IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Identifying Competitors from Large Unstructured Dataset Using Naïve Bayes Classifier and Apriori Algorithm

A.A. Kushwah[1*], Y.C. Kulkarni[2]

[1] Dept. of Information Technology, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India
[2] Dept. of Information Technology, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India

[*]Corresponding Author: ankita.kushwah92@@gmail.com, Tel.: +91 8806693543

Available online at: www.ijcseonline.org

Accepted: 17/Jun/2018, Published: 30/Jun/2018

***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 justify 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 sectors. In this paper, we state novel description of the competitiveness between two items, based on the market sector. This system estimation of competitiveness uses customer reviews of different domains, a plentiful source of information. This system shows an efficient approach for evaluating competitiveness in large review datasets and finding the top-k competitors. Our experiments are based on a corpus of Yelp.in, TripAdvisor.com, and Amazon customer reviews which states that the proposed methodology can extract comparative relations more precisely. In this paper, we state an efficient framework for the classification of reviews of mainstream domain using k-means clustering and Naïve Bayes algorithm. This system evaluates the competitiveness of two items from frequent item set to find top-k competitor using Apriori algorithm.

***Keywords*—**Data mining, Web mining, Information Search and Retrieval, K_means clustering, Naïve Bayes Classifier, Rule mining**.**

I. Introduction

Competitive intelligence initially classifies the potential risk and chances by gathering the data on context to handle the manager in making tactical decisions for an organization. Many organization recognizes the significance of competitive intelligence in risk management and decision support system. They also spend a great quantity 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][6]

Distinguishing potential risks is essential for firms to evaluate the data on their contestants' products and tactics. A company can examine the comparative weakness and strengths of its own product depends on the contestants' products and tactics. Then the company plan new products and promote to balance of its competitors. Conventionally, the information on the contestants is derived from press releases, like analyst reports and trade journals and competitors' websites and news sites. Tactlessly, this information is typically created through company that fabricates the product. As a result, the capacity of accessible information is partial and its objectivity is doubtful. The unavailability of adequate data on competitors highly confines the competence of competitive intelligence [1][2][7][8].

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, analysing news articles, searching the web, take a part in business conventions, etc. While the company summarizing sources have truncated the search efforts. The company also built some business relationship information available, because of their restricted resources and variances in criteria. A company can endure a scalability issue and deliver partial data [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 instance, 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 just questioning on the web. But, it is easy to classify mainstream domains where such facts are tremendously uncommon, such as jewellery, hotels, furniture and restaurant. Inspired through these limitations, we present a new description of the competitiveness between two items on the basis of market sectors. The organization of the paper is designed as follows:

Paper organised as follows

- Section I contains the introduction of competitive Intelligence.
- Section II contains the related work.
- Section III contains the problem statement.
- Section IV contains the proposed work of the system.
- Section V contains experimental results.
- Section VI contains conclusion and future work.

II. RELATED WORK

This explores the numerous methodologies applied to mine competitors with orientation of consumer lifetime value, relationship, review and activities using data mining methods. According to the web evolution, resulting extensive usage of various applications like e-commerce and other service-oriented applications. This wide-ranging usage of web application has delivered on massive amount of data at one’s disposal. The data is the idea that exists in its raw form resulting in information for advanced processing. The organization handled the essential challenges of extracting very valuable information from this massive amount of information. This hypothesis has directed for the conception of data mining. Mining competitiveness from a specified thing or item is the most influenced aspect of the thing or item which fulfils consumers requirements and this can be extracted from the data that can be stored in database. This portion gives two types of literatures such as competitor mining and unstructured data management.

In unstructured data, the data is accumulated from the web are sometimes semi-structured and unstructured. The semi-structured data are in the structured of XML, JSON etc. The unstructured data are in different structure and it is not fall under any predefined category. When handling various customers in the market of business sector will have difficulty for supporting the rising cost created interactions among people. Though, all customer data inserted in the database the subsequent records will deliver a comprehensive profile of these customers and their relation with one another. This conception led to a valuable source for business that is used to analysed customer data, customer requirements, and customer gratification levels.

The hypothesis of the data mining helps to use transaction data to gain better knowing of customer and successfully uncover concealed knowledge in business sector into the procedure of competitor mining. In paper [9], [10], author discussed that data mining is a tactic to support companies in emerging is a most helpful policy to reunite the competitors in market sector. In the conception of data warehousing, it is most helpful in business for convening business dispersed heterogeneous information and delivering unified suitable information access method. The data mining methods are used to convert uncover facts into manifest facts. Web data techniques are enormously flexible in competitor mining. Consequently, best competitive policy for effective exploitation of web data for well-timed decision support. The customer information is collected through a numerous method for competitor mining which is frequently unstructured. Though, most data mining techniques are handle structures data. Consequently, in data mining method is not taken into account and abundant valuable service information is lost. Structured systems are consisting of data and computing activity which is pre-arranged and defined. Unstructured systems consisting of generally full of textual data or information. It not contains any pre-determined structured form of data. The extraction of web information is done at the record level or data unit level. The data records are as a single data unit while concluding the additional stages to extract complete data units within data records. Record level extraction basically classify data regions containing all records and formerly partitioning the data regions into single records. An extraction of structured data from web pages has been broadly studied. Previous work based on constructed packages were originates problematic to manage and be utilized to different websites.

The semi-automatic technique called as wrapper induction [13] was proposed to face this issue. This technique demands some labelled pages in target domain as input to execute the induction. Therefore, they still have drawback for large scale application. To overcome above limitations fully automatic technique have been developed.

In paper [14], author focused on the problem of unsupervised web data extraction. It uses a fully automatic data extraction tool called viper. This tool is used to extract and partition data showing frequent structures out of single Web pages with greater accurateness by classifying tandem repeats by means of visual context data. Though, this method lacks performance in rare datasets.

In the conception of competitor mining, previous work is based on the utilized text information to collect comparative facts between two things or items. However, the comparative facts are based on assumptions which may not always exist. Competitor classification is mentioned as identification process through which competitor of main firm are categorized based on relevant similarities [11].

In paper [15], author proposed an automatic method that uncover competing companies from public information sources. In this methodology data is creeped from text and use transformation-oriented learning to gain suitable data normalization which merges structured and unstructured data systems. It also uses probabilistic data modelling to denote models of linked information and accomplished in autonomously uncover competitors. To classify the competitor Bayesian network methodology is used. Author also discussed on the iterative graph reconstruction process for implication in relational data. It shown to lead towards the perfection in performance. The author used machine learning algorithms and probabilistic approaches to identify competitor. Author also confirm the system consequences and arrange it on web as powerful analytical tool for individual and official investor. But, the methodology has many problems like finding alliances and market demands using the machine learning approach.

In paper [16], [17], author discussed on competitiveness between two things using various domains and manage several flaws of earlier work. Author also describes the arrangements of things in various different dimension attributes space. But, this methodology deal with several problems like identify the top_k competitor of given thing and manage structured data.

In paper [12], [18], author designed a new online metrics for participant relationship predicting. This concept is based on the content, firm and also website to measure online isomorphism. The isomorphism is a concept of competing firms. This firms represents the each other under mutual market services. To classify competitiveness concluded different analysis to find predictive models for classify competitors based on online metrics which are highly expert to those for the usage of offline data. This methodology is used to combined online and offline metrics to improve the predictive performance. This method also performs ranking process with reflections of possibility. Many works in the same process in the literature have discussed the requirement of the correctness of classification and delivered hypothetic framework for that. Specified the predicted isomorphism between competing firms, the procedure of competitiveness classification through pairwise analysis of comparisons between important and targeted firm is well originated. The part of analysis is a group of firms since competitor relationship is seen as a distinct communication between the groups.

In paper [19], author have recommended frameworks for manual classification of frame work. These manual frameworks are very costly over large for classification of competitor number over large number of important and targeted firms.

In paper [20], author proposed a technique of mining competitiveness data with respect to an entity. The entity such as product or item or thing or company or person, from the web. Author also present an algorithm called as CoMiner mainly used to develop and support for specified domain. But, the effort for further domain is still challenging.

III. PROBLEM STATEMENT

A lot of experts were showed the experiments on competitor survey and item feature extracting information. The issue of inevitably extracting the information is corelated to the user given possibly have two forms such as structured and unstructured. Managing unstructured information in the web source everlastingly create many challenges. The extracted information should be renewed into structure form are identified. In the previous work, performance is based on more computation time. Because of this issue, we depict the competitiveness between two items is presented on the basis of market sectors in minimum computation time by using the algorithm K_means, Naïve Bayes and Apriori yields to the least computation.

IV. PROPOSED SYSTEM

The tactical importance of distinguishing and observing business competitors is an inevitable research, which inspired by numerous business challenges. Monitoring and finding firm's competitors have studied in the previous work. Data mining is the best way of managing such enormous information's for mining competitors. Item reviews form online offer rich information about clients' opinions and interest to get over-all idea about competitors. Though, it is usually problematic to know all reviews in dissimilar websites for competitive products and acquire perceptive suggestions manually.

In this paper, we describe the suitable meaning of the competitiveness between two items, on the basis of market sectors. Our estimation of competitiveness uses customer reviews, a plentiful source of knowledge that is available in a large range of domains. we introduce the efficient methods for finding competitiveness in large review datasets and finding the top-k competitors of an item. To find competitiveness in the item K_means algorithm and Naïve Bayes classifier method is used. In this paper, the main role is of datasets, as we are taken large unstructured dataset. Our experiments are based on the online reviews on the corpus of Amazon.com, Yelp.in, and Tripadvisor.com website. In this paper we describe the five different mainstream domains like: {restaurant, e-commerce, shopping, financial services, health and medical}. All the review information is categorized with respect to mainstream domain using k_means clustering algorithm. After clustering, the review dataset is preprocess where elimination of stop words and elimination of special symbol operation are takes place. After preprocessing, feature extraction operation takes place. While evaluating feature extraction results, chi-square algorithm is used to find the number of co-occurrences of attribute in each review. This chi-square results are further process to naive bayes classifier. Naïve Bayes shows results into classification. To finding top-k competitor Apriori algorithm takes place. We are finding top-k competitor on the basis of support factor of 2% and confidence.

A. Classification on multiple domains using naïve bayes algorithm.

4.1 Read Dataset: In the first step, user read the dataset of all domains. The dataset contains reviews of all five domains and these reviews are unstructured reviews.

4.2 K-means Clustering: In this methodology, the unstructured reviews are classified into a particular category. The K_means clustering is an unsupervised learning algorithm. This algorithm is used when working is based on unlabelled data. Here, unlabelled describes the information not labelled with categories or groups. The key point is to find categories in the information. Here, groups or categoery is represented by the variable (k). The working of this algorithm is repetitively allocated each datum point into the one K categories on the basis of attributes or features. This data points are group based on attribute match. In this paper, we are implemented the logic of the k_means clustering algorithm.

4.3 Data Preprocessing: In this step, the cluster information is going to get pre-process. The main goal of stop-words that it should be eliminated from a text which create the text appearance as heavier and low importance for examining. Removing stop-words decreases the dimensionality of term. Several common words in text files or text documents like pro-nouns, prepositions, articles, special symbol etc. that does not give the sense of the document. These are the words which consider as a stopwords. For instance, a, an, the, with, etc. This procedure is used eliminating the stop words of clustered information and removing the special symbols. Using preprocessing operation clustered data is pre-process.

4.4 Chi-Square: The pre-process information is filtered reviews. From these reviews, we selected the unique attributes of all domains and create train and test file. Chi-square method identify the attribute count of each review. It provides the rank to the attributes that shows that how many times the attribute occur in the review. The train and test file are used for testing the outcome of source.

4.5 Naïve Bayes: Naïve Bayesian algorithm signifies the supervised learning procedure as well as numerical method used for the classification. This algorithm is easy and simple probabilistic classifier on the basis of Bayesian theorem with accurate independence assumption. This method used to calculate classification result on the basis of categories i.e. domains and attribute. The chi-square results are further process to the naïve bayes algorithm. In this paper, this algorithm identifies the results on the basis of multiple domains. The formula of naïve bayes is follow as:

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

Where,
P(A) is the prior probability of categoery.
A= Name of categoery i.e. the predicted answer.
P(B) is the prior probability of attribute.
B= (B1, B2, B3……Bn)
P(A/B) is the posterior probability of categoery and specified attribute.
P(B/A) is the likelihood which is the probability of predictor specified category.

B. For finding Top-k competitor on single domain:

4.6 Rule mining: Association rule mining is used for finding the top-k competitor. Association rule mining is very popular and suitable algorithm for identifying relations between variables in large databases. For example, the rule like: {sandwich, coffee} => {burger} finds the data of supermarket. This indicates that if any client or customer buys sandwich or coffee together then that customer is also likely to buy burger. For identifying this number of combinations association rule mining is used. For finding top-k competitor, the reviews are taken from restaurant domain. The reviews are unstructured reviews that are taken from yelp.in and TripAdvisor.com websites. For identifying the interesting occurrences rule from all possible rules support and confidence factor is used. In this paper, this algorithm identifies the results on the basis of single domain. The formula of support and confidence is follow as:

1) Support(X) = number of combinations which holds the itemset X (attributes) /total number of combinations.

2) Confidence (X->Y) = Support (X, Y)/ Support (X).
It signifies the likelihood of item Y being purchased when item X is purchased.
To find best competitor by using Apriori Algorithm:

$$1)\ Support(x) = \frac{Number\ of\ occurances}{Total\ frames\ action}$$

$$2)\ Confidance\ \{x \rightarrow y\} = \frac{Support\{x,y\}}{Support\{x\}}$$

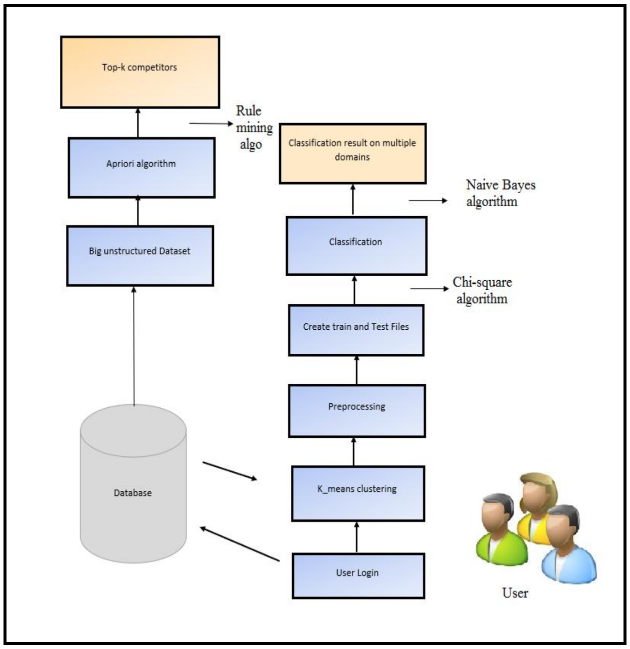


Fig 1: Architecture of proposed system

V. Methodology

i. Algorithm 1: Chi-squared Algorithm

1) Calculate the chi-squared statistic x^2.

2) Determine the number of degrees of freedom (df)of the statistic. This depends on the particular expected distribution but is usually n-1 (where, n is the number of categories).

3) Select a confidence level, usually either 2% or 95% or 99%.

4) Determine the critical value of the x^2 -distribution with (df)degrees of freedom and the confidence level chosen above. Essentially, this is defined as the value x at which the portion of the chi-squared distribution below x is at least the desired confidence level.

5) Compare the chi-squared statistic to the critical value. If it is below the critical value, the null hypothesis is not rejected. If it is above the critical value, the null hypothesis is rejected, and the expected distribution must be wrong.

Chi-square Algorithm:

$$x_c^2 = \sum \frac{(O_i - E_i)}{E_i}$$

Where,

O= observed value

E=Executed value

i= i is the i-th position in the category

c= degrees at freedom.

ii. Algorithm 2: Naive Bayes

Step 1: we have categorized the probability of attributes favouring to domains.
Step 2: probability of attribute present in the review (p>0.5) and for absent(p<0.5).
Step 3: Create a dataset for all attributes. For instance, attributes like, tax, sandwich etc.
Step 4: From the dataset we have obtained the frequency table for each attribute.
Step 5: For each frequency table we find the likelihood for each cases.
Step 6: Now, apply Naive Bayesian formula to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

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

Where,
B = (B1, B2, …………………, Bn)
B = (Tax, Sandwich………………)
A = Yes / No
Here, 'Yes' represent that attribute is present in domains namely: restaurant, finance, ecommerce, shopping, financial, health and medical, others; (only one domain).
And 'No' represent that attribute is present in other domain.
Classification results also holds the information about actual and predicted classification that are follow as:
1)True positive: if the outcome of prediction value is p and the actual value is also p then it is known as true positive (Tp).
2)False Positive: if the actual value is n then it is known as false positive (Fp).

3)Precision: Precision is a ratio of exactness and quality.
Precision = tp / (tp+fp)
4)Recall: It is a measure of completeness and quality.
Recall = tp/ (tp / fn)
5) F-measure: it combines the precision and recall which is the harmonic mean of precision and recall.
F-measure = 2*Tp / (2*Tp)+Fp+Fn.

iii.Algorithm 3: Apriori Algorithm

Step 1: Generate a candidate set or candidate table C_1.In this step, we enlist all the items which is in frequent itemset and minimum support is .2%.
Step 2: Find the support count of item in frequent item set. In this step, first iteration is completed. This is the 1st item-set.
Step 3: Find attributes which follow the minimum support and Generate the set or table L_1 which satisfies the minimum support criteria.

Step 4: Generate another candidate set C_2: In this step, we need to find support count for these attributes together. In this step, second iteration is completed. This is the 2st item-set.

Step 5: Generate the set L_2: In this step, set C_2 contains only those items set which have support value 2 and greater than 2. From previous set, all the elements which is less than support count is eliminated and rest of the items are to be returned. And L_2 table is generated after satisfying the minimum criteria.

Step 6: Generate the next candidate set: C_3: In this step, we create 3 frequent itemsets. In this step we check the Combinations of items in frequent item-set. Third iteration is completed. This is the 3rd item-set.
Step 7: Now, apply the Apriori priority which says that if the itemset is frequent it means it sub-itemset is also frequent. This states that these itemsets are only frequent if the subset of these items is already there in L_2.

Step 8: Generate the set L_3 which contains item set of C_3 table.

Step 9: Generate the candidate set called C_4. In this step, we have to check the combination if the combination is not containing in set L_3. Then from C_3 item-set these 3 item-set have been found.
Step 10: Now, find out association rules from these frequent item set. In order to find the associations, the concept of confidence is come into picture. First generate the types of association rules from these frequent itemset. The association rule may be a combination of these frequent itemsets. Generate the rules for C_3 set as C_3 set contains all rules that present in frequent item set.
Step 11: Reverse all rules.
Step 12: if confidence is 75% some strong rules are selected whose value is 100% or greater than 75%.

VI. Experimental Results

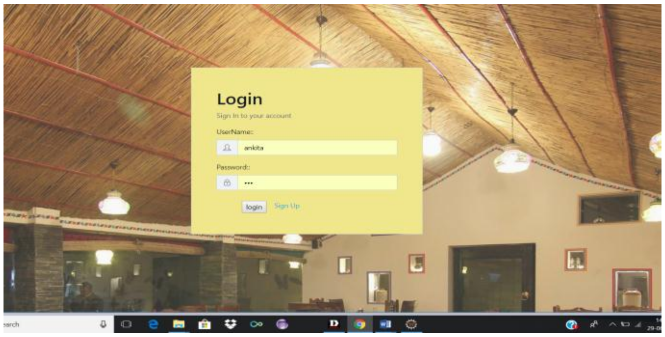
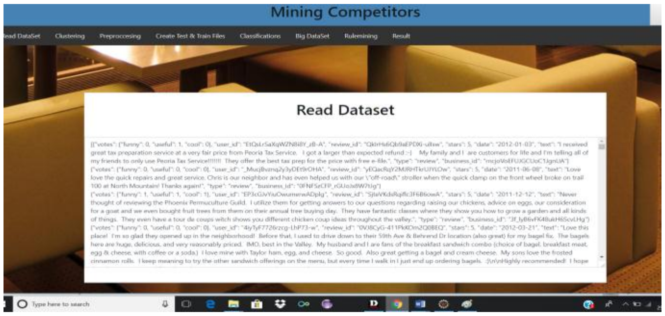


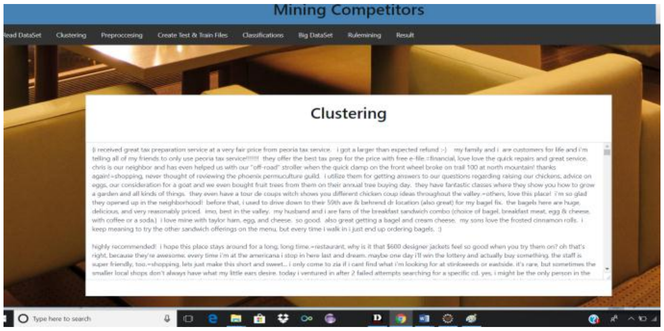
Fig 2: Login Page



Read Dataset

```
[{"votes": {"funny": 0, "useful": 1, "cool": 0}, "user_id": "EtQsLrSaXqWZN8i8Y_zB-A", "review_id": "QkIrHs6Qb9aEPDXG-uIlsw", "stars": 5, "date": "2012-01-03", "text": "I received
great tax preparation service at a very fair price from Peoria Tax Service.  I got a larger than expected refund :-)   My family and I  are customers for life and I'm telling all of
my friends to only use Peoria Tax Service!!!!!!!  They offer the best tax prep for the price with free e-file.", "type": "review", "business_id": "mcjoVoEFUXGCUoC1JgnUA"}
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "_MucjBvanq2y3yDEt9rOHA", "review_id": "yEQacRqY2MJRHTkrUJYLOw", "stars": 5, "date": "2011-06-08", "text": "Love
love the quick repairs and great service. Chris is our neighbor and has even helped us with our \"off-road\" stroller when the quick clamp on the front wheel broke on trail
100 at North Mountain! Thanks again!", "type": "review", "business_id": "0FNFSzCFP_rGUoJx8W7tJg"}
{"votes": {"funny": 1, "useful": 1, "cool": 1}, "user_id": "EP3cGJvYiuOwumerwADplg", "review_id": "SjtaVKdsRqlfc3F6B6owA", "stars": 5, "date": "2011-12-12", "text": "Never
thought of reviewing the Phoenix Permuculture Guild.  I utilize them for getting answers to our questions regarding raising our chickens, advice on eggs, our consideration
for a goat and we even bought fruit trees from them on their annual tree buying day.  They have fantastic classes where they show you how to grow a garden and all kinds
of things.  They even have a tour de coups witch shows you different chicken coup ideas throughout the valley.", "type": "review", "business_id": "3f_lyB6vFK48ukH6ScvLHg"}
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "4iyTyF7726rzcg-LhP73-w", "review_id": "0VJ8CyG-411PkKOm2Q0BEQ", "stars": 5, "date": "2012-03-21", "text": "Love this
place!  I'm so glad they opened up in the neighborhood!  Before that, I used to drive down to their 59th Ave & Behrend Dr location (also great) for my bagel fix.  The bagels
here are huge, delicious, and very reasonably priced.  IMO, best in the Valley.  My husband and I are fans of the breakfast sandwich combo (choice of bagel, breakfast meat,
egg & cheese, with coffee or a soda.)  I love mine with Taylor ham, egg, and cheese.  So good.  Also great getting a bagel and cream cheese.  My sons love the frosted
cinnamon rolls.  I keep meaning to try the other sandwich offerings on the menu, but every time I walk in I just end up ordering bagels.  :)\n\nHighly recommended!  I hope
```

Fig 3: Read Dataset Page

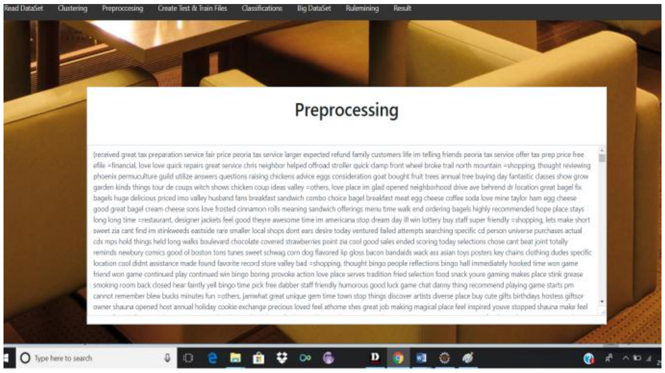


Clustering

```
[i received great tax preparation service at a very fair price from peoria tax service.  i got a larger than expected refund :-)   my family and i  are customers for life and i'm
telling all of my friends to only use peoria tax service!!!!!!!  they offer the best tax prep for the price with free e-file.=financial, love love the quick repairs and great service.
chris is our neighbor and has even helped us with our "off-road" stroller when the quick clamp on the front wheel broke on trail 100 at north mountain! thanks
again!=shopping, never thought of reviewing the phoenix permuculture guild.  i utilize them for getting answers to our questions regarding raising our chickens, advice on
eggs, our consideration for a goat and we even bought fruit trees from them on their annual tree buying day.  they have fantastic classes where they show you how to grow
a garden and all kinds of things.  they even have a tour de coups witch shows you different chicken coup ideas throughout the valley.=others, love this place!  i'm so glad
they opened up in the neighborhood!  before that, i used to drive down to their 59th ave & behrend dr location (also great) for my bagel fix.  the bagels here are huge,
delicious, and very reasonably priced.  imo, best in the valley.  my husband and i are fans of the breakfast sandwich combo (choice of bagel, breakfast meat, egg & cheese
with coffee or a soda.)  i love mine with taylor ham, egg, and cheese.  so good.  also great getting a bagel and cream cheese.  my sons love the frosted cinnamon rolls.  i
keep meaning to try the other sandwich offerings on the menu, but every time i walk in i just end up ordering bagels.  :)

highly recommended!  i hope this place stays around for a long, long time.=restaurant, why is it that $600 designer jackets feel so good when you try them on? oh that's
right, because they're awesome. every time i'm at the americana i stop in here last and dream. maybe one day i'll win the lottery and actually buy something. the staff is
super friendly, too.=shopping, lets just make this short and sweet... i only come to zia if i cant find what i'm looking for at stinkweeds or eastside. it's rare, but sometimes the
smaller local shops don't always have what my little ears desire. today i ventured in after 2 failed attempts searching for a specific cd. yes, i might be the only person in the
```

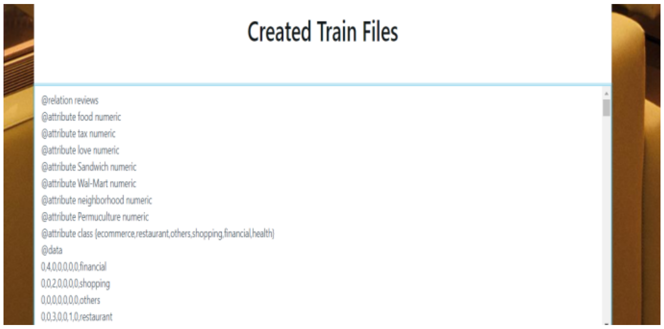
Fig 4: Clustering Page



Preprocessing

```
[received great tax preparation service fair price peoria tax service larger expected refund family customers life im telling friends peoria tax service offer tax prep price free
efile =financial, love love quick repairs great service chris neighbor helped offroad stroller quick clamp front wheel broke trail north mountain =shopping, thought reviewing
phoenix permuculture guild utilize answers questions raising chickens advice eggs consideration goat bought fruit trees annual tree buying day fantastic classes show grow
garden kinds things tour de coups witch shows chicken coup ideas valley =others, love place im glad opened neighborhood drive ave behrend dr location great bagel fix
bagels huge delicious priced imo valley husband fans breakfast sandwich combo choice bagel breakfast meat egg cheese coffee soda love mine taylor ham egg cheese
good great bagel cream cheese sons love frosted cinnamon rolls meaning sandwich offerings menu time walk end ordering bagels highly recommended hope place stays
long long time =restaurant, designer jackets feel good theyre awesome time im americana stop dream day ill win lottery buy staff super friendly =shopping, lets make short
sweet zia cant find im stinkweeds eastside rare smaller local shops dont ears desire today ventured failed attempts searching specific cd person universe purchases actual
cds mps hold things held long walks boulevard chocolate covered strawberries point zia cool good sales ended scoring today selections chose cant beat joint totally
reminds newbury comics good ol boston tons tunes sweet schwag corn dog flavored lip gloss bacon bandaids wack ass asian toys posters key chains clothing dudes specific
location cool didnt assistance made found favorite record store valley bad =shopping, thought bingo people reflections bingo hall immediately hooked time won game
friend won game continued play continued win bingo boring provoke action love place serves tradition fried selection food snack youre gaming makes place stink grease
smoking room back closed hear faintly yell bingo time pick free dabber staff friendly humorous good luck game chat danny thing recommend playing game starts pm
cannot remember blew bucks minutes fun =others, jamwhat great unique gem time town stop things discover artists diverse place buy cute gifts birthdays hostess giftsor
owner shauna opened host annual holiday cookie exchange precious loved feel athome shes great job making magical place feel inspired youve stopped shauna make feel
```

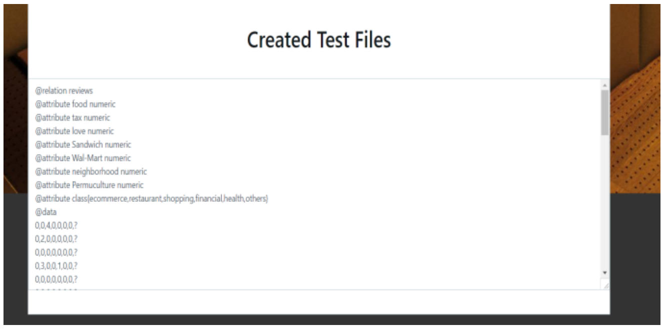
Fig 5: Preprocessing page



Created Train Files

```
@relation reviews
@attribute food numeric
@attribute tax numeric
@attribute love numeric
@attribute Sandwich numeric
@attribute Wal-Mart numeric
@attribute neighborhood numeric
@attribute Permuculture numeric
@attribute class {ecommerce,restaurant,others,shopping,financial,health}
@data
0,4,0,0,0,0,0,financial
0,0,2,0,0,0,0,shopping
0,0,0,0,0,0,0,others
0,0,3,0,0,1,0,restaurant
```

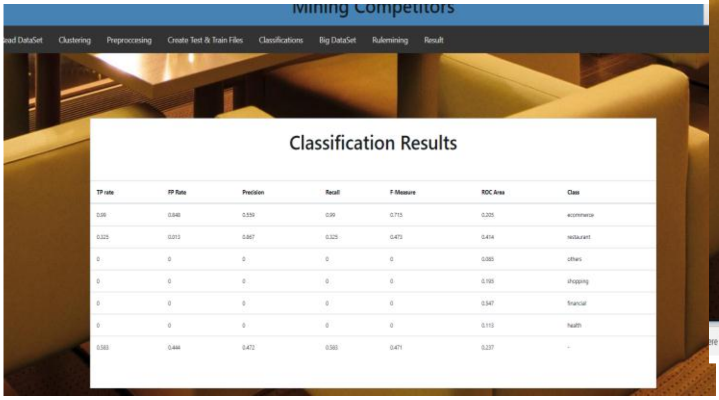
Fig 6: Train file using chi-square



Created Test Files

```
@relation reviews
@attribute food numeric
@attribute tax numeric
@attribute love numeric
@attribute Sandwich numeric
@attribute Wal-Mart numeric
@attribute neighborhood numeric
@attribute Permuculture numeric
@attribute class{ecommerce,restaurant,shopping,financial,health,others}
@data
0,0,4,0,0,0,0,?
0,2,0,0,0,0,0,?
0,0,0,0,0,0,0,?
0,3,0,0,1,0,0,?
0,0,0,0,0,0,0,?
```

Fig 7: Test file using chi-square



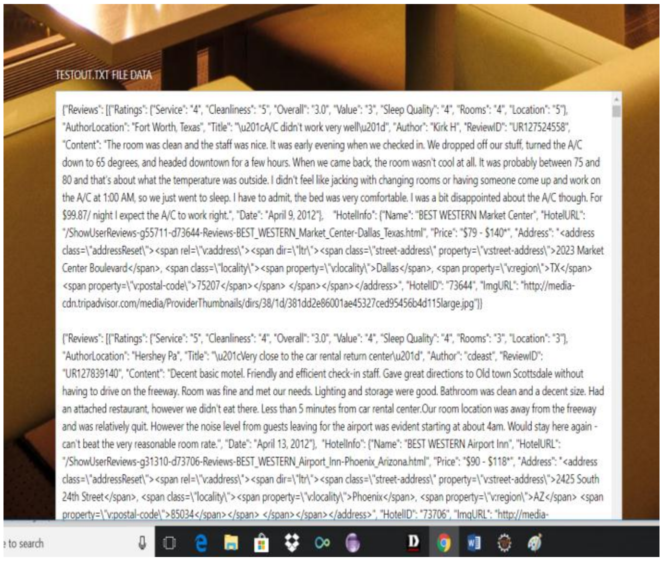
Classification Results

TP rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.99	0.848	0.559	0.99	0.715	0.205	ecommerce
0.325	0.013	0.867	0.325	0.473	0.414	restaurant
0	0	0	0	0	0.085	others
0	0	0	0	0	0.195	shopping
0	0	0	0	0	0.547	financial
0	0	0	0	0	0.113	health
0.583	0.444	0.472	0.583	0.471	0.237	-

Fig: 8 Classification Result

Table 1: Classification Result

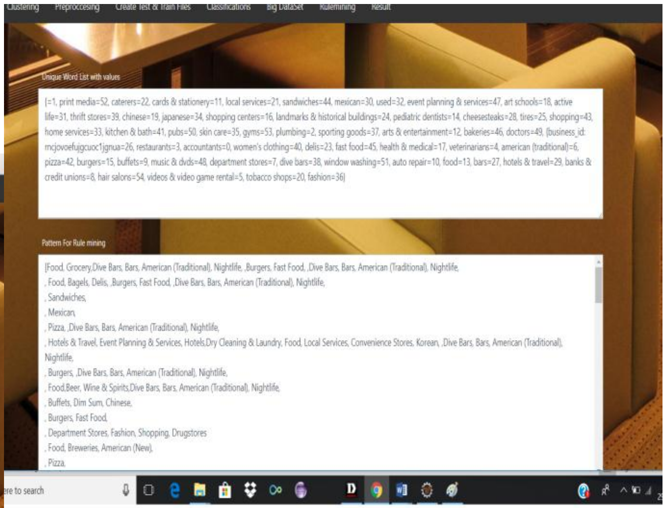
Tp Rate	Fp Rate	Precision	Recall	F-measure	ROC area	Class
0.99	0.848	0.559	0.99	0.715	0.205	e-commerce
0.325	0.013	0.867	0.325	0.473	0.414	Restaurant
0	0	0	0	0	0.195	Shopping
0	0	0	0	0	0.547	Financial
0	0	0	0	0	0.113	Health and Medical
0	0	0	0	0	0.085	Others
0.583	0.444	0.472	0.583	0.471	0.237	Average Weighted Score



```
{"Reviews": [{"Ratings": {"Service": "4", "Cleanliness": "5", "Overall": "3.0", "Value": "3", "Sleep Quality": "4", "Rooms": "4", "Location": "5"},
"AuthorLocation": "Fort Worth, Texas", "Title": "\u201cA/C didn't work very well\u201d", "Author": "Kirk H", "ReviewID": "UR127524558",
"Content": "The room was clean and the staff was nice. It was early evening when we checked in. We dropped off our stuff, turned the A/C
down to 65 degrees, and headed downtown for a few hours. When we came back, the room wasn't cool at all. It was probably between 75 and
80 and that's about what the temperature was outside. I didn't feel like jacking with changing rooms or having someone come up and work on
the A/C at 1:00 AM, so we just went to sleep. I have to admit, the bed was very comfortable. I was a bit disappointed about the A/C though. For
$99.87/ night I expect the A/C to work right.", "Date": "April 9, 2012"},    "HotelInfo": {"Name": "BEST WESTERN Market Center", "HotelURL":
"/ShowUserReviews-g55711-d73644-Reviews-BEST_WESTERN_Market_Center-Dallas_Texas.html", "Price": "$79 - $140*", "Address": "<address
class=\"addressReset\"><span rel=\"v:address\"><span dir=\"ltr\"><span class=\"street-address\" property=\"v:street-address\">2023 Market
Center Boulevard</span>, <span class=\"locality\"><span property=\"v:locality\">Dallas</span>, <span property=\"v:region\">TX</span>
<span property=\"v:postal-code\">75207</span></span> </span></span></address>", "HotelID": "73644", "ImgURL": "http://media-
cdn.tripadvisor.com/media/ProviderThumbnails/dirs/38/1d/381dd2e86001ae45327ced95456b4d115large.jpg"}}

{"Reviews": [{"Ratings": {"Service": "5", "Cleanliness": "4", "Overall": "3.0", "Value": "4", "Sleep Quality": "4", "Rooms": "3", "Location": "3"},
"AuthorLocation": "Hershey Pa", "Title": "\u201cVery close to the car rental return center\u201d", "Author": "cdeast", "ReviewID":
"UR127839140", "Content": "Decent basic motel. Friendly and efficient check-in staff. Gave great directions to Old town Scottsdale without
having to drive on the freeway. Room was fine and met our needs. Lighting and storage were good. Bathroom was clean and a decent size. Had
an attached restaurant, however we didn't eat there. Less than 5 minutes from car rental center.Our room location was away from the freeway
and was relatively quit. However the noise level from guests leaving for the airport was evident starting at about 4am. Would stay here again -
can't beat the very reasonable room rate.", "Date": "April 13, 2012"},  "HotelInfo": {"Name": "BEST WESTERN Airport Inn", "HotelURL":
"/ShowUserReviews-g31310-d73706-Reviews-BEST_WESTERN_Airport_Inn-Phoenix_Arizona.html", "Price": "$90 - $118*", "Address": "<address
class=\"addressReset\"><span rel=\"v:address\"><span dir=\"ltr\"><span class=\"street-address\" property=\"v:street-address\">2425 South
24th Street</span>, <span class=\"locality\"><span property=\"v:locality\">Phoenix</span>, <span property=\"v:region\">AZ</span> <span
property=\"v:postal-code\">85034</span></span> </span></span></address>", "HotelID": "73706", "ImgURL": "http://media-
```

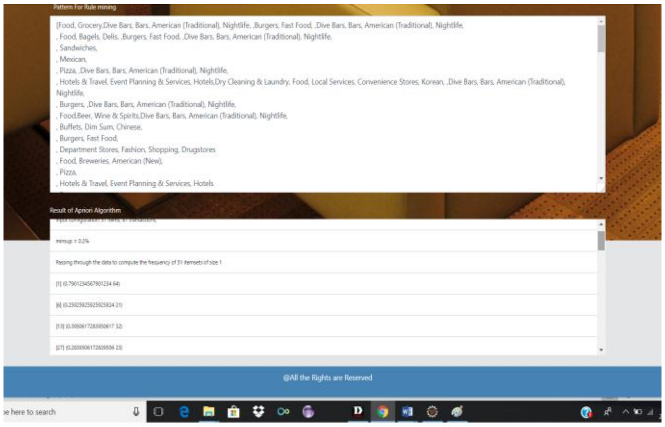
Fig 9: Big Dataset page



```
{=1, print media=52, caterers=22, cards & stationery=11, local services=21, sandwiches=44, mexican=30, used=32, event planning & services=47, art schools=18, active
life=31, thrift stores=39, chinese=19, japanese=34, shopping centers=16, landmarks & historical buildings=24, pediatric dentists=14, cheesesteaks=28, tires=25, shopping=43,
home services=33, kitchen & bath=41, pubs=50, skin care=35, gyms=53, plumbing=2, sporting goods=37, arts & entertainment=12, bakeries=46, doctors=49, {business_id:
mcjovoefujgcuoc1jgnua=26, restaurants=3, accountants=0, women's clothing=40, delis=23, fast food=45, health & medical=17, veterinarians=4, american (traditional)=6,
pizza=42, burgers=15, buffets=9, music & dvds=48, department stores=7, dive bars=38, window washing=51, auto repair=10, food=13, bars=27, hotels & travel=29, banks &
credit unions=8, hair salons=54, videos & video game rental=5, tobacco shops=20, fashion=36}
```

```
[Food, Grocery,Dive Bars, Bars, American (Traditional), Nightlife, ,Burgers, Fast Food, ,Dive Bars, Bars, American (Traditional), Nightlife,
, Food, Bagels, Delis, ,Burgers, Fast Food, ,Dive Bars, Bars, American (Traditional), Nightlife,
, Sandwiches,
, Mexican,
, Pizza, ,Dive Bars, Bars, American (Traditional), Nightlife,
, Hotels & Travel, Event Planning & Services, Hotels,Dry Cleaning & Laundry, Food, Local Services, Convenience Stores, Korean, ,Dive Bars, Bars, American (Traditional),
Nightlife,
, Burgers, ,Dive Bars, Bars, American (Traditional), Nightlife,
, Food,Beer, Wine & Spirits,Dive Bars, Bars, American (Traditional), Nightlife,
, Buffets, Dim Sum, Chinese,
, Burgers, Fast Food,
, Department Stores, Fashion, Shopping, Drugstores
, Food, Breweries, American (New),
, Pizza,
```

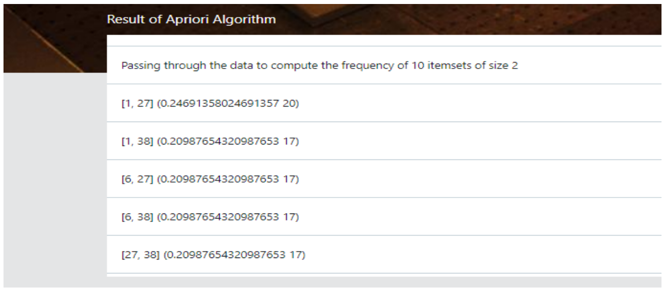
Fig 10: Results of Rule Mining page-1



```
[Food, Grocery,Dive Bars, Bars, American (Traditional), Nightlife, ,Burgers, Fast Food, ,Dive Bars, Bars, American (Traditional), Nightlife,
, Food, Bagels, Delis, ,Burgers, Fast Food, ,Dive Bars, Bars, American (Traditional), Nightlife,
, Sandwiches,
, Mexican,
, Pizza, ,Dive Bars, Bars, American (Traditional), Nightlife,
, Hotels & Travel, Event Planning & Services, Hotels,Dry Cleaning & Laundry, Food, Local Services, Convenience Stores, Korean, ,Dive Bars, Bars, American (Traditional),
Nightlife,
, Burgers, ,Dive Bars, Bars, American (Traditional), Nightlife,
, Food,Beer, Wine & Spirits,Dive Bars, Bars, American (Traditional), Nightlife,
, Buffets, Dim Sum, Chinese,
, Burgers, Fast Food,
, Department Stores, Fashion, Shopping, Drugstores
, Food, Breweries, American (New),
, Pizza,
, Hotels & Travel, Event Planning & Services, Hotels
```

```
Input configuration: 51 items, 81 transactions,
minsup = 0.2%
Passing through the data to compute the frequency of 51 itemsets of size 1
[1] (0.7901234567901234 64)
[6] (0.25925925925925924 21)
[13] (0.3950617283950617 32)
[27] (0.2839506172839506 23)
```

Fig 11: Results of Rule Mining page-2



Result of Apriori Algorithm

Passing through the data to compute the frequency of 10 itemsets of size 2

[1, 27] (0.24691358024691357 20)

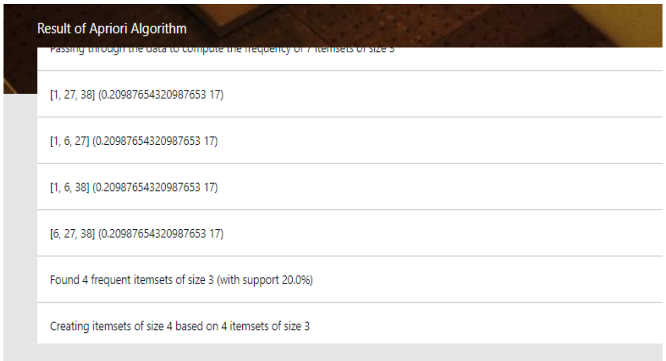
[1, 38] (0.20987654320987653 17)

[6, 27] (0.20987654320987653 17)

[6, 38] (0.20987654320987653 17)

[27, 38] (0.20987654320987653 17)

Fig 12: Results of Rule Mining page-3



Result of Apriori Algorithm

Passing through the data to compute the frequency of 7 itemsets of size 3

[1, 27, 38] (0.20987654320987653 17)

[1, 6, 27] (0.20987654320987653 17)

[1, 6, 38] (0.20987654320987653 17)

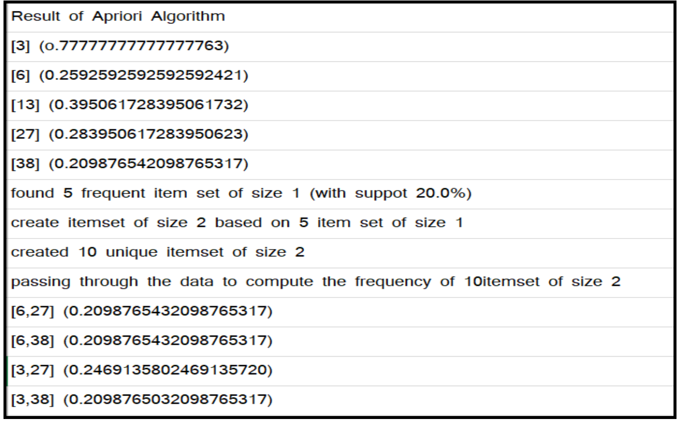
[6, 27, 38] (0.20987654320987653 17)

Found 4 frequent itemsets of size 3 (with support 20.0%)

Creating itemsets of size 4 based on 4 itemsets of size 3

Fig 13: Results of Rule Mining page-4

Table 2: Rule Mining Results



Result of Apriori Algorithm
[3] (o.77777777777777763)
[6] (0.2592592592592592421)
[13] (0.395061728395061732)
[27] (0.283950617283950623)
[38] (0.209876542098765317)
found 5 frequent item set of size 1 (with suppot 20.0%)
create itemset of size 2 based on 5 item set of size 1
created 10 unique itemset of size 2
passing through the data to compute the frequency of 10itemset of size 2
[6,27] (0.2098765432098765317)
[6,38] (0.2098765432098765317)
[3,27] (0.2469135802469135720)
[3,38] (0.2098765032098765317)

VII. CONCLUSION AND FUTURE WORK

We concluded the definition of competitiveness which is applicable for domains, overcoming limitations of earlier issues. 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. We presented the definition of competitiveness between two items and identifying the top-k competitors from large unstructured datasets. The efficiency of our model was tested via an experimental evaluation on real datasets from different domains. At last, the Naïve Bayes and Apriori yielded slightest calculation time when competing at others. Our experimental result is only on the basis of feature extraction outcomes. The results for finding competitiveness of the two item, feature extraction procedure is used. Subsequently, this development also works with more accuracy for efficient results.

This strategy is also used to apply in different workspace like market sector, business organization. It can also be developed into an android application. Additionally, this strategy is also compatible with KNN approach for comparing competitiveness between multiple domains with computation time.

VIII. ACKNOWLEDGMENT

We might want to thank the analysts and also distributors for making their assets accessible. We additionally appreciative to the commentator for their significant recommendations furthermore thank the collage powers for giving the obliged base and backing.

IX. REFRENCES

[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.

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

[10] S. Bao, R.Li,Y.Yu,andY.Cao, "Competitorminingwiththeweb," IEEE Trans. Knowl. Data Eng., 2008.

[11] G. Pant and O. R. L. Sheng, "Avoiding the blind spots: Competitor identification using web text and linkage structure," in ICIS, 2009.

[12] D. Zelenko and O. Semin, "Automatic competitor identification from public information sources," International Journal of Computational Intelligence and Applications, 2002.

[13] 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.

[14] K. Lerman, S. Blair-Goldensohn, and R. McDonald, “Sentiment summarization: evaluating and learning user preferences,” in ACL, 2009, pp. 514–522.

[15] C.-T. Ho, R. Agrawal, N. Megiddo, and R. Srikant, “Range queries in olap data cubes,” in SIGMOD, 1997, pp. 73–88.

[16]Y.L Wu, D. Agrawal, and A. ElAbbadi, “Using wavelet decomposition to support progressive and approximate range-sum queries over data cubes,” in CIKM, ser. CIKM ’00, 2000, pp. 414–421.

[17] D. Gunopulos, G. Kollios, V. J. Tsotras, and C. Domeniconi, “Approximating multi-dimensional aggregate range queries over real attributes,” in SIGMOD, 2000, pp. 463–474.

[18] M. Muralikrishna and D. J. DeWitt, “Equi-depth histograms for estimating selectivity factors for multi-dimensional queries,” in SIGMOD, 1988, pp. 28–36.

[19] N. Thaper, S. Guha, P. Indyk, and N. Koudas, “Dynamic multidimensional histograms,” in SIGMOD, 2002, pp. 428–439.

[20] K.-H. Lee, Y.-J. Lee, H. Choi, Y. D. Chung, and B. Moon, “Parallel data processing with mapreduce: a survey,” AcM sIGMoD Record, vol. 40, no. 4, pp. 11–20, 2012.

[21] S.B¨ orzs¨onyi, D. Kossmann, and K. Stocker, “The skyline operator,” in ICDE, 2001.

Authors Profile

Ms. A.A. Kushwah pursed Bachelor of ngineering from Dr. Babasaheb Ambedkar, College of Engineering and Research, Nagpur University, Nagpur, India, in year 2015. She is a student of Bharti Vidyapeeth (Deemed to be University), College of Enieering, Pune. She is Currently Pursuing Master of Technology (M-tech) in Information Technology Department.

Prof. Y.C. Kulkarni pursed M.E.(Computer Enginerring from Bharati vidyapeeth (Deemed to be University) College of Engineering, Pune. Her main research work focuses on Software Engineering, Data Mining, Information Search and Retrival etc. She has 25 years of teaching experience and 4 years of Research Experience.