Volume No:4, Issue No:10 (March-2019)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Finding Top-K Competitors from Large Unstructured Datasets

K. Akhil
Department of Computer Science and Engineering
NS Raju Institute of Technology,
Visakhapatnam, Andhra Pradesh-531173, India.

T. Sai Sekhar
Department of Computer Science and Engineering
NS Raju Institute of Technology,
Visakhapatnam, Andhra Pradesh-531173, India.

V. Savinay
Department of Computer Science and Engineering
NS Raju Institute of Technology,
Visakhapatnam, Andhra Pradesh-531173, India.

G. Rajasekharam
Department of Computer Science and Engineering
NS Raju Institute of Technology,
Visakhapatnam, Andhra Pradesh-531173, India.

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 paper, we present a formal defination 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.

Index Terms ***-Data mining, Web mining, Information Search and Retrieval, Electronic commerce***

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 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 [2-4] and query session data or query topic models. 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, (iii) according to their co occurrences in query sessions [5], and (iv) based on their similarity in the topic distribution space. 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 [6] that takes a user location

Cite this article as: K. Akhil, T. Sai Sekhar, V. Savinay & G. Rajasekharam, "Finding Top-K Competitors from Large Unstructured Datasets", International Journal of Research in Advanced Computer Science Engineering, Volume 4 Issue 10, 2019, Page 14-19.

Volume No:4, Issue No:10 (March-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

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 or geo-documents. Furthermore, 53% of Bing's mobile searches in 2011 were found to have a local intent. We apply a random walk with restart (RWR) process [7] 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 [8]; 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 [9-10]. 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 [11]. 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) [12-14] for RWR search to compute the location-aware suggestions.

2 Defining Competitiveness

The typical user session on a review platform, such as Yelp, Amazon or Trip Advisor, 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 and 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 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. For instance, the first segment includes 100 customers who are interested in parking and wi-fi, while the second segment includes 50 customers who are only interested in parking.

2.1 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-

Volume No:4, Issue No:10 (March-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

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 2 I, we want to identify the k items from I that maximize. A naive algorithm would compute the competitiveness between i and every possible candidate. The complexity of this brute force method is clearly which can be easily dominated by the power set factor and, as we demonstrate in our experiments, is impractical for large datasets. One option could be to perform the naïve computation in a distributed fashion. Even in this case, however, we would need one thread for each of the *n*2 pairs. This is far from trivial, if one considers that *n* could measure in the tens of thousands. In addition, a naïve Map Reduce 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.

3. Experimental Evaluation

In this section we describe the experiments that we conducted to evaluate our methodology. All experiments were completed on an desktop with a Quad-Core 3.5GHz Processor and 2GB RAM.

3.1 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 rooms 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.

3.3 Preliminary investigation

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. Preliminary investigation begins. The activity has three parts:

- Request Clarification
- Feasibility Study
- Request Approval

3.4 Request clarification

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires. Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network (LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

4. FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is

Volume No:4, Issue No:10 (March-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- Operational Feasibility
- Economic Feasibility
- Technical Feasibility

4.1 Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

4.2 Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

4.3 Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and Web Logic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

5. Result and Discussions

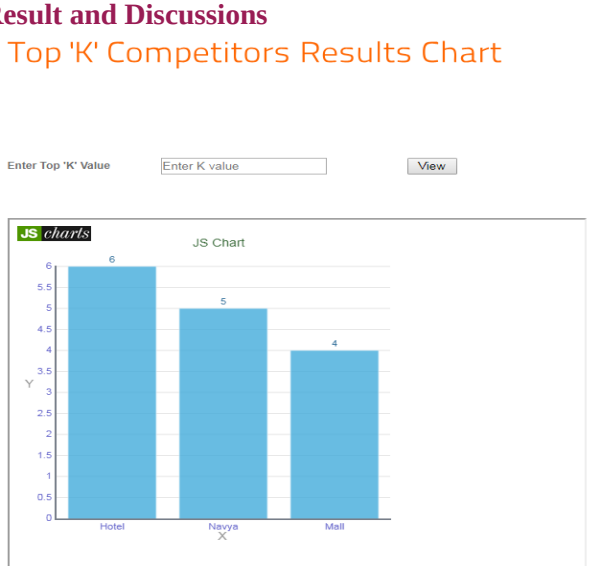


Fig. 1Top k Competitors Results

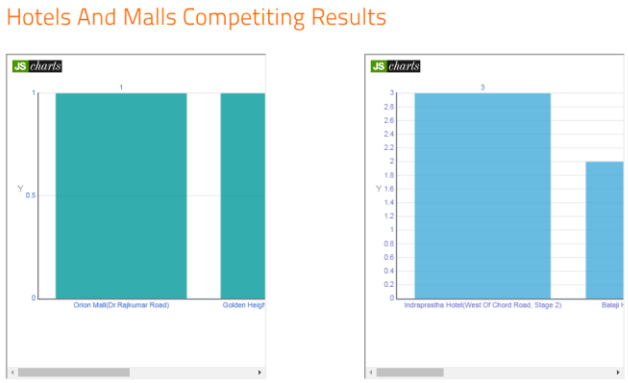
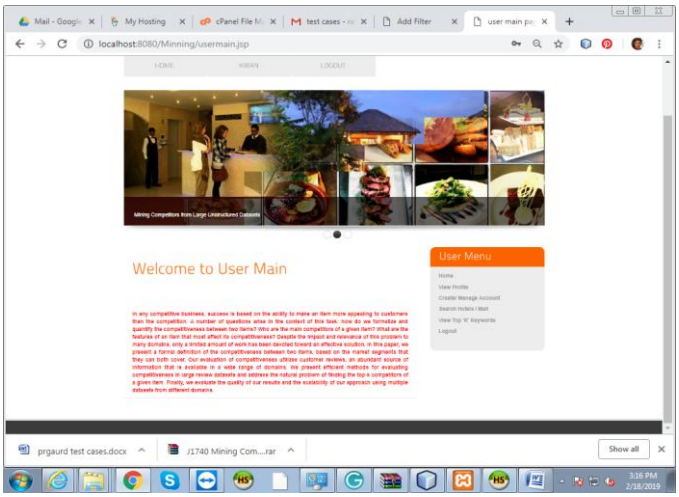


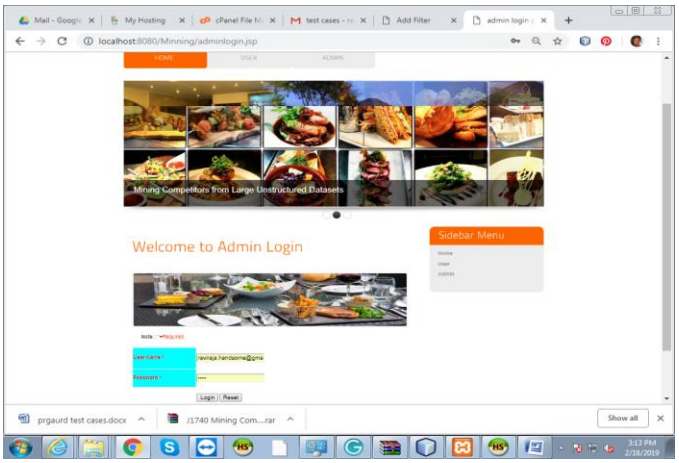
Fig. 2 Hotel and Malls Competiting Results

MAIN INTERFACE

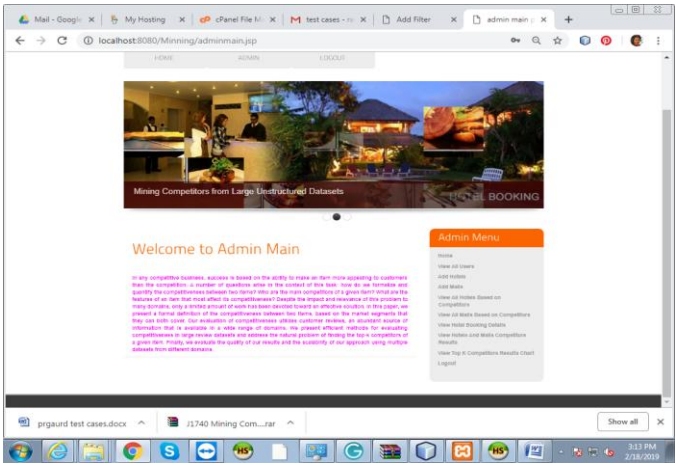


Volume No:4, Issue No:10 (March-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

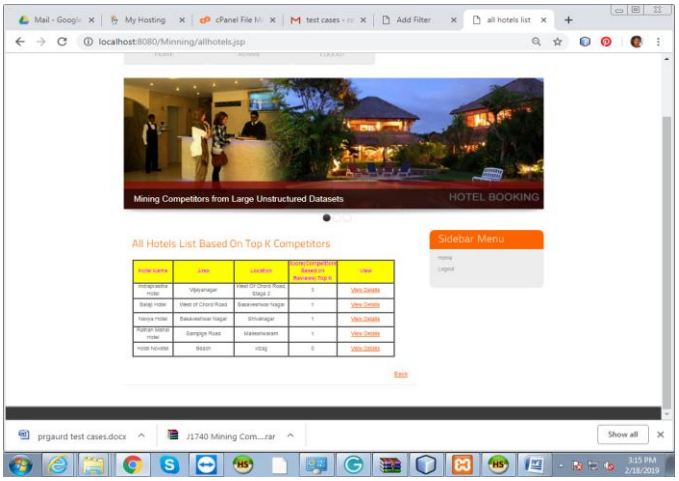
ADMIN LOGIN MODULE



ADMIN PAGE



ALL HOTELS LIST BASED ON TOP K COMPETITORS



6. Conclusion

We presented a formal dentition 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 san end-to-end methodology for mining such information from large datasets of customer reviews. Based on our competitiveness dentition, 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.

7. References

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

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

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

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

Volume No:4, Issue No:10 (March-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

[5] 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.

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

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

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

[9] 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.

[10] 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.

[11] E. Marrese-Taylor, J. D. Vel´asquez, 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.

[12] 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.

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

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