MINING COMPETITORS FROM LARGE UNSTRUCTURED DATASETS

Mr. N. SUBRAMANYAM[1], K. VYSHNAVI MALA[2]

[1]Assistant Professor, Dept of MCA, Audisankara Institute of Technology (AUTONOMOUS), Gudur (M), Tirupati (Dt), AP

[2]PG Scholar, Dept of MCA, Audisankara Institute of Technology (AUTONOMOUS) Gudur (M), Tirupati (Dt), AP

ABSTRACT:

In any competitive business, success is based on the ability to make an item more appealing to customers than the competition. A number of questions arise in the context of this task: how do we formalize and quantify the competitiveness between two items? Who are the main competitors of a given item? What are the features of an item that most affect its competitiveness? Despite the impact and relevance of this problem to many domains, only a limited amount of work has been devoted toward an effective solution. In this project, we present a formal definition of the competitiveness between two items, based on the market segments that they can both cover. Our evaluation of competitiveness utilizes customer reviews, an abundant source of information that is available in a wide range of domains. We present efficient methods for evaluating competitiveness in large review datasets and address the natural problem of finding the top-k competitors of a given item. Finally, we evaluate the quality of our results and the scalability of our approach using multiple datasets from different domains. In this project, we propose C- Miner, an algorithm which uses a data mining technique called frequent sequence mining to discover block correlations in storage systems. C-Miner runs reasonably fast with feasible space requirement, indicating that it is a practical tool for dynamically inferring correlations in a storage system.

1. INTRODUCTION

Nowadays, huge amounts of sequential information are stored in databases (e.g. stock market data, biological data and customer data). Discovering patterns in such databases is important in many domains, as it provides a better understanding of the data. For example, in international trade, one could be interested in discovering temporal relations between the appreciations of currencies to make trade decisions. Various methods have been proposed for mining patterns in sequential databases such as mining repetitive patterns, trends and sequential patterns. Among them, mining sequential patterns is probably the most popular set of techniques. The marketing and management community have focused on empirical methods for competitor identification as well as on methods for analyzing known competitors. Extant research on the former has focused on mining comparative expressions (e.g. “Item A is better than Item B”) from the Web or other textual sources. Even though such

expressions can indeed be indicators of competitiveness, they are absent inmany domains. For instance, consider the domain of vacation packages (e.g. flight-hotel-car combinations). In this case, item shave no assigned name by which they can be queried or compared with each other. Further, 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, jewelry, hotels, restaurants, and furniture. Motivated by these Shortcomings, we propose a new formalization of the competitiveness between two items, based on the market segments that they can both cover [1].

- This paper 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.

- 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.

- In this paper, it is argued that data mining is an approach to assist companies in developing more effective strategies to meet the competitions in the market. Data warehousing is useful and accurate for assembling a business’ dispersed heterogeneous data and providing unified convenient information access technique.

- Data mining technology can be used to transform hidden knowledge into manifest knowledge. A competitor mining from web data system is extremely flexible. Therefore, one of the best competitivestrategies is the successful utilization of web data for timely decision support.

- Information extraction from web pages is an active research area. Researchers have been developing various solutions from all kinds of perspectives to provide the comparative report. Many web information extraction systems rely on human users to provide marked samples so that the data extraction rules could be learned.

- Because of the supervised learning process, semi- automatic systems usually have higher accuracy thaneach type.fully automatic systems that have no human intervention. Semi-automatic methods are not suitable for large-scale web applications that need to extract data from thousands of web sites.

- Also, web sites tend to change their web page formats frequently, which will make the previous generated extraction rules invalid, further limiting the usability of semi-automatic methods. That’s why many more recent works focus on fully or nearly fully automatic solutions.

- In the paper, 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.

• Web information extraction can be at the record level or data unit level. The former treat each data record as a single data unit while the latter go one step further to extract detailed data units within the data records [10]. Record level extraction method generally involves identifying the data regions that contain all the records, and then partitioning the data regions into individual records. Structured data extraction from Web pages has been studied extensively. Early works on manually constructed wrappers were found difficult to maintain and be applied to different Websites, because they are very labor intensive.

2. LITERATURE SURVEY

This paper builds on and significantly extends our pre- liminary work on the evaluation of competitiveness [30]. To the best of our knowledge, our work is the first to address the evaluation of competitiveness via the analysis of large unstructured datasets, without the need for direct comparative evidence. Nonetheless, our work has ties to previous work from various domains.

Managerial Competitor Identification: The management literature is rich with works that focus on how managers can manually identify competitors. Some of these works model competitor identification as a mental categorization process in which managers develop mental representations of competitors and use them to classify candidate firms [3], [6], [31]. Other manual categorization methods are based on market- and resource-based similarities between a firm and candidate competitors [1], [5], [7]. Finally, managerial competitor identification has also been presented as a sense- making process in which competitors are identified based on their potential to threaten an organization's identity [4].

Competitor Mining Algorithms: Zheng et al. [32] identify key competitive measures (e.g. market share, share of wallet) and showed how a firm can infer the values of these measures for its competitors by mining (i) its own detailed customer transaction data and (ii) aggregate data for each competitor. Contrary to our own methodology, this 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.

Doan et al. explore user visitation data, such as the geo-coded data from location-based social networks, as a potential resource for competitor mining [33]. While they report promising results, the dependence on visitation data limits the set of domains that can benefit from this approach.

Pant and Sheng hypothesize and verify that competing firms are likely to have similar web footprints, a phenomenon that they refer to as online isomorphism [34]. Their study considers different types of isomorphism between two firms, such as the overlap between the in-links and out- links of their respective websites, as well as the number of times that they appear together online (e.g. in search results or new articles). Similar to our own methodology, their approach is geared toward pairwise competitiveness. However, the need for isomorphism features limitsits applicability to firms and make it unsuitable for items and domains where such features are either not available or extremely sparse, as is typically the case with co-occurrence data. In fact, the sparsity of co-occurrence data is a serious limitation of a significant body of work [8], [10], [11], [35] that focuses on mining competitors based on comparative expressions found in web results and other textual corpora. The intuition is that the frequency of expressions like "Item A is better than Item B" "or item A Vs.

Item B" is indicative of their competitiveness. However, as we have alreadydiscussed in the introduction, such evidence is typically scarce or even non-existent in many mainstream domains. As a result, the applicability of such approaches is greatly limited. We provide empirical evidence on the sparsity of co-occurrence information in our experimental evaluation.

Finding Competitive Products: Recent work [36], [37], [38] has explored competitiveness in the context of product de- sign. The first step in these approaches is the definition of a dominance function that represents the value of a product. The goal is then to use this function to create items that are not dominated by other, or maximize items with the maximum possible dominance value. A similar line of work [39], [40] represents items as points in a multidimensional space and looks for subspaces where the appeal of the item is maximized. While relevant, the above projects have a completely different focus from our own, and hence the proposed approaches are not applicable in our setting.

Skyline computation: Our work leverages concepts and techniques from the extensive literature on skyline computation [24], [25], [41]. These include the dominance concept among items, as well as the construction of the skyline pyramid used by our CMiner algorithm. Our work also has ties to the recent publications in reverse skyline queries [42], [43]. Even though the focus of our work is different, we intend to utilize the advances in this field to improve our framework in future work.

3. PROPOSED WORK

The significance of this project is to help the customer to view the products as their convenience. In this project customer can write their views and also can check reviews whether it's good or bad. 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

The figure illustrates the competitiveness between three items i, j and k. Each item is mapped to the set of features that it can offer to a customer. Three features are considered in this example: A, B and C. Even though this simple example considers only binary features (i.e. available/not available), our actual formalization accounts for a much richer space including binary, categorical and numerical features. The left side of the figure shows three groups of customers g1, g2, and g3. Each group represents a different market segment. Users are grouped based on their preferences with respect to the features. For example, the customers in g2 are only interested in features A and B. We observe that items is and k are not competitive, since they simply do not appeal to the same groups of customers. On the other hand, j competes with both i (for groups g1 and g2) and k (for g3). Finally, an interesting observation is that j competes for 4 users with i and for 9 users with k. In other words, k is a stronger competitor for j, since it claims a much larger portion of its market share than I. 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 thesechallenges via a highly scalable framework for top-k computation, including an efficient evaluation algorithm and an appropriate index.

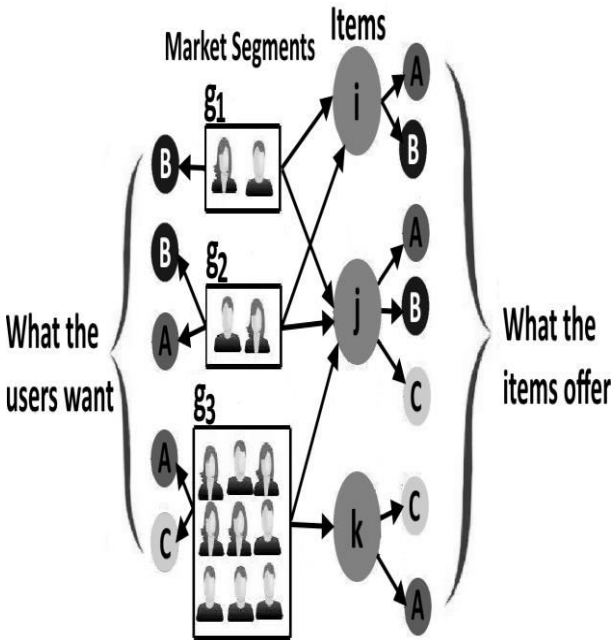

Fig: A (simplified) example of our competitiveness paradigm

The figure illustrates the competitiveness between three items is, j and k. Each item is mapped to the set of features that it can offer to a customer. Three features are considered in this example: A, B and C. Even though this simple example considers only binary features (i.e. available/not avail- able), our actual formalization accounts for a much richer space including binary, categorical and numerical features. The left side of the figure shows three groups of customers g1, g2, and g3. Each group represents a different market segment. Users are grouped based on their preferences with respect to the features. For example, the customers in g2 are only interested in features A and B. We observe that items i and k are not competitive, since they simply do not appeal to the same groups of customers. On the other hand, j competes with both i (for groups g1 and g2) and k (for

g3). Finally, an interesting observation is that j competes for 4 users with i and for 9 users with k. In other words, k is a stronger competitor for j, since it claims a much larger portion of its market share than i.

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.

•Aformal 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.

DEFINING COMPETITIVENESS

The typical user session on a review platform, such as Yelp, Amazon or TripAdvisor, consists of the following steps:

1) Specify all required features in a query.

2) Submit the query to the website's search engine and retrieve the matching items.

3) Process the reviews of the returned items and make a purchase decision.

In this setting, items that cover the user's requirements will be included in the search engine's response and will compete for her attention. On the other hand, non-covering items will not be considered by the user and, thus, will not have a chance to compete. Next, we present an example that extends this decision-making process to a multi-user setting. Consider a simple market with 3 hotels i, j, k and 6 binary features: bar, breakfast, gym, parking, pool, wi-fi. Table 1 includes the value of each hotel for each feature. In this simple example, we assume that the market includes 6 mutually exclusive customer segments (types). Each segment is represented by a query that includes the features that are of interest to the customers included in the segment. Information on each segment is provided in Table 2. For instance, the first segment includes 100 customers who are interested in parking and wi-fi, while the second segmentincludes 50 customers who are only interested in parking.

FINDING THE TOP-K COMPETITORS

Given the definition of the competitiveness in Eq 1, we study the natural problem of finding the top-k competitors of a given item. Formally:

Problem 1. [Top-k Competitors Problem]: We are presented with a market with a set of n items I and a set of features

F. Then, given a single item i∈ I, we want to identify the k items from I that maximize CF (i, ·).

A naive algorithm would compute the competitiveness between i and every possible candidate. The complexity of this brute force method is clearly Θ (2|F| × n2 × log K), which can be easily dominated by the powerset factor and, as we demonstrate in our experiments, is impractical for large datasets. One option could be to perform the naive computation in a distributed fashion. Even in this case, however, we would need one thread for each of the n2 pairs. This is far from trivial, if one considers that n could measure in the tens of thousands. In addition, a naive MapReduce implementation would face the bottleneck of passing everything through the reducer to account for the self-join included in the computation. In practice, the self- join would have to be implemented via a customized technique for reduce-side joins, which is a non-trivial and highly expensive operation [23].

These issues motivate us to introduce CMiner, an efficient exact algorithm for Problem 1. Except for the creation of our indexing mechanism, every other aspect of CMiner can also be incorporated in a parallel solution.

First, we define the concept of item dominance, which will aid us in our analysis:

4. CONCLUSION

We presented a formal definition of competitiveness be- tween two items, which we validated both quantitatively and qualitatively. Our formalization is applicable across domains, overcoming the shortcomings of previous ap- proaches. 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 computation- ally 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 experimentalevaluation 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.

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

[2] R. Deshpande and H. Gating on, "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 Dissertations, 2007.

[5] M. Bergen and M. A. Peteraf, "Competitor identification and com- petitory 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, "Combiner: An effective algorithm for mining competitors from the web," in ICDM, 2006.

[9] Z. Ma, G. Pant, and O. R. L. Sheng, "Mining competitor relation- ships 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] V. Srikanth. "A NOVEL METHOD FOR BUG DETECTION TECHNIQUES USING INSTANCE SELECTION AND FEATURE SELECTION" v srikanth | INTERNATIONAL JOURNAL OF INNOVATIVE ENGINEERING AND MANAGEMENT RESEARCH, 08 DECEMBER. 2017. https://www.ijiemr.org/public/uploads/paper/976_approvedpaper.pdf

[12] V. Srikanth. "SECURED RANKED KEYWORD SEARCH OVER ENCRYPTED DATA ON CLOUD" v srikanth | INTERNATIONAL JOURNAL OF INNOVATIVE ENGINEERING AND MANAGEMENT RESEARCH, 08 Febraury.

2018. http://www.ijiemr.org/downloads.php?vol=Volume-7&issue=ISSUE-02

[13] V. Srikanth. "WIRELESS SECURITY PROTOCOLS (WEP,WPA,WPA2 & WPA3)" v srikanth | Journal of Emerging Technologies and Innovative Research (JETIR), 08 mAY. 2019. https://www.jetir.org/papers/JETIRDA06001.pdf

[14] V. Srikanth, et al. "Detection of Fake Currency Using Machine Learning Models." Deleted Journal, no. 41, Dec. 2023, pp. 31–38. https://doi.org/10.55529/ijrise.41.31.38.

[15] V. Srikanth, et al. "A REVIEW ON MODELING AND PREDICTING OF CYBER HACKING BREACHES." 25 Mar. 2023, pp. 300–305. http://ijte.uk/archive/2023/A-REVIEW-ON-MODELING-AND-PREDICTING-OF-CYBER-HACKING-BREACHES.pdf.

[16] V. Srikanth, "DETECTION OF PLAGIARISM USING ARTIFICIAL NEURAL NETWORKS." 25 Mar. 2023, pp. 201–209. http://ijte.uk/archive/2023/DETECTION-OF-PLAGIARISM-USING-ARTIFICIAL-NEURAL-NETWORKS.pdf.

[17] V. Srikanth, "CHRONIC KIDNEY DISEASE PREDICTION USING MACHINELEARNINGALGORITHMS." 25 January. 2023, pp. 106–122. http://ijte.uk/archive/2023/CHRONIC-KIDNEY-DISEASE-PREDICTION-USING-MACHINE-LEARNING-ALGORITHMS.pdf.

[18] Srikanth veldandi, et al. "View of Classification of SARS Cov-2 and Non-SARS Cov-2 Pneumonia Using CNN". journal.hmjournals.com/index.php/JPDMHD/article/view/3406/2798.

[19] Srikanth veldandi, et al. "Improving Product Marketing by Predicting Early Reviewers on E-Commerce Websites." Deleted Journal, no. 43, Apr. 2024, pp. 17–25. https://doi.org/10.55529/ijrise.43.17.25.

[20] Srikanth veldandi, et al."Intelligents Traffic Light Controller for Ambulance." Journal of Image Processing and Intelligent Remote Sensing, no. 34, July 2023, pp. 19–26. https://doi.org/10.55529/jipirs.34.19.26.

[21] Veldandi Srikanth, et al. "Identification of Plant Leaf Disease Using CNN and Image Processing." Journal of Image Processing and Intelligent Remote Sensing, June 2024, https://doi.org/10.55529/jipirs.44.1.10. [22] N. Thaper, S. Guha, P. Indyk, and N. Kouda, "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," AcM sIGMoD Record, vol. 40, no. 4, pp. 11–20, 2012.

Mr. N. SUBRAMANYAM has received his M.C.A in Computer Application from SV University in 2006 and MTech degree in Computer science from JNTU, Anantapur in 2022. He has been dedicated to the teaching field from the last 12 years. His research areas included CNN Deep learning. He is currently working as Assistant Professor in Audisankara Institute of Technology (AUTONOMOUS) Andhra Pradesh, India.

K. VYSHNAVI MALA has pursuing her MCA from Audisankara institute of Technology (AUTONOMOUS), Gudur, Affiliated to JNTUA in 2024. Andhra Pradesh, India.