Medical Search Engine: Automatic Data Driven Approach for Medical Knowledge Extraction

S. Francis Shamili
ME Computer Science and Engineering
Dhanalakshmi Srinivasan Engineering College
Perambalur, India

R. Aarthy
Assistant professor, Computer Science and Engineering
Dhanalakshmi Srinivasan Engineering College
Perambalur, India

***Abstract*— Incomplete data is one major kind of multi dimensional dataset that has missing values in its dimension. It is difficult to retrieve information from this type of dataset when it becomes large. Finding top K dominant (TKD) values in this type of dataset is a challenging procedure. Traditional algorithms like Skyband based algorithm and Upper bound based algorithms are used to enhance this process. Skyband based algorithm finds the complete data from the incomplete dataset, but it is not designed to find the TKD query. This algorithm is suffered from long running time and huge storage space. This paper proposes Map-Reduced Enhanced Bitmap Index Guided Algorithm (MRBIG) for dealing with aforementioned issues. The Map-Reduce framework evolved when data became too large for machines to process. Map-Reduce framework has two fundamental functions called Mapper and Reducer, which are used to distribute huge tasks between multiple nodes to make them faster. In Map-Reduce framework, each task is split into different chunks based on internal patterns and gets distributed between nodes. The Mapper receives the data fragment then find the complete data from the incomplete dataset. The Mapper results are aggregated to calculate the TKD using the Reducer. By using the bitmap indexing approach, the results discovery will be efficient and easier to handle in most cases. For the interactive functioning of the medical search engine application Artificial Intelligence and Machine Learning techniques are used. This application will provide clinical guidelines and descriptions from doctors and also allow the doctors to maintain the medical history of patient for the future reference. Automatic data driven approach allow users to search the disease and the application pushes out a list of procedures for treatment which is more efficient than the traditional system.**

Keywords*— *Top K Search, Incomplete Data, Bit Map Index Construction, Map-Reduce Framework, Find Top K value.

I. INTRODUCTION

Big data is a term for records sets which might be so big or complex that traditional information processing application software programs are insufficient to address them. Challenges encompass seize, garage, evaluation, facts curation, search, sharing, transfer, visualization, querying, updating and statistics privateness.

The term "massive information" regularly refers surely to the usage of predictive analytics, consumer behavior analytics, or certain different advanced statistics analytics strategies that extract value from records, and seldom to a specific size of information set. "There is no doubt that the quantities of information now available are indeed huge, but that's no longer the maximum applicable function of this new information surroundings." Analysis of information sets can find new correlations to "spot commercial enterprise trends, prevent diseases, fight crime and so on." Scientists, commercial enterprise executives, practitioners of medicine, marketing and governments alike often meet difficulties with big information-units in regions such as Internet search, finance, urban informatics, and business informatics. Scientists come across boundaries in e-Science work, together with meteorology, genomics, connectomics, complex physics simulations, biology and environmental research.

Data sets grow swiftly - in component due to the fact they're more and more collected with the aid of cheap and several data-sensing mobile devices, aerial (remote sensing), cameras, microphones, radio-frequency identification (RFID) readers and tags and wireless sensor network devices. The internationals technological in keeping with-capita capability to save records has roughly doubled each forty months for the reason that Nineteen Eighties; as of 2012, every day 2.Five exabytes (2.Five×1018) of facts are generated. One question for large corporations is figuring out who should personal large-information tasks that have an effect on the complete corporation.

Relational database management structures and computing device records- and visualization-programs often have difficulty coping with big records. What counts as "big information" varies depending on the competencies of the customers and their equipment, and increasing competencies make massive records a moving goal. "For some corporations, dealing with loads of gigabytes of data for the primary time may also trigger a need to reconsider facts control options. For others, it is able to take tens or loads of terabytes earlier than statistics length will become a sizeable attention**.**

DATA MINING

Data mining (the analysis step of the "Knowledge Discovery in Databases" procedure, or KDD), a area at the intersection of laptop technology and data, is the technique that attempts to discover knowledge in massive information units. It makes use of methods at the intersection of artificial intelligence, machine learning, facts, and systems. The usual purpose of the data mining is to extract important knowledge from a data set and transform it into an comprehensible structure for in addition, It includes database and information protection elements, records preprocessing, model and inference concerns, interestingness metrics, complexity issues, put up-processing of discovered structures, visualization, and on line updating.

TOP K SEARCH

Generally, data mining (every now and then referred to as facts or expertise discovery) is the technique of studying information from distinct perspectives and summarizing it into useful facts - information that can be used to growth sales, cuts charges, or both. It lets in customers to research data from many exceptional dimensions or angles, categorize it, and summarize the relationships recognized. Technically, data mining is the process of finding reciprocity or patterns among lots of fields in large relational databases. Given a set S with d dimensional objects top k dominating queries ranks these objects base on the number of objects in S dominated by o, and returns k objects that dominates maximum number of objects. The TKD query identifies the most frequently occurrence of objects, and is a powerful decision making tool used to rank objects in real life applications. Here an incomplete dataset is taken where some objects face the missing of attribute values in some dimensions, and study the problem of TKD query and processing over incomplete data. A TKD query on incomplete data returns k objects that dominates the maximum number of objects from a given incomplete data set. TKD queries on incomplete data share a few similarities with the skyline operator over incomplete information, because they each are based on the identical dominance definitions. Here mentioned that TKD queries on incomplete data have some benefits, i.e., its output is controllable by using a parameter k, and therefore, it's far invariable to the scale of incomplete dataset in specific dimensions. In addition to emphasize the dominance relationship definition on incomplete data, is absolutely meaningful. The Improved Bitmap Index Generation (IBIG) set of rules by employing the bitmap compression strategies and the binning techniques for improving the efficiency for area inside the TKD query over incomplete facts.

II. RELATED WORK

P. Shirisha, et al.[1] Proposed a Top-k query on multi dimensional data to share some similarities with the operator over multi dimensional data. Top-k queries on multi-dimensional data have some advantages, i.e its output is controllable by a parameter k, and hence, it is invariable to the scale of multi dimensional data set in different dimensions. The top-k dominating inquiry delivers k data objects which dominate the huge number of objects in a dataset. This question is an important tool for decision making because it affords data analysts an intuitive way for finding enormous objects. In addition, it merges the advantages of top-k and skyline queries without sharing their disadvantages: (i) the results size can be controlled, (ii) no ranking features want to be particular by way of users, and (iii) the result is impartial of the scales at specific dimensions. Despite their significance, top-k dominating queries have now not received good enough attention from the research. In this paper, we design Similarity Search Algorithm that apply on indexed multi-dimensional data and fully exploit the characteristics of the problem. Experiments on Movie Lens datasets demonstrate that our algorithms significantly overcome a past skyline-based approach, while our results on real datasets show the meaningfulness of top-k dominating queries.

Guohui Li, et al.[2] Implemented an Top-k dominating query assimilates the advantages of top-k and skyline queries and eliminates their limitations. It uses a ranking function to rank points just as in top-k query, and uses the dominating relationship as in skyline query. This query returns the k records with the higest domination scores from the dataset. The existing works explored the top-k dominating query in the certain databases. Uncertain data are inevitable in many applications due to various factors such as limitations of measuring equipments, delays in data updates and data randomness and incompleteness. Recently, query processing over uncertain data has attracted much attention. In many real application scenarios, the collected uncertain data are produced in a streaming fashion, such as financial data trackers, sensor networks, environmental surveillance, and location based services, etc. We first formally define the problem of continuous probabilistic top-k dominating (PTOPK) query processing over uncertain data streams based on a count-based sliding window model. Based on the observation that PTOPK does not change dramatically in consequent sliding window and most uncertain data objects not in PTOPK cannot be inserted in PTOPK in a certain time period, an efficient postponed examination algorithm (PEA) is implemented. With PEA, the scores calculation for some uncertain data objects not in PTOPK can be postponed and the computation cost can be saved. Extensive experiments have been conducted to demonstrate the efficiency of our approaches.

Christos Kalyvas, et al.[3] Implemented the living in the Information Age allows almost everyone have access to a large amount of information and options to choose from in order to fulfill their needs. In many instances, the quantity of information available and the rate of alternate may hide the highest quality and truly desired answer. This reveals the necessity of a mechanism so one can highlight the first-class alternatives to select amongst every possible scenario. Based on this the skyline query turned into proposed which is a decision aid mechanism, that retrieves the cost for- money options of a dataset by means of identifying the objects that gift the finest combination of the traits of the dataset. This paper surveys the state-of-the-art techniques for skyline query processing, the numerous variations of the initial algorithm that were The algorithm may require a large number of passes until the complete skyline is computed and eventually terminate as at the end of each pass the size of the temporary file will be decreased. BNL works well if the size of the resulted skyline is small and in best case fits into the window which will result in the termination of the algorithm in one iteration. BNL algorithm cannot compute skyline points progressively. Its performance is very sensitive to the number of dimensions and to the proposed to solve similar problems and the application-specific approaches that were developed to provide a solution efficiently in each case. Aditionally in each section a taxonomy is outlined along with the key aspects of each algorithm and its relation to previous studies. underlying data distribution. Especially, it is good for up to five dimensions for a uniform distribution but its performance degrades if the distribution tends towards an anti-correlated distribution.

Zhen Chen, et al.[4] Proposed an Network monitoring system has been the core function of network management, network fault diagnosis, and network security. In addition to actual-time firewalls and intrusion detection structures, traffic-archiving systems are critical to network forensic. An Internet Traffic- Archiving System (ITAS) captures packet or flow records for subsequent analysis and processing. Such systems have many important applications. The first commercial database product was published to implement a bitmap index. It uses a sequence of bits to identify and denote the presence or absence of an item in the indexed data. With bitmap indexing, logical operations, such as AND, OR, NOT, and XOR, can be used to answer complex queries. Bitmap index once was designed for scientific data and database, which is usually generated by scientific instruments or scientific simulation. Scientific data are extremely large and without further modification change. Bitmap index databases solve the problem on how to quickly identify a small amount of selected data in a mass of scientific data, while the traditional relational database is not suitable for this work. The technologies utilized in bitmap index databases are bitmap indexing, bitmap compression, and type. Efficient indexing of network packets or flow is central to traffic archiving systems. Indexing of traffic data has the following characteristic: (1) Large volume of data: The number of index messages is massive, even for brief periods. (2) High rates of incoming data: To keep up with the rate of packet influx, systems must be highly efficient. (3) Fixed data structure: The index information for each network packet has a fixed format with a fixed length. (4) Appending without modification: Network packet index information will increase only. Once the information is generated, it can’t be changed. (5) High redundancy: Data items on a same network are frequently repeated.

Xixian Han, et al.[5] Proposed a novel algorithm called TDEP is to utilize sorted lists built for each attribute with low cost to return top-k dominating result on massive data efficiently. In many applications including multi criteria decision making, top-k dominating query is a practically useful tool to return k tuples with the highest domination scores in a potentially huge data space. The present algorithms, either requiring indexes built at the specific characteristic subset, or incurring high I/O price or memory value, cannot manner top k dominating query on huge information efficiently. Through evaluation, it is found that TDEP can be divided into two stages: developing phase and shrinking phase. In every section, TDEP retrieves the looked after lists in round-robin style and maintains the applicants until the prevent circumstance is satisfied. The theoretical evaluation is supplied for the execution conduct in phases. An efficient approach is developed to compute the domination rankings of tuples with the obtained applicants handiest. Besides, TDEP adopts early pruning to reduce the quantity of candidate tuples maintained appreciably. The extensive experimental results, conducted on synthetic and real-life data sets, show the significant performance advantage of TDEP over the existing algorithms. The main differences of TDEP from DA lie that (D1) TDEP develops a novel method to compute domination score of a terminating tuple which only involves candidates obtained already rather than re-retrieving the sorted column files in disk; (D2) TDEP proposes early pruning to discard the unnecessary candidate tuples directly. Although TDEP performs in a similar way as DA to retrieve the sorted lists in round-robin fashion, it is obvious that TDEP will outperform DA (verified by theoretical analysis and the experimental result depicted later). Its reason is that by D1, TDEP retrieves fewer tuples than DA, which incurs a lower I/O cost, and by D2, TDEP maintains fewer candidate tuples in memory, which reduces memory cost significantly.

III. EXISTING METHODOLOGIES

Many databases in the world such as governmental and non-governmental contain missing values (MVs) in their attribute values. MVs is a value for attribute that was lost in the recording process. There are various reasons for their lost, such as manual data entry errors, equipment fail and incorrect measurements. The process of preparing clean data usually requires a preprocessing stage in which the data is prepared and cleaned, in order to be useful for the knowledge extraction process. The simplest way of dealing with MVs is to delete the record that contains them from the data set. However, this method is not practical when the data contains a large number of records with MVs which make bias during the inference. MVs make data analysis difficult. The occurrence of MVs can also lead to serious problems for researchers. In fact, unsuitable handling of the MVs in data analysis may found bias and can result in ambiguous conclusions being drawn from a research study, and can also limit the generalizability of the research findings.

A. Threshold Algorithm with no (or limited) Sorted/Random Access:

There are several applications (e.g., merging of information for multiple web services) where the standard access methods for top-k algorithms, i.e., sorted access and random access, are either unavailable, or are extremely expensive. Several important algorithmic variants have been developed for these problem variants, e.g., TA-SORTED, NRA, TA-ADAPT, STREAM-COMBINE and so on, and this portion of the seminar will be dedicated to discussing these variants.

B. Adaptive Algorithms:

Another set of algorithms that are assuming increasing importance are top-k algorithms that adapt the rate at which they consume data from each sorted stream. Algorithms such as QUICK COMBINE shall be discussed in this context.

C. Approximation Algorithms:

An important class of top-k algorithms is approximation algorithms. Such algorithms have the ability of stopping early and producing a set of results that are close to true top-k results. Measures of approximation will be discussed, e.g., the usual IR measures of precision and recall, as well as distance measures between ranked lists such as Kendall Tau and Spearman's footrule measures. Several variations of top-k algorithms that stop early yet approximate the true top-k results shall be discussed.

D. Probabilistic Algorithms:

Instead of deterministic approximation guarantees, an important class of top-k algorithms are probabilistic in nature, i.e., they stop early by aggressively pruning the set of top-k candidates, but in the process are able to give probabilistic guarantees for the objects that were rejected from consideration.

E. Algorithms using Views:

Like traditional query processing, top-k query processing can also benefit from the existence of materialized results of previously executed topk queries. There have been several efforts in this context, e.g., the PREFER algorithm (SIGMOD01) as well as the recent LPTA algorithm (linear programming based TA, VLDB06), and we shall briefly discuss these systems here.

IV. HOSPITAL RECOMMENDATION USING MAP REDUCE WITH TOP K FINDING

Propose a Medical Search Engine system that can automatically provide high quality knowledge triples extracted from the noisy question-answer pairs, and at the same time, estimate expertise for the doctors who give answers on these Q&A websites. The main component in the proposed MKE system belongs to the topic of truth discovery .Truth discovery methods can automatically estimate source reliability (doctor expertise) from the data without any supervision, and incorporate such estimated source reliability into the aggregation of noisy multi-source information. Information retrieval systems use various approaches to rank query answers. Users are more concerned about the most important i.e., top-k query answers in the potentially huge answer space. Different emerging applications demands for competent support for top-k queries. For example, in the context of the Web, the effectiveness and efficiency of meta search engines, which combine rankings from different search engines, are highly related to efficient rank aggregation methods. Similar applications are present in the perspective of information retrieval and data mining. Most of these applications compute queries that involve joining and aggregating multiple inputs to provide users with the top-k results. Proposed top-k query is to retrieve best answers from a potentially very large result set. The top-k queries searching require a system able to "rank" objects. In proposed Medical-based Personalized Recommendation System, top-k query is used to rank the doctor and hospitals records. Records are ranked according to both user and system ratings so that user can get accurate recommendations.

A. Bitmap Index Generation

The bitmap index creates separate columns for every cost in vi. Having vi gives a complete representation of present numbers in every dimension amongst all items. The bitmap index table would be initialized to 0 and then based on each item we would modify the values in the bitmap index table as described below.

- For every missing value, remove the fields in the corresponding row with none changes. So every vi with a missing value stays as all 0s.
- For every number we've in vi, insert number 1 inside the corresponding row, and all of the following right rows.

B. Map-Reduce Algorithm

Map-Reduce Algorithm uses the following three main steps:

1. Map Function
2. Shuffle Function
3. Reduce Function

1) Map Function

Map Function is step one in Map-Reduce Algorithm. It takes input Datasets and divides them into smaller sub-data. Then perform required computation on each sub-data in parallel.

This step plays the subsequent sub-steps:

1. Splitting
2. Mapping

- Splitting step takes input Data Set from Source and divide into smaller Sub-Data Sets.
- Mapping step takes those smaller Sub-Data Sets and perform required movement or computation on each Sub-Data Set.

2) Shuffle Function

It is the second step in Map-Reduce Algorithm. Shuffle Function is also known as "Combine Function".

It performs the following two sub-steps:

1. Merging
2. Sorting

It takes a listing of outputs coming from "Map Function" and performs those sub-steps on every and every key-cost pair.

- Merging step combines all key-values which have equal keys (this is grouping key-fee pairs by means of comparing "Key"). This step returns <Key, List<Value>>.
- Sorting step takes data from Merging step and type all key-values through use of Keys. This step additionally returns <Key, List<Value>> output however with taken care of key-price pairs.

3) Reduce Function

It takes list of <Key, List<Value>> sorted pairs from Shuffle Function and perform reduce operation as shown below.

In the above map reduce flow:

1. The data may be divided into n variety of chunks relying upon the amount of facts and processing capacity of individual unit.

2. Next, it's miles handed to the mapper functions. Please observe that everyone the chunks are processed simultaneously at the equal time, which embraces the parallel processing of statistics.

3. After that, shuffling happens which ends up in aggregation of comparable patterns.

4. Finally, reducers combine them all to get a consolidated output.

5. This algorithm embraces scalability as relying on the size of the enter statistics, we are able to hold growing the variety of the parallel processing gadgets.

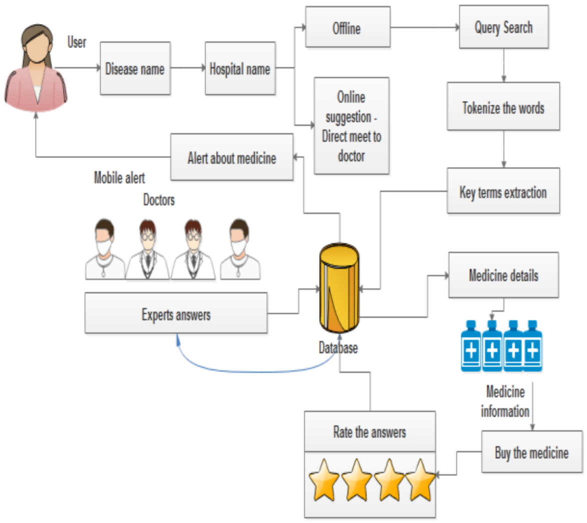

Fig 1: Architecture for Proposed Work

ACKNOWLEDGMENT

Top-k queries returns top components from a dataset and it is extremely useful in different real world applications. The top-k dominating inquiry delivers k data objects which dominate the highest number of objects in a dataset. This query is a crucial device for selection guide since it presents facts analysts an intuitive way for finding good sized objects. In order to further reduce the cost of score computation, here present BIG algorithm, which employs the upper bound score pruning, the bitmap pruning and fast bitwise operations based on the bitmap index to enhance the score computation and improve query overall performance as a consequence. Map-Reduce framework has two main functions namely Mapper and Reducer, which are used to separate huge tasks between multiple nodes to make them faster. Mapper results are aggregated to calculate the final result using the Reducer. The contribution from Upper Bound Based algorithms consists of scoring methods, borrowed by MRBIG and BIG algorithm with a scoring evaluation of the fields independently. Having a scoring method may be considered as persisted paintings on UBB algorithms. Scores help to retrieve desired top-k items in a significantly faster manner. It is worth mentioning that finding top-k dominance for incomplete data is a field that has not been fully addressed. With emerging big data, MRBIG can be utilized well to design recommender systems and provide an approximately real-time solution to TKD query processing. Here a web based application with dynamic data generation process is to implement with Top K query processing. Admin can update the data dynamically. Then the dynamic data collections are transferred to find the Top K result for user query. This will also implement using hospital related dataset and find the hospital with high recommendation value.

REFERENCES

[1] Min Chen, Jun Yang, Yixue Hao, Shiwen Mao, and Kai Hwang. "A 5G cognitive system for healthcare." Big Data and Cognitive Computing 1, no. 1 (2017): 2.

[2] Min Chen, Shiwen Mao, and Yunhao Liu. "Big data: A survey." Mobile networks and applications 19, no. 2 (2014): 171-209.

[3] Hossain, M. Shamim, and Ghulam Muhammad. "Healthcare big data voice pathology assessment framework." iEEE Access 4 (2016): 7806-7815.

[4] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. "Disease prediction by machine learning over big data from healthcare communities." Ieee Access 5 (2017): 8869-8879.

[5] Min Chen, Yujun Ma, Yong Li, Di Wu, Yin Zhang, and Chan-Hyun Youn. "Wearable 2.0: Enabling human-cloud integration in next generation healthcare systems." IEEE Communications Magazine 55, no. 1 (2017): 54-61.

[6] Yin Zhang, Meikang Qiu, Chun-Wei and Atif Alamri, "Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data" in IEEE Systems Journal · August 2015.