ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

National Symposium On Emerging Trends In Computing & Informatics, NSETCI 2016, 12th July 2016,Rajagiri School of Engineering & Technology, Cochin, India.

AN EFFECTIVE APPROACH IN LARGE DATASETS- SINGLE INSTANCE STORAGE

***Catherine Mathew[1] and Varghese S Chooralil[2].**

1. Department of Computer Science and Engineering, Rajagiri School of Engineering and Technology Kochi, India
2. Department of Computer Science and Engineering, Rajagiri School of Engineering and Technology Kochi, India

Manuscript Info

Key words:
Data mining
Data Cleaning,
Deduplication,
Principle Component Analysis,
Single Instance Storage Framework

Abstract

SIS (Single Instance Storage) Framework is used in combining data from multiple sources into one comprehensive and easily manipulated database. The primary aim of SIS Framework is to provide a business with analytics results from data mining.SIS is designed to provide an architecture that will make social data accessible and useful to users. The deduplication process is finding duplicate records or redundant data when comparing with one or more database or data sets .This information is too costly to acquire because of which SIS process getting more attention nowadays. In data cleaning process removing redundant records in a single database is a difficult step, because outcomes of large data processing or data mining may get greatly influenced by duplicates data. As the database size increasing day by day the matching processes complexity becoming one of the major challenges for SIS Framework. The basic steps in implementing SIS include Blocking, Selection and classification. Semantic similarity is used for Blocking. The selection consists of: Sample selection, redundancy removal. The intermediate Subsets is given to the classifier after feature selection using Principle Component Analysis (PCA). Classification is done to efficiently identify the most ambiguous data in the training set.

.

Introduction:-

With the rapid development of internet, demand of online data analysis becomes key role in all areas. Sentiment analysis and Opinion mining involves the study of opinions and its related concepts such as sentiments, evaluations, attitudes and emotions. In IT business field databases play an important role. Many operations and decisions are carried out on the basis of outputs of databases. Therefore quality of information depends on the quality of data, implicitly methods which are used to store and to retrieve the data from database Ratnaraja Kumar (2016). In single-instance storage process we identify references in data records which refer to the same real world entity Ratnaraja Kumar (2016)**.** It is one of the crucial steps in data cleaning process. In collective single-instance storage we want to find types of real world entities in a set of records which are related. It is a generalization of single-instance storage.

The main difference between SIS and deduplication is SIS-based system is a new version of the file and store a new copy of the file each day. A data deduplication system would store only a few records of the database that had modified from the previous night's backup Ratnaraja Kumar (2016).

The motivation of this work was from the fact that when many a times data lack a unique or global identifier And also data are neither controlled nor defined in a continues manner in a different data sources. Storage-based data deduplication reduces the amount of storage space needed for a given set of files. The scope of the SIS framework lies in number of different methods and techniques for reducing repeated occurrences of identical data down to a single (or at least fewer) instances. The objective of SIS framework lie in simple, concise view on one or more selected areas, in support of the decision process .This concise view is constructed by integrating multiple, heterogeneous data sources .These data sources contains historical data that spans a much longer time horizon than operational databases. The goal of SIS framework is to work with large data sets like in data warehouse.

This Paper includes the related works on identifying the duplicate records using SIS framework and their related techniques, the proposed approach, Architecture & Methodology, Results and discussions, Conclusion and References.

Related Works and Background Knowledge:-

This section reviews some of the previous work in this field. The earliest approach to data reduction was data compression, which searches for repetitive strings of information within a single file. The other approach was SIS, which reduces the amount of storage by recognizing when files are repeated. As a result, users are concerned with the integrity of their data. The various methods of SIS data all employ slightly different techniques. However, the integrity of the data will ultimately depend upon the design of the SIS system, and the quality used to implement the algorithms. SIS process involves removing copies of files.

In the paper Scaling Up All Pairs Similarity Search, the author states that if a large collection of sparse vector data with a high dimensional space is given, this research investigate the problem of finding all possible pairs of vectors whose similarity score is above a given threshold .Roberto J. Bayardo (2007) An optimization and novel indexing strategies solves the problem stated above. Without depending on extensive parameter tuning or approximation methods, a simple algorithm is proposed by an author based on above strategies Roberto J. Bayardo (2007) The approach proposed by an author is efficient than previous state-of-the-art approach to handle a variety of data sets with large speedup and wide setting of similarity thresholds .Merit of this paper: Given a large collection of sparse vector data in a high dimensional space, the paper investigate the problem of finding all pairs of vectors whose similarity score(as determined by a function such as cosine distance) is above a given threshold Roberto J. Bayardo (2007) The paper proposes a simple algorithm based on novel indexing and optimization strategies that solve this problem without relying on approximation methods or extensive parameter tuning. The approach efficiently handles a variety of datasets across a wide setting of similarity thresholds, with large speedups over previous state-of-the-art approaches Roberto J. Bayardo (2007)

The Proposed Approach:-

The basic Framework of SIS consist of Blocking, Comparison ,Selection and Classification Guilherme Dal Bianco(2015) First, a strategy is employed to identify the blocking threshold, and thus produce the candidate pairs. The Comparisons stage has to be done in two-stages. In its first stage, SIS produces small balanced subsamples of candidate pairs and its second stage the redundant pairs are identified .Once identified the dataset is improvised using feature selection using PCA After which classification is done using SVM classifier. The Minimum False Pair and True pairs are identified and duplicates record are removed.

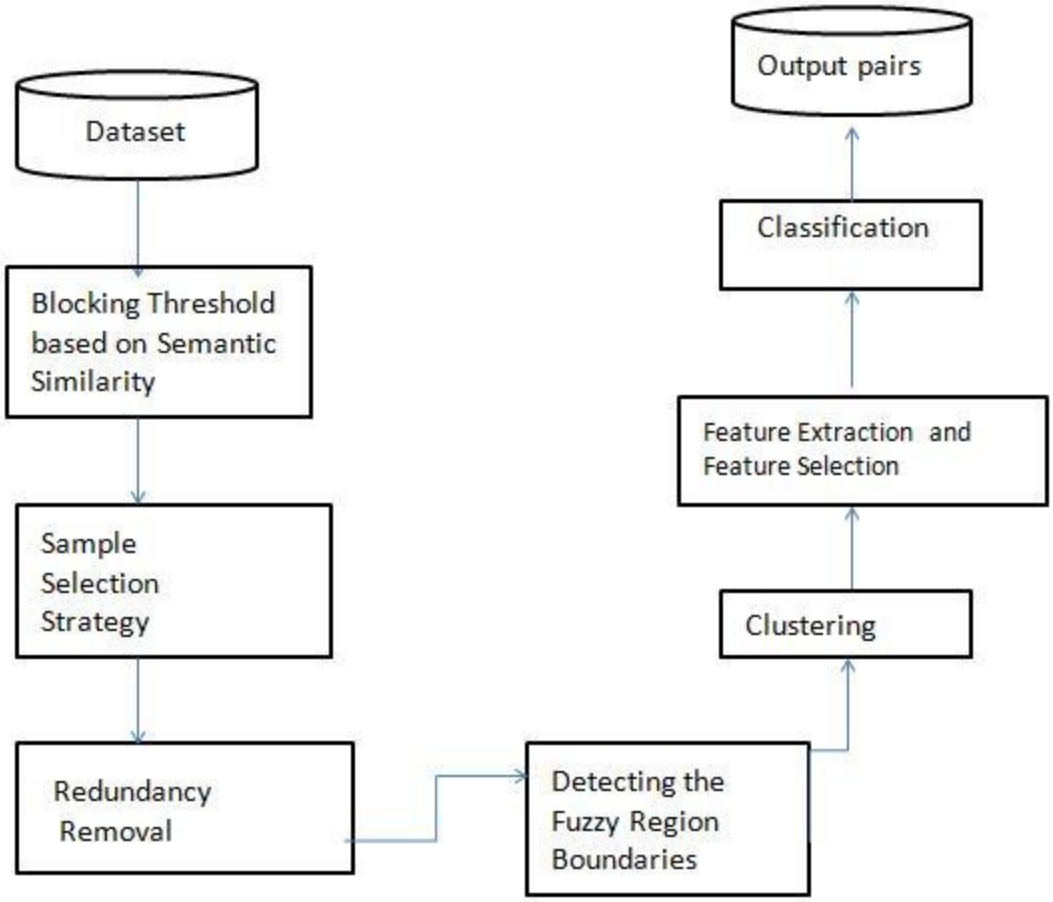


Fig. 1. Proposed design

Workflow of Single Instance Storage Framework:-
(i) The dataset is deployed into database for deduplication process.

(ii) The avoid duplicate pairs we select fixed the threshold. These thresholds are used to split the dataset. This sets for consider the separate blocking.

(iii) The next stage is to fix the some attribute based on that fixed attribute to analysis each and every block.

(iv) In Fuzzy region boundaries detection section use a previous remaining matching pair are taken to the fuzzy region identification. This identification gets some pairs.

(v)The candidate subset so formed is done clustering with fuzzy c-mean algorithm into two clusters MFP and MTP.

(vi) The dataset so obtained is further improved using Feature Extraction and Feature Selection

(vii) These result pairs again enter into the classification stage. This stage picks the accurate pairs.

Redundant removal based fetches the matching labeled pair classification finally detect the duplicate

Architecture & Methodology:-

The overall system architecture will clearly give an idea of each module. The basic work flow is shown in Fig 1

A. Blocking:-

Using semantic similarity Peipei Li, Haixun(2014) the similarity has been calculated. Once calculated the number of matching pairs represents a small subset of the dataset. The threshold that is matched must have less matching pairs than the total number of records in the dataset. A global threshold is assigned as initial threshold. Now the entire dataset is matched to create a set of candidate pairs. The candidate pairs are selected and sorted using their similarity values to produce a ranking.

B. Comparison:-

The ranking so created in the blocking phase is given as input to sample selection. The ranking is divided into levels based on similarity value of candidate pairs .Within each level, random selection of candidate pairs are done to create samples. This strategy prevents non-matching pairs dominating the samples Ratnaraja Kumar (2016)The candidate pairs in each levels are brought together incrementally using rule based active sampling R. M. Silva(2014).

Rule Based Sampling:

In this we have a training set say D,a unlabelled set U={U_1 , U_2 , U_3 , U_4 , $U_{5.....}$, U_n }.U is consider as set of informative documents that we thought to compose into the training set D.The rules obtained for these documents U_1 , U_2 , U_3 , U_4 , $U_{5.....}$, U_n are taken into consideration.

Case 1:Suppose we add U_i to training set D ,provide U_i belongs to U ,then the no: of rules for document in U that shares feature value with U_i wll increase or remain unchanged.

Case 2:If no: of rules for documents in U don't share feature values in U_i will remain unchanged.

So in other words we can say that the no: of rules derived from each document in U can be used to determine an approximate of the amount of redundant information between documents already in D and those for document U

C. Fuzzy Region Boundaries:-

In this phase, the training set is created by the selection phase which is used to detect the fuzzy region boundaries R. M. Silva(2014). This region is detected by using manually labeled pairs which are selected by from each level. The pairs labeled by the user may result in MTP and MFP. Sometimes this MTP and MFP are far from users expected position R. M. Silva(2014).So MTP and MFP are assuming to be defined within fuzzy region boundaries. The similarity value of MTP and MFP identifies alpha and beta values. Then the fuzzy region is formed by all the candidate pairs within a similarity values between alpha and beta values R. M. Silva(2014).

D. Clustering:-

Here we divide the subset candidates into two sets by labeling them into MFP and MTP .Fuzzy Cmeans is used in clustering.

E. Selection:-

The candidate sets obtained after processing are further selected so as it is believed to be redundant, relevant and free from noise. Feature extraction is done using the following concept.

1. Increase the no: of occurrence within a document
2. Increase the rarity of the term in the collection.

During feature selection the pattern are represented as feature vectors. The idea of feature selection using PCA Ashish Singhalt Proceeding (2011) is to find a good subset of features. For dimensionality reduction, project the data onto a lower dimensional subspace.

F. Classification:-

The dataset so obtained is given to SVM classifier to predict the correctness of the predicted data. The system architecture further explains the work flow of SIS architecture.

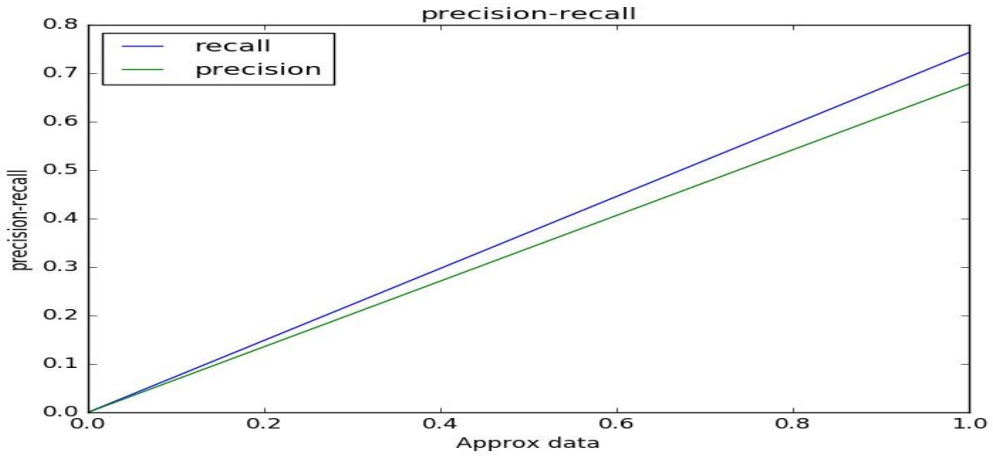
Results and Discussions:-

My paper discusses on removing the duplicate and redundant data from backups like data warehouses. The paper discuses a novel idea in the data cleaning process which is primarily stage in Data mining. Once the proposed work is carried out on data set, the result Figure 2 you obtain will contain only a single occurrence (instance) of the data in the entire data set. In other words only a single occurrence of the data will be stored in the data warehouse.

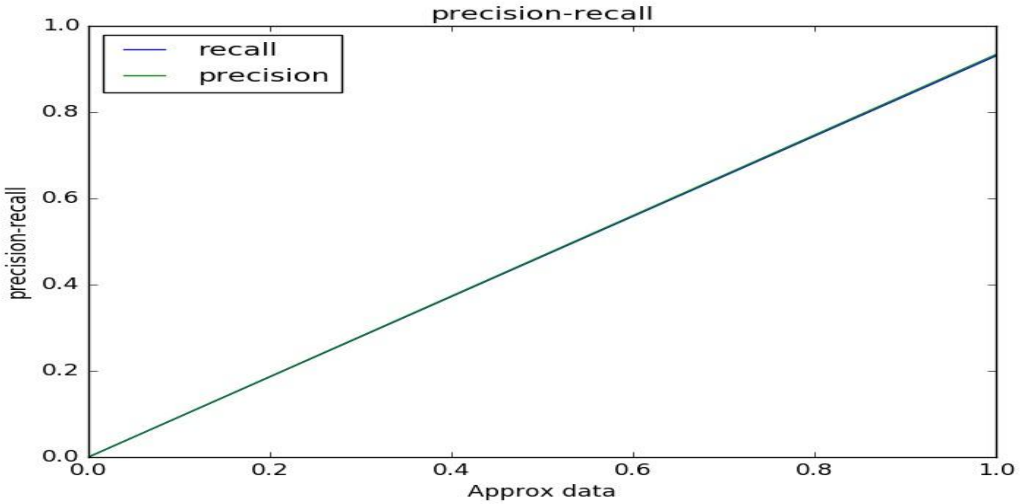
The paper proposes a new advance novel approach of SIS framework for finding large scale deduplication. The proposed SIS framework [Figure 1] is able to select a very small, non-redundant and informative set of examples with high effectiveness for data warehouses. In the second stage a rule-based active sampling strategy, which requires no training set (as required in classifier committees), is incrementally applied to the selected balanced pairs to reduce redundancy. The performance analysis is done through precision-Recall Figure 3b.The Precision recall increases as the no of data increases due to which we can claim the work to be scalable.

id	pairs	class
1	books/mk/bancilhonDK92/VelezBD92:The O2 Object Manager An Overview.:Fernando VÃ©lez Guy Bernard Vineeta Darnis:2002-01...	mtp
2	books/idea/taniar2003/RahayuFT03:Database-Driven Product Catalog System.:J. Wenny Rahayu Andrew Flahive David Taniar:200...	mfp
3	books/crc/tucker97/Cardelli97:Type Systems.:Luca Cardelli:2008-02-24 2208-2236 1997 books/crc/tucker1997 The Computer Scien...	mtp
4	books/sp/dittrichDB91/BuchmannCV91:Handling Constraints and their Exceptions An Attached Constraint Handler for Object-Orient...	mtp
5	conf/3dim/X05:Foreword.::2005-11-24 2005 conf/3dim/2005 3DIM http //doi.ieeecomputersociety.org/10.1109/3DIM.2005.47 db/c...	mtp
6	books/sp/Ahn04:Least Squares Orthogonal Distance Fitting of Curves and Surfaces in Space:Sung Joon Ahn:2004-12-07 Lecture N...	mtp
7	books/mk/ZanioloCFSSZ97:Advanced Database Systems.:Carlo Zaniolo Stefano Ceri Christos Faloutsos Richard T. Snodgrass V. S. ...	mtp
8	books/idea/encyclopediaDB2005/HainautHHRE05:CASE Tools for Database Engineering.:Jean-Luc Hainaut Jean Henrard Jean-Marc...	mfp
9	books/duv/Schoning93:Anfrageverarbeitung in Komplexobjekt-Datenbanksystemen:Harald SchÃ¶ning:2002-01-03 Deutscher Unive...	mtp
10	books/kl/snodgrass95/SooJS95:An Algebra for TSQL2:Michael D. Soo Christian S. Jensen Richard T. Snodgrass:2008-05-06 db/boo...	mtp
11	books/cu/Appel1998c:Modern Compiler Implementation in C:Andrew W. Appel:2004-01-12 Cambridge University Press 1998 0-521-...	mtp
12	conf/3dim/SilvaBB03:Enhanced Robust Genetic Algorithms for Multiview Range Image Registration.:Luciano Silva Olga Regina Pereir...	mtp
13	books/bc/tanselCGSS93/LeungM93:Stream Processing Temporal Query Processing and Optimization.:T. Y. Cliff Leung Richard R. M...	mtp

Fig. 2. Classification Output: Non-redundant Pairs



a) During Manual Labeling in candidate pairs



b) Clustering: For Labeling in candidate pairs

Fig. 3. Comparison of Recall and Precision

Conclusion

The Single Instance storage strategy deals with the idea of removing the manual labeling effort of users while working on datasets. The Sample Selection stage selects small sub samples randomly of candidate pairs where as in the next stage removes redundancy sub samples that are incrementally analyzed Vishnu Priya Paramasivam(2015)
The candidate pairs are driven through Feature Extraction and Feature Selection phases and then given to classifier for prediction. The main advancement of this work in the area of de-duplication is the idea of removing manual labeling effort and introducing the idea of Feature Extraction and Feature selection.
Advantages of Proposed work:
Mostly these data warehouses, datasets are searched by SIS framework, Once duplicate records are discarded, the search quality can be enhanced. This architecture ensures time saving and improved productivity. The SIS framework aims at identifying entities that are potentially the same in data repository. The quality of SIS framework lies in its ability to identify and remove duplicate records as well as efficiency in saving storage space.
.

References:-

1. **Mr. J. Ratnaraja Kumar(2016)** A Survey Study for Deduplication in Large Scale Data International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 2.
2. **Guilherme Dal Bianco, Renata Galante, Marcos Andr_e Goncalves, Sergio Canuto, and Carlos A. Heuser(2015)** A Practical and Effective Sampling Selection Strategy for Large Scale DeduplicationIEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 9, SEPTEMBER.
3. **Vishnu Priya Paramasivam (2015),Department of Computer Science K. Ramakrishnan College of Technology** Two Stage Sampling Selection Strategy for Large Scale Deduplication.
4. **Peipei Li, Haixun Wang, Kenny Q. Zhu, Zhongyuan Wang, Xuegang Hu, and Xindong Wu(2014), Fellow, IEEE** A Large Probabilistic Semantic Network Based Approach to Compute Term Similarity
5. **Ashish Singhalt (2011)Proceedings of the American Control Conference** Arlington, VA.Matching Patterns from Historical Data Using PCA and Distance Similarity Factors
6. **Roberto J. Bayardo(2007) Google, Inc.Proc. of the 16th Intl Conf. on World Wide Web** Scaling Up All Pairs Similarity Search , Ban_, Alberta, Canada, 131-140.
7. **R. M. Silva, M. A. Goncalves, and A. Veloso(2014)**, A two-stage active learning method for learning to rank, J. Assoc. Inform. Sci. Technol. vol. 65, no. 1, pp. 109128..