**REDUCING SEARCH SPACE IN BIG DATA**

PRIYA SHARMA 1MS12CS144

SWEETY KUMARI 1MS12CS110

SURABHI KUMARI 1MS12IS110

SATHVE V G 1MS12IS117

***LITERATURE SURVEY***

**1. INTRODUCTION**

Many existing data mining algorithms search interesting patterns from transactional databases of precise data. However, there are situations in which data are uncertain. Items in each transaction of these probabilistic databases of uncertain data are usually associated with existential probabilities, which express the likelihood of these items to be present in the transaction. When compared with mining from precise data, the search space for mining from uncertain data is much larger due to the presence of the existential probabilities. This problem is worsened as we are moving to the era of Big data. Furthermore, in many real-life applications, users may be interested in a tiny portion of this large search space for Big data mining. Without providing opportunities for users to express the interesting patterns to be mined, many existing data mining algorithms return numerous patterns—out of which only some are interesting. This paper is about handling large dataset, which is not possible with existing traditional clustering algorithms. In this paper we have used SNN (SHARED NEAREST NEIGHBOUR) algorithm and FUZZY-C algorithm which is an innovative approach for clustering which is one of the most efficient clustering algorithms which can handle most of the issues related to clustering, like, it can generate clusters of different sizes, shapes and densities. This paper is about handling large dataset, which is not possible with existing traditional clustering algorithms. Our algorithm greatly reduces the search space for Big data mining of uncertain data, and returns only those patterns that are interesting to the users for Big data analytics**.**

**2. Literature Review**

Speed is the most important thing in today’s generation. Even milliseconds count and make a difference. That is why the paper we are talking about is to reduce search space to give appropriate precise results to customers. Technology.

The search paper on which we are doing project can have impact on society too .Eg:- By giving precise accurate information when searching for a particular blood group. It can save lives. So this paper plays very important role in today’s

The other papers referred were about reducing time in signals which is effective only in large scale so we didn’t implement it. The other paper was about traffic in social networks in which again speed is a primary concern but not in social impacts.

The other paper is about grid analysis to reduce volumes of data which again is useful when done on very large scale only. The other paper is about bandwidth in mobile networks .There are many techniques advancing every second for bandwidth in mobiles .So we should keep up in the changing trends which will change before implementing one generation

[1] **Traffic Signal Phase and Timing Estimation From Low-Frequency transit Bus**

By DataS. Alireza Fayazi, Ardalan Vahidi, Grant Mahler, and Andreas Winckler

Published by IEEE

Abstract:

It takes a sample of bus routes in San FrancisCo to study the time spent on traffic signals and predict the cycle of the traffic signals to save idling time , save fuel,lower emmisions and intersection collision avoidance. It uses GPS and next bus to track data and filter the once having lower velocity for better results.

[2] **Efficient Algorithms for Social Network Coverage and Reach**

By Deepak Puthal, Surya Nepal, Cecile Paris, Rajiv Ranjan and Jinjun Chen

Published by IEEE

Abstract:

To use vertex cover solution can help choose a set of nodes to cover the network which prolong the network life time and to use vertex reach problem can be applied to “spread of influence” problems in social networks.

[3] **Grid Analytics: How Much Data Do You Really Need?**

By Stanley E. McHann, Jr.

Published by IEEE

Abstract:

With Increase in Technology leads to increase in volume of data flowing into system which inturn will need to be stored and managed. One strategy is to collect all of the data possible and figure out what to do with the data at later date.

[4] **Performance Optimization of Big Data in Mobile Networks**

By Abhinandan Ramaprasath, Anand Srinivasan, Chung-Horng Lung

Published by IEEE

Abstract:

This paper is about bandwidth optimization algorithm based on cache coherency where the user data transfer is optimized without compromising the user expectation or the need for service providers to expand their capacity. Traditional caching algorithms store the timestamp of the file that is requested allowing the browser to check if the file was modified after. Our algorithm takes advantage of this by constructing a Differential file to avoid downloading a failed file again.

[5] **Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data**

By Carson Kai-Sang Leung Richard Kyle MacKinnon Fan Jiang

Published by IEEE.

Abstract:

Search space for users is always large and mostly contains unnecessary data irrevelant to their search. So this paper tries to narrow the search results by making it relevant.

Output only the uncertain data

Output data that satisfies the minimum threshold.

Output only those data which are valid and frequent.

Again these data are mapped and reduced using hadoop and big data technologies to get the final search space which is relevant to the user search.

The algorithm has ‘x’ objects and ‘t’ transactions. So each probability is calculated and added to form the final extrinsic probability. These objects are filtered based on constraints , uncertainity and frequent patterns and later fed into hadoop using map-reduce techniques to get final results.

[6] **Scalable Approximation of Kernel Fuzzy c-Means**

By Zijian Zhang,Timothy C. Havens

Published by IEEE

Abstract:

Fuzzy-c follows a clustering algorithm where all similar data points are grouped together .Each group is called a cluster. For N objects k cluster centers are selected randomly in beginning. All data points are assigned to nearest cluster center.

The output will be a partition matrix U ∈ {0, 1} N×k, a matrix of Nk values. Each element uik is the membership of vector xi in cluster k; and the partition matrix element uik = 1 if xi belongs to cluster k and is 0 otherwise.

First data is sampled using incremental clustering algorithm. The main idea of this algorithm is to take the cluster centers V t−1 ∈ HKt−1, and project them into HKt as meta-vectors. Let the partitions Ut and Uα denote fuzzy partition values of Xt and A, respectively. Let wXt and wα be the weights of Xt and A of the partition. Then finally they are added to get final result. This algorithm is done for one data chunk at a time .So its complexity is O(n\*n) where n is size of data chunk whereas Kernel clustering algorithm has complexity O(N\*N) where N is size of the chunk.

[7] **Introduction to kNN Classication**

Abstract:

This paper focus on using knn-algorithm.Its is a clustering algorithm where data is classified based on a constraint.

A random integer k is taken from the user. Then all the other points nearest to this selected point are classified into one group

Euclidian distance is used to measure and classify the points. Cross-validation is done by introducing random noise points and checking efficiency.

Can’t be used for all data types and too slow.

Condensed nearest neighbor is used to improve efficiency of knn.

Check each point to be the correct class and add it to a new database otherwise discard it.

Using KNN with k = 1 check if it’s an absorbing point or not.

Add to database if it is not absorbing point. This greatly increases efficiency.

[8] **An approach towards shared nearest neighbor clustering algorithm**

By Levent Ertöz, Michael Steinbach, Vipin Kum

By University of Minnesota

Abstract:

Jacard coefficient and cosine similarity is used instead of Eucilidian distance so that only common attributes are selected for shared nearest neighbor.

The Jaccard coefficient between two points is equal to the total number of attributes they are sharing divided by attributes in either vector.

The cosine similarity between two points is equal to the dot product of two vectors divided by the individual norms of two vectors.

First, a similarity matrix is constructed based on weight of the points, then the threshold is applied and some points are eliminated.

Let i and j be the two points. Then, the strength of link between i and j is now defined as: Strength (i, j) = ∑ (k+1-m)(k+1-n)

Noise points are eliminated by removing the link of weak points and keeping the strong ones.

Can be used for Earth Science data for sea level pressure.

**3. CONCLUSION**

This project basically discuss about different clustering and searching algorithms like Shared nearest neighbor and fuzzy-c and it describes how searching can be made more efficient using clustering in case of a large data sets where traditional searching and clustering algorithms fail.

**4. REFERENCES**

*[1] Traffic Signal Phase and Timing Estimation From Low-Frequency Transit Bus*

*By DataS. Alireza Fayazi, Ardalan Vahidi, Grant Mahler, and Andreas Winckler*

*Published by IEEE*

*[2] Efficient Algorithms for Social Network Coverage and Reach*

*By Deepak Puthal, Surya Nepal, Cecile Paris, Rajiv Ranjan and Jinjun Chen*

*Published by IEEE*

*[3]Grid Analytics: How Much Data Do You Really Need?*

*By Stanley E. McHann, Jr.*

*Published by IEEE*

*[4]Performance Optimization of Big Data in Mobile Networks*

*By Abhinandan Ramaprasath, Anand Srinivasan, Chung-Horng Lung*

*Published by IEEE*

*[5] Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data*

*By Carson Kai-Sang Leung\* Richard Kyle MacKinnon Fan Jiang*

*Published by IEEE.*

*[6]Scalable Approximation of Kernel Fuzzy c-Means  
 By Zijian Zhang,Timothy C. Havens*

*Published by IEEE*

*[7] Introduction to kNN Classication*

*[8] An Approach Towards The Shared Nearest Neighbor (SNN) Clustering Algorithm  
 By Levent Ertöz, Michael Steinbach, Vipin Kum, By University of Minnesota*