Multimedia Big Data clustering challenges for Novel Metaknowledge-based Processing Technique

Miss. Sima Sangle [1], Prof. Gajanan Chakote [2]
[1, 2] Department of Computer Science
[1] ME student, Mastyodari Shikshan Santha's College of Engineering and Technology
[2]Assistant Professor, Mastyodari Shikshan Santha's College of Engineering and Technology

Abstract- *Previous investigation has challenged us with the task of exibiting relational models between text-based data and then clustering for predictive survey using Golay Code technique. We focus on a novel approach in multimedia datasets to withdraw metaknowledge. Our alliance has been an on-going task of studying the relational patterns based on metafeatures extracted from metaknowledge in multimedia datasets between datapoints. Those choosed are notable to outfit the mining technique we appeal, Golay Code algorithm. In this research paper we condence findings in optimization of structured and unstructured multimedia data in order to be processed in 23-bit Golay Code for cluster recognition of metaknowledge representation for 23-bit representation* .

Keywords- Big Multimedia Data Processing and Analytics; Information Retrieval Challenges; Content Identification, Meta- feature Extraction and Selection; Metalearning System; 23-Bit Meta-knowledge template; Enowledge Discovery, Golay Code.

I. INTRODUCTION

The latest centre of attentions has been centered on content withdraw and finding knowledge of Big Data Analytics [1][4][5]. We have complete investigation for Big Data clustering using mining and cluster techniques to challenge the occurrence of Big Data evolving the 23-bit metaknowledge template. Our ongoing analysis centred on meaning of datacube - is an assign of the data that is typical of data - each dimension of the cube - which can exceed 2 and even 3 dimensions. This can theoretically permit us to intersect particular datapoint as a outcome of an deliberate analysis. The multi-dimensional platform for convergence towards predictive analysis given us by 23-question Golay code templatess.

II. LITIURATURE SURVEY

The one of the main characteristic of big data is to execute computation on data present in GB and PB (petabyte) and even on exa-byte (EB) with the computational process. The dissimilar sources heterogeneous, vast and data having different characteristics of data content in big data. So system make used of parallel computing, it's a correspondent programming support and software to capably examine and mine the entire data in various format are the target focus of big data process to transform in quantity to quality. Map Reducer is batch orientated parallel processing of data. There are some short come and performance gap with relational data base. To the performance and increase the nature of large data Map Reducer has used data mining algorithm and machine learning. Currently processing of big data relay on parallel computing technique like Map Reducer supply cloud computing as a good platform big data for community as service. The mining algorithm used in this are , including locally weighted linear regression, k-Means, linear support vector machines, logistic regression, Gaussian discriminant analysis, the individualistic variable analysis, expectation maximization, naive Bayes, and back-propagation neural networks .

Data mining algorithm obtain the optimizes result it perform computing on large data. By increasing performance and appropriate algorithm are process in parallel programming which is applied to number of machine learning algorithm which is based on Map Reducer frame work .With the machine learning we can state that the process can be change to summation operation. Summation operation can be perform on subset of data separately and accomplish simply on Map Reducer programming. Reducer node collect all the processed data and collect into summation. Ranger et al. Proposed application of Map Reducer to hold up parallel programming and multiprocessor system which include three different data mining algorithm K-means ,linear regression principal component analysis. In paper [3] the Map Reducer mechanism in Hadoop execute the algorithm in single-pass, query based and iterative frame work of Map Reducer, distributing the data between number of nodes in parallel processing algorithm that the Map Reducer approach for large data mining by examining standard data mining task on mid size clusters. Polarimetries and sun[4]. In this they proposed a mutual distributed aggregation (DisCo) frame work for pre-processing of practically and collaborative technique. The performance in Hadoop it is and open source Map Reducer project show that DisCo have ideal which is accurate and can analyse and process enormous data.

Therefore the large data set are can be divided into small subset and that subset can be assign to various number of machine in Mapper the data is process by the mapper it perform operation on it. To took up the poor analysed capabilities and the week analysed software which are traditional Hadoop system. In detail integration which give the data for the computation in parallel processing model that make use of full Hadoop. It is beyond their limits for processing it. Increase of big data application has increase in the areas where the data is generated more and more which can't be handled by the normal software. The most important challenge in Hadoop is to process the Big Data and to get the valuable information from that large data sets. An extraction of information about diseases and symptoms from hadoop data sets using data mining. There valuable data obtained can be used for the future measure.

III. CHALLENGES OF BIG MULTIMEDIA DATA

Data retrieval from establishment for the justification of mining and analysis can show arduous and is one of the demanding feature of multimedia data clustering. The writer of the Unified Framework for Representation, Analysis of Multimedia Content for Correlation and Prediction [3], Paul and Singh, highlighted some common provocation, which demonstrate accurate when procceding towards any survey binary representation of 0/l

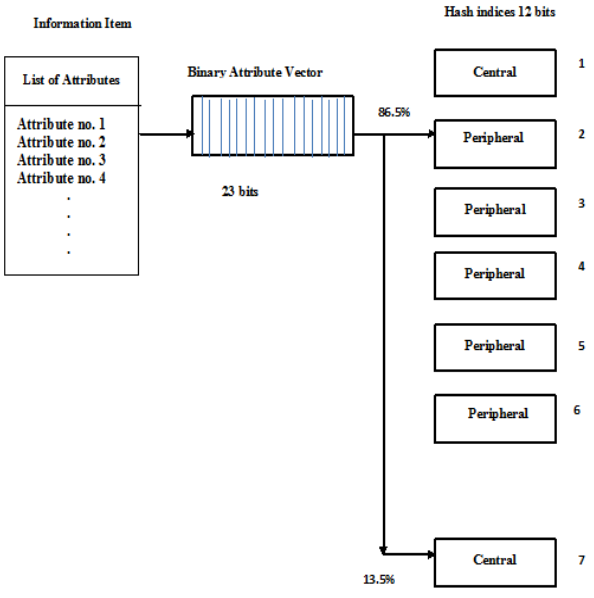

Figure 1: Golay Code 23 bit processing

The outcome of the Golay code index lookup is represented by allocating a cluster label to the media record .

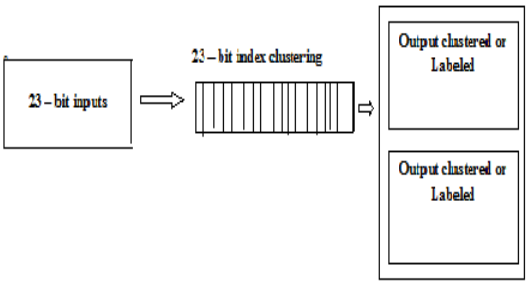

Figure 2: Golay Code 23 bit index lookup and label resolution of clusters.

IV. METAKNOWLEDGE REPRESENTATION AND STRUCTURED DATA

We put in our methodology to refine the removal of meta knowledge depiction from structured data, prior to appealing the alike to unstructured multimedia data. The classification algorithms we used are summarized in the table below, Table l.

Decision trees were very successful in the validation of choosing meta knowledge ascribe used to represent the bits of 23-bit record. DMRT was useful to derive threshold in meta-knowledge single attribute representations.

Num.	Meta-knoqledge and Meta feature discovery algorithms	
	Algorithm	Benefits
1	Scatter Plots (Correlation Coefficient)	> A quick if the "average" pattern is linear, curved or random. > If the trend is positive or negative association. Strenth of relationship Identification of group of outlier (X , Y)
2	Statistical Relationship	> Statiatical relationbship in variation of possible values of X and X. > Regression equation to describe the "best" line through the data and to predict Y based on X. >If the linear relationship anticipated then describe the strength and direction.
3	Decision Tree	> A fast learning curve, easy to test model and effective way to find " Terminal" node. > Purposeful test of quasi model by over-fitting to evaluate by adding/removing attributes. > Assistance in deriving threshold driven question to qualify answer Yes/No [1 , 0]
4	DMRT	>Descriminative classifier making validation of threshold setting in questions.
5	GLM	>A determination of attributes(questions) impact and contribution to binary outcome.

Table 1: Applied Algorithm for meta-knoqledge feature discovery used in R

The series of algorithms appeal give advice to decide properness and strength of metaknowledge . It also assists with declaration about values (thresholds of boundaries) of meta-feature template questions. These questions (as bits l–23) are to be used as a base to show the removed metaknowledge in binary 23-bit word on the input of Golay Code processing.

The key benefit of this proposal is to implant in processing a greedy loopback to diminish the mistake of specific algorithms. For exemplar, specific GLM, then the AUC (value minimization) is the target of deriving subset of metaknowledge attributes that process can be embedded into greedy algorithm processing.

V. METAKNOWLEDGE REPRESENTATION SEMANTIC ONTOLOGY (UNSTRUCTURED DATA)

Appealing the same methodology on unstructured data of multimedia file is much small successful as no structured data ascribe with particulars degrees of self-determination are depicted or uncomplicatly removed to erects data cubes. Therefore, naturally, we originate a operation to obtain such meta knowledge use semantic knowledge, so that multimedia can be classify. In order to do that, a more general classification is done, i.e. not based on particular data merit depiction. It is based on the comprehension hold in the media file represented semantically. This revolve out to be a additional effectual and correct method to be functional in Golay Code processing when applied to multimedia file, and consequently appliedto Big Data.In order to test a result to represent metakn owledge using semanticcharacterization of source files (hmtl, ms word, .pdf, jpeg etc.), we created a sample collection base d on files obtained from [8], which is focused on the financia l industry. It was easier to utilize the financial industry definition in terms of semantic structure and phrase.First the semantic and generic definition was derived.T herefore,each file identified as having a semantic element match in its content or notin order to construct the datacube of metaknowledge.

Such semantic element presence is then scored as 1 for present and ofor not present

and consequently processedwith Golay Code. Theorder of attri butes within the data

cube corresponds to theorder of the most generic semantic ele ments to be placed first on the23 bit record,

followedby the most

important ones to distinguish the records in clusters. In this cas e, wemean the first corresponds to the lowest bits of the 23bit i nput Golay Code record.

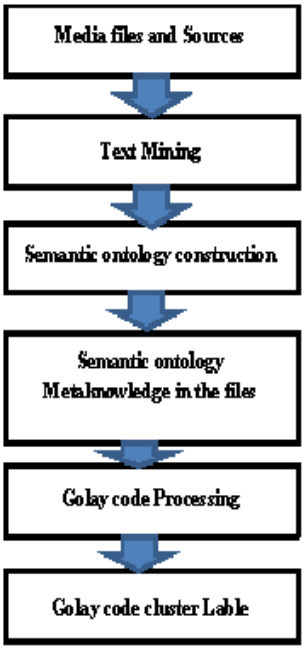

Figure 3: Derivation of Semantic Meta-knowledge per media files

The semantic metaknowledge is derived and extracted from medie files and place on the 23-bit record of Golay code algorithm.

VI. CONCLUSION

Big data is collection of complex data sets, An Data mining and seclusion protection framework for big data has been proposed. Data mining permit to traverse main knowledge and privacy protection allow to provide the anonymous data to the user. The framework is union of accessing mined data and Privacy preservation mechanism. System processes all the data gathered from various sources. Through this system we get expected information when the user enters the disease name or disease symptoms. All the data related to application users query accordingly is provided to the user real-time. User enters the keyword to the system and system provide the related information regarding to the keyword.

ACKNOWLEDGMENT

We would like to take this opportunity to express my sincere gratitude to my Project Guide Prof. Chakote Gajanan for his encouragement, guidance, and insight throughout the research and in the preparation of this dissertation. He truly exemplifies the merit of technical excellence and academic wisdom.

REFERENCES

[1] Zaiane , Osmar R., et al.”Mining Multimedia data”. Proceedings of the 1998 conference of the Centre for Advanced Studies on Collaborative research. IBM Press, 1998

[2] Paul, S. Nissi, and Y. Jayanta Singh. “Unified Framework for representation, analysis of multimedia content for correlation and prediction.” Emerging Trends and Applications in Computer Science (IUCETACS) 2013 !st International Conference on. IEEE, 2013.

[3] Huang, Tiejun , Yonghong Tian ,Wen Gao, and Jian Lu. “Mediaprinting: Identifying multimedia content for digital rights management.” (2007): 1-1..

[4] Chen, Chao, and Mei- Ling Shyu. ”Clustering - based binary - class classification for imbalanced data sets.” In Information Reuse and Integration (IRI) , 2011 IEEE International Conference on , pp. 384-389. IEEE , 2011.

[5] Kamran Kowsari “Investigation of FuzzyFind Searching with Golay Code Transformations”. M. Sc. Thesis, The George Washington University , Department, Department of Computer Scioence , 2014.

[6] R enviroment : cran.org

[7] Matlab Group , MACSYMA reference manual , 1974 , MIT.

[8] Confusion matrix : http://en. Wikipedia.org/wiki/confusion-matrix.s