Mining Competitor from Large Unstructured Data sets using C Miner

Shivangi Sane
Department of Computer Engineering
PES Modern College Of Engineering

Prof. Dr. Mrs. S. A.Itkar
Department of Computer Engineering
PES Modern College Of Engineering

Abstract—In Present World's competitive business, the achievements is totally in light of capacity to make things more engaging to clients than the oppositions. Data Mining approach for identifying and monitoring firm's competitors and to solve number of question like standardizing and quantifying the competitiveness between two items, finding the main contenders of given item And the features of item are effective or not, finding competitiveness between two items based on the market segments that they can both cover is calculated here. By obtaining a wide source of information and customer reviews,the new formalization of the competitiveness between two items based on market segments that they can both cover, is made. Proficient strategies for assessing aggressiveness in extensive audit datasets and address the regular issue of finding the best k contenders of a given thing. Our assessment of aggressiveness uses client surveys, a plenteous well-spring of data that is accessible in an extensive variety of spaces. We display proficient strategies for assessing intensity in huge survey datasets and address the characteristic issue of finding the best k competitors of a given item.

Index Terms—Data mining, Web mining, Information Search and Retrieval.

I. INTRODUCTION

Focused insight at first arranges the potential hazard and chances by gathering the data about the setting to deal with the supervisor in settling on strategic choices for an association. Much association perceives the hugeness of aggressive insight in big business chance administration and choose emotionally supportive network. They likewise put a lot of cash in aggressive insight. The principal centrality of client decisions, e.g., in connection with new item extension systems. These strategies are extensively insisted in promoting research. Generally client decisions are assessed through conjoint investigation utilizing on the web or paperpencil study. However, this sort of decisions can exceptionally cost with reference to time and money.A Long queue of research has shown the key significance of recognizing and observing a company's rivals. Propelled by this issue, the showcasing and the executives network have concentrated on experimental techniques for contender distinguishing proof just as on strategies for investigating known contenders .Extant research on the previous has concentrated on mining near articulations (for example item A is superior than Item B") from the Web or other literary sources. In spite of the way that such verbalizations can without a doubt be markers of force, they are absent in various territories. For instance, consider the region of journey packs (e.g flight-lodging ,vehicle mixes). For this circumstance, things have no doled out name by which they can be addressed or stood out from each other. Further, the repeat of printed comparable confirmation can move fundamentally across over spaces. For example, when seeing brand names at the firm measurement for example "Google versus Yahoo" or "Sony versus Panasonic", similar examples can be found by just questioning the web. Be that as it may, it is anything but difficult to distinguish standard spaces where such proof is very rare for example, shoes, adornments, lodgings, eateries, and furniture. Supported by these inadequacies, we propose another formalization of the aggressiveness between two things, in view of the market sections that they can both cover.

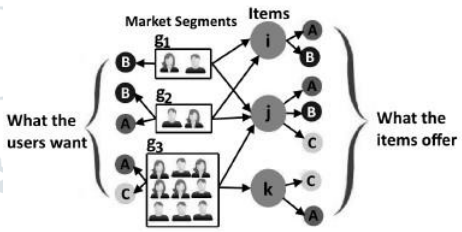

Fig. 1. A (simplified) example of our competitiveness

A. OBJECTIVE

1) To evaluate competitiveness between items in large review data sets.
2) To evaluate the quality and stability of proposed approach.

II. REVIEW OF LITERATURE

Our work has connections to past work from different domains."Web impressions of firms: Utilizing on the web isomorphism for contender identification" this paper by G. Gasp and O. R. Sheng is another online measurements is utilized which depends on the substance, in-connections and outlinks of firm's sites to gauge the nearness of online isomorphism just as reveal its utility in anticipating contender relationships.[1]

Identifying customer preferences about tourism products using an aspect-based opinion mining approach is the another work done which uses opinions available on the web as reviews.[2] This uses similarity concept. There are some other methods like using a map-reduce techniques and processing parallel data which is in massive amount[3] and another approach is mining competitor relationships from online news[4] etc.

Recognizing client inclinations about the travel industry items utilizing an angle based on inparticle mining approach is the another work done which utilizes assessments accessible on the web as reviews.[2] This uses comparability

idea. There are some different strategies like utilizing a outline methods and preparing parallel information which is in huge amount[3] and another methodology is mining contender connections from online news[4] and so on.

Administrative Competitor Identification: The administration writing is rich with works that attention on how chiefs can physically distinguish contenders. A portion of these works show contender identification as a psychological arrangement process in which chiefs create mental portrayals of contenders and use them to characterize applicant firms [5], [6], [7]. Other manual arrangement strategies depend on market-and asset based likenesses between a firm and hopeful contenders [8], [9], [10]. At long last, administrative contender identification has additionally been displayed as a sense making procedure in which contenders are identified dependent on their capability to compromise an associations character [11]

Contender Mining Algorithms: Zheng et al. [12] distinguish key aggressive measures (for example piece of the pie, offer of wallet) and indicated how a firm can gather the estimations of these measures for its rivals by mining (I) its own definite client exchange information what's more, (ii) total information for every contender. As opposed to our own system, this approach isn't suitable for assessing the aggressiveness between any two things or ms in a given market. Rather, the creators accept that the arrangement of contenders is given what's more, in this way, they will probably figure the estimation of the picked measures for every contender. Furthermore, the reliance on value-based information is a confinement we don't have. Doanatl. investigate client appearance information, for example, the geocoded information from area based social systems, as a potential asset for contender mining [13]. While they report promising results, the reliance on appearance information restricts the arrangement of spaces that can benefit from this methodology. Gasp and Sheng conjecture and confirm that contending rms are likely to have comparative web impressions, a wonder that they allude to as online isomorphism [2]. Their examination considers diffierent sorts of isomorphism between two firms, for example, the cover between the in-connections and outlinks of their separate sites, just as the occasions that they seem together on the web (for example in query items or new articles).

Finding Competitive Products: Recent work has investigated intensity in the setting of item structure. The first venture in these methodologies is the definition of a predominance work that speaks to the estimation of an item. The objective is then to utilize this capacity to make things that are not overwhelmed by other, or amplify things with the most extreme conceivable strength esteem. A comparable profession speaks to things as focuses in a multidimensional space and searches for subspaces where the intrigue of the thing is augmented. While significant, the above ventures have a totally difierent center from our own, and henceforth the proposed methodologies are not material in our setting.

Contender distinguishing proof is a key errand for administrators keen on checking their focused landscape, shoring up their barriers against likely aggressive invasions, and arranging focused assault and reaction systems. It is an important antecedent to the undertaking of contender investigation, and the beginning stage for examining the elements of aggressive system (Smith et al., 1992). Previously one can evaluate the relative qualities and shortcomings of opponents, or track aggressive moves and countermoves, one should initially distinguish the focused set and build up an exact feeling of the space in which key associations are probably going to happen. The motivation behind this paper is to give a lot of tractable structures for contender ID what's more, rival examination that encourage expansive natural filtering. To illuminate our structures, we obtain from Peteraf and Bergen's (2001) structure for contender examination. Their work obtains from Chen's (1996) model of contender examination, adjusting his develops to our motivations by illustration on the advertising writing on buyer conduct (Levitt, 1960; Nedungadi, 1990; Peter and Olson, 1993, Mowen what's more, Minor, 1995). In particular, we bring into sharp center the job of client needs in characterizing the commercial center to indicate how a more prominent acknowledgment of client needs can grow consciousness of what sneaks on the aggressive skyline

III. SYSTEM ARCHITECTURE/ SYSTEM OVERVIEW

The system is using C Miner Algorithm for Comparison amongst hotels (Products). our System proposes a new concept of the comparison between two items, based on the market segments that they can both consist. Market reviews are used in application.Application describes a method for computing all the segments in a given market based on mining large review data sets(hotel).Comparison of products allows us to operationalize our definition of competitiveness and address the problem of finding the top-k competitors of an item in any given market.

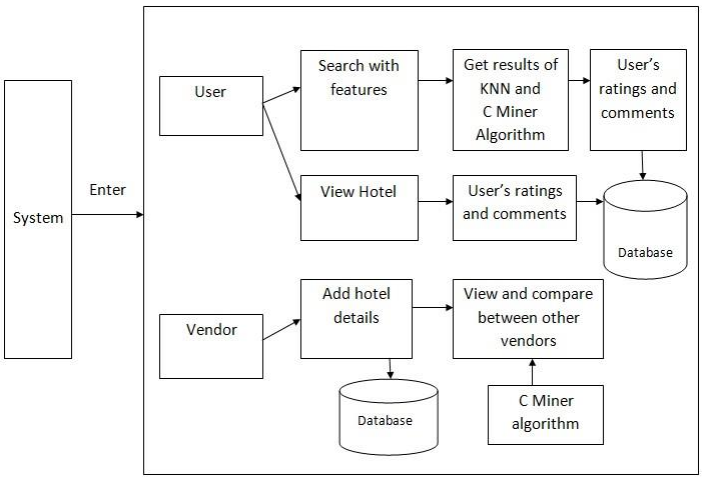

Fig. 2. Proposed System Architecture

A. Explanations:-

Module 1 - User (Customer)

Users first register to the system and search the hotel with features then get top k hotels according to C Miner Algorithms Result.

Module 2 - Vendor (hotel Owner)

Vendors first register and login into system after vendor Add own Hotel and vendor also show comparison results according to C Miner.

B. Advantage :-

1) Our work is the first to address the evaluation of competitiveness via the analysis of large unstructured data sets, without the need for direct comparative evidence.
2) A formal definition of the competitiveness between two items, based on their appeal to the various customer segments in their market
3) A formal philosophy for the recognizable proof of the distinctive kinds of clients in a given market, just as for the estimation of the level of clients that have a place with each sort.
4) A highly salable framework for finding the top-k competitors of a given item in very large data sets

C. Algorithms

1) C Miner Algorithm

Input: Set of items I, Item of interest i I, feature space F, Collection Q 2F of queries with non-zero weights, skyline pyramid DI, int k

Output: Set of top-k competitors for i

Steps

1 TopK masters(i)
2 if (k —TopK—) then
3 return TopK
4 end if
5 k ← $k|TopK|$
6 $LB \leftarrow 1$
7 $X \leftarrow GETSLAV\ ES(TopK;DI)DI[0]$
8 $while(|X|! = 0)do$
9 $X \leftarrow UPDATETOPK(k;LB;X)$
10 $if(|X|! = 0)then$
11 $TopK \leftarrow MERGE(TopK;X)$
12 $if(|TopK| = k)then$
13 $LB \leftarrow WORSTIN(TopK)$
14 $endif$
15 $X \leftarrow GETSLAV\ ES(X;DI)$
16 $endif$
17 $endwhile$
18 $returnTopK$
19 $RoutineUPDATETOPK(k,LB,X)$
20 $localTopK \leftarrow$
22 $up(j) \leftarrow P_{q \in Q}\, pq * v_{i,jq}$
23 for every q Q do
24 maxV ← $v^{q}_{i,j}$
25 for every item j ∈ Xdo
26 $up(j) \leftarrow up(j)$- maxV + p(q) $v^{q}_{i,j}$
27 if (up(j) ¡ LB) then
28 X ← $X\,j$
29 $else$ 30 $low(j) \leftarrow low(j)$+ p(q) $v^{q}_{i,j}$
33 LB ← $WORSTIN(localTopK)$
34 $endif$
35 $endif$
36 $endfor$
37 $if(|X|k)then$
38 $break$
39 $endif$
40 $endfor$
41 $foreveryitemjXdo$
42 $foreveryremainingqQdo$
43 $low(j) \leftarrow low(j)$+ p(q) $v^{q}_{i,j}$
44 end for
45 localTopK:update(j; low(j))
46 end for
47 return TOPK(localTopK)

D. Mathematical Model

We define competitiveness between i and j in a market with a feature subset F as follows

$$C_F i, j = \sum_{q \epsilon 2^F} p(q) * V_{i,j}^{\,q} \ldots\ldots\ldots\ldots (1)$$

where $C_F i,j$ represents the probability that the two items are included in the consideration set of random user.

p(q) represents the percentage of users represented by each query q

V_{i,q_j} is a pairwise coverage of a query that includes binary, categorical , ordinal or numeric features.

Pairwise coverage of feature query

P(q)=$\frac{Freq(q,R)}{\sum_{q\in 2^F} Freq(q',R)}$ .....(2)

$V_{i,j}^{f}$= f[i] * f[j] (Binary Features) .....(3)

$V_{i,j}^{f}$=min(f[i],f[j]) (numeric features) .....(4)

$V_{i,j}^{f}=\frac{min(f[i],f[j])}{|v^f|}$ (ordinal features) .....(5)

CMiner uses the following update rules for the lower and upper bounds for a candidate j:

low(j) ← low(j) + p(q) * $v_{i,j}^{q}$ .....(6) up(j)←

up(j)- p(q) * $v_{i,j}^{q}$ + p(q) $*v_{i,j}^{q}$

.....(7)

By expanding the sequences and using the initial values low(j) = 0 and up(j) = CF(i, i), we can re-write the bounds:

low$^{m}(j)=\sum_1^m p(q_m) * v_{i,j}^{q_m}$ .....(8)

up$^{m}(j)$ =C$_f(i,j)$ .....(9)

where lowm(j) and upm(j) are the values of the bounds after considering the mth query qm. We can then define a recursive function T(j) = up(j) − low(j) as follows:

T(j) ← T(j) $-p(q) * v_{i,j}^{q}$.....(10)

T$^{m}(j)=C_f(i,j)-\sum_{l=1}^{m} p(ql) * v_{i,j}^{q}$.....(11)

E. HARDWARE AND SOFTWARE REQUIREMENTS

Hardware Requirements

1) Processor - Intel i5 core
2) Speed - 1.1 GHz
3) RAM - 2GB
4) Hard Disk - 40 GB
5) Key Board - Standard Windows Keyboard
6) Mouse - Two or Three Button Mouse
7) Monitor - SVGA
8) Floppy Drive - 44 Mb Software Requirements

1) Operating System - XP, Windows7/8/10
2) Coding language - Java, MVC, JSP, HTML, CSS etc
3) Software - JDK1.7
4) Tool - Eclipse Luna
5) Server - Apache Tomcat 8.0
6) Database - MySQL 5.0

IV. SYSTEM ANALYSIS AND RESULT

Our experiments include data sets which were collected for the purposes of this project were intentionally selected from different domains to portray the cross-domain applicability of our approach.The Data sets are made by adding the different hotels of Pune City with different features.In addition to the full information on each item in our data sets, we also collected the full set of reviews that were available on the online sources like websites of hotels. The highly-cited method by CMiner Algorithms is used to convert each review to a vector of opinions, where each opinion is defined as a feature-polarity combination (e.g. service,Room Available,Free Wifi-service).

Evidence on Comparative Methods :

	Cooccurrences (m-sec)	comparative (m-sec)
hotel With Restaurants	1.7	1.2
hotel Without Restaurants	0.09	1

Comparison between Algorithms

Algorithms Name	Execution time in m-sec
Naive-Bayes Algorithms	1.7
C Miner	0.09

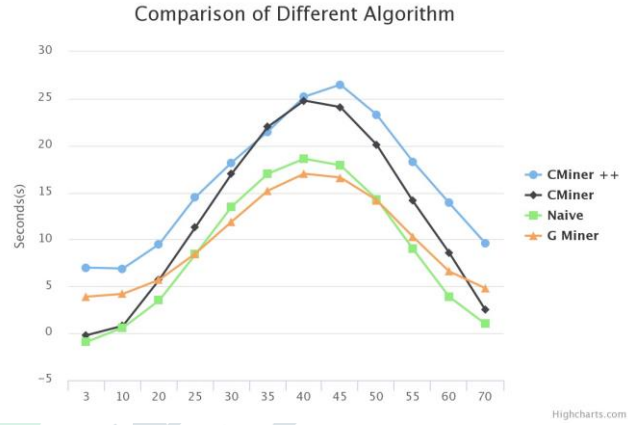

Fig. 3. Result after comparing different algorithms

V. CONCLUSION

Our displayed work formalizes meaning of aggressiveness between two things, which we have approved both quantitatively and subjectively. Our formalization is pertinent crosswise over areas, defeating the deficiencies of past methodologies. We consider various components that have been to a great extent disregarded before, for example, the situation of the things in the multi-dimensional element space and the inclinations and suppositions of the clients. Our work acquaints an end-with end approach for mining such data from expansive datasets of client surveys. In light of our intensity definition, we tended to the computationally difficult issue of finding the best k contenders of a given thing. The proposed system is effective and appropriate to areas with extremely expansive populaces of things.

REFERENCES

[1] G. Pant and O. R. Sheng, Web footprints of firms: Using online isomorphism for competitor identification, *Information Systems Research, vol. 26, no. 1, pp. 188209*, 2015.

[2] black E. Marrese-Taylor, J. D. Velasquez, 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. 182191,*, 2013.

[3] blackK.-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. 1120*, 2012.

[4] blackS 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.

[5] blackB. H. Clark and D. B. Montgomery, Managerial Identification of Competitors, *Journal of Marketing.*, 1999.

[6] blackJ. F. Porac and H. Thomas, Taxonomic mental models in competitor definition, *The Academy of Management Review*, 2008.

[7] blackF. Porac and H. Thomas, Taxonomic mental models in competitor definition, *Academy of Management Review, vol. 15, no. 2, pp. 224240*, 1990.

[8] M. E. Porter, Competitive Strategy: Techniques for Analyzing Industries and Competitors. *Free Press*, 1980.

[9] M. Bergen and M. A. Peteraf, Competitor identification and competitor analysis: a broad-based managerial approach, *Managerial and Decision Economics*, 2002.

[10] M.-J. Chen, Competitor analysis and interfirm rivalry: Toward a theoretical integration, *Academy of Management Review*, 1996

[11] W. T. Few, Managerial competitor identification: Integrating the categorization, economic and organizational identity perspectives, *Doctoral Dissertaion*, 2007

[12] Z. Zheng, P. Fader, and B. Padmanabhan, From business intelligence to competitive intelligence: Inferring competitive measures using augmented site-centric data, *Information Systems Research, vol. 23, no. 3-part-1, pp. 698720*, 2012.

[13] 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. 283289.

[14] R. Li, S. Bao, J. Wang, Y. Yu, and Y. Cao, Cominer: An effective algorithm for mining competitors from the web, in *ICDM*, 2006.

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