International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-6, Special Issue-3, April 2018 E-ISSN: 2347-2693

A Review on Mining Large Unstructured Datasets to Find Top-K Competitors

B.Lasya Reddy[1*], Shaik Salam[2]

[1*]Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College , JNTUA, Tirupathi, India
[2] Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College , JNTUA, Tirupathi, India

*Corresponding Author: lasyareddy09@gmail.com , Tel.:9010330306

Available online at: www.ijcseonline.org

Abstract— Now-a-days in any business field we are hearing about the word 'competition'. So, by competitive analysis we can analyze the competitors and can assess the strengths and weakness of a competitor. Competition is necessary in marketing to know which companies are primary competitors and also know which company is competing with itself. So by this we make our products, services and marketing stands out well in business. Competitiveness between two items can be defined based on market segments that they can both cover. Competitiveness is evaluated in large review datasets and address the problem of finding top-k competitors. For evaluating of competitiveness, it utilizes customer reviews which are abundantly available in wide range of domains. There are so many efficient methods for addressing the problem of finding top-k competitors in terms of scalability, accuracy.

Keywords— Data Mining, Competitor Mining, Competitors, Information search and retrieval

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 otherhand, 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 'A' since it is competing for 14 users with 'A' where as with 'C' it is competing for '4' users and with 'D' for 12 users. So, in this market 'B' is highly competitive.

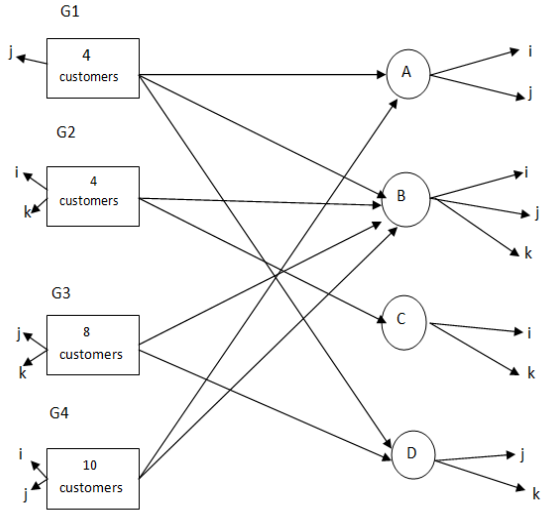

Figure 1: Example for competitiveness model

II. LITERATURE REVIEW

The paper [3] developed an automatic system that discovers companies which are in competition from public information sources. In this the data is extracted and also uses transformation learning techniques to get appropriate data normalization which combines structured and unstructured sources uses probabilistic models to represent the unlinked data and succeeds in discovering competitors. The paper also introduced iterative graph reconstruction process and also used machine learning algorithms for finding competitors. But this technique has a problem of finding market demands.

The paper[1],[6] presented a formal definition of competitiveness between two items. In this authors have used many domains and also handled the problems in pre vious approaches. In this author consider the items are positioned in multi-dimensional feature space and also considers the opinions and preferences of users. However, this technique has addressed the problem of finding top-k competitors of a given items.

The paper [11] verifies that competing products are likely to have similar web footprints a phenomenon that refers to online isomorphism. In this they consider different types of isomorphism between two firms such as overlap between the in-link and out-link of respective websites. But the need for isomorphism feature limits its applicability to products and makes it unsuitable for items and domains where such features are not available (or) extremely sparse.
The paper [2] has suggested the frameworks for manually identifying of competitors. Due to large and newly emerging of companies, it is time consuming for us to find competitors manually.

The paper [4] accomplishes a task for mining competitors with respect to an entity. Here entity refers to person, product (or) a company. The paper proposed an algorithm called "CoMiner" which first extracts the comparative items of input entity and rank them according to comparability. But CoMiner was developed for supporting a specific domain and effort for further domains is still challenging.

The paper [5] proposes ranking methods for finding competition information. In this they proposed effective techniques for finding competitors.

The paper [10] proposes a graphical model for visualizing and extracting relationships between products from the customer reviews. With the interdependencies between the products helps the business organization to discover risks and marketing strategies.

The paper [12] proposes an extension of database by using skyline. Because of skyline, dominated items can be found out.

III. PROBLEM STATEMENT

A. Top-K Competitors Problem

It is the problem of finding top-k competitor of a given item. This problem presents computational challenge especially in presence of large datasets with hundreds (or) thousands of items. This problem can be addressed by using an efficient algorithm.

Let us consider a market with set of I items namely $i_1,i_2,i_3\ldots\ldots\ldots i_n$. Each item in the set is have attributes $a_1,a_2,a_3\ldots\ldots\ldots a_n$. In this we assume those attributes as features of an item in the given set of items. The value of each attribute is represented based on the features of an item, since we have different types of features like binary, categorical, ordinal and numeric features. Now we need to select k items where k is a positive integer.

For selecting the top most items prior to that we need to find the dominated items in the given set. An item can be dominated by another item if it has a better value when we compared the features of an item. For finding dominated items we need to construct the skyline pyramid for entire set of items I. skyline filters out the set of interesting points from a large dataset of points. By construction of skyline it reduces the consideration of items.

IV. CONCLUSION

Data mining is very important in finding patterns, forecasting and discovering the knowledge in different business domains. For improving businesses, competitor information is necessary to the user. So, competitor mining is one way for analyzing competitors for selected items.

REFERENCES

[1] Valkanas, George, Theodoros Lappas, and Dimitrios Gunopulos."Mining Competitors from Large Unstructured Datasets." IEEE Transactions on Knowledge and Data Engineering vol.29,Issue 9,pp 1971-1984, 2017.

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

[3] D. Zelenko and O. Semin, "Automatic competitor identification from public information sources," *International Journal of ComputationalIntelligence and Applications*, 2002.

[4] R. Li, S. Bao, J. Wang, Y. Yu, and Y. Cao, "Cominer: An effective algorithm for mining competitors from the web," in *ICDM*, pp. 948–952 2006.

[5] R. Li, S. Bao, J. Wang, Y. Liu, and Y. Yu, "Web scale competitor discovery using mutual information," in *ADMA*, pp. 798–808 2006.

[6] T. Lappas, G. Valkanas, and D. Gunopulos, "Efficient and domain invariant competitor mining," in *SIGKDD*, pp. 408–416, 2012,

[7] Q. Wan, R. C.-W. Wong, and Y. Peng, "Finding top-k profitable products," in ICDE, vol.24,Issue 10,pp 1774-1788 2011.

[8] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," ser. WSDM '08.

[9] 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. 182–191, 2013.

[10] K. Xu, S. S. Liao, J. Li, and Y. Song.Mining comparative opinions from customer reviews for competitive intelligence.Decis. Support Syst., 2011.

[11] G. Pant, and O. R. L. Sheng. Mining competitor relationships from online news: A network-based approach. Electronic Commerce Research and Applications vol.10,Issue.4, pp 418-427 2011.

[12] S. Borzs ¨onyi, D. Kossmann, and K. Stocker, "The skyline operator," ¨in ICDE, 2001.

[13] Kumar, B. Senthil, and Nisha Joseph. "A Review on Competitor Mining and Unstructured Dataset Handling Techniques." Journal of Network Communications and Emerging Technologies (JNCET) vol.7, no. 9,pp22-26 2017.

Authors Profile

Ms. B.Lasya Reddy pursed Bachelor of Technology from S.V College of Engineering, Tirupathi in 2015 and currently pursuing Master of Technology from Sree Vidyanikethan Engineering College,Tirupathi.

Mr Shaik salam pursed Bachelor of Engineering from Marthwada University, Aurangabad in 1990 and Master of Engineering from Sathyabama University, Chennai in year 2008. He is currently pursuing Ph.D from Acharya Nagurjuna University,Guntur. He is currently working as Associate Professor in Department of Computer Science and Engineering in Sree Vidyanikethan Engineering College, Tirupati since 1997. He has published more than 10 research papers in Data Mining. His main research work focuses on Data Mining. He has 20 years of teaching experience and 4 years of Research Experience.

