International Journal of Computer Science and Engineering Communications
Vol.3, Issue 3, 2015, Page.1154-1158
ISSN: 2347–8586
www.scientistlink.org

Recent Query Processing techniques and trends under Monogram Text mining

B. Sruthi *, S.R. Srivathsan*, R. Kavitha**

* UG Scholar, ** Assistant Professor-II
School of Computing, SASTRA University, Thanjavur, India
sharikha2009@gmail.com, goksriva@gmail.com*, kavitha_r@cse.sastra.edu***

ABSTRACT: The proliferation in the amount of data in the web has lead to the need for better understanding of the queries thereby leading to question answering systems. Retrieving desired results from vast amount of data is exhausting. Hence lots of processing techniques are involved to arrive at the relevant answers. In this paper all the processes involved in a question answering system for monogram text mining are being discussed and some of existing techniques have been compared based on the corresponding domains.

Keywords: *Query processing, NLP, classification, filtering and mining.*

1.Introduction

Knowledge and Information are different dimensions of data. Information is processed data whereas knowledge is information that is modeled to be useful. To get knowledge you need some cognitive and analytical ability while for information you do not need cognitive ability [1]. However due to extravagant amount of data in the web, users are looking for information rather than knowledge. The output of a question answering system may be a single line or a paragraph and the system itself may look to be simpler than a search engine. As for the proverb “Appearances are deceptive” suits best for this situation, a question answering system may look simpler from outside but it takes lot of processing. Several complex steps are involved in a question answering system like query processing, classification, filtering and mining.

2.Query Processing

As the size of the information increases, there is need to improve the efficiency of the techniques and algorithms involved in query processing. Query processing is a technique to obtain desired information in a reliable and predictable manner. To obtain the results within the time frame it is important to optimize the query. Generally using NLP, queries are processed and keywords are extracted. Then the system tries to understand the question using those keywords and in turn it looks for a specific answer in the right place. Various query processing techniques that are used in different domains are discussed here. Hybrid query processing engines (HyPE) is extended to support operator-stream-based scheduling and heuristic that utilize the inter-device parallelism and an optimization heuristic called Probability-based outsourcing are introduced [2]. In HyPE efficient processing device utilization involves waiting-time-aware response time, threshold-based outsourcing, throughput and roundrobin. Query optimization approaches can be improved to serialize a set of queries to an operator stream. To serve the changing workload of query access pattern, a lightweight Adaptive Multi-Route Index (AMRI) [3] approach which employs a bitmap time-partitioned design is used. This method incorporates migration strategies so that it can accept old and new query. Among the available index assessment method CDIA using highest count compression is better at improving throughput. AMRI uses both synthetic and real data and can adapt to systems with heavy route fluctuations. Compression methods are customized to handle the search benefit relationships between indices in AMR and hence problem of index tuning in AMRs are handled effectively. [4] Propose an architecture that provides secure storage infrastructure by combining cloud based WBANs with statistical modeling techniques. Periodic sending of sensor data to medical database is no longer needed because WBAN with cloud services provide optimized query processing.

This combination provides better individual query latency for real- time diagnosis as well as energy efficient processing. The proposed architecture and its query mechanism can be extended to provide better performance. The problem with the kOLS methods is that they are applicable only to the Euclidean distance hence they are not sufficient in metric space. To overcome these problems Metric index structures based solutions are employed on the dataset. These structures makes use of pruning rules and reuse techniques and optimally estimate the score to support different types of data. A flexible and extensible algorithm called EB is proposed in [5]. This algorithm does not consider the representations of the object and as long as the similarity between the objects is maintained and the triangular inequality is satisfied, this algorithm is applicable. The efficiency of the algorithm can be improved by devising better pruning rules and optimality score estimations. Query processing plays a vital role in relational databases because users generally access the data in the DB by posting queries. These queries in turn must be processed effectively to extract user desired results. In [6]Query processing techniques can be classified along different design dimensions such as query model, data access methods, implementation level, data and query uncertainty and ranking functions. Specialized algorithms are employed to process a particular relational operation.

3.Classification

Generally information is stored as text and almost all the sources have unstructured text as their knowledge base. Text classification plays a vital role to convert these large chunks of unstructured data into a predefined (structured) format. Classification is also used to predict membership for data instances. When a new observation comes into picture it is classified based on certain attributes. There exist different classification algorithms which help to find the correct membership for the given data but these algorithms have different characteristics and hence it is important to understand the advantage and disadvantages of each algorithm so that it helps to identify the suitable algorithm depending on the complexities and various criteria’s.

Classification is done in order to retrieve specific information in a speculated timeframe. Different classifiers are used to categories wide variety of information. K-Nearest Neighbors, Naive Bayes, Decision tree, Decision Rule, SVM are some of the commonly used classification algorithms. There have been several studies done comparing various algorithms. In [6] an overview of few algorithms is provided and a comparative study was conducted based on several criteria like time complexity, principal and performance. They have classified the documents in three ways, supervised, unsupervised and semi supervised methods. The quality of the data set affects the efficiency of the algorithm. Each algorithm have their own advantages and disadvantages, but KNN proved to be the best based on time complexity, then neural network, then comes Naive Bayes. Information gain and chi square performs well in feature selection. Since KNN has been proved to be more efficient based on time complexity it has been implemented in [8] a recommender system.

KNN was implemented along with the Euclidean distance. It implemented a unique way of classifying the user, based on the click streams and mapping them to a particular class in the data mart. It was more straightforward, consistent and transparent. A method for improving the accuracy of data mining classification algorithms [9] by Nikolaos et.al proposed a new method called CL.E.D.M for classifying data. Classification through ELECTRE and Data Mining trains the data mining algorithms and extracts best decision rules. ELECTRE is the method used on decision analysis. The method's performance was tested using three well known DBs and compared with data mining algorithms (decision tree algorithm C4.5, rule-based algorithms CN2 and CL). This method proved to be more accurate. An idea to create a classification algorithm to cater pure numerical data or a mixed data set was suggested as future work. However different algorithms may suit best for different linear based on the application. Hem et.al in [10] proposed a novel approach combining decision tree and ID3 where in decision tree was used for model classification and prediction followed by ID3 to divide the attributes into groups and if the information gain was not good enough, they were further divided. It increased the quality of solution and classified the data more accurately. According to the observations made by Geetika, the K-nearest-neighbor method suffers severely from what is called the *curse of dimensionality* [11] i.e. as the value of K increases, the accuracy of the method deteriorates.

If there seems to be any variation in the data set, decision tree method may tend to choose false positive information. Naive Bayes method proved to be more efficient than decision tree in a particular field where all the input attributes where significant. Another algorithm C4.5 best fits where the data are categorized and it is said to be more sensitive to input data and its accuracy could be increased using feature selection along with the algorithm.

4.Filtering

Data cleaning is one of the important processes when it comes to a question answering system. While retrieving information lot of noisy data may get along with the expected result and hence filtering has to be done to remove all those unwanted data. This becomes the key to improve the accuracy of any QA system. Wie Deng et.al have revised a well known algorithm call SVD and proposed a new method of filtering [12]. They overlapped the community of users with the cluster and fed it as feedback to SVD++. They have also used a 'difference matrix' as the input for their own algorithm called Difference-SVD. On comparing their algorithm with the existing SVD, difference-SVD proved to be more efficient based on time complexity and its accuracy could be increased with larger data set. Though it works well for collaborative filtering it is not suitable for content based filtering.

Filtering can also be done using quantity and quality aspects. This has been done by Martin et.al in [13] for human content filtering in twitter. He had conducted 239 surveys from the users to suggest which person to follow for better post content qualitative cues were more of a help for recommendation. However there was a drawback that based on the way the data sets were provided, the user results varied. Combining both the quantitative and qualitative cues was suggested in future works.

More and more novel approaches being made by research scholars, Vidyavathi et.al has proposed a filtering method in [14]. This was implemented for feature selection method and mutual information maximization was used for that. Preselected components were clustered for selection. Then the evaluation was done using SMV classifier. The prediction accuracy was more for smaller number of features. This is not more efficient under time complexity as it does lot of preprocessing to check for redundant data and hence it cannot be used in real time.

Hong et.al has considered the collaborative filtering method for social streams in [15]. Similarities between the users were estimated, then the value of the article. Both were then collaborated to find out the value of an article for a particular user class. Another approach based on memory, reading speed, processing speed and several other factors were discussed. This is one of the conventional ways of filtering since it does not use the keywords unlike any other earlier methods.

One of the methods used in collaborative filtering is the layered approach. Nikos et.al in [16] for paper recommendation using Mendeley data set. Based on the user ratings and the number of views for the papers existing in a library they were classified. A system by Manouselis & Costopoulou was used for filtering. It facilitated a more detailed and informed evaluation of such systems. This was tested only for a closed data set and need not be as efficient for a dynamic data set.

Another approach for collaborative filtering by Lingfeng et.al in [17]. An intuitive and efficient query strategy using classes like users, ratings and item set. The algorithm 1 identified the user community and interest groups and normalized using Gaussian distribution. Expectation maximization algorithm was used to find the maximal likelihood and Bayesian method for active learning process. The algorithm 2 was used for sorting the topics in descending order. It reduced the complexity of item selection procedure.

5. Monogram text mining

Mining is the final step that leads to a more focused answer for the query. From all the earlier processes the output could be a document or relevant links from the web or a set of data from the repository. It all comes to the mining process at the end. There are several mining techniques and algorithms which have been proved to be efficient in specific applications. There are different areas where mining can be implemented. Either the contents from the web can be mined or from a database or from a text document retrieved by a web crawler. Sumaiya Kabir et.al in [18] have proposed a new method for mining more relevant information from a huge pile of relevant and irrelevant data from the web using artificial agents to build a library that relates entities if the desired result was found in the RDF database or not. This was an ontology based method using a single agent.

In order to make the system more complex and efficient multiple agents can be used considering several other factors like inter agent communication. It is very predominant to use ontology based approach for web mining due to its efficiency. Ziang Li et.al in [19] have proposed an ontology based web mining approach to improve the accuracy of unemployment rate prediction. Initially Ontology construction was done followed by wrapper based feature selection. And a comparative study was conducted to identify the best data mining method among several methods like NNs and SVRs and v-SVR (RBF kernel) prediction proved to be the most effective method. The limitation of developing a more generic method to identify the related event pairs motivated Cao et.al in [20] to develop a three phased approach. Initially identifying the casual relationship then extracting the event arguments and finally measuring the casual association. They have used lexico-syntactic patterns for the first phase and dependency argument mapping rules for the second phase. More accurate due to the use of local dependency tree and has a better performance.

6.Conclusion

Accuracy and performance are the two major factors to be considered while constructing a question answering system. The system may not even find an answer for a given query but it is important that the system does not retrieves irrelevant information. In this paper we have discussed some of the existing methods and techniques involved in building a Question Answering System (QAS). We conclude by saying that based on the requirements and application of the QAS the most suited methods can be chosen to make the system more accurate and efficient. Implementation of monogram text mining for a specific domain is further in progress.

Table.1: Comparative study of various algorithms

ALGORITHM	ADVANTAGES	DISADVANTAGES
1.DECISION TREE (Classification)	• Understandable • Knowledge can be extracted and represented in form of rules Rules can be generated easily • Reduce problem complexity • Robust	• Highly unstable with respect to minor perturbation • Training time expensive Information about one class is distributed throughout tress. • May suffer over fitting • Trees can be complex
2.KNN (Classification)	• As 'k' value increases, noise decreases • Easy to implement Provides high performance	• Classification time is long • Difficult to find optimal value of 'k' Suffers from curse of dimensionality
3.ID3 (Classification)	• Used to produce decision tree	• Does not necessarily provide optimal solution • Can over fit to training set and not suitable for handling continues variables
4.NAIVE BAYES (Classification)	• Takes into account prior information about given problem • Simple structure, easy to implement and compute. • Can handle incomplete data set • Efficient output and can be used for large amount of data	• Shows poor performance when features are correlated
5.MAUT (Multiple Attribute Utility Theory) (Filtering)	• handle the trade- offs among multiple objectives • Ease of use	

References

[1] Thorndike, "E. L., Intelligence and its measurement", A symposium. Journal of educational Psychology, pp.123-147.

[2] Sebastian Breb, Norbert Siegmund, Max Heimel, Michael Saecker , Tobias Lauere, Ladjel Bellatreche f, Gunter Saakea, 2014 "Load-aware inter-co-processor parallelism in database query processing", Data & Knowledge Engineering,Volume 93, pp. 60–79.

[3] Karen Works, Elke A. Rundensteiner, Emmanuel Agu, 2013 "Optimizing adaptive multi- route query processing via time-partitioned indices", Adaptive query processing, Journal of Computer and System Sciences, Volume 79, pp. 330–348.

[4] Ousmane Diallo, Joel J.P.C. Rodrigues, Mbaye Sene , Jianwei Niu , 2014 "Real-time query processing optimization for cloud-based wireless body area networks", Elsevier, Information Sciences, Volume 284, pp. 84–94.

[5] Yunjun Gao, Shuyao Qi , Lu Chen , Baihua Zheng , Xinhan Li, 2015 "On efficient k-optimal- location-selection query processing in metric spaces", Elsevier Inc, Information Sciences, Volume 298, pp. 98–117.

[6]Bamnote G.R, Agrawal S. S,2013, "Introduction to Query Processing and Optimization", Volume 3, Issue 7.

[7] Vandana Korde, Namrata Mahender C, 2012, "Text classification and classifiers: A survey", Text Classification.

[8] Adeniyi D. A, Wai Z, Yongquan Y, 2009, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method".

[9] Nikolaos Mastrogiannis, BasilisBoutsinas, IoannisGiannikosa, 2009, "A method for improving the accuracy of data mining classification algorithms", Computers & Operations Research.

[10] Hem Jyotsana Parashar, Singh Vijendra, and Nisha Vasudeva, 2012, "An efficient classification approach for data mining", Machine Learning and Computing.

[11] Geetika, 2012, "A Survey of Classification Methods and its Applications".

[12] Wei Deng, Rajvardhan Patil, Lotfollah Najjar, Yong Shi, Zhengxin Chen, 2014, "Incorporating Community Detection and Clustering Techniques into Collaborative Filtering Model", Procedia Computer Science, Vol 31, pp. 66–74.

[13] Martin Chorley J, Gualtiero Colombo B, Stuart Allen M, Roger Whitaker M, 2015, "Human content filtering in Twitter: The influence of metadata", Human-ComputerStudies,Volume 74, pp.32–40.

[14] Vidyavathi B. M, Ravikumar C. N, "A novel hybrid filter feature selection method for data mining", Volume 3, Number 3.

[15] Hong Yan, Taketoshi Ushiama, 2014, "Effective browsing technique based on behavioral collaborative filtering on social streams", Procedia Computer Science, Volume 35, pp. 1702 – 1710.

[16] Nikos Manouselis, Katrien Verbert, 2013, "Layered evaluation of multi-criteria collaborative filtering for scientific paper recommendation", Procedia Computer Science, Volume 18, pp. 1189 – 1197.

[17] Lingfeng Niu, Jianmin Wu, Yong Shi, 2011, "Second-order Mining for Active Collaborative Filtering", Procedia Computer Science, Volume 4, pp. 1726–1734.

[18] Sumaiya Kabir, Shamim Ripon, Mamunur Rahman and Tanjim Rahman, 2014, "Knowledge- Based Data Mining Using Semantic Web", IERI Procedia, Volume 7, pp. 113 – 119.

[19] Ziang Li, Wei Xu, Likuan Zhang, Raymond Lau Y. K, 2014 "An ontology-based Web mining method for unemployment rate prediction", Decision Support Systems, Volume 66, pp. 114–122. [20] Ya-nan Cao, Peng Zhang, Jing Guo, Li Guo, 2014, "Mining Large-scale Event Knowledge from Web Text", Procedia Computer Science Volume 29, pp. 478–487.