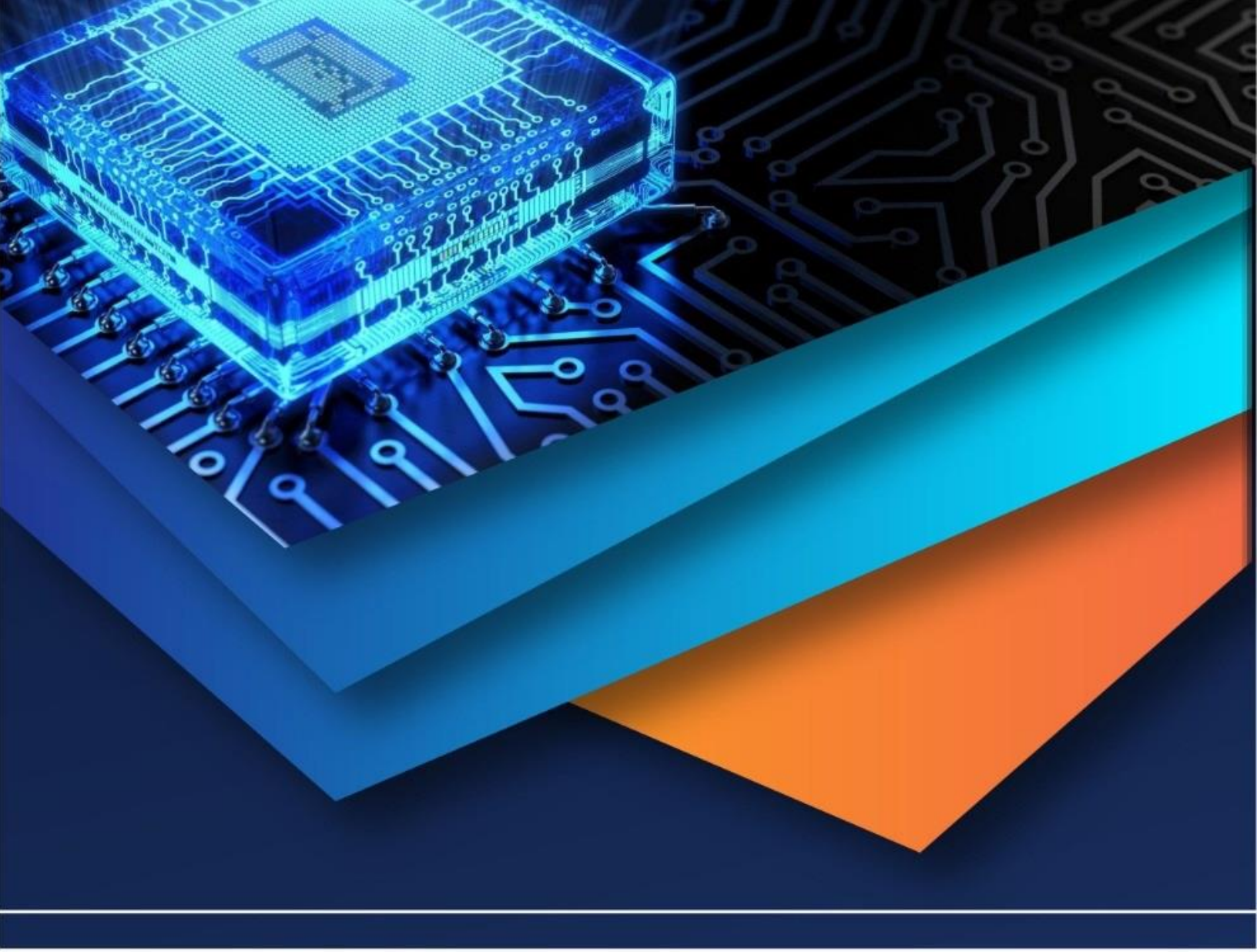
Fig. 2: Accuracy Graph

V. CONCLUSION

A formal definition of competitiveness was presented between two items, which validate both quantitatively and qualitatively. The system formalization is applicable to all domains, overcoming the deficiencies of previous approaches. This system considers a number of factors that have been largely overlooked in the past, such as the position of elements in the multidimensional features space and user preferences and opinions. The system proposes a C 4.5 algorithm for better accuracy to find the top competitors. The work introduces a methodology for end-to-end for the extraction of such information from large datasets of customer feedback. The proposed system solves the problem, to find the top competitors in other domain by considering the features of a particular domain. Results show that the proposed system is more accurate, efficient and applicable on all domains.

REFERENCES

[1] b, yu, p.s ,ding, x, liu "a holistic lexicon-based approach to opinion mining", [2008]
[2] abbasi, a, chen, h, salem "a sentiment analysis in multiple languages: feature selection for opinion classification in web forums", [2008]
[3] wang, f, l, chen, qi, l, "comparison of feature-level learning methods for mining online consumer reviews", [2012]
[4] zhan, j, loh, h.t, liu, y "gather customer concerns from online product reviews – a text summarization approach ", [2009]
[5] ruigu. jin, jian, and ping ji, "identifying comparative customer requirements from product online reviews for competitor analysis", [2016]
[6] k. xu, s. s. liao, j. li, and y. song, “mining comparative opinions from customer reviews for competitive intelligence”, [2011]
[7] S. Lawrence, K. Dave, and D. Pennock “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews” WWW, [2003]
[8] X. Ding and B. Liu, The Utility of Linguistic Rules in Opinion Mining” SIGIR [2007]
[9] X.-Yan Zhu, F. Jing, L. Zhuang and L. Zhang, “Movie Review Mining and Summarization” [2006]
[10] S. Kim and E. Hovy, “Automatic Identification of Pro and Con Reasons in Online Reviews” [2006]



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY