DOI: 10.21917/ijsc.2015.0146

EFFECTIVE SUMMARY FOR MASSIVE DATA SET

A. Radhika
Department of Computer Science and Engineering, B.S. Abdur Rahman University, India
E-mail: radhika.a@bsauniv.ac.in

Abstract

The research efforts attempt to investigate size of the data increasing interest in designing the effective algorithm for space and time reduction. Providing high-dimensional technique over large data set is difficult. However, Randomized techniques are used for analyzing the data set where the performance of the data from part of storage in networks needs to be collected and analyzed continuously. Previously collaborative filtering approach is used for finding the similar patterns based on the user ranking but the outcomes are not observed yet. Linear approach requires high running time and more space. To overcome this sketching technique is used to represent massive data sets. Sketching allows short fingerprints of the item sets of users which allow approximately computing similarity between sets of different users. The concept of sketching is to generate minimum subset of record that executes all the original records. Sketching performs two techniques dimensionality reduction which reduces rows or columns and data reduction. It is proved that sketching can be performed using Principal Component Analysis for finding index value.

Keywords:

Collaborative Filtering, Sketching Technique, Principal Component Analysis

1. INTRODUCTION

Big Data refers to datasets whose sizes are beyond the ability of typical database software tools to capture, store, manage and analyze. There is no explicit definition of how big a dataset should be in order to be considered Big Data. Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. In fields as diverse as pharmacology, finance, fraud detection, and intelligence analysis, better analysis and decision making can be facilitated by taking into consideration large amounts of heterogeneous data from many sources in many formats, and degrees of structure, and update rates. Pouring this data together often yields new insights and interesting cross-connections not readily apparent when considering the various data sets in isolation. Such "mash-ups" can provide the basis for operational decision making in complex and dynamic domains, support new forms of online collaboration, and help manage risks in complex markets. Without the ability to reach scale, potential Semantic Web adopters turn to cloud computing technologies such as map/reduce, not fully understanding the tradeoffs between the two technologies and, in particular, the limitations of map/reduce processing for handling graph structured or linked data.

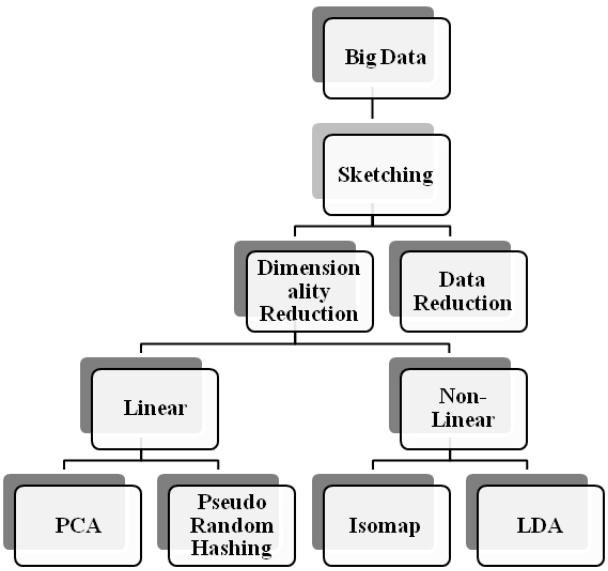

Fig.1. Architecture of Big Data

2. LITERATURE SURVEY

Bachrach and Porat [1] have proposed low complexity and low time complexity using collaborative filtering for finding the similar patterns between the users regarding the items consumed. To reduce the space complexity and time complexity, sketching methods are incompatible, forcing a choice between low running time or a small sketch size using Pseudo Random Hashing algorithm.

Joseph et al., [2] describe recommender systems uses collaborative filtering to measure the user experience. The experience is collected in the form of metrics which is based on rating from a large set of items. These collaborative filtering is expanded as content-based and knowledge-based approaches. The error is measured by the average error across in which it is frequently measured. Average error measures miss the features that are most important to user satisfaction in many applications of recommenders.

Hebrich and Yoram [3] have proposed similar rankings of items are analyzed by using the ranking correlation coefficient. Collaborative filtering method allows finding approximate ranking with high accuracy and confidence. One of the dis-advantages is maintaining real world collaborative system to handle large volume of datasets. These fingerprints are extremely short, much shorter than compression techniques allow, but only allow specific computations on the data. Empirical analysis of collaborative filtering is done by using min-wise independent families of hashes.

Y. Bachrach et al., [4] have proposed estimating the similarity among massive data sets is a central problem. To

avoid this problem Odd Sketch a compact binary sketch for estimating datasets is used. The Odd sketches provide space efficient estimator for sets which are having high similarity. The theoretical analysis of the quality estimator provides reliability to find the similarity. Odd sketching approach is used to estimate the Jaccard similarity of different data sets. Bloom filter is similar to binary sketch with one hash function .The Odd Sketch, a compact binary sketch for estimating the Jaccard similarity of two sets.

Konstan et al., [5] have proposed traditional collaborative filtering systems the amount of increases with the number of participants which produces high quality recommendations. New recommender systems can effectively produce high quality recommendations. For this Item-based techniques were used to identify the relationship among different users. Techniques such as item-item similarities and different techniques experimentally evaluate results and comparison is made by k-nearest neighbor approach. Item-based algorithm provides better performance and better quality.

Robert et al., [6] have stated predominant approach for collaborative filtering is neighborhood item sets, the algorithm used is k-nearest neighborhood. It detects the past user-item relationships based on collaborative filtering. The literature lacks a rigorous way to derive these weights. In this work they showed how the interpolation weights can be computed as a global solution to an optimization problem that precisely reflects their role. Data mining methods have been applied in various domains like cross language information retrieval.

Daniel et al., [7] proposed optimal algorithm is generated for finding the distinct elements in data sets. Finding the probability of datasets with setting the midstream for reducing both space and time complexities. F0 algorithm uses rough estimator to reduce the space bound by means of probability. Each values that algorithm maintains is being index hashed based on the least significant bit of its hashed values. The algorithm fails if it is used for larger datasets.

Xiaoyuan and Taghi [8] have stated collaborative filtering is one of the most successful recommender techniques. Shortcomings of memory-based CF algorithms include their dependence on user ratings, decreased performance when data are sparse, new users and items problems, and limited scalability for large datasets, and so forth. Clustering CF algorithms make recommendations within small clusters rather than the whole dataset, and achieve better scalability.

Pavan and Tirthapura [9] have proposed one of the most significant successes of research on data stream processing has been the efficient estimation of the frequency moments of a stream in one-pass using limited space and time per item. Each element in the stream is a single item and the algorithm needs to process this item efficiently, both with respect to time and space.

Chan and Kung [10] have proposed a hierarchical algorithm for image retrieval by sketch. The application scenario is that the user inputs a rough sketch depicting the prominent edges or contours of objects and wishes to retrieve database images that have similar shapes. Observation states that this hierarchical algorithm and another method described in complement each other. The fusion of the two methods is also an immediately focus of future work.

Liu and Guan [11] have proposed internet has become an essential part of the daily life for billions of users worldwide, using a large variety of network services and applications every day Principle Component Analysis (PCA) is the best-known spatial detection method for the coordinated low-profile traffic anomalies.

Yan et al., [12] have proposed Min-wise hash is a widely-used hashing method for scalable similarity search in terms of Jaccard similarity, while in practice it is necessary to compute many such hash functions for certain precision, leading to expensive computational cost yet it has a provably slightly smaller variance in estimating pair wise Jaccard similarity estimation of pair wise Jaccard similarity and more accurate results of approximate nearest neighbor search.

Feldman and Sohler [13] have proposed Euclidean distances from the n rows of an n×d matrix A to any compact set that is spanned by k vectors.

Gadepally and Kepner [14] have proposed the ability to collect and analyze large amounts of data is a growing problem within the scientific community. Dimensional Data Analysis (DDA) is a proposed technique that allows big data analysts to quickly understand the overall structure of a big dataset, determine anomalies.

Ranjit Jeba Thangaiah et al., [15] proposed method advocates an adaptive aggregation strategy using gain ratio and clustering methods of candidate features.

Shriram.R and Vijayan Sugumaran [16] has proposed a method for information retrieval for a query expressed in a native language. It uses heuristic method for categorization of documents in terms of relevance.

3. SKETCHING

The sketching is a technique which improves the effectiveness of the available dataset space and the quality of the results. In this project it is characterized the correct need of sketching and formulate the optimization problem for the given dataset. Most data mining algorithms have super-linear complexity. Deploying mining algorithms on very large data sets, most likely will result in terrible performance. Hence there is a need to reduce the dimensionality of the data in a systematic way, constructing short “data summaries” effectively reduce data volume.

The complete enumeration of the sketching and its algorithm are explained in the implementation chapter as below. The sketching is mainly categorized into two types as shown in the Fig.2.

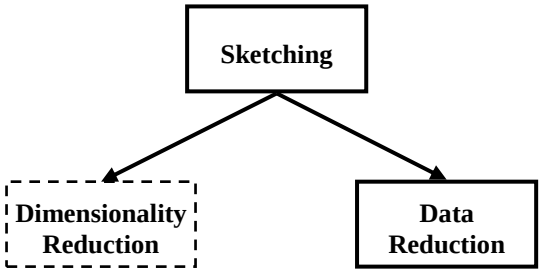

Fig.2. Sketching techniques

3.1 DIMENSIONALITY REDUCTION

The dimension reduction technique maps the distance between observations from original high dimensional space into a low level dimensional space. In the presence of many of features, select the most relevant subset of (weighted) combinations of features.

Feature Selection: $X_1, \ldots, X_m \rightarrow X_{k1}, \ldots, X_{kp}$

Dimensionality Reduction: $X_1, \ldots, X_m \rightarrow f_1(X_1, \ldots, X_m), \ldots f_p(X_1, \ldots, X_m)$

The above stated feature selection from set X to X_m and the dimensionality reduction possess the subset of function X as shown above. The dimensionality reduction focuses on exploratory data analysis and data visualization. Two approaches to reduce number of features.

- One of the features is to select the salient features by some criteria.
- Second feature is to obtain a reduced set of features by a transformation of all features (PCA).

These dimensionality reduction are of two types which is simplified by the below diagram as shown in the Fig.3.

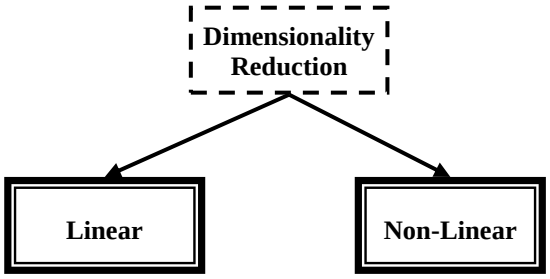

Fig.3. Dimensionality Reduction

3.1.1 Linear:

The linear method is a linear combination of features in which the data will be discovered for visualization. The linear method has many algorithm for dimensionality reduction namely Principal Component Analysis, Maximum Variance Subspace etc. This linear method have been detail explained in the below chapter. Linear approach PCA finds the subspace linear projections of input data. They compute the covariance matrix, compute its eigen vectors, finds the reduced size of the matrix and finally transformation takes place for the eigen vectors.

Similarly Pseudo Random Hashing takes the pair-wise distances and gives a mapping for the projection of seed value. It finds an embedding that preserves the inter-point distances, equivalent to MDS when those distances are Euclidean and equation (1) states the formula for finding centering matrix.

For example:

$$\text{Centering matrix}: P^e = I - \frac{1}{n} ee^T$$

P^eX: substract the row mean from each row

XP^e: substract the column mean from each column

$$X = \begin{pmatrix} 1 & 3 & -2 \\ 0 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix} \Rightarrow \text{row mean} = \begin{pmatrix} 1 & 2 & 0 \end{pmatrix}$$

$$\Rightarrow P^e X = \begin{pmatrix} 0 & 1 & -2 \\ -1 & 0 & 1 \\ 1 & -1 & 1 \end{pmatrix}$$

$$D = \left(\left\| x_i - x_j \right\|^2 \right)_{ij} : \text{distance matrix}$$

$$\Rightarrow P^e D P^e = -2\left((x_i - \mu) \bullet (x_j - \mu) \right)_{ij}$$

If Euclidean distance is used in constructing D, it is equivalent to PCA. The dimension in the embedded space is d, if the rank equals to d. If only the first p eigen values are important (in terms of magnitude), we can truncate the eigen-decomposition and keep the first p eigen values only.

3.1.2 Non-Linear:

Each data may not be summarized by linear combination of features. Many data sets contain essential nonlinear structures that invisible to MDS. MDS preserves all inter-point distances and may fail to capture inherent local geometric structure. Resorts to some nonlinear dimensionality reduction approaches. Kernel methods, Depend on the kernels, most kernels are not data dependent, Manifold learning, Data dependent kernels.

4. EXISTING SYSTEM

General similarity based techniques summarized the datasets, but also limit the functionality of the big data because fewer operations are supported over the similarity techniques. Massive data requires efficient algorithms to maintain the time and storage. The Naïve approach is a commonly used summarization. It has some draw backs such as high running time and it requires more space to store the data. Naïve approach collects the entire dataset of different records and finds the similarity between the records. If the record need to be updated each and every time it is not possible as it may need to handle billions of records and represent knowledge about the user's similarity pattern. To find the approximate similarity pattern pseudo random hashes have been used as it computes similarity measures between any two users.

4.1 DISADVANTAGES OF EXISTING SYSTEM

There are two major problems in the above Big Data summarization. First, it is difficult to manage high dimensional data, storing the data becomes complex process. Second, the different sketching techniques fail over for simple analytics querying operation. Finally, besides data storing and retrieving for streaming datasets cannot perform querying in simple analytics.

5. PROPOSED SYSTEM

The proposed technique is by combining two methodologies Principal component analysis and Pseudo-random hashing. The Principal component analysis is used for data reduction,

summarization of data with many variables by a smaller set of derived variables. It takes a data matrix of *n* objects by *p* variables, which may be correlated, and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original *p* variables. The first *k* components display as much as possible of the variation among objects. The algorithm for dimensionality reduction objects are represented as a cloud of *n* points in a multidimensional space with an axis for each of the *p* variables; the centroid of the points is defined by the mean of each variable.

Algorithm for Dimensionality reduction

BEGIN

Step 1: *Import the dataset*

Step 2: *Read the file by calling it as object name*

Step 3: *Bind the variables in the dataset*

Step 4: *Find the summary and correlation of the variable for the object created*

Step 5: *Plot the values obtained*

END

Dimension reduction is done by two ways:

I. Random projection

II. Sketching

5.1 RANDOM PROJECTION

High dimensional data mostly used Random projection for clustering. Random projection has been shown to have promising theoretical properties. In practice, however, results in highly unstable in clustering performance. Empirical results shown in random projection that the proposed approach achieves better. To gain insights into the performance improvement obtained by our ensemble method, we analyze and identify the influence of the quality and the diversity of the individual clustering solutions for performance.

For example:

l = 7, k = 4 (projection size)

Choose projection (1, 2, 5, 7)

Input Sequence: TAGACATCCGATT

Output: Bu

5.2 SKETCHING

The concept of sketching is to generate minimum subset of record that executes all the original records. Sketching performs two techniques dimensionality reduction which reduces rows or columns and data reduction. Evaluations of substantial improvements in the utilization of the available sketching space and the quality of the resulting approximation error guarantees. Dimensionality reduction provide necessary and sufficient conditions for multi-query sketch sharing that guarantee the correctness of the resulting sketch-based estimators. The difficult optimization problem of determining sketch-sharing configurations that are optimal (e.g., under a certain error metric for a given amount of space). Optimal sketch sharing typically gives rise to NP-hard questions and novel heuristic algorithms for finding effective sketch-sharing configurations in practice. More concretely, the key contributions of the work can be summarized.

States the algorithm for Jaccard similarity estimation in set:

BEGIN

Step 1: *J is a join set, with available memory M*

Step 2: *Find the union operation among the join set*

Step 3: *Find the connected attributes vector value*

Step 4: *Select the edges with more median value*

Step 5: *Compute the memory*

END

Algorithm

Input: S: an adjacency matrix representing a collection of items.

Input: k: a natural number for similarity estimation.

Output: S_k: The correlation factor of the items.

INPUT : set S_1 and S_2

1. Let S_1 and S_2 be two sets and size in bits n
2. Merge S_1 and S_2 to obtain $S_1 \cup S_2$
3. For each x € S_1 and x € S_2 we know everything on its membership in S_1 and S_2
4. Compute $S_1 \cap S_2$ ¥ x € S_1, S_2
5. **Function** union (S1,S2)
6. Complete variance S= $S_1 \cup S_2$/ $S_1 \cap S_2$

Emit(S), where S is the reduced set.

The input is considered as the complete set of items S and the subsets of the items S_1 and S_2. Obtaining the similar item set using the UNION and INTERSECTION operation for the item set S shown in the Jaccard similarity identifies similarity between datasets S_1 and S_2 and does grouping to obtain the reduced set.

For example:

Consider three sets A = {1,2,3,4} B = {2,3,5,7} and C = {2,4,6}. The Jaccard similarity is found by dividing intersection of the sets and the union of the sets. In the above example the only element present in all the sets is 2. The union of the sets is the total number of elements present in all the three sets which is 11. Thus the Jaccard similarity for the above example is 1/11.

6. METHODOLOGY

Principal components analysis (PCA) is a very popular technique for dimensionality reduction. Given a set of data on *n* dimensions, PCA aims to a linear subspace of dimension *d* lower than *n* such that the data points lie mainly on this linear subspace. Such a reduced subspace attempts to maintain most of the variability of the data. The linear subspace can be specified by d orthogonal vectors that form a new coordinate system, called the 'principal components'. The principal components are orthogonal, linear transformations of the original data points, so there can be no more than *n* of them.

However, the hope is that only $d < n$ principal components are needed to approximate the space spanned by the *n* original axes. The most common definition of PCA, due to for a given set of data vectors x_i, $i \in 1 \ldots t$, the *d* principal axes are those

orthonormal axes onto which the variance retained under projection is maximal. In order to capture as much of the variability as possible, let us choose the principal component, denoted by U_1, to have maximum variance. Suppose that all centered observations are stacked into the columns of an n x t matrix X, where each column corresponds to an n-dimensional observation and there are t observations. Let the principal component be a linear combination of X defined by coefficients (or weights)$w = [w_1 \ldots w_n]$.

In matrix form:

$$U_1 = w^T X$$

$$\text{var}(U_1) = \text{var}(w^T X) = w^T S w$$

where, S is the n x n sample covariance matrix of X. Clearly var(U_1) can be made arbitrarily large by increasing the magnitude of w. Therefore, we choose w to maximize $w^T S w$ while constraining w to have unit length.

$$\max w^T S w$$

$$\text{subject to } w^T w = 1.$$

Objective of PCA is to rigidly rotate the axes of this p-dimensional space to new positions (principal axes) that have ordered such that principal axis 1 has the highest variance, axis 2 has the next highest variance and axis p has the lowest variance. Covariance among each pair of the principal axes is zero. PCA uses Euclidean Distance calculated from the p variables as the measure of dissimilarity among the n objects. PCA derives the best possible k dimensional ($k < p$) representation of the Euclidean distances among objects.

In a first step the transform matrix and the measurement ensemble (in the algorithm called compression matrix) are generated given the data and the type of measurement ensemble. The, for the compression matrix, relevant information is the type of measurement ensemble how do we want to measure Together they form the compression matrix by describing how many columns to select (randomly) and what these should look like (from what type of matrix to select them from).

In forward selection, we start with zero attributes and then start to pick the attributes with the highest statistical significance. After picking the first attribute, we next select the best second attribute and find the one conferring the most significant improvement in the cross-validation check. The set of attributes will be grown until significant improvements found. One issue of the "forward selection" approach is that it may miss "grouped features". The attributes are ranked according to the variation in the data that they explain. PCA is designed to model linear variability's in high-dimensional data. However, many high dimensional data sets have a nonlinear nature. In these cases the high-dimensional data lay on or near a nonlinear manifold (not a linear subspace) and therefore PCA cannot model the variability of the data correctly. One of the algorithm as shown below is designed to address the problem of nonlinear dimensionality reduction is Kernel PCA. In Kernel PCA, through the use of kernels, principle components can be computed efficiently in high-dimensional feature spaces that are related to the input space by some nonlinear mapping.

Algorithms for Principal Component Analysis

Input X: an adjacency matrix representing a collection of items.

Input: u: a natural number

Output: X_k, Y_k: vector of hub and authority scores for each tuples.

BEGIN

Step 1: *Recover basis: Calculate XX>= Pt i= Ψ> i and let U = eigenvectors of XX>corresponding to the top d Eigen values.*

Step 2: *Encode training data: Y = U>X where Y is a d £ t matrix of encodings of the original data.*

Step 3: *Reconstruct training data: X = UY = UU>X.*

Step 4: *Encode test example: y = U>x where y is a d-dimensional encoding of x.*

Step 5: *Reconstruct test example: x = Uy= UU>x.*

END

For example consider the below equation for estimating eigen value.

$$\Psi(q_j) = \lambda_j$$

$\lambda_j - j^{\text{th}}$ Eigen value where, j = 1,2,…….,m used as the variance probe. We have to arrange the Eigen value in descending order where λ_1 is the highest value and λ_m is the lowest value.

Defining a set of elements,

$$a = \{a_1, a_2, a_3..., a_m\}^T = \{xq_1, xq_2, xq_3,\ldots,xq_m\}^T$$

$$a = Q^T x$$

The above equation represents the orthogonal matrix of elements with Q transpose.

$$\sum_{j=1}^{m} a_j q_j$$

The above equation represents the synthesis q_j's are the basic vector for synthesis.

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.

The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. From a set of N correlated descriptors, we can derive a set of N uncorrelated descriptors (the principal components). Each principal component (PC) is a suitable linear combination of all the original descriptors. PCA reduces the information dimensionality that is often needed from the vast arrays of data in a way so that there is minimal loss of information.

Mathematically, PCA relies on the fact that most of the descriptors are interrelated and these correlations in some instances are high. It results in a rotation of the coordinate system in such a way that the axes show a maximum of variation (covariance) along their directions. The data manipulation involves decomposition of the data matrix X into two matrices a and q. The two matrices a and q are orthogonal. The matrix a is usually called the loadings matrix, and the matrix q is called the scores matrix.

The eigenvectors of the covariance matrix constitute the principal components. The corresponding eigen values give a hint to how much "information" is contained in the individual components. The loadings can be understood as the weights for

each original variable when calculating the principal component. The matrix contains the original data in a rotated coordinate system.

The mathematical analysis involves finding these new "data" matrices *a* and *q*. The dimensions of a (i.e. its rank) that captures all the information of the entire data set of *A* (i.e. number of variables) is far less than that of *X* (ideally 2 or 3). One now compresses the N dimensional plot of the data matrix *X* into 2 or 3 dimensional plot of *a* and *q*.

The first principal component accounts for the maximum variance (eigen value) in the original dataset. The second, third (and higher order) principal components are orthogonal (uncorrelated) to the first and accounts for most of the remaining variance. A new row space is constructed in which to plot the data, where the axes represent the weighted linear combinations of the variables affecting the data. Each of these linear combinations is independent of each other and hence orthogonal.

The data when plotted in this new space is essentially a correlation plot, where the position of each data point not only captures all the influences of the variables on that data but also its relative influence compared to the other data.

6.1 THE R ENVIRONMENT

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Among other things it has an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either directly at the computer or on hardcopy, and a well developed, simple and effective programming language (called '*S*') which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed most of the system supplied functions are themselves written in the S language.) The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software. *R* is very much a vehicle for newly developing methods of interactive data analysis.

7. TABLE COMPARISONS

Table.1. The Dataset size of 1 GB and Execution time

Dataset size in MB	Execution time of PCA (seconds)	Execution time of Pseudo Random (seconds)
100	28	33
200	36	51
300	53	67
500	61	77
500	77	98
600	84	121
700	96	132
800	109	143
900	120	169
1024	138	180

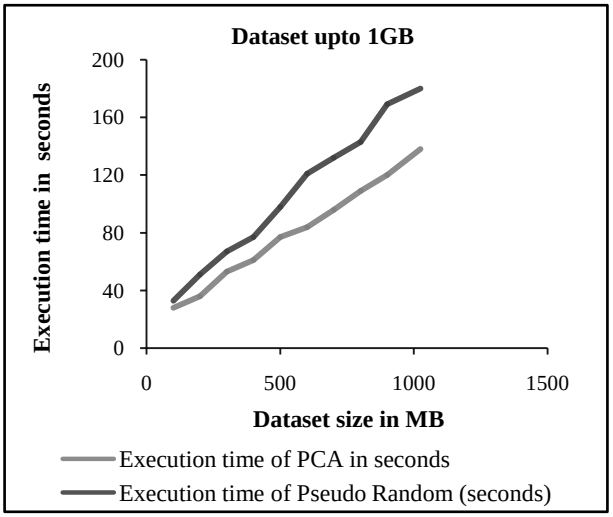

Fig.1. Dataset size of 1 GB and Execution time

There is increase in dataset the execution time increases, when compared with two algorithms the execution time significantly vary for each second which is as shown in the above figures. The graphical representation of the results illustrates how the execution increases with respect to time.

7.1 METRICS:

The metrics used to plot the graph is size of the dataset and the execution time. The efficiency is calculated using the formula:

$$\text{Efficiency} = \frac{\text{Size of the dataset in MB}}{\text{Execution time in seconds}}$$

Metrics for item similarity:

$$\text{Distance Similarity} = (\text{No. of items in the set1}) \cup \left(\frac{(\text{No. of items in set2})}{\text{No. of items in the dataset}} \right)$$

8. CONCLUSION

The instant method for sketching massive datasets, based on principal component analysis. It is focused on principal component analysis and examined the dimensionality reduction in detail, the same technique can be used for recommender systems on minimal value elements under several methods. This approach is thus a general technique for speeding up computations by summarizing the datasets.

REFERENCES

[1] Yoram Bachrach and Ely Porat, "Sketching for Big Data Recommender Systems Using Fast Pseudo-random Fingerprints", *Proceedings of the 40th International Conference on Automata, Languages, and Programming – Volume Part II*, pp. 459-471, 2013.

[2] Joseph A. Konstan and John Riedl, "Recommender systems from algorithms to user experience", *User Modeling and User-Adapted Interaction*, Vol. 22, No. 1-2, pp. 101-123, 2012.

[3] Yoram Bachrach and Ralf Hebrich, "Fingerprinting Ratings for Collaborative Filtering: Theoretical and Empirical Analysis", *Proceedings of the 17th International Conference on String Processing and Information Retrieval*, pp. 25-36, 2010.

[4] Yoram Bachrach and Ralf Herbrich, "Fingerprinting Ratings for Collaborative Filtering-Theoretical and Empirical Analysis". *String Processing and Information Retrieval,Vol. 6393*, pp. 25-36, 2010.

[5] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Item Based Collaborative Filtering Recommendation Algorithms", *Proceedings of the 10th International Conference on World Wide Web*, pp. 285-295,2001.

[6] Robert M. Bell and Yehuda Koren, "Improved Neighborhood based Collaborative Filtering", *Proceedings of KDD Cup and Workshop*, pp. 7-14, 2007.

[7] Daniel M. Kane, Jelani Nelson and David P. Woodruff, "An Optimal Algorithm for the Distinct Elements Problem", *Proceedings of the 29th ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database systems*, pp. 41-52, 2010.

[8] Xiaoyuan Su and Taghi M. Khoshgoftaar, "A survey of Collaborative Filtering Techniques", *Advances in Artificial Intelligence*, Vol. 2009, pp. 1-19, 2009.

[9] A. Pavan and S. Tirthapura, "Range-Efficient Counting Of Distinct Elements In A Massive Data Stream", *SIAM Journal on Computing*, Vol. 37, No. 2, pp. 359-379, 2007.

[10] Y. Chan and Y. Kung, "A Hierarchical Algorithm for Image Retrieval by Sketch", *IEEE First Workshop on Multimedia Signal Processing*, pp. 564-569, 1997.

[11] Y. Liu, L. Zhang and Y. Guan, "Sketch-based Streaming PCA Algorithm for Network-wide Traffic Anomaly Detection", *Proceedings of IEEE 30th International Conference on Distributed Computing Systems*, pp. 807-816, 2010.

[12] Jianqiu Ji, Jianmin Li, Shuicheng Yan, Qi Tian and Bo Zhang, "Min-Max Hash for Jaccard Similarity", *Proceedings of IEEE 13th International Conference on Data Mining*, pp. 301-309, 2013.

[13] D. Feldman, M. Schmidt and C. Sohler, "Turning Big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering", *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1434-1453, 2013.

[14] V. Gadepally and J. Kepner, "Big Data Dimensional Analysis", *Proceedings of IEEE High Performance Extreme Computing Conference*, pp. 1-6, 2014.

[15] P. Ranjit Jeba Thangaiah, R. Shriram and K. Vivekanandan, "Adaptive hybrid methods for Feature selection based on Aggregation of Information gain and Clustering methods", *International Journal of Computer Science and Network Security*, Vol. 9, No. 2, pp. 164-169, 2009.

[16] R. Shriram and Vijayan Sugumaran, "Cross Lingual Information Retrieval Using Data Mining Methods", *Proceedings of Americas Conference on Information Systems*, 2009.