Mining Competitor From Large Unstructured Datasets using LDA & K Means Algorithm

M.G.Bandal[1], S.S.Jadhav[2], S.R.Jambhulkar[3] , Prof. M.S.Dighe[4]
Department of Computer Engineering
[1,2,3,4] Bhivrabai Sawant Polytchnic, Wagholi, Pune

Abstract- *For discovering our competitors from unstructured datasets the data mining concept is used. This approach is provided by recognizing the competitiveness between the two businesses. The competitiveness may influenced by the facility provided by the any type of business. In this paper, we are uses the customer point of view in terms of reviews & ratings & other sources of detail from web for the evaluation of competitiveness*

Keywords- Data Mining, Competitor Analysis, Contenders.

I. INTRODUCTION

The main concept used for discovering competitor is data mining, which referred as process of sorting through large datasets with an overall goal to extract information from dataset and transform the information into comprehensible structure for further use. Data mining is the excellent to observing and identifying the competitor from types of businesses [1].

In this paper, the auther proposes the approach to ranking the competitor which is intimate for business. Auther represent an approach of graph theoretic measure & machine learning methodology. To conclude competitor co-relation methodology take collection of a news stories that controlled by company & classify company quotation in news stories. It design directed or weighted network [2]

In paper [3] [4], Authers presented a ceremonial definition of competitiveness. These take place between two items. In this paper, importance is given to the opinion of users in an multidimensional feature space by considering position of item.

Auther in [5], Give idea of framework for manual recognization of competitor. a large number of focal, target firms & over time these parameters get the manual nature of framework very costly

PROPOSED SYSTEM

A formal definition of the competitiveness between two items, based on their appeal to the various customer segments in their market. Our approach overcomes the reliance of previous work on scarce comparative evidence mined from text. A formal methodology for the identification of the different types of customers in a given market as well as for the estimation of the percentage of customers that belong to each type. A highly scalable framework for finding the top-k competitors of a given item in very large datasets

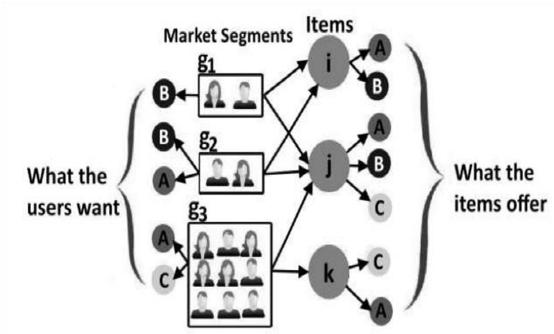

Fig -1: Example of competitiveness between business

ADVANTAGE OF PROPOSED SYSTEM

- It proposed framework is efficient and applicable to domains with very help to provide the facility that customer want.
- The Large populations of items
- It help to study the market.
- To develop the strategy for organizational growth, by knowing the facilities provided by the competitors

II. ALGORITHM INFORMATION

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA)-It is used for topic modeling that creates a life style document of the users

$$p(w/d) = p(w/z)p(z/d)$$

Here terms as:-

z= denotes as set of life Styles i.e. “n” no of habits.

w=denotes as Activity i. e daily work that we perform.

d=denotes as document

To complete the life style we have to perform n number of activities

It is also worth noting that since our system uses unsupervised learning algorithms to recognize activities and the topic model to discover life styles.

K Means Algorithm

Let X = {x1,x2,x3,……..,xn } be the set of data points and V = {v1,v2,…….,vc } be the set of centers

Steps:

Randomly select ‘c’ cluster centers

Calculate the distance between each data point from the cluster centers Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

Recalculate the new cluster center using:

Mean of the existing cluster ,that mean considered as a new cluster center of each cluster.

.Recalculate the distance between each data point and new obtained cluster centers.

If no data point was reassigned then stop, otherwise repeat from step 3 because it was iterative process.

III. MODULES IN SYSTEM

Load data

In this module, we are using HOTEL Review Dataset using this dataset only we find the Top-K Competitors. Loading our data must be in unstructured format. The data contains details like Hotel-Name, Review, Service, type etc. From the Document Dataset we are going to find Top-K competitors. We have to mine the data and then only we analyses our data.

Mining data

In this module we are mining our unstructured data ,for that we use the Stanford postage. A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. Using this POS Tagger we mined our data and into generated main word for each review document dataset

Frequent words

Datasets are run through the POS tagger, lemmatizer, and dependency parser of the Stanford Cornella .This results in all sentences having a set of lemmas.The stemming process was carried out and done foor each files intyhe document dataset.Then all words are collected and stored as Word bank

Aspect weight

We had found frequent words by Using NLP process form the loaded Document dataset. The Frequent words were stored as a whole world bank. Then the Frequency for each word in the document dataset was evaluated and shown. From that Frequency obtained, then we will obtain Valid Frequency from the set we already analyzed. Thus we obtain the Aspect Weight for the Words.

Finding Competitors

In this module, we are using HOTEL Review Dataset using this dataset only we find the Top-K Competitors. Loading our data must be in unstructured format. The data contains details like Hotel-Name, Review, Service,type etc. From the Document Dataset we are going to find Top-K competitors. We have to mine the data and then only we analyses our data.

IV.FLOW DIAGRAM

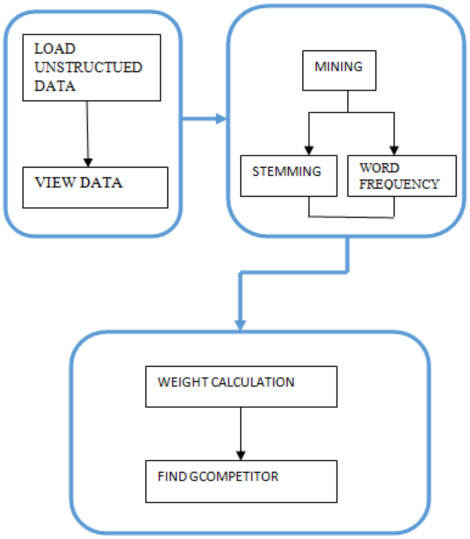

Fig -2: FLOW OF SYSTEM

V. CONCLUSION

Data mining has intimate with respect to finding the examples, gauging, disclosure of learning & so forth in various business areas.

In these we are finding our competitor using the Latent Dirichlet Allocation is used for topic modeling that creates a life style document of the users & K Means Algorithm

To enhance such business or giving proper competitor to the business to the client require

REFERENCES

[1] http://en.m.wikipedia.org/wiki/Data_mining

[2] http://en.m.wikipedia.org/wiki/Data_mini Ma, G. Pant, and O. R. L. Sheng, “Mining competitor relationships from online news: A network-based approach,” Electronic Commerce Research and Applications, 2011ng

[3] Lappas , Theodoros, George Valkanas, and Dimitrios Gunopulos. "Efficient and domain-invariant competitor mining." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

[4] Valkanas, George, Theodoros Lappas, and Dimitrios Gunopulos. "Mining Competitors from Large Unstructured Datasets." IEEE Transactions on Knowledge and Data Engineering (2017).

[5] Bergen, Mark, and Margaret A. Peteraf. "Competitor identification and competitor analysis: a broad-based managerial approach." Managerial and decision economics 23.4-5 (2002): 157-169..