Mining Large Unstructured Datasets to Find Top-K Competitors

Md. Irfan[1], H. Ateeq Ahmed[2]

[1]PG Scholar, Dept. Of CSE, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, A.P.

[2]Assistant Professor, Dept. Of CSE, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, A.P.

ABSTRACTL: In the current competitive business scenario, there is a need to analyze the competitive features and factors of an item that most affect its competitiveness.The evaluation of competitiveness always uses the customer opinions in terms of reviews, ratings and abundant source of information's from the web and other sources. This paper develops an augmented competitor mining using product reviews. The itemsets are analyzed for selecting the relevant features. Using the c-miner, the frequent items are discovered and then represented by skyline operators. However, if all customer data is inserted into a database, the resulting records will provide a detailed profile of these customers and their interactions with one another, and will be an important resource for businesses that wish to probe customer data, customer needs, and customer satisfaction levels. Experimental analysis has shown the efficiency of the proposed algorithm

I.INTRODUCTION

Data mining is the popular area which facilitates for improvement in business by mining user requirements and user references to get information about products (or) services and mine the competitors of a specific business. From past decades of research has demonstrated the importance of identifying competitors of an item (or) a product. Mainly marketing and management community have focused very much on identifying competitors. Item reviews from online provides the information about customer opinions and from that we can get general idea about competitors[13].

Our competitiveness paradigm is based on the following observation that competitiveness between two items is based on whether they compete for attention and business of same group of customers [1] for example a user is trying to pick a restaurant for dinner and he has a limited budget and only interested in continental food and also has idea of location that should be nearer to beach. So, only those restaurants that satisfy these criteria will compete for user‘s attention. On the other hand, the restaurants which are not having continental food and also very expensive are not competitors for this particular user and they don‘t have chance to compete [6].

The fig. 1 illustrates competitiveness between four items [1] A, B, C, D and these items are mapped to features that they are offering to the users. Three features are considered in this example i,j,k. G1, G2, G3, G4 are different group of customers and they are grouped based on their preferences. For example the customers in G3 are only interested in j and K. In this we can say that B is competitive with items A, C, D. Here B is highly competitive with Asince it is competing for 14 users with A whereas with C it is competing for 4 users and with D for 12 users. So, in this market B is highly competitive.

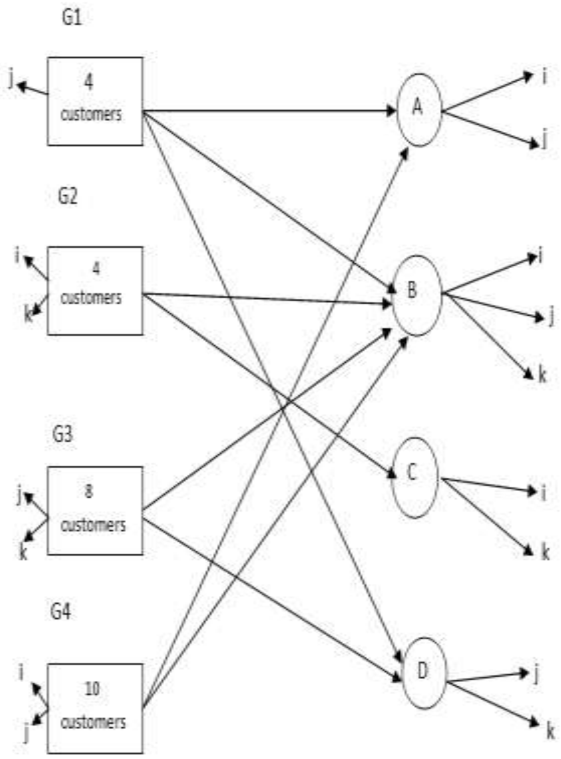

Figure 1: Example for competitiveness model

This example illustrates the ideal scenario, in which we have access to the complete set of customers in a given market, as well as to specific market segments and their requirements. In practice, however, such information is not available. In order to overcome this, we describe a method for computing all the segments in a given market based on mining large review datasets. This method 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. As we show in our work, this problem presents significant computational challenges, especially in the presence of large datasets with hundreds or thousands of items, such as those that are often found in mainstream domains. We address these challenges via a highly scalable framework for top-k computation, including an efficient evaluation algorithm and an appropriate index. Our work makes the following contributions:

- 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.

II. RELATED WORK

This section presents the prior work suggested in competitor mining. Authors in [5] developed an automatic system that discovers competing companies from public information sources. In this system data is crawled from text and it uses transformation oriented learning to obtain appropriate data normalization, combines structured and unstructured information sources, uses probabilistic modeling to represent models of linked data, and succeeds in autonomously discovering competitors. Bayesian network for competitor identification technique is used. The authors also introduced the iterative graph reconstruction process for inference in relational data [6], and shown that it leads to improvements in performance. To find the competitors, the authors used machine learning algorithms and probabilistic approaches. They also validate system results and deploy it on the web as a powerful analytic tool for individual and institutional investors. However, the technique has many problems like finding alliances and market demands using the machine learning approach. In the paper [7], authors presented a formal definition of the competitiveness between two items. Authors used many domains and handled many shortcomings of previous works. In this paper, the author considered the position of the items in the multi-dimensional feature space, and the preferences and opinions of the users. However, the technique addressed many problems like finding the top-k competitors of a given item and handling structured data. Authors in [8] proposed a new online metrics for competitor relationship predicting. This is based on the content, firm links and website log to measure the presence of online isomorphism, here the Competitive isomorphism, which is a phenomenon of competing firms becoming similar as they mimic each other under common market services [9]. Through different analysis they find that predictive models for competitor identification based on online metrics are largely superior to those using offline data. The technique is combined the online and offline metrics to boost the predictive performance. The system also performed the ranking process with the considerations of likelihood. Several works in the same strategy in literature have discussed the need for accurate identification of competitors and provided theoretical frameworks for that. Given the expected isomorphism between competing firms, the process of competitor identification through pair-wise analysis [10] of similarities between focal and target firms is well founded. The unit of analysis is a pair of firms since competitor relationship is seen as a unique interaction between the pair. Authors in [11] have suggested frameworks for manual identification of competitors. The manual nature of these frameworks makes them very costly for competitor identification over a large number of focal and target firms, and over time [12]. In the paper[13], authors attempts to accomplish a novel task of mining competitive information with respect to an entity , the entity such as a company, product or person from the web. The authors proposed an algorithm called “CoMiner”, which first extracts a set of comparative candidates of the input entity and then ranks them according to the comparability, and finally extracts the competitive fields. But the CoMiner [14] specifically developed to support for specific domain. However the effort for the further domains is still challenging. Authors in [15] have proposed ranking methods to give the competitor in a ranked way. They have used data from location based social media. Authors proposed the use of Page-Rank model and it’s variant to obtain the Competitive Rank of firms. However mining competitors from the social media developed many privacy related issues.

III.PROBLEM DEFINITION

In e-commerce application it is very difficult to identify the competition among the product. In market the executionand comparative analysis of the product is going to achieve on the basis of manual comparative analysis of the reviewscomments and on the basis of that evaluation of product will done. So we need to provide the atomized way to make allprocess efficient and give the results on single click.

IV. IMPLEMENTATION AND PROPOSED MODEL

This section presents the working of our proposed model. The proposed model composes of four phases and it’s explained as follows:

a) Administrator phase:

Administrator phase is the first step who takes the responsibility of the administering the user’s activities and upload the items like hotels, recipes and cinemas etc. They also check the profile details, customer queries and their interests using C- miner. C-miner falls under the class of frequent sequence mining. It operates on discovering correlations in data blocks. Each data item is mapped with the blocks. In similar way, the search operation is processed in sequence to sequence mapping. By doing so, the frequently searched item are extracted and sorted to top k itemsets.

b) Customer phase:

Customer phase is the second step which operates on requirements of customers. Based on the queries given by the customers, the results will be displayed. For each item, the database is created. Relied upon the customer requirements, the requested data item is displayed.

c) C-miner algorithm:

This step depicts how the top k itemsets are retrieved. C-miner algorithm is used to discover the top k competitors for given item. It iterates like skyline pyramid which reduces the dimensionality. Each item is mapped with its corresponding blocks. The items are selected on basis of correlation score. Therefore, during competitor mining process, unstructured data is not taken into account and much valuable service information is lost. Structured systems are those where the data and the computing activity is predetermined and well- defined. Unstructured systems are those that have no predetermined form or structure and are usually full of textual data.

d) Skyline operator:

The skyline operator is performed for the coordinate points of data items. It helps to determine the subset of points which dominates the other set of points. Generally, the skyline function is given as sky (I). Given the skyline Sky (I) of a set of items I and an item i ∈ I, let Y contain the k items from Sky (I) that are most competitive with i. Then, an item j ∈ I can only be in the top-k competitors of i, if j ∈ Y or if j is dominated by one of the items in Y.

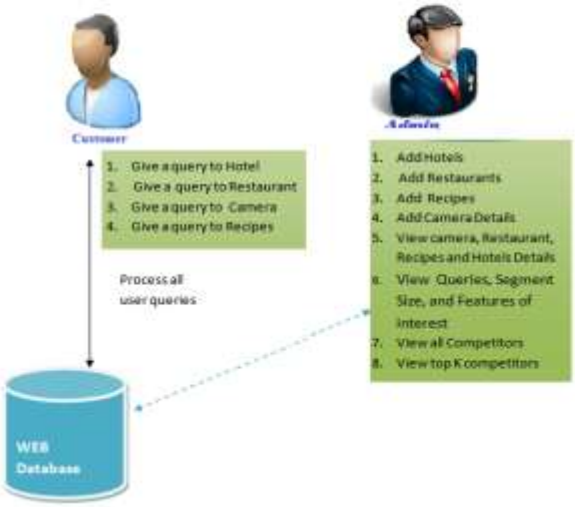
Customer
Admin
1. Give a query to Hotel
2. Give a query to Restaurant
3. Give a query to Camera
4. Give a query to Recipes
1. Add Hotels
2. Add Restaurants
3. Add Recipes
4. Add Camera Details
5. View camera, Restaurant, Recipes and Hotels Details
6. View Queries, Segment Size, and Features of interest
7. View all Competitors
8. View top K competitors
Process all user queries
WEB Database

Figure 2: Proposed architecture

V.CONCLUSION

Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Machine learning algorithms are widely used in various applications. Every business related application uses data mining techniques. To improve such business or providing appropriate competitors for the business to the user need the support of web mining techniques. The competitor mining is one such a way to analyze competitors for the selected items. In this paper, we propose an enhanced c-miner algorithm which arranges the unstructured data into structured data. Based on the customer reviews, the itemset are selected. Since the aim is to analyze the competitors, the relevant features are picked from the reviews. With the help of c-miner and skyline approach, the information is represented and then top k item is extracted from the available resources. Experimental analysis has shown the efficiency of the proposed algorithm.

VI.REFERENCES

[1] Ding, X., Liu, B., Yu, P.S., 2008. A holistic lexiconbased approach to opinion mining. In: Proceedings of the WSDM'08.

[2] Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. 26 (3), 12:1–12:34

[3] Chen, L., Qi, L., Wang, F., 2012. Comparison of feature-level learning methods for mining online consumer reviews. Expert Syst. Appl. 39 (10), 9588–9601.

[4] Zhan, J., Loh, H.T., Liu, Y., 2009. Gather customer concerns from online product reviews – a text summarization approach. Expert Syst. Appl. 36 (2 Part 1), 2107–2115

[5] Jin, Jian, Ping Ji, and Rui Gu. "Identifying comparative customer requirements from product online reviews for competitor analysis." Engineering Applications of Artificial Intelligence 49 (2016): 61-73.

[6] Saxena, Prateek, David Molnar, and Benjamin Livshits. "SCRIPTGARD: automatic contextsensitive sanitization for largescale legacy web applications." Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011.

[7] Ghamisi, Pedram, Jon Atli Benediktsson, and Johannes R. Sveinsson. "Automatic spectral–spatial classification framework based on attribute profiles and supervised feature extraction." IEEE Transactions on Geoscience and Remote Sensing 52.9 (2014): 5771-5782.

[8] Petrucci, Giulio. "Information extraction for learning expressive ontologies." In European Semantic Web Conference, pp. 740-750. Springer, Cham, 2015.

[9] Gentile, Anna Lisa, Ziqi Zhang, Isabelle Augenstein, and Fabio Ciravegna. "Unsupervised wrapper induction using linked data." In Proceedings of the seventh international conference on Knowledge capture, pp. 41-48. ACM, 2013.

[10] K. Simon and G. Lausen, “ViPER: Augmenting Automatic Information Extraction with Visual Perceptions,” Proc. 14th ACM Int’l Conf. Information and Knowledge Management, pp. 381-388, 2005

[11] Zelenko, Dmitry, and Oleg Semin. "Automatic competitor identification from public information sources." International Journal of Computational Intelligence and Applications 2.03 (2002): 287-294.

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

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

[14] Pant, Gautam, and Olivia RL Sheng. "Web footprints of firms: Using online isomorphism for competitor identification." InformationSystems Research26.1 (2015): 188-209.

[15] 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.

About Authors:

Md. Irfan is current pursuing M.Tech in CSE. dept., Dr.K.V.Subba Reddy Institute of Technology, Kurnool, AP**,** India.

H. Ateeq Ahmed, Assistant Professor in Dept. Of CSE, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, AP**,** India.