Extracting Top Competitors from Unorganized Data

M. Senthamil Selvi, P. V. Kavitha, J. Angel Ida Chellam

Abstract: ***In a competitive business, success factor is based on the ability to make an item more interesting to customers than competition. An E-Commerce application allows the user to view the items and their features along with the option of commenting about the item and can also view comments of other customer. From the large reviews, it is difficult for a customer to make a decision. With the set of items in existing market, competitiveness should be evaluated using the reviews so that manufacturing item is not dominated by other existing items. The proposed novel approach defines the competitiveness between two items based on market segments. A "CMiner" algorithm is used to find the top competitors of a given item using the result of Item dominance. This method improves the quality of the result when compared to previous competitor ranking models based on probability value.***

Keywords: Customer reviews, Competitor mining, Data mining, Firm analysis, Information Search and Retrieval, Item Dominance.

I. INTRODUCTION

Data mining is a process which mines usable and important data from a large raw dataset. It is the popular area of the research which facilitates the business improvement process such as mining user preferences, opinion about the product or services and the competitions of specific business.

A. Competitor Mining

Organization must operate within a competitive industry environment. Understanding the pros and cons of the competition is critical to making sure the business survives and grows. Competitor analysis is the process of getting information about important competitors and use that information to predict competitor behaviors while making best buying decision . A competitor analysis is a critical part of a company marketing plan. A company can launch what makes their product or service different and what attributes make the customers to show interest towards their target market with the help of competitive analysis. There is a need to completely assess business competition on a regular basis even to run a small business.

Competitors are distinguished as follows,

1) Primary Competition: In business marketing domain, a product that is same or own product in the same geographic area is a direct competitor, which means they are either targeting the same customers or have a similar product or both.

However, customers purchase depends on different varieties of price points, locations, service levels, and product features. But they will not choose the same mix of these options in every comparison. They will realize as many options as they can to fulfill their need, which may include looking at a different service model or a different product altogether. Competition plays a key factor in determining the strongest markets for your business solutions.

2) Secondary Competition: Companies offering dissimilar or additional products in relation to their product or service are known as indirect competitors. These competitors may offer a high or low end version of their product, or sell something similar to thoroughly different customers. If you are selling Ariel detergent powder, a secondary competitor might be a Surf Excel detergent powder. Another example is the producer of eyeglasses who competes indirectly with contact lens manufacturers.

3) Tertiary Competition: This category includes businesses that are tangentially related to yours, and really comes in handy when they are looking to expand their product catalog. These could be related products and services that are likely, as well as businesses that may be useful to partner with further down the line. For instance, if they sell jewelers, a tertiary competitor may sell gems and stones.

4) Replacement competition: A replacement competitor is someone could do something instead of choosing product. These are the most challenging competitors to identify. Assuming the average 8-year-old girl is reading a book instead of playing a mobile game. You have to be a bit of an anthropologist and study your customers to determine what they consider as replacement competition for your products and services.

B. Review Analysis

Online reviews are the most significant information hub that allow consumers to search for elaborative and reliable information by sharing past user experiences. Product review reveals about consumer views of product usage. Reviews written by other customers describe the usage experience and perspective of customer with similar needs.

Revised Manuscript Received on March 08, 2016.

M. Senthamil Selvi, Professor & Head, Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore

P. V. Kavitha, Assistant Professor (Sr.Gr), Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore

J. Angel Ida Chellam, Assistant Professor (Sr.Gr), Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore

Online customer reviews focuses on different aspects of their business, including products, services, purchase decisions etc., The product review analysis helps in understanding consumer interests and provides a marketing intelligence about the type of products which the consumers are showing interest towards purchase.

Michael LeBoeuf coined that, "A satisfied customer is the best business strategy of all", Amazon is always maintaining high and good impact on customer satisfaction in the market domain. One of the main reasons why customers are attracted towards online e-commerce websites is the presence of reviews/opinions about the product, which undoubtfully helps customers to know about every detail/features of the product before buying a product. Since, consumers do not have the freedom to physically inspect and check the product while shopping online, product reviews are the ones they can trust in order to buy a product.

1. Competitiveness Definition

Online reviews have become a standard part for the purchasing process of people.

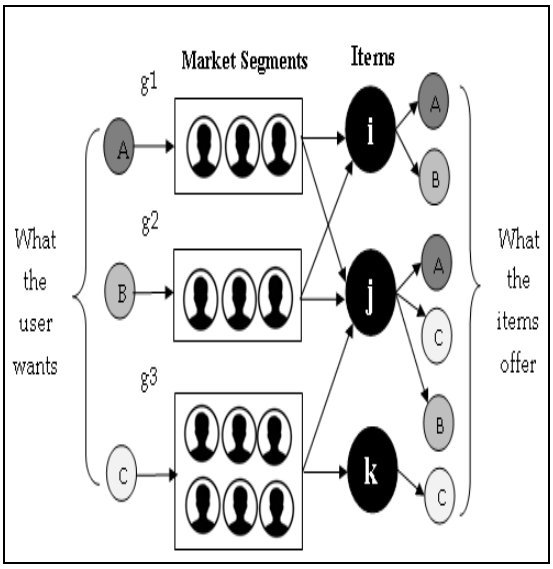

Fig. 1 Competitor Paradigm

The above mentioned diagram clearly explains about what are the features (A, B, C) the items offer and what the user wants by simply forming the market segment and the items with the list of features.

Each group of customers deals with unique market segment. Customers are clustered based on their preferences with respect to the features/properties by analyzing all group of customers in a given market. In practical, such information is not available. In general, all the market segments are evaluated from large review datasets. Each segment is characterized by a query that contains the number of customer's interest shown towards the items and also the features of customer's interest. The competitiveness between the items depends on the number of customers represented by each query and also how much both the items cover their features in that query.

II. RELATED WORK

Mark Bergen and Margaret explains a broad based managerial approach that compares firms based on their capabilities to meet market needs for evaluating competitive threats. This framework presents a two stage such as Competitor Identification and Competitor Analysis [1].

Rui Li and Shenghua Bao explained the "CoMiner" algorithm that mines a set of comparative candidates of the input features, prioritizes them according to the comparability and then mines the competitive features. It calculates the product's overall ranking score from the directed weighted graph derived for the reviews. [2], [4], [3]. Latest trends have shown that large numbers of customers are switching to E-shopping. This work boons a feature based ranking technique that extracts millions of customer reviews. The product features are identified to analyze the frequencies and relative usage.

Kunpeng Zhang and Yu Cheng explain the "Aspect based opinion mining technique" extracts the customer preferences about the product. It analyzes opinions about product features using predefined rules and patterns. Noun and Noun phrases are identified using Part-of-speech tagging and Syntax Tree parsing [5]. Doan Thanhn Nam and Eepeng Lim explains about the value of business stores using user visitation data which are now easily available from location based social media. In this method, two assumptions are considered to rank stores of the same type and the competitions between rank stores closer to each other. In this store ranking framework, an undirected graph is constructed to calculate the store's ranks based on competitive probability values [6].

Bushra Anjum and Chaman Lal Sabharwal explain about the "Product Ranking Algorithm" which calculates the entropy measure of product reviews. Reviews are subjective opinions about a product or the service by the users. It explores a hybrid approach such as Entropy, Bilinear and statistical measures that analyze and rank products heterogeneous customer data. The ranking of the product is based on text reviews, QA data and star rating of products [7].

III. PROPOSED WORK

A. Problem Statement

Research has demonstrated the strategic importance of identifying and monitoring competitors of a firm. Extracting competitors from online reviews are the most important job for competitive analysis. Marketing and management community have focused on extracting comparative expressions from the web sources. These expressions does not produce more accurate results and it supports for limited domains only. An efficient algorithm 'CMiner' is used to find top-K competitors of a given item based on market segments.

B. Overview of the Project

In this proposed work, data is collected from the online reviews of a customer for a particular product using which number of customers interested in each feature of a product is identified. The competitive score of each item is calculated to determine the competitiveness. It includes pairwise coverage and feature probability. The pairwise coverage is used to identify the features that are satisfied by each item and the

feature probability is used to identify the number of customers interested in particular feature from large market segment of customers. With the help of competitive score, items are arranged in an order to find item dominance. This is helpful for reducing time in the process of competitor identification. The output of skyline is given to CMiner algorithm along with the list of items and their features. Then, the CMiner algorithm is applied to identify top-K competitor of a given item for user specified k value.

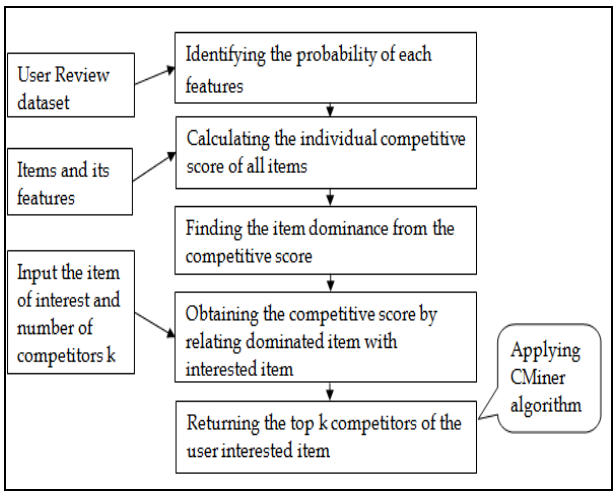

Fig. 2 Flow Diagram of extracting top competitors from unorganized data

C. Module Description

This project deals with the user reviews to find competitive score of the items which helps in identifying top k competitors of a given item. The modules of the project work are explained as follows:

1. Calculating competitive score
   1.1 Pairwise coverage
   1.2 Feature probability
2. Finding item dominance using competitive score
3. Identifying top K competitors using CMiner algorithm

1. Calculating Competitive Score

In order to find the competitor of an item, there is a need to calculate the score of an individual item. It includes two paradigms i.e., user required features and number of customers recommending the same feature. Competitive score is calculated for individual item and also for competitiveness between the two items. Feature Probability P(f) is the percentage of users represented by particular feature 'f' that belongs to set of features F and $V^{f}_{i,i}$ be the pairwise coverage of all feasible feature values 'f' that can be covered by both items (i & j) or by particular item (i or j).

(i) Competitiveness C_F (i, i) shows the probability of item incorporated in the attention of set of users. This depends on the recommendation system in which item with the highest competitive score is recommended for the user. Competitiveness score of individual item is defined as follows.

$$C_F(i, i) = \sum_{f \in F} P(f) \times v^{f}_{i,i} \quad (1)$$

(ii) Competitiveness C_F (i, j) shows the probability of two items included in the attention of set of users. This score helps in finding strong competitor of a given item. Competitiveness between two items is defined as follows.

$$C_F(i, i) = \sum_{f \in F} P(f) \times v^{f}_{i,j} \quad (2)$$

Competitive score calculation includes two primary components. They are pairwise coverage and feature probability.

1.1 Pairwise coverage

Pairwise coverage $v^{f}_{i,j}$ of a feature is defined as feasible feature values 'f' that can be covered by both items i and j or by the individual item i or j. In this case, the binary features can be fully covered(1) or not covered(0) for feature values 'f'.

(i) The pairwise coverage of a binary feature f for an individual item either i or j is computed as,

$$v^{f}_{i,i} = f(i) \quad (3)$$

(ii)The pairwise coverage of a binary feature f for both items i and j are computed as,

$$v^{f}_{i,j} = f(i) \times f(j) \quad (4)$$

1.2 Feature Probability

Feature probability estimation process requires the customer opinion reviews on a particular feature. Feature probability P(f) is computed as follows,

$$P(f) = \frac{\text{frequency}(f,R)}{\sum_{f \in F} \text{frequency}(f,R)} \quad (5)$$

Feature probability is found by dividing the frequency of each feature f in large review dataset R and the total sum of frequency of all possible features.

2. Finding item dominance using competitive score

Item dominance is a structure consists of all the items with its individual competitive score. Competitive score of individual items are arranged in an order which represent the item dominance. While calculating competitive score, all the features and their probability are considered as the main item dominance factors. The item with the highest competitive score dominates all other items. This approach is also called as skyline which refers the item not dominated by any other item. Competitor identification approach becomes easier by using item dominance, which reduces the number of items to identify the top k competitors greatly.

The main purpose of item dominance is to reduce the time to find top k competitor, because itself provides a result when the user interested item is dominated by k items. When required number of competitors is not achieved, then the item dominance is given as input to the CMiner algorithm.

3. Identifying top Competitor using CMiner algorithm

The CMiner algorithm is performed to obtain the topk competitors of user interested item. Once item dominance is done, the items which are dominating the user interested item i are identified. If user specified k items are obtained directly from item dominance result, those are added in the top k competitor list and the process gets concluded. Otherwise the CMiner algorithm performs the following process.Initially, the dominating items identified from item dominance are stored in the database as topk and k value is decreased accordingly. The items dominated by i are stored separately to find remaining competitors. And for that each item, the competitive score is calculated by relating it with i for all features. If the competitive score is less than lower bound, the item gets eliminated. Otherwise the item is added in top-k database. A task called 'Pruning' is performed using the predefined lower bound value to reduce the number of items for competitor identification. This process is repeated until k number of competitors of i are obtained in top-k database.

IV. EXPERIMENTAL RESULT

This methodology is experimentally worked out on the Desktop with a Intel core i7 8th gen processor and 8GB RAM.

A sample hotel dataset is taken to implement this novel approach. The review dataset of hotel was taken from Booking.com. It includes 35,000 reviews on 100 hotels with their features. The features were the services offered by the hotels. For this approach, we have selected 7 features from the reviews datasets Market segment were formed from the review datasets which indicates the number of customers preferring each feature.

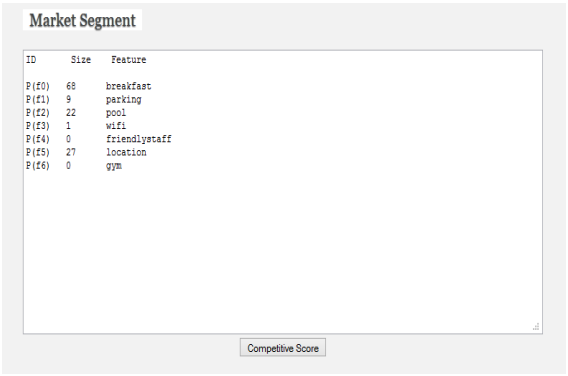

Fig. 3 Finding Market Segment

The individual competitive score of each hotel was calculated to form Item Dominance. Our experiment provides the user friendly interface as the user can able to give their interested item and number of competitors to be retrieved.

Fig. 4 Input to top k

The proposed algorithm "CMiner" takes the user inputs and market segments into consideration to find the top competitors of user specified item. This includes Item dominance as a major factor for producing better results and reducing time complexity.

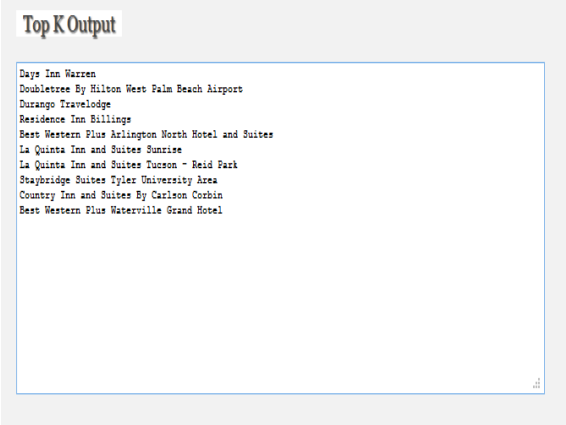

Fig. 5 Listing Top K Competitors

The existing ranking model calculates the probability of each feature and ranks the items using the values computed. This method does not produce exact competitors of a given item.

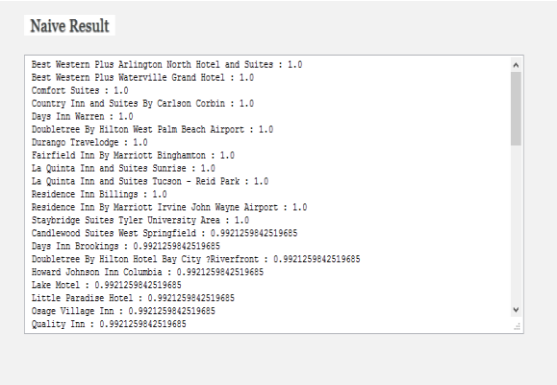

Fig. 6 Naïve Bayes result

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The result of comparison of existing Naïve Bayes model and CMiner algorithm shows that CMiner produces better results in finding top k competitors. The accuracy between two algorithms are presented in the graph.

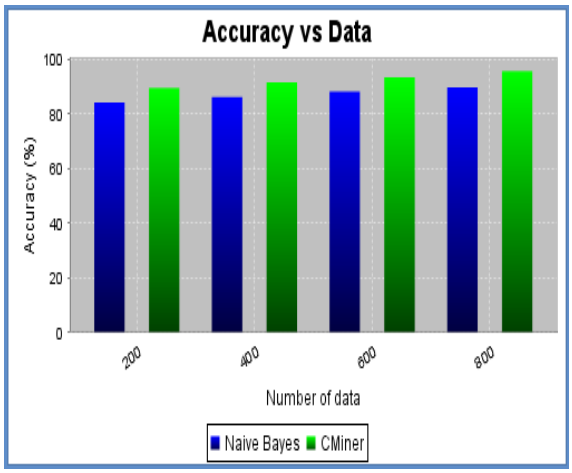

Fig. 7 Accuracy Comparison

IV. CONCLUSION

The proposed model is mainly focused on calculating competitive score of all items and obtaining top k competitors using CMiner algorithm. Our work addresses the evaluation of competitiveness by analyzing large datasets without the need for comparative evidence. By identifying the competitiveness, business organizations not only found their competitors but also get benefited by satisfying user needs. This method incorporates the text reviews of items to find top competitors. In our future work, ratings of the product will also be included to extend the quality of result and to make it more efficient.

REFERENCES

1. George Valkanas, Theodoros Lappas, and Dimitrios Gunopulos," Mining Competitors from Large Unstructured Datasets", IEEE Transactions on Knowledge and Data Engineering, 1041-4347 (c) 2016.
2. M. Bergen and M. A. Peteraf, "Competitor identification and competitor analysis: a broad-based managerial approach," Managerial and Decision Economics, 2002.
3. R. Li, S. Bao, J. Wang, Y. Yu, and Y. Cao, "Cominer: An effective algorithm for mining competitors from the web," in ICDM, 2006.
4. Kunpeng Zhang,Ramanathan Narayanan," Voice of the Customer: Mining Online Customer Reviews for Product-Feature Based Ranking".
5. 5.Kunpeng,Zhang,Yu,Cheng,Wei-Kang,Liao,Alok,Choudhary,"Mining Millions of Reviews : A Technique to Rank Products Based on Importance of Reviews".
6. 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.
7. T.-N. Doan, F. C. T. Chua, and E.-P. Lim, "Mining business competitiveness from user visitation data," in International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer, 2015, pp. 283–289.
8. Bushra Anjum,Chaman Lal Sabhaewal," An Entropy Based Product Ranking Algorithm using reviews and Q&A data".