A SURVEY ON EFFICIENT DECENTRALIZED CREWING ALGORITHM IN SEO

[1]Janani P,[2]Banuchandar J
[1]PG scholar, [2] PG scholar
[1]Depatertmernt of CSE,[2]Department of CSE
[1]M.Kumarasamy College of Engineering Karur, India,
[2]M.Kumarasamy College of Engineering ,Karur ,India

ABSTRACT -- Crewing of search engine queries has involved important concern in recent years. Many search engine appliances such as query recommendation require query crewing as a pre-requisite to job properly. Certainly, crewing is needed to unlock the true value of query logs. However, crewing search queries effectively is reasonably challenging, due to the elevated diversity and arbitrary input, processing, storage, transmission cost by users. In this crewing method, which is capable of crewing dynamic and distributed data sets. Crewing nodes continuously assist through decentralized rumor-based communication to maintain summarized views of the data set. We modify crew for execution of the partition-based and density-based crewing methods on the summarized views, and also offer enhancements to the basic algorithm. Search queries are typically tiny and confusing in terms of user requirements. Many diverse queries may refer to a single concept, while a single query may cover many concepts. Crewing gives good quality results but is computationally reasonably exclusive. We Proposed a Decentralized Crewing Frame Work for Search Engines. Coping with dynamic data is made possible by gradually adapting the crewing model. This paper presents a decentralized crewing approach using search engine results might themselves be used to discover query similarity. We propose a decentralized crew similarity metric for diverse queries based on the ranked URL results returned by a search engine for queries. This is used to develop a very efficient and accurate algorithm for crewing queries. Our experimental results demonstrate more accurate crewing performance is efficiently, better scalability and robustness, scalable transmission cost of our approach against known baselines.

I.INTRODUCTION:

As of today, the indexed web contains at least 30 billion pages. In fact, the overall web may consist of over 1trillion unique URLs, more and more of which is being indexed by search engines every day. Out of this morass of data, users typically search for the relevant information that they want by posing search queries to search engines. The information explosion on the Internet has placed high demands on search engines. Yet people are far from being satisfied with the performance of the existing search engines, which often return thousands of documents in response to a user query. Many of the returned documents are irrelevant to the user's need. The precision of current search engines is well under people's expectations. In order to find more precise answers to a query, a new generation of search engines or question answering systems have appeared on the Web. The problem that the search engines face is that the queries are very diverse and often quite vague and/or ambiguous in terms of user requirements. Many different queries may refer to a single concept, while a single query may correspond to many concepts. To organize and bring some order to this massive unstructured dataset, search engines crew these queries to group similar items together. To increase usability, most commercial search engines, such as Google, Yahoo!, Bing, and Ask also augment their search facility through additional services such as query recommendation or query suggestion. These services make it more convenient for users to issue queries and obtain accurate results from the search engine, and thus are quite valuable. From the search engine perspective, effective crewing of search queries is a necessary pre-requisite for these services to function well. Due to all of these reasons, crewing of search engine queries has attracted significant attention in recent years. However, existing prevalent crewing methods, such as K Means cannot assure good results in such a diverse environment. There are several challenges posed by the unique nature of the environment. The primary issue is to determine how to measure similarity between queries. To enable more precise information retrieval, a representative and accurate descriptor is indispensable for computing the similarity between queries. In this paper, we propose a new approach to query crewing using user logs. It is one of the weight calculations. If users clicked on the same documents for different queries, then the queries are similar. Our preliminary results are very encouraging many queries that we consider similar are actually crewed together using our approach. In addition, we notice that many similar questions would have been grouped into different crews by traditional crewing approaches.

In this paper, a Distributed decentralized Crewing algorithm is proposed and instantiated with three popular partition based and density-based and distance based crewing methods. We first introduce a basic method in which nodes gradually build a summarized view of the data set by continuously exchanging information on data items and data representatives using gossip-based communication. Gossiping is used as a simple, robust and efficient dissemination technique, which assumes no predefined structure in the network. The summarized view is a basis for executing weighted versions of the crewing algorithms to produce approximations of the final crewing results.

Next the text summarization is an important component of Web search engines. Search engines can potentially return a large number of search results (URLs) for a given query, so in addition to each result, a small snippet of text is included that attempts to summarize the contents of the page behind the URL. Users scan these snippets to understand which URLs contain the most relevant results for their queries. Therefore, providing effective summaries via key concepts can increase the overall user satisfaction. Therefore, describing queries only by content-based keywords or purely through click-through data is not always accurate for search engine query clustering. Based on this, we also define a new similarity metric that can be used in any distance based crewing algorithm. Due to the diversity of queries and the curse of dimensionality, current crewing algorithms have high computational cost. We also propose an efficient crewing algorithm to reduce computational cost. We compare the query crewing results of our approach with several existing state of the art methods and show that our algorithm provides good cohesion, separation on crewed queries and significantly reduced runtime.

II. LITERATURE REVIEW

This paper presents a hierarchical crewing method named RACHET (Recursive Agglomeration of Crewing Hierarchies by Encircling Tactic) for analyzing multi-dimensional dispersed data. Crewing of multidimensional information is a vital step in many fields including data mining [6], statistical data analysis [1, 12], pattern recognition and image processing [7], and business applications [2]. Hierarchical crewing based on a dissimilarity measure is perhaps the most common form of crewing. It is an iterative process of merger (agglomeration) or splitting (partition) of crews that creates a tree structure called a dendrogram from a set of data points. Present crewing approaches do not offer a solution to the dispersed hierarchical crewing trouble that meets all these requirements. Most crewing approaches are limited to the centralized data situation that requires bringing all the data together in a single, centralized warehouse. For large datasets, the diffusion cost becomes excessive. If centralized, crewing massive centralized data is not feasible in practice using existing algorithms and hardware. We present a crewing algorithm named RACHET that is especially suitable for very large, high-dimensional, and horizontally dispersed datasets. RACHET builds a global hierarchy by merging crewing hierarchies generated locally at each of the dispersed data sites. Its time, space, and transmission costs are at most O(n) (linear) in the size of the dataset. This includes only the complexity of the transmission and agglomeration phases and does not include the complexity of generating local crewing hierarchies. Finally, RACHET's summarized description of the global crewing hierarchy is sufficient for its accurate visual representation that maximally preserves the proximity between data points. RACHET provides a comparable quality solution to the distributed hierarchical crewing problem while being scalable with both the number of dimensions and the number of data sites. Results on the small real-world UCI ML data sets indicate that RACHET can provide a more effective crewing solution than the solution generated by the centralized crewing. RACHET algorithm is scalable to such sizes of the problem because it transforms a large problem into a set of small sub problems with cumulative computational cost much less than the aggregate problem. This algorithm having some limitations that result are not accuracy.

A large number of previous works have demonstrated that cooperative spectrum sensing (CSS) among multiple users can greatly improve detection performance. However, when the number of secondary user is large, the sensing overheads will likely be intolerable if all SUs participate in CSS. In this paper, we proposed a fully decentralized CSS scheme based on recent advances in consensus theory and unsupervised learning technology. In cognitive radio networks, secondary users (SUs) opportunistically exploit spectrum holes left by primary users to improve utilization of the wireless spectrum. To determine the temporal or/and spatial spectrum holes, a main challenge is reliable and efficient spectrum sensing. Compared with the single-SU spectrum sensing methods, cooperative spectrum sensing (CSS) among multiple SUs can be employed to tackle the problem of hidden primary receivers [3] and wireless channel fading [4] by exploiting multi-user spatial diversity [5]. A number of centralized CSS techniques have been proposed, including methods of hard combination, soft combination and quantized combination [13]. To further reduce sensing overheads and improve the network scalability and robustness, there has been a recent focus on the design of decentralized CSS schemes. Unsupervised learning is receiving much attention in the field of pattern recognition and machine learning. The main idea is "learning without a teacher", which inherently has something in common with decentralized network operation. Consensus theory is initially bio-inspired, and it plays a vital role in the field of distributed control and decision making, multi-agent cooperation and decentralized data fusion in wireless sensor networks. The proposed scheme mainly consists of two phases the crewing phase and sensing phase both of which are implemented in an ad hoc manner. The crewing phase aims at crewing the SUs with potentially best detection performance and the sensing phase carries out distributed data fusion of the sensing outcomes of the SUs in the crew. Numerical results show that the proposed scheme achieves detection performance comparable to that of the OCS scheme with significantly reduced sensing overheads. This algorithm results are not accuracy.

This paper considers the distributed K-means crewing problem where the data and computing resources are distributed over a large P2P network. It offers two algorithms which produce an approximation of the result produced by the standard centralized K-means crewing algorithm. The first is designed to operate in a dynamic P2P network that can produce crewing. The second algorithm uses uniformly sampled peers and provides analytical guarantees regarding the accuracy of crewing on a P2P network. K-means crewing is a well-known and well-studied exploratory data analysis technique. The standard version assumes that all data are available at a single location. However, if data sources are distributed over a large-scale Peer-to-Peer (P2P) network, collecting the data at a central location before crewing is not an attractive and practical option. This algorithm can easily adapt to dynamic P2P network where existing nodes drop out and new nodes join in during the execution of the algorithm and the data in network changes. However, it is difficult to analyze that the algorithm and performance guarantees are experimentally verified. Our second algorithm works by taking a uniform random samples of nodes from the network. This algorithm results are not accuracy.

In a large network of computers or wireless sensors, each of the components has some data about the global state of the system. Much of the system's functionality such as message routing, information retrieval and load sharing relies on modeling the global state. Computing global data mining models decision tree, k means crewing in large distributed systems may be very costly due to the scale of the system and due to communication cost, which may be high. The cost further increases in a dynamic scenario when the data changes rapidly. In this paper we describe a two step approach for dealing with these costs. First, we describe a highly efficient local algorithm which can be used to monitor a wide class of data mining models. Then, we use this algorithm as a feedback loop for the monitoring of complex functions of the data such as its k-means crewing. The theoretical claims are corroborated with a thorough experimental analysis. Local algorithms are one of the most efficient of algorithms developed for distributed systems. Local algorithms are in-network algorithms in which data is never centralized but rather computation is performed by the peers of the network. At the heart of a local algorithm there is a data dependent criteria dictating when nodes can avoid sending updates to their neighbors. An algorithm is generally called local if this criteria is independent with respect to the number of nodes in the network. Therefore, in a local algorithm, it often happens that the overhead is independent of the size of the system. Primarily for this reason, local algorithms exhibit high scalability. These include association rule mining, facility location, outlier detection, L2 norm monitoring [10], classification [9], and multivariate regression [8]. In all these cases, resource consumption was shown to converge to a constant when the number of nodes is increased. Still, the main problem with local algorithms, thus far, has been the need to develop one for every specific problem. This algorithm computational overhead is low.

CONCLUSION

To extract maximum value from search queries, search engines must develop efficient approaches for generating more precise and multi-functional crews of similar queries. However, most prevalent crewing methods all suffer from certain limitations on crewing this highly diverse set of queries. In this paper we first identified the necessity of an effective and efficient distributed crewing algorithm. Dynamic nature of data demands a continuously running algorithm which can update the crewing model efficiently, and at a reasonable pace. We introduced SE Crew, a general fully decentralized crewing algorithm, and instantiated it for partition-based and density-based crewing methods. The proposed algorithm enabled nodes to gradually build a summarized view on the global data s et, and execute weighted crewing algorithms to build the crewing models. Adaptability to dynamics of the data set was made possible by introducing an age factor which assisted in detecting data set changes updating the crewing model. Our experimental evaluation and comparison showed that the algorithm allows effective crewing with efficient transmission costs, while being scalable and efficient SE Crew can be customized for other crewing types, such as hierarchical or grid-based crewing. To achieve this, representatives can be organized into a hierarchy, or carry statistics of approximate grid cells. Further discussion of these algorithms is deferred to future work.

REFERENCES:

1. M.R. Anderberg, Cluster Analysis and Applications, Academic Press: New York, 1973.

2. R. Brachman, T. Khabaza,W. Kloesgen, G. Piatetsky-Shapiro, and E. Simoudis, "Mining business databases," Communications of ACM, vol. 39, no. 11, pp. 42–48, 1996.

3. Cabric D, Mishra S M, Brodersen R. Implementation issues in spectrum sensing for cognitive radios. In: Michael B M, ed. Proceedings of 38th Asilomar conference on Signals, Systems, and Computers. Pacific Grove, USA, 2004. 772–776.

4. Sahai A, Tandra R, Mishra S M, et al. Fundamental design tradeoffs in cognitive radio systems. In: Milind B, Anant S, Kitti H, eds. Proceedings of the First International Workshop on Technology and Policy for Accessing Spectrum. Boston, USA, 2006. 1–6

5. Duan D L, Yang L Q, Principe J C. Cooperative diversity of spectrum sensing for cognitive radio systems. IEEE Trans Signal Process, 2010, 58: 3218–3227

6. U. Fayyad, D. Haussler, P. Stolorz, and R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, MIT Press: Cambridge, MA, 1996.

7. K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press: New York, 1990.

8. K. Bhaduri and H. Kargupta, "An Efficient Local Algorithm for Distributed Multivariate Regression in Peer-to-Peer Networks," in *Proceedings of SDM'08*, Atlanta, Georgia, 2008, pp. 153 – 164.

9. P. Luo, H. Xionga, K. Lu, and Z. Shi, "Distributed classification in peerto- peer networks," in *Proceedings of SIGKDD'07*, San Jose, California, 2007, pp. 968–976.

10. R. Wolff, K. Bhaduri, and H. Kargupta, "Local L2 Thresholding based Data Mining in Peer-to-Peer Systems," in *Proceedings of SDM'06*, Bethesda, Maryland, 2006, pp. 428–439.

11. S. Bandyopadhyay, C. Giannella, U. Maulik, H. Kargupta, K. Liu, and S. Datta, "Clustering Distributed Data Streams in Peer-to-Peer Environments," *Information Science*, vol. 176, no. 14, pp. 1952–1985, 2006.

12. L. Kaufman and P. Rousseeuw, Finding Groups in Data, John Wiley and Sons: New York, 1989.

13. Miia M, Marja M, Aarne M. Cooperative spectrum sensing using quantized soft decision combining. In: Souhir D, Joao P M, eds. Proceedings of 4th international Conference on Cognitive Radio Oriented Wireless Networks and Communications. Hannover, Germany, 2009. 1–5

14. I. Karthika, P. Gokulraj, S. Saravanan" Prediction of sales using Big data analytics" Journal Of Advances In Chemistry Vol 12, No 20.

15.I.Karthika, K.P.Porkodi," Automatic Monitoring And Controlling Of Weather Condition Using Big Data Analytics" , International Journal of Advanced Research in Computer and Communication Engineering Vol. 6, Issue 1, January 2017.