CS226 Literature Survey

Submitted By: Group #10

Gowtham Tumati (862186477)
Lovepreet Dhaliwal (862188799)
Sudip Bala (862188812)
Abhishek Ayachit (862188073)

1.Related Work

1.1 Goal

Big Data comes with a huge volume of data which is not always readily available. The flight information data that is historical is not easily available to the developers to analyze and process it. According to Samet Ayhan et. al, majority of the time is spent in collecting and cleaning the data[1]. Not just [1], but more such people have spent a lot of time collecting the raw data and then spent a great amount of labor to process such huge amount of data. One such research was done by Rajendra Akerkar. In this research, analytics on aviation data was the goal but to achieve that, the data required was historical which was the aviation data from 1980s[2].

1.2 Data Cleaning and Integration

The paper written by Nan Tang described different methods to clean the data. The methods provided included, Constraints based RDF cleaning, Master data, Interactive Cleaning. The Big RDF Data cleaning is used to clean the big data, which uses an existing database engines or the distributed systems. The main point of the paper is that it describes the different methods for cleaning the data and proposes the one that is mentioned above[3]. To integrate the data from multiple data sources were unified to create an uneven schema which can be converted into a global schema for all the data sources. To compile the data from multiple sources, big data integration techniques are discussed which can cope up with the huge volume of data coming at a great velocity and from a variety of sources. The paper also addressed the need of the Big Data Integration techniques over traditional record linkage techniques. The paper also addressed the handling of the incomplete and redundant data as the flight safety anomalies and erroneous data needs to be handled carefully. Another aspect of cleaning the flight data is to reduce the data generated from all the sensors. The main drawback about the paper is that it failed to mention the actual technical details about the techniques that can be used on the big data[4].

1.3 Search Techniques for Big Data

Unlike the traditional databases, the amount of time required to access the data in the big data is very high. Vatsal Jatakia et. al provides different techniques that can be used in order to query in the big data. These techniques include Exhaustive Search, Beacon-Guided Searching, Nearest Neighbour Method, Genetic Algorithm Implementation. Among all this, considering the advantages and disadvantages of all the searching techniques, the last one, i.e., Genetic Algorithm or GA shows the best results as it is scalable, flexible and it is robust and to the noisy evaluation functions. Another important point is that any part of the algorithm can be changed and customized[5].

Another technique that can be used to search data fast is to use indexing on the big data. Pooja Anand and Dr. Sandeep Maan in the paper [6] mentioned 2 indexing techniques that can be used with the big data. These techniques are, Generalized Inverted Index or GIN and Generalized Search Tree (GiST) can be used in order to create indexes on the big data. GiST is based on the B-tree and R-tree. It has the same implementation for indexing and retrieval as the R-tree for those based on R-tree and as B-tree for those based on B-tree. On the other hand, GIN uses custom fields as indexes. GIN is useful when for one key, many values need to be mapped in an index. When range comparison operations and equality is required Generalized Search Tree Indexes used. While both the techniques are good with full-text search implementation of the index, GIN are better for indexing array values and GiST are better for geometric data types. In conclusion, GIN is best for searching on the static data and GiST are better for searching on dynamic data. The paper gives details about these two techniques but failed to compare these techniques with other methods and compares only these two methods.

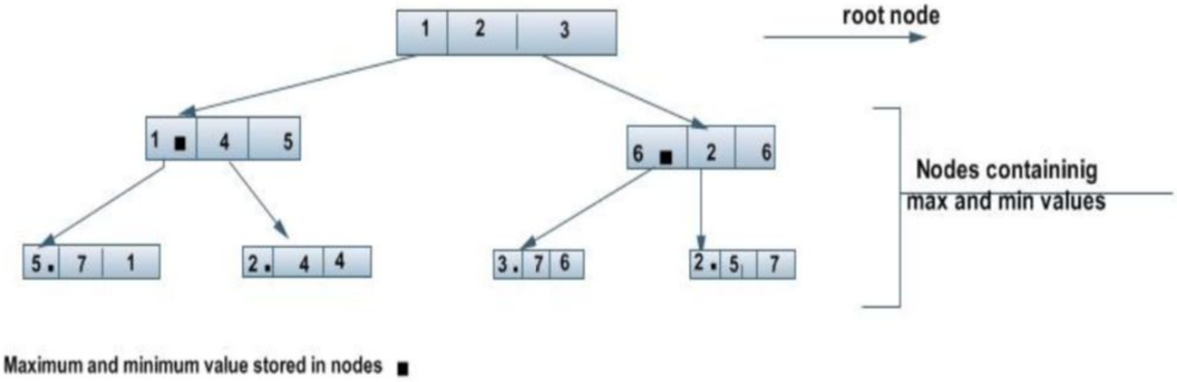

Fig 1: GiST[6]

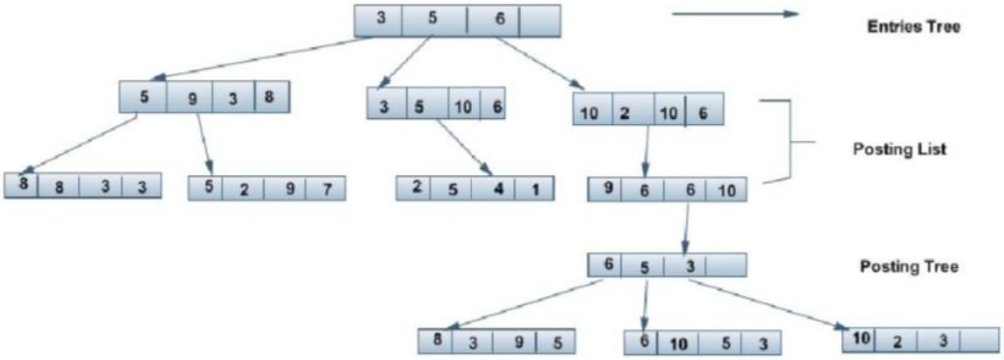

Fig 2: GIN

The paper [7] provides a different method for indexing to create a search engine on the big data. The indexes were created along with the Page Rank which can be used on category based search engine. A new way of Key-hash indexing is used to create hash-key folders by using total unique search term in the search engine. The advantage of this method is that when there are concurrent requests, the detch time of the result does not change. The method also works well with distributed architecture. The paper also evaluates the proposed method by evaluating the method.

1.4 Big Data Analytics and Data mining

After successfully integrating and cleaning the data, to get inferences from the big data, different data mining techniques need to be used. Different techniques can be used for different use cases. One such use case is mentioned in the [2]. The use case is to identify the patterns of the passengers in the aviation industry. To achieve this, the paper mentions Bayesian network as it is flexible with noise free as well as incomplete datasets. Bayesian Maps are used to in supplying the missing information which uses Bayes theory in order to find the certainty factors.

Another method is experimented and evaluated in [8]. It uses C5.0, CART, CHAID to build a decision tree, dividing datasets into subsets and using chi-square to identify the optimal split respectively. It divides the original dataset into sets, training and testing datasets. To evaluate the model, accuracy, recall and specificity is calculated.

1.5 Big Data Visualization

Data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. There are different ways of visualizing different kinds of data. A few of these type are discussed and compared with respect to visualizing the big data in [9]. The techniques used to visualize the data includes following: Line Chart, Bar Chart, Pie Chart, Table, Scatter Plot, Bubble Chart, Parallel Coordinates, Tree Map, etc. While comparing these when it comes to visualizing the big data, all the techniques are compared on the basis of Volume, Velocity and Variety. According to [9], all the techniques mentioned above except Line Chart and bar Chart are suitable for big volume of data. For high variety of data, all the techniques except Pie Chart and Tree Map are suitable. And finally, for the high velocity of the data, all the techniques except Tree Map are suitable.

References

[1] S. Ayhan, J. Pesce, P. Comitz, D. Sweet, S. Bliesner and G. Gerberick, "Predictive analytics with aviation big data," 2013 Integrated Communications, Navigation and Surveillance Conference (ICNS), Herndon, VA, 2013, pp. 1-13.

[2] Rajendra Akerkar , “ANALYTICS ON BIG AVIATION DATA: TURNING DATA INTO INSIGHTS”, International Journal of Computer Science and Applications, Vol. 11, No. 3, pp. 116 – 127, 2014

[3] N. Tang, "Big RDF data cleaning," *2015 31st IEEE International Conference on Data Engineering Workshops*, Seoul, 2015, pp. 77-79.

[4] Burmester G., Ma H., Steinmetz D., Hartmannn S. (2018) Big Data and Data Analytics in Aviation. In: Durak U., Becker J., Hartmann S., Voros N. (eds) Advances in Aeronautical Informatics. Springer, Cham

[5] V. Jatakia, S. Korlahalli and K. Deulkar, "A survey of different search techniques for big data," *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, 2017, pp. 1-4.
[6]Pooja Anand , Dr. Sandeep Maan, “A Study on Big Data with Indexing Technique for Searching and Retrieval of Data Fastly”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, Issue 1, January 2018
[7] N. Ragavan, "Efficient key hash indexing scheme with page rank for category based search engine big data," *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Srivilliputhur, 2017, pp. 1-6.
[8] A. Lukáčová, F. Babič and J. Paralič, "Building the prediction model from the aviation incident data," *2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, Herl'any, 2014, pp. 365-369.
[9] Ajibade, Samuel & Adediran, Anthonia. (2016). An Overview of Big Data Visualization Techniques in Data Mining. 4. 105-113.