MINING COMPETITORS FROM LARGE UNSTRUCTURED DATASETS

[1]RESHMA DUDEKULA, [2]SUREKHA SOMU, [3]SUCHARITHA RAJA, [4]NAGASREE V, [5]MADHURI KONIREDDY, [6]M.SUBBA REDDY
[12345] B. Tech Student, [6]Assistant Professor
DEPARTMENT OF CSE
SVR ENGINEERING COLLEGE, NANDYAL.

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. Along line of research has demonstrated the strategic importance of identifying and monitoring a firm's competitors. Motivated by this problem, 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) from the Web or other textual sources. Even though such expressions can indeed be indicators of competitiveness, they are absent in many 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

Users often have difficulties in expressing their web search needs; they may not know the keywords that can retrieve the information they require [1]. Keyword suggestion (also known as query suggestion), which has become one of the most fundamental features of commercial Web search engines, helps in this direction. After submitting a keyword query, the user may not be satisfied with the results, so the keyword suggestion module of the search engine recommends a set of m keyword queries that are most likely to refine the user's search in the right direction. Effective keyword suggestion methods are based on click information from query logs [2], [3], [4], [5], [6], [7], [8] and query session data [9], [10], [11], or query topic models [12]. New keyword suggestions can be determined according to their semantic relevance to the original keyword query. The semantic relevance between two keyword queries can be determined (i) based on the overlap of their clicked URLs in a query log (ii) by their proximity in a bipartite graph that connects keyword queries and their clicked URLs in the query log [5], [6], [7], [8], (iii) according to their co occurrences in query sessions [13], and (iv) based on their similarity in the topic distribution space [12]. However, none of the existing methods provide location aware keyword query suggestion, such that the suggested keyword queries can retrieve documents not only related to the user information needs but also located near the user location. This requirement emerges due to the popularity of spatial keyword search that takes a user location and user-supplied keyword query as arguments and returns objects that are spatially close and textually\ relevant to these arguments. Google processed a daily average of 4.7 billion queries in 20111, a substantial fraction of which have local intent and target spatial web objects (i.e., points of interest with a web presence having locations as well as text descriptions) or geo-documents (i.e., documents associated with geo-locations). Furthermore, 53% of Bing's mobile searches in 2011 were found to have a local intent.2 To fill this gap, we propose a Location-aware Keyword query Suggestion (LKS) framework. We illustrate the benefit of LKS using a toy example. Consider five geo-documents d1–d5 as listed in Figure 1(a). Each document di is associated with a location di:_ as shown in Figure 1(b). Assume that a user issues a keyword query kq = \seafood" at location _q, shown in Figure 1(b). Note that the relevant documents d1–d3 (containing \seafood") are far from _q. A location aware suggestion is \lobster", which can retrieve nearby documents d4 and d5 that are also relevant to the user's original search intention. Previous keyword query suggestion models (e.g., [6]) ignore the user location and would_sh", which again fails to retrieve nearby relevant documents. Note that LKS has a different goal and therefore differs from other location-aware recommendation methods .

The first challenge of our LKS framework is how to effectively measure keyword query similarity while capturing the spatial distance factor. In accordance to previous query suggestion approaches LKS constructs and uses a keyword-document bipartite graph (KD-graph for short), which connects the keyword queries with their relevant documents as shown in Figure 1(c). Different to all previous approaches which ignore locations, LKS adjusts the weights on edges in the KD-graph to capture not only the semantic relevance between keyword queries, but also the spatial distance between the document locations and the query issuer's location _q. We apply a random walk with restart (RWR) process [22] on the KD-graph, starting from the user supplied query kq, to find the set of m key- word queries with the highest semantic relevance to kq and spatial proximity to the user location. RWR on a KD-graph has been considered superior to alternative approaches [7] and has been a standard technique employed in various (location-independent) keyword suggestion studies.

The second challenge is to compute the suggestions efficiently on a large dynamic graph. Performing keyword\ suggestion instantly is important for the applicability of LKS in practice. However, RWR search has a high computational cost on large graphs. Previous work on scaling up RWR search require pre-computation and/or graph segmentation part of the required RWR scores are materialized under the assumption that the transition probabilities between nodes (i.e., the edge weights) are known beforehand. In addition, RWR search algorithms that do not rely on pre-computation accelerate the computation by pruning nodes based on their lower or upper bound scores and also require the full transition probabilities. However, the edge weights of our KD-graph are unknown in advance, hindering the application of all these approaches. To the best of our knowledge, no existing technique can accelerate RWR when edge weights are unknown a priori (or they are dynamic). To address this issue, we present a novel partition-based algorithm (PA) that greatly reduces the cost of RWR search on such a dynamic bipartite graph. In a nutshell, our proposal divides the keyword queries and the documents into partitions and adopts a lazy mechanism that accelerates RWR search. Pam and the lazy mechanism are generic techniques for RWR search, orthogonal to LKS, therefore they can be applied to speed up RWR search in other large graphs. In summary, the contributions of this paper are: _ We design a Location-aware Keyword query Suggestion (LKS) framework, which provides suggestions that are relevant to the user's information needs and can retrieve relevant documents close to the query issuer's location. _ We extend the state-of-the-art Bookmark Coloring Algorithm (BCA) [28] for RWR search to compute the location-aware suggestions.

II. LITERATURE SURVEY

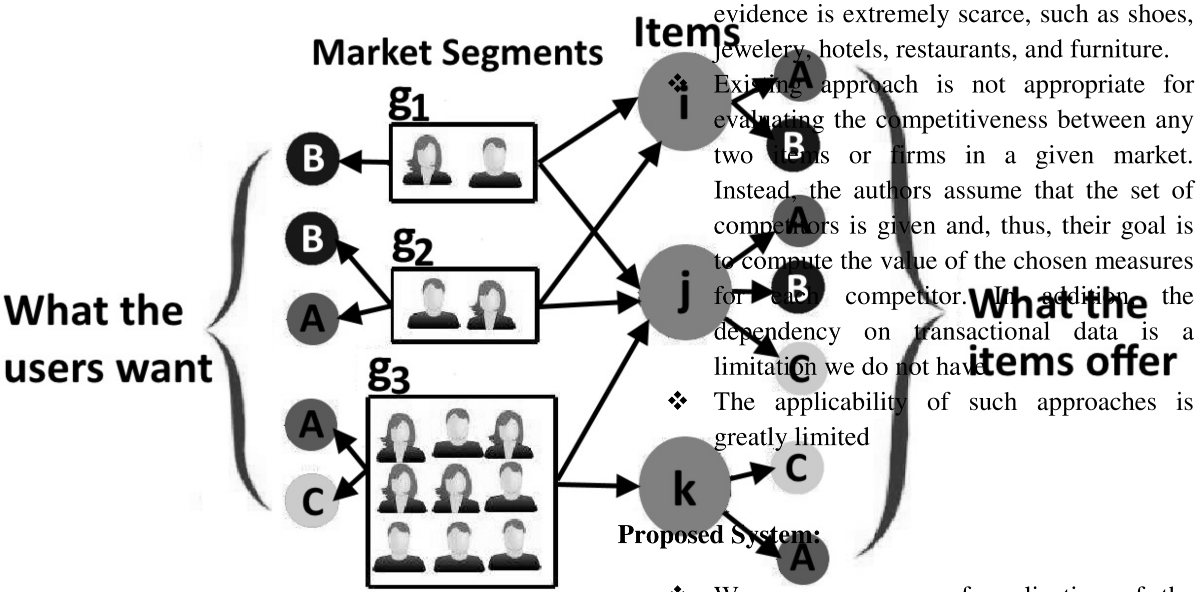
developed an automatic system that discovers companies which are in competition from public information sources. In this the data is extracted and also uses transformation learning techniques to get appropriate data normalization which combines structured and unstructured sources uses probabilistic models 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

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.

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.

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.

III. SYSTEM DESIGN AND ANALYSIS



Existing System:

- ❖ 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. Other manual categorization methods are based on market- and resource-based similarities between a firm and candidate competitors.
- ❖ Zheng et al. 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.

Disadvantages Of Existing System:

- ❖ 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.

Advantages Of Proposed System:

- 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.
- 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.

IV. IMPLEMENTATION

Admin

In this module, admin has to login with valid username and password. After login successful he can do some operations such as view all user, their details and authorize them , Add hotels(Hotel name, Location, Area name, Item name, item price, item description, item image, no. Of rroms available, Room Charge Distance from Location), Add malls(Mall name, location, area name, mall description, mall specialization ,mall image, Distance from Location) , View all hotel details with rank, Comments , view all mall details with rank, comments, View all hotel booking details and payment details, view hotels and mall rank result chart, view top k searched keywords in chart .

User

In this module, there are n numbers of users are present. User should register before doing some operations and also add your location while registration . After registration successful he can login by using valid user name and password and location. After Login successful he will do some operations like view profile details, Create and manage account, search nearest neighbor hotels and malls from your location and view details, GMap, give comment, Book hotels, show top K searched keywords.

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 gave a comprehensive analysis of the competitor mining algorithms with its advantages and drawbacks. Finally, the CMiner++ yielded least computation time when comparing others. The most important features and process are not considered in the all baseline algorithms. This can be improved in the further researches

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 networkbased 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.