AN ANALYSIS OF K-QUERY PROCESING FOR UNSTRUCTURED DATA

V SANGEETHA
Research Scholar
Department of Computer Science and Engineering
Shri Jagdish Prasad Jhabarmal Tibrewala University

DR. PRASADU PEDDI
CO-GUIDE
Department of Computer Science and Engineering
Shri Jagdish Prasad Jhabarmal Tibrewala University

DR. MARAM ASHOK
Department of Computer Science and Engineering
Principal
Malla Reddy Institute Of Engineering and Technology

Abstract

The search operation of keyword query is challenge for ordinary users to search vast amount of data, the ambiguity of keyword query makes it is difficult to effectively give result to keyword queries. To overcome this complex problem, in this paper we propose an approach that automatically diversifies & search the keywords from XML data. The collection of digital data is growing at an exponential rate. Data originates from wide range of data sources such as text feeds, biological sequencers, internet traffic over routers, through sensors and many other sources. To mine intelligent information from these sources, users have to query the data. Indexing techniques aim to reduce the query time by preprocessing the data. Diversity of data sources in real world makes it imperative to develop application specific indexing solutions based on the data to be queried. Data can be structured i.e., relational tables or unstructured i.e., free text. Moreover, increasingly many applications need to seamlessly analyze both kinds of data making data integration a central issue.

Keywords*: unstructured, Data Analytics,*

Introduction

Top-K is a vital part of our data mining that transcends our fundamental constraints. Utilizing medications, eliminating toxins from the blood, and delivering bile and blood proteins to aid in assimilation are some fundamental aspects of Top-K. Damage to the Top-K can occur for a variety of reasons, such as alcohol misuse, excess weight, viral hepatitis, etc.; some diseases would result in recognised complications and may lead to Top-K transplant. There is a chance of preserving the Top-K from serious problems if the Top-K disease is identified at an early stage.

To search the information is essential activity of our lives. Web search engines are widely used for searching textual documents, images, and videos. There are also large collections of structured and semi-structured data both comes on web and enterprises, like relational databases, XML, data extracted from text documents, work flows, etc. In Traditional days, to access the resources, users have to learn structured query languages, they also need to access data of each individual application domain. By creating the databases more search able will increases the information amount that the user may access and also have ability to gain results of searching more efficient as compared to searching of keywords on textual documents, and also increases the usability of databases and make powerful impact on people's lives.

The main aim of diversification is to minimizing the user's dissatisfaction by balancing relevance of search results. Now a days diversification of search results on unstructured documents is a very big

problem, diversification of search results over structured databases has much less attention. Keyword queries over structured data are offering a target for diversification. Single interpretation of a keyword query can't satisfy the users, and there may possibility that multiple interpretations may overlapping the results. The main challenge here is to give users a quick and efficient result of a keyword query in the available database, to enable user to effectively select the interpretation.

Top-k Query Processing

Data mining and its uses are widely discussed in every other area of construction, science, medicine, and administration. The term "data mining" refers to the process of recovering information or anticipated results from massive amounts of data. One of the significant problems that unexpectedly developed in today's society is the Top-K issue. Because the early indicators are not very logically novel, the significance and reality of the contamination will be revised in matured sorts out only. Therefore, a fundamental step in caring for the disease is the distinctive proof of Top-K issue in its early stages. The Top-K problems might range from simple hypersensitivity to dangerous Top-K cirrhosis.

Database as well as information retrieval systems allow users to rank query answers. Such ranking is typically based on some scoring function. The data object score acts as a valuation for that object according to its characteristics. For example, price, year of manufacturing, number of miles driven, etc of car objects in a automobile database, or number of occurrence of query pattern in a given document. Data objects can be evaluated by a single attribute or multiple attributes that contribute to the total object score. Thus, ranking enables access to the query answers in the order of their relevance. In many application domains, end-users are more interested in the most important (top-k) query answers in the potentially much larger answer space.

There are many different techniques and algorithms for relational data that can be classified as data mining. The underlying assumption behind clustering and classification is the a-priori existence of a model of which the actual data is just an observed instance. Association rules, on the other hand, are data-centric, and patterns that emerge do not have to be combined to derive a complete model. Furthermore, the amount of data that has been subjected to association-rule mining is several orders of magnitude larger than the amount of data normally used in classification and clustering.

Multi Top-k Queries over Uncertain Data

Most of the data in the real world is incomplete and particular to one query is not shared all the data. So Tao Chen and Lei Chen proposed sharing among multiple top-k queries over uncertain data streams based on the frequency upper bound of each top-k queries. The main goal of this work is to share the computation queries among the possible and satisfy the frequency semantics at the same time. First it is performed single query with probability tuples and time based sliding window with size which is based on the top-k probability. Sharing queries is very challenging for the uncertain top-k queries with different frequency upper bounds and different k values. First grouping based on the same

frequency bound is called inter grouping. The intra grouping queries with the same frequency upper bound but different k values. Two types of the work used for similarity search. One is combine group with different frequency upper bonds and another one is sharing among queries between groups after combination. Dynamic programming solution is used to find optimal solution in sharing between groups and overlapping sub problems. Greedy algorithm is used in single queries, it is not a optimal solution but it is more efficient than dynamic program in term of space and time.

Unstructured Data Analytics

AaaS

IBM stated that AaaS is the service which can be used to perform analysis of unstructured data. IBM introduced AaaS platform that allows companies to submit data which can be structured or semi structured or unstructured format to perform analysis.

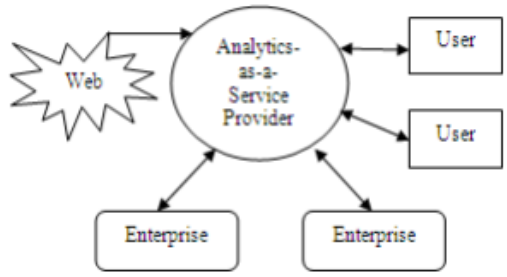


Figure: Analytics-As-A-Service Model Proposed By IBM

IBM soon realized the problem that are faced by AaaS:

- Definition of Service Level Agreement
- Quality of Service examining strategy
- Pricing
- Data management
- Processing models

AaaS is a next level of SaaS that allows companies to use analytics as a service.

Unstructured Data Mining

Without data mining, meaningful information cannot be extracted from big data. As already stated, existing methodologies are designed to process schema oriented data and cannot be used for unstructured data, Researchers are trying to develop techniques that can perform data mining on unstructured data in real time.

K-NN Algorithm

This is a parameter-free algorithm. Classification and regression are performed using the K-NN method. KNN is useful for problems involving categorization and relapse prediction. However, it is more often used in contractual matters in the corporate world. The fact that solving most scientific problems requires making a decision may explain why characterization models are so popular. For instance, whether or not a client has been worn down, whether or not client X should be targeted for advanced engagements, and so on. These inquiries are becoming more astute and have direct ties to a practical manual. Here, we'll talk about K-closest neighbours (KNN), another popular clustering method. Our primary focus will be on the mechanics of the computation, specifically how the info parameter influences the yield/forecast. There are two elements to the data set: a training dataset and a test dataset. The algorithm was trained using 70% of the data and then tested using the remaining 30%.

Table: k-NN model accuracy with different K values and label

Value of K	Liver disease (No/Yes)	Label (No)	Label (Yes)	Accuracy of k-NN Model (%)
		1	2	
1	1	88	36	61.45
	2	33	22	
2	1	97	27	67.03
	2	32	23	
3	1	91	33	60.33
	2	38	17	
4	1	95	29	60.89
	2	41	14	
5	1	98	26	62.01
	2	42	13	

Table: k-NN model accuracy with different K values

K	Accuracy
1	61.45
2	67.03
3	60.33
4	60.89
5	62.01

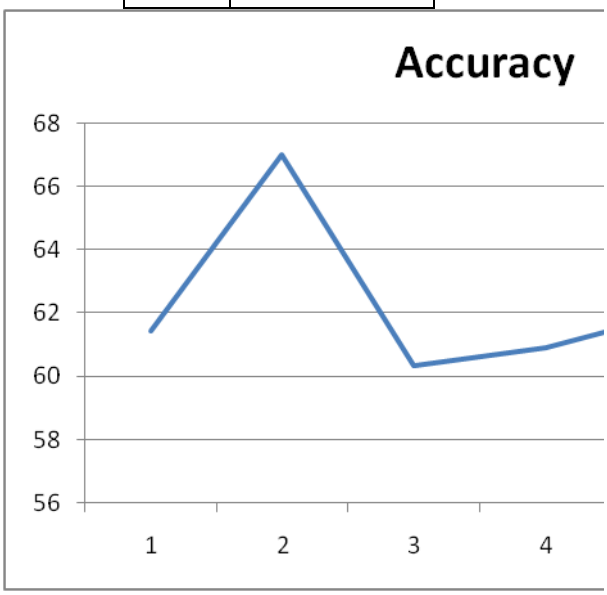


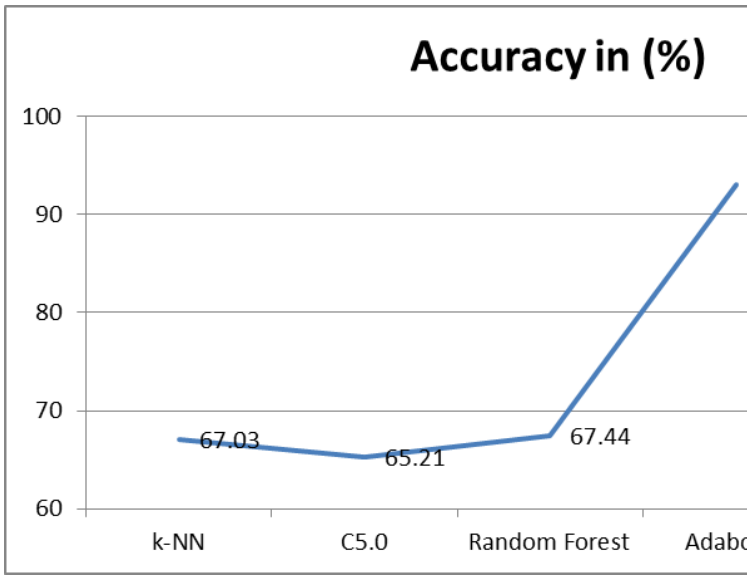
Figure: Graph between k-NN model accuracy with different K values

Random Forest algorithm

The random forest is a method used to sort data. A large number of decision trees make up this system. Conventional decision trees have a few drawbacks, including a lack of variation and the ability to reduce overfitting. All of these issues may be solved by using Random Forest. Overfitting will reduce accuracy. When compared to other methods, random forests always perform better. To build a classifier, Random Forest uses a fusion of many unrelated basic classifiers. So that you may better grasp how the Random Forest algorithm works, let's look at a concrete example. An announcement from the bank seeking new executives.

Table: Accuracy of k-NN, Random Forest, Adaboost and C5.0 algorithms

Name of Algorithm	Accuracy in (%)
k-NN	67.03
C5.0	65.21
Random Forest	67.44
Adaboost	93.07



Conclusion

From the current scenario, it is clear that Big Data will stay for long. Data is

generated at exponential rate, though big data has its own advantages but the challenge is how to analyze and perform operation on data mining**.** Big data is evolving over time and its size is getting doubled in every two years. Previously all tools were developed for schema oriented data. Though big data offers many opportunities but extracting data is new challenge. Data mining is the process of using computational methods to mine large data sets in order to find patterns, establish connections, and solve problems. In order to better prepare for the future, data mining tools are becoming more popular. Clustering, categorization, backtracking, and visit configuration ageing are only some of the other data analysis and preparation procedures that are a part of data mining. So, a classification technique may handle a broader range of data than either regression or association. This is the fundamental motivation for the ever-increasing prevalence of categorization. One of the most important and precise uses of AI is data mining. In this way, a large amount of everyday data may be efficiently organised, allowing for critical analysis that may ultimately aid the quest for basic authority.

References

1. *Y. Chen, W. Wang, Z. Liu, and X. Lin, "Keyword search on structured and semi-structured data," in SIGMOD Conference, 2009, pp. 1005–1010.*
2. *B. Wang, S. Yu, W. Lou, and Y. T. Hou, "Privacy-preserving multi keyword fuzzy search over encrypted data in the cloud," in IEEE INFOCOM, 2014.*
3. *V.Kaltsa, K. Avgerinakis, A. Briassouli, I. Kompatsiaris and M. Strintzis, "Dynamic texture recognition and localization in machine vision for outdoor environments," Computers in Industry, vol. 98, 2018.*
4. *Ai-Qin Mu1,2, De-Xin Cao1, (2009), "A Modified Particle Swarm Optimization Algorithm", Natural Science Vol.1, No.2, 151-155.*
5. *Alex S. Befelerand Adrian M. diBisceglie, (2002), "Hepatocellular Carcinoma: Diagnosisand Treatement", Gastroenterology.*
6. *E. Manogar and S. Abirami, "A DataStudy on Deduplication for Optimized Techniques Storage", 2014 InternationalSixth Conference on Advanced Computing, IEEE 2014, pp.161-166.*
7. *Gopala Krishna Murthy Nookala, Bharath Kumar Pottumuthu, (2013), "Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5.*
8. *ZhanhuaiQinluHe, Li and Zhang, "Data Techniques", 2010 International Future formation Conference, IEEE 2010, pp. 430-433.*
9. *ShwetaKharya, (2012), "Using Data Mining Techniques for Diagnosis of Cancer Disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2.*
10. *Kotsiantis. S.B, (2004), Increasing the Classification Accuracy of Simple Bayesian classifier, AIMSA, PP. 198- 207.*
11. *Milan kumarai, SunilaGodara, (2011), comparative study of data mining Classification methods in cardiovascular Disease Prediction, International journal of Computer Science and technology, vol 2, Issue2, June 2011, page no 304- 308.*
12. *Omar s. soliman, Eman abo Elhamd, (2014), Classification of Hepatities c virus using modified particle Swarm optimization and Least Squares Support Vector Machine, International Journal of Scientific & Engineering Research, volume 5, Issue3, March -2014 122.*