International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 25th Dec 2018. Link :http://www.ijiemr.org/downloads.php?vol=Volume-07&issue=ISSUE-13

Title: **EFFICIENT KEYWORD-AWARE REPRESENTATIVE TRAVEL ROUTE FRAMEWORK**

Volume 07, Issue 13, Pages: 726-732

Paper Authors

[1]**D.PRIYANKA,**

[2]**S.KRISHNA REDDY**

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

MINING COMPETITORS FROM LARGE UNSTRUCTURED DATASETS

[1]PUSHPALATHA, [2]DR.NAZIMUNISA

[1]Mtech student, Sree Dattha Institute of Engineering and Science

[2]Professor, Sree Dattha Institute of Engineering and Science

ABSTRACT

In the present world Completive business, the achievement is totally in light of the capacity to make a thing more engaging clients than the opposition. Huge information is a trendy expression that is utilized for expansive size information which incorporates organized information, semi-organized information and unstructured information. The span of huge information is large to the point, that it is almost difficult to gather process and store information utilizing conventional database administration framework and programming methods. In this way, huge information requires diverse methodologies and devices to break down information. The way toward gathering, putting away and breaking down expansive measure of information to discover obscure examples is called as large information investigation. Here we show a formal meaning of the intensity between two things, in light of the market fragments that they can both cover. Our assessment of aggressiveness uses client surveys, a plenteous wellspring 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 contenders of a given thing. At long last, we assess the nature of our outcomes and the versatility of our approach utilizing various datasets from various areas.

1. INTRODUCTION

The strategic importance of detecting and observing business competitors is an inevitable research, which motivated by several business challenges. Monitoring and identifying firm's competitors have studied in the earlier work. Data mining is the optimal way of handling such huge information's for mining competitors. Item reviews form online offer rich information about customers' opinions and interest to get a general idea regarding competitors. However, it is generally difficult to understand all reviews in different websites for competitive products and obtain insightful suggestions manually. In the earlier works in the literatures, many authors analyzed such big customer data intelligently and efficiently . For example, a lot of studies about online reviews were stated to gather item opinion analysis from online reviews in different levels. However, most researchers in this field ignore how to make their findings be seamlessly utilized to the competitor mining process. Recently, a limited number of researches were noted to utilize the latest development in artificial intelligence (AI) and data mining in the ecommerce applications. These studies help designers to understand a large amount of customer requirements in online reviews for product improvements. But, these discussions are far from sufficient and some potential problems. These have not been

fully investigated such as, with product online reviews, how toconduct athorough competitor analysis. Actually, in a typical scenario of a customer-driven new product design (NPD), the strengths and weakness are often analyzed exhaustively for probable opportunities to succeed in the fierce market competition.

Data extraction from site pages is a dynamic research range. Scientists have been creating different arrangements from a wide range of viewpoints to give the similar report. Many webdata extraction frameworks depend on humanclients to give stamped tests with the goal that the information extraction principles could be scholarly. Due to the managed learning process, self-loader frameworks for the most part have higher exactness than completely programmed frameworks that have no human intercession. Self-loader techniques are not appropriate for substantial scale web applications that need to remove information from a large number of sites.

Additionally sites tend to change their site page designs much of the time, which will make the past created extraction rules invalid, additionally constraining the ease of use of self-loader strategies. That is the reason numerous later workconcentrate on completely or about completely programmed arrangements.

2. PROBLEM DEFINITION

This examination gives the different philosophies actualized to mine rivals withreference to client lifetime esteem, relationship, conclusion and conduct utilizing information mining procedures. The web development has brought about boundless utilization of numerous applications like internet business and other administration situated applications. This shifted utilization of web applications has given a tremendous measure of information available to one. Information is the information that exists in its crude shape bringing about data for additionally preparing. With enormous measure of information, associations confronted the critical test of separating exceptionally valuable data from them. This has prompted the idea of information mining. Mining contender's of a given thing, the most impacted factor of the thing which fulfills the client need can be removed from the information that is commonly put away in the database. This area gives two sorts of literary works, for example, contender mining and unstructured information administration.

GOALS

- We present a formal definition of the competitiveness between two items, based on the market segments that they can both cover.
- We evaluate the quality of our results and the scalability of our approach using multiple datasets from different domains.
- Our work is the first toaddress the evaluation of competitiveness via the analysis of large unstructured datasets, without the need for direct comparative evidence.

ALGORITHMS:

The CMiner Algorithm:

Next, we present CMiner, an exact algorithm for finding the top-k competitors of a given item. Our algorithm makes use of

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

the skyline pyramid in order to reduce the number of items that need to be considered.

Given that we only care about the top-k competitors, we can incrementally compute the score of each candidate and stop when it is guaranteed that the top-k have emerged. The pseudocode is given in Algorithm 1. Discussion of CMiner: The input includes the set of items I, the set of features F, the item of interest i, the number k of top competitors to retrieve, the set Q of queries and their probabilities, and the skyline pyramid DI. The algorithm first retrieves the items that dominate i, via masters(i) (line 1). These items have the maximum possible competitiveness with i. If at least k such items exist, we report those and conclude (lines 2-4). Otherwise, we add them to TopK and decrement our budget of k accordingly (line 5). The variable LB maintains the lowest lower bound from the current topk set (line 6) and is used to prune candidates. In line 7, we initialize the set of candidates X as the union of items in the first layer of the pyramid and the set of items dominated by those already in the TopK.

Algorithm 1 CMiner

Input: Set of items $\mathcal{I}$, Item of interest $i \in \mathcal{I}$, feature space $\mathcal{F}$, Collection $\mathcal{Q} \in 2^{\mathcal{F}}$ of queries with non-zero weights, skyline pyramid $\mathcal{D}_{\mathcal{I}}$, int k
Output: Set of top-k competitors for i

1: $TopK \leftarrow masters(i)$
2: **if** ($k \leq |TopK|$) **then**
3: **return** $TopK$
4: **end if**
5: $k \leftarrow k - |TopK|$
6: $LB \leftarrow -1$
7: $\mathcal{X} \leftarrow$ GETSLAVES$(TopK, \mathcal{D}_{\mathcal{I}}) \cup \mathcal{D}_{\mathcal{I}}[0]$
8: **while** ($|\mathcal{X}|$!= 0) **do**
9: $\mathcal{X} \leftarrow$ UPDATETOPK$(k, LB, \mathcal{X})$
10: **if** ($|\mathcal{X}|$!= 0) **then**
11: $TopK \leftarrow$MERGE$(TopK, \mathcal{X})$
12: **if** ($|TopK| = k$) **then**
13: $LB \leftarrow$WORSTIN$(TopK)$
14: **end if**
15: $\mathcal{X} \leftarrow$GETSLAVES$(\mathcal{X}, \mathcal{D}_{\mathcal{I}})$
16: **end if**
17: **end while**
18: **return** $TopK$

19: **Routine** UPDATETOPK$(k, LB, \mathcal{X})$
20: $localTopK \leftarrow \emptyset$
21: $low(j) \leftarrow 0, \forall j \in \mathcal{X}$.
22: $up(j) \leftarrow \sum_{q \in \mathcal{Q}} p(q) \times V^q_{j,j}, \forall j \in \mathcal{X}$.
23: **for every** $q \in Q$ **do**
24: $maxV \leftarrow p(q) \times V^q_{i,i}$
25: **for every item** $j \in \mathcal{X}$ **do**
26: $up(j) \leftarrow up(j) - maxV + p(q) \times V^q_{i,j}$
27: **if** ($up(j) < LB$) **then**
28: $\mathcal{X} \leftarrow \mathcal{X} \setminus \{j\}$
29: **else**
30: $low(j) \leftarrow low(j) + p(q) \times V^q_{i,j}$
31: $localTopK.update(j, low(j))$
32: **if** ($|localTopK| \geq k$) **then**
33: $LB \leftarrow$WORSTIN$(localTopK)$
34: **end if**
35: **end if**
36: **end for**
37: **if** ($|\mathcal{X}| \leq k$) **then**
38: **break**
39: **end if**
40: **end for**
41: **for every item** $j \in \mathcal{X}$ **do**
42: **for every remaining** $q \in \mathcal{Q}$ **do**
43: $low(j) \leftarrow low(j) + p(q) \times V^q_{i,j}$
44: **end for**
45: $localTopK.update(j, low(j))$
46: **end for**
47: **return** TOPK$(localTopK)$

This is achieved via calling GETSLAVES(T opK, DI). In every iteration of lines 8-17, CMiner feeds the set of candidates X to the UPDATETOPK() routine, which prunes items based on the LB threshold. It then updates the TopK set via the MERGE() function, which identifies the items with the highest competitiveness from TopK∪ X . This can be achieved in linear time, since both X and T opK are sorted. In line 13, the pruning threshold LB is set to the worst (lowest) score among the new T opK. Finally, GETSLAVES() is used to expand the set of candidates by including items that are dominated by those in X .

3. PROBLEM SOLUTION

DISADVANTAGES:

- The frequency of textual comparative evidence can vary greatly across domains. For example, when comparing brand names at the firm level (e.g. "Google vs Yahoo" or "Sony vs Panasonic"), it is indeed likely that comparative patterns can be found by simply querying the web. However, it is easy to identify mainstream domains where such evidence is extremely scarce, such as shoes, jewelery, hotels, restaurants, and furniture.
- Existing approach is not appropriate for evaluating the competitiveness between any two items or firms in a given market. Instead, the authors assume that the set of competitors is given and, thus, their goal is to compute the value of the chosen measures for each competitor. In addition, the dependency on transactional data is a limitation we do not have.
- The applicability of such approaches is greatly limited

PROPOSED SYSTEM:

- We propose a new formalization of the competitiveness between two items, based on the market segments that they can both cover.
- 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.

4. CONCLUSION

We presented a formal definition of competitiveness between two items, which we validated both quantitatively and qualitatively. Our formalization is applicable across domains, overcoming the shortcomings of previous approaches. We consider a number of factors that have been largely overlooked in the past, such as the position of the items in the multi-dimensional feature space and the preferences and opinions of the users. Our work introduces an end-to-end methodology for mining such information from large datasets of customer reviews. Based on our competitiveness definition, we addressed the computationally challenging problem of finding the top-k competitors of a given item. The proposed framework is efficient and applicable to domains with very large populations of items. The efficiency of our methodology was verified via an experimental evaluation on real datasets from different domains. Our experiments also revealed that only a small number of reviews is sufficient to confidently estimate the different types of users in a given market, as well the number of users that belong to each type.

REFERENCES

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

[2] R. Deshpand and H. Gatingon, "Competitive analysis," Marketing Letters, 1994.

[3] B. H. Clark and D. B. Montgomery, "Managerial Identification of Competitors," Journal of Marketing, 1999.

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

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

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

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

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

[9] 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.

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

[11] S. Bao, R. Li, Y. Yu, and Y. Cao, "Competitor mining with the web," IEEE Trans. Knowl. Data Eng., 2008.

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

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

[14] 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.

[15] 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.

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

[17] E. Marrese-Taylor, J. D. Velasquez, F. Bravo-Marquez, and Y. Mat- ´ suo, "Identifying customer preferences about tourism products using an aspect-based opinion mining approach," Procedia Computer Science, vol. 22, pp. 182–191, 2013.

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

[19] Y.-L. Wu, D. Agrawal, and A. El Abbadi, "Using wavelet decomposition to support progressive and approximate range-sum queries over data cubes," in CIKM, ser. CIKM '00, 2000, pp. 414–421.

[20] 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.

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

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

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

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

[25] D. Papadias, Y. Tao, G. Fu, and B. Seeger, "An optimal and progressive algorithm for skyline queries," ser. SIGMOD '03.

[26] G. Valkanas, A. N. Papadopoulos, and D. Gunopulos, "Skyline ranking a la IR," in ` ExploreDB, 2014, pp. 182–187.

[27] J. L. Bentley, H. T. Kung, M. Schkolnick, and C. D. Thompson, "On the average number of maxima in a set of vectors and applications," J. ACM, 1978.

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

[29] A. Agresti, Analysis of ordinal categorical data. John Wiley & Sons, 2010, vol. 656.

[30] T. Lappas, G. Valkanas, and D. Gunopulos, "Efficient and domaininvariant competitor mining," in SIGKDD, 2012, pp. 408–416.

[31] J. F. Porac and H. Thomas, "Taxonomic mental models in competitor definition," Academy of Management Review, vol. 15, no. 2, pp. 224–240, 1990.

[32] 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. 698–720, 2012.

[33] 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. 283–289.

[34] G. Pant and O. R. Sheng, "Web footprints of firms: Using online isomorphism for competitor identification," Information Systems Research, vol. 26, no. 1, pp. 188–209, 2015.

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

[36] Q. Wan, R. C.-W. Wong, I. F. Ilyas, M. T. Ozsu, and Y. Peng, ¨ "Creating competitive products," PVLDB, vol. 2, no. 1, pp. 898– 909, 2009.

[37] Q. Wan, R. C.-W. Wong, and Y. Peng, "Finding top-k profitable products," in ICDE, 2011.

[38] Z. Zhang, L. V. S. Lakshmanan, and A. K. H. Tung, "On domination game analysis for microeconomic data mining," ACM Trans. Knowl. Discov. Data, 2009.

[39] T. Wu, D. Xin, Q. Mei, and J. Han, "Promotion analysis in multidimensional space," PVLDB, 2009.

[40] T. Wu, Y. Sun, C. Li, and J. Han, "Region-based online promotion analysis," in EDBT, 2010.

[41] D. Kossmann, F. Ramsak, and S. Rost, "Shooting stars in the sky: an online algorithm for skyline queries," ser. VLDB, 2002.

[42] A. Vlachou, C. Doulkeridis, Y. Kotidis, and K. Nørvag, "Reverse ˚ top-k queries," in ICDE, 2010.

[43] A. Vlachou, C. Doulkeridis, K. Nørvag, and Y. Kotidis, "Identifying the most influential data objects with reverse top-k queries," PVLDB, 2010.

[44] K. Hose and A. Vlachou, "A survey of skyline processing in highly distributed environments," The VLDB Journal, vol. 21, no. 3, pp. 359–384, 2012.