www.ioirp.com
International Journal of Innovative Research in Computer Science and Engineering (IJIRCSE)
ISSN: 2394-6364, Volume – 1, Issue – 12. November 2015

Two Stage Sampling Selection Strategy for Large Scale Deduplication

Vishnu Priya Paramasivam
Department of Computer Science and Engineering
K.Ramakrishnan College of Technlogy
Tiruchirappalli,India
vishnupsivam@gmail.com

***Abstract*— Deduplication is the task of identifying which objects are potentially the same in a data repository. The data deduplication task has attracted a considerable amount of attention from research community in order to provide effective and efficient solutions. In very large scale, it is difficult to select reduced set of pairs and label large number of informative pairs because it requires experts. In this paper, we propose a two stage sampling selection strategy (T3S) that selects a reduced set of pairs to tune the deduplication process in large datasets.T3S works on following two stages. In first stage, we propose to select balanced subsets. In second stage, an active selection strategy is invoked to remove the redundant pairs in the subset created in the first stage to produce even smaller and more informative training set. Then apply classification approaches to efficiently identify the most ambiguous in the training set. Finally, to investigate the performance analysis and checking for previous output.**

Keywords— Deduplication, Signature-based deduplication, Fuzzy region.

I. INTRODUCTION

A dramatic growth in the generation of huge amount of data or information from a wide range of sources. The sources maybe mobile devices, media streaming and social networks. This has opened opportunities for the emergence of several applications such as shopping websites, digital libraries and media streaming. Those applications required good quality of data to provide efficient services. Data storage as well as data quality can be degraded due to the presences of repeating or redundant data in the repository. The duplicates may occur as misspelling, abbreviations, conflicting data and redundant entities. If the repository contains large number of replicated data, then search or recommendation may not produce results as expected by the end user. The ability to check a collected object already exists in the storage is an important task to improve data quality. That is the data quality can be achieved by detecting and removing duplicates. Record Deduplication aims at identifying which objects are potentially the same in the data repository [5]. The removal of repeating data is an old problem, but still it receives the potential attention from database community because of its difficulties. In large data repository or storage the occurrences of those kind of redundancy makes a daunting process of retrieval and reduce the efficiency of storage. To handle this by means of Deduplication.

Deduplication is the data compression technique for eliminating duplicate copies of repeating data. This data Deduplication task has considerable amount of attention from research community in order to provide effective and efficient solutions. Deduplication Promises to reduce the transfer and storage of redundant data, which optimizes network bandwidth and storage capacity.

II. TWO STAGE SAMPLING SELECTION STRATEGY(T3S)

Deduplication process takes three steps: Blocking, Comparison and Classifications. In Blocking, reduces the number of comparisons by grouping together pairs that can share common features [7], that is to put together all the records in the same block which has common attribute values. In comparison, to found out the similarity between pairs belonging to same block. The similarity can be performed by using some similarity functions [2] [8] [7] such as Jaccard, cosine and so on. Finally, in classification phase identifies which is

matching pairs and non-matching pairs [6]. This can be done by means of global threshold. In Deduplication the blocking and classification phases [1] [3] depend on user to configure the process. In this situation, the researchers proposed [4], FS-Dedup framework to find close to optimum configuration. FS-Dedup was demonstrated to be more effective than manually tuned methods, while still reducing labeling efforts. However, the resulting subsamples may still be composed of redundant pairs, with negative impacts in the labeling effort

In proposed system, introduce an innovative procedure over previous method for reducing the redundancy in the subsamples. A two stage sampling selection strategy (T3S) is introduced to select the reduced set of pairs. It is able to select a very small, non-redundant and informative set of pairs with high effectiveness for large scale datasets. In first stage, a strategy to produce balanced subsets of candidate pairs for labelling. In second stage, to remove redundancy in the subset that is created in the first stage to produce smaller and more informative pairs. In blocking phase initial or global threshold is identified by using Sig-Dedup. Signature based Deduplication [4], is used to maps the dataset strings into a set of signatures by using inverted index. For each and every token in the record, same substring produce same signature. And the threshold value has to be set for generating candidate pairs. The initial threshold is defined in order to minimize the number of lost matching pairs. The main purpose of this threshold is to define how many tokens are indexed by the sorted record. After defining the global threshold value for blocking process, the entire dataset is matched to create the set of candidate pairs.

Now introduce T3S, to produce samples and to select random pairs within the set of candidate pairs. First stage of T3S adopts the concept of levels to allow each sample to have similar diversity among the subsamples. The ranking created in blocking stage is divided into 10 level (0.0-0.1, 0.1-0.2,.....and0.9-1.0) by using similarity value. Then randomly select the candidate pairs from each level as per defined size. It is possible to produce a balanced set of pairs to be used in T3S next stage to remove redundancy. The second stage is integrated with FS-Dedup to eliminate the redundant data to select the fuzzy region by using active fuzzy region selection. The fuzzy region boundary is fixed by means of MTP and MFP. Then remove redundant data within the fuzzy region. The classification [6] step aims at categorizing the candidate pairs belonging to the fuzzy region as matching or non-matching. Two classifiers are used T3S-NGram and T3S-SVM.

III. TWO STAGE SAMPLING SELECTION STRATEGY(T3S) WORKING

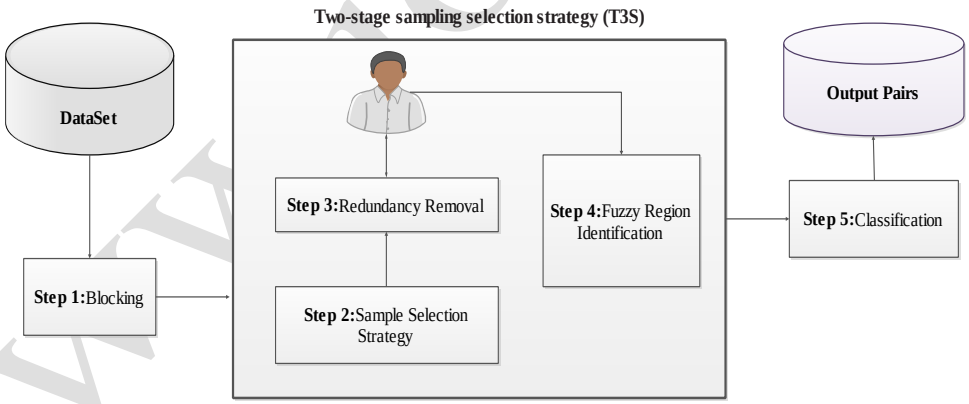

Figure 1.1 System Architecture

Figure 1.1 describes the system architecture. A typical de duplication method is divided into three main phases: Blocking, Comparison, and Classification. 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-stage sampling selection strategy (T3S). In its first stage, T3S produces small balanced subsamples of candidate pairs. In the second stage, the redundant information is selected in the subsamples and removed by means of a rule-based active sampling .This will not previously labelled training set.

Following this, will describe how these two steps work together to detect the boundaries of the fuzzy region. Finally, two classification approaches are introduced in which it is configured by using the pairs manually labeled in the two stages.

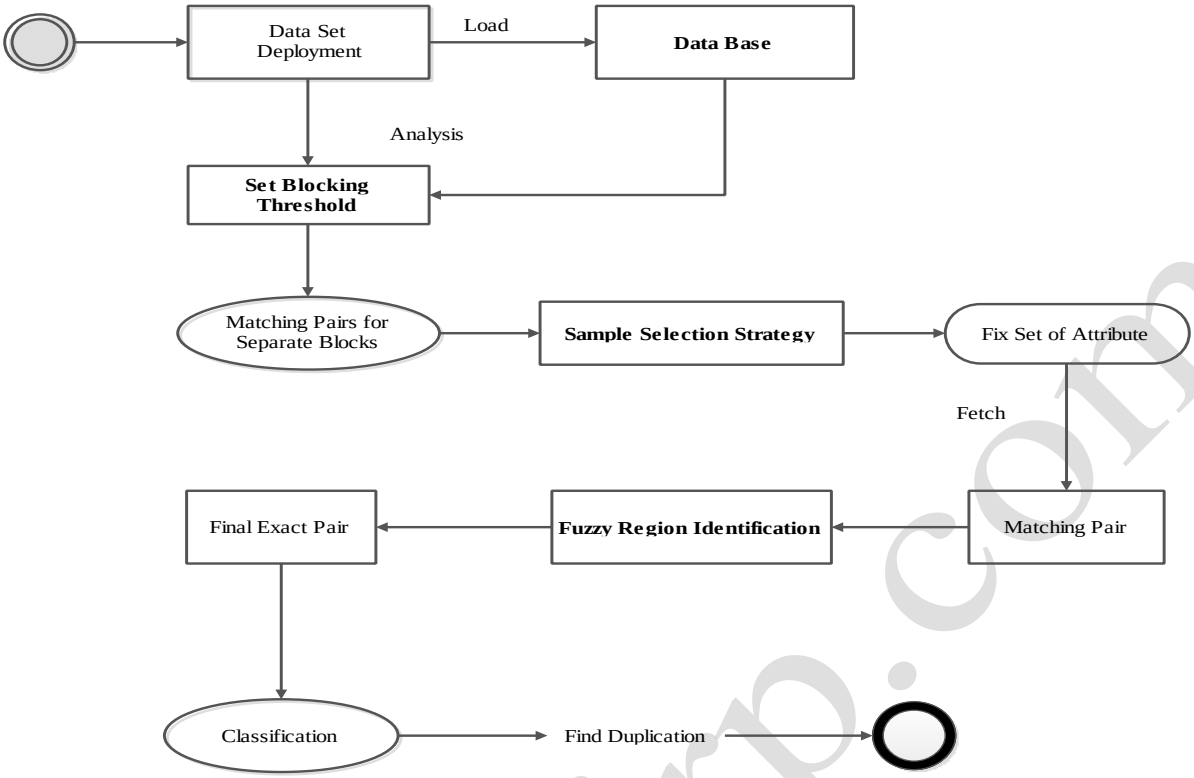

Figure 1.2 Workflow execution

Figure 1.2 defines the overall workflow of T3S algorithm. The following steps will demonstrate the working of T3S algorithm:

(i) The dataset is deployed into database for deduplication process.
(ii) The admin to avoid the duplicate pairs by to fixed the blocking threshold. These thresholds are used to split the dataset. This sets for consider the separate blocking.
(iii) The next stage for admin fix the some attribute based on user details, on that fixed attribute to analysis each and every block.
(iv) And then compare two lists which give a fixed attributes and user details. It supposes match the both details fetch the corresponding pairs.
(v) In that section use a previous remaining matching pair are taken to the fuzzy region identification. This identification gets some pairs.
(vi) These result pairs again enter into the classification stage. This stage picks the accurate pairs. This process used to avoid the redundancy.
(vii) Redundant removal based fetches the matching labeled pair classification finally detect the duplicate.

IV. COMPONENTS

The various components used here are: (A) Identifying Approximate indexing, (B) Selecting the Samples, (C) Redundancy Removal, (D) Detecting the fuzzy region and Classification. The working of each of the components is elaborated in the following section:

A. Identifying Approximate Indexing

First and foremost the admin has to avoid the duplicate pairs by fixing the blocking threshold. The approximate blocking threshold is determined by using Sig-Dedup filters. This blocking threshold is called as initial threshold. The purpose of this threshold is to define how many token are indexed by the sorted record.

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 such dataset. After defining the global threshold, it will be taken as a initial threshold value for the blocking process. Then the entire dataset is matched to create a set of candidate pairs. At the end, matched candidate pairs are to be sorted using their similarity values to produce a ranking.

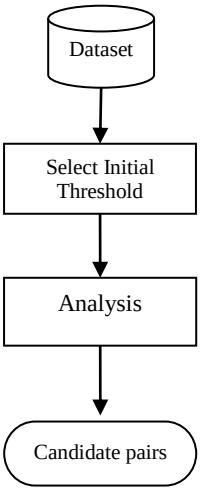

Figure 1.3 Identifying Approximate Indexing

B. Selecting the Sample

In this phase, sample selection strategy will be helpful to produce balanced subsamples of candidate pairs. The concept of levels allows each sample to have a similarity diversity for the full set of pairs. The ranking created by blocking step is divided into 10 levels by using similarity value of each candidate pairs. Inside each level, random selection of candidate pairs are taken to create samples with a defined size.

This fragmentation produces levels, composed of different matching patterns to prevent non-matching pairs dominating the sample. It is possible to produce a balanced set of pairs to be used in T3S next stage to remove redundancy in the information randomly selected.

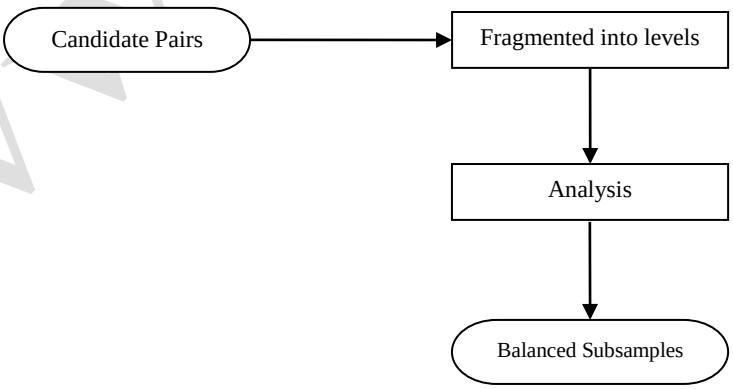

Figure 1.4 Selecting the Sample

C. Redundancy Removal

The sample selection stage produces balanced subsamples as a input to this phase. However several pairs are selected inside each level which are composed of redundant information and does not help to increase the training set diversity. In this stage, T3S aims at incrementally removing the non-informative pairs by using

SSAR. The purpose of SSAR is to select the label which has only the most informative pairs required to maximize the training size diversity while minimizing labeling effort.

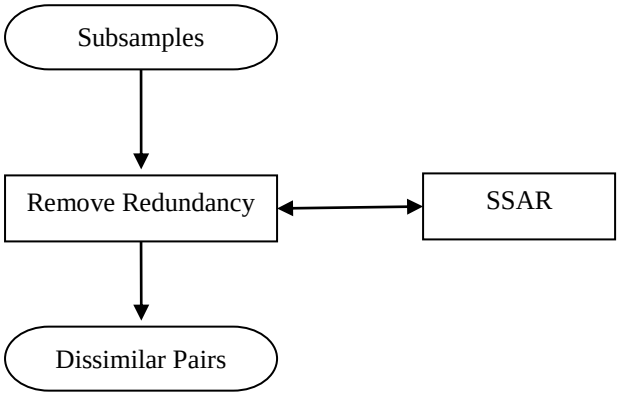

Figure 1.5 Selecting the Sample

SSAR may select pairs with similar information at the border of the levels by producing a training set with redundant information. Finally, the training set become more informative and only the most dissimilar pairs are selected by SSAR.

D. Detecting the fuzzy region and Classification

In this phase, the training set is created by the two stages of T3S which is used to detect the fuzzy region boundaries. This region is detected by using manually labelled pairs which are selected by using SSAR from each level.

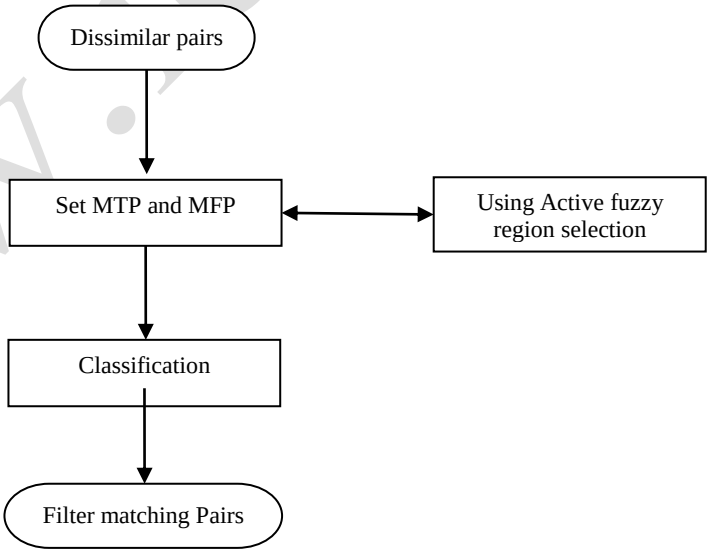

Figure 1.6 Detecting the fuzzy region and Classification

The pairs labelled by the user may result in MTP and MFP. Sometimes this MTP and MFP are far from users expected position. 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.

Once the boundary value is defined the pairs belonging to fuzzy region are sent to classification step. And this step is used to categorize the candidate pairs within the fuzzy region as matching or non-matching.

IV. CONCLUSION

We have Existing T3S, a two-stage sampling strategy aimed at reducing the user labeling effort in large scale de duplication tasks.T3S is able to considerably reduce user effort while keeping the same or a better effectiveness. In the first stage, T3S selects small random subsamples of candidate pairs in different fractions of datasets. In the second, subsamples are incrementally analyzed to remove redundancy. To evaluated T3S with synthetic and real datasets and empirically showed that, in comparison with four baselines, T3S is able to considerably reduce user effort while keeping the same or a better effectiveness. To investigate whether is performance analysis and checking for previous output.

With the continuous and exponential increase of the number of users and the size of their data, data deduplication becomes more and more a necessity for cloud storage providers. By storing a unique copy of duplicate data, cloud providers greatly reduce their storage and data transfer costs. The deduplication method best suited to protect data in cloud. This process DE-duplicates data both across backups and within backups and does not require any knowledge of the backup data format. The content similarity is used for the de-duplications process and filtering the de-duplication content, to help an analyst in cloud with similar content by designating of Data duplication can be easily removed by the content similarity algorithm.

References

[1] Bellare K., Iyengar S., Parameswaran A. G. (2012), and Rastogi V., 'Active sampling for entity matching', *in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 1131–1139.

[2] Chaudhuri S., Ganti V., and Kaushik R. (2006), 'A primitive operator for similarity joins in data cleaning', *in Proc. 22nd Int. Conf. Data Eng*., p. 5.

[3] Christen P. (2008), 'Automatic record linkage using seeded nearest neighbour and support vector machine classification', *in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 151–159.

[4] Dal Bianco G., Galante R., Heuser C. A., and Gonalves M. A. (2013), 'Tuning large scale deduplication with reduced effort', *in Proc. 25th Int. Conf. Scientific Statist. Database Manage*, pp. 1–12.

[5] Elmagarmid A., Ipeirotis P., and Verykios V. (2007), 'Duplicate record detection: A survey', *IEEE Trans. Knowl. Data Eng., vol. 19, no. 1*, pp. 1–16.

[6]Sarawagi S. and Bhamidipaty A. (2002), 'Interactive deduplication uses active learning', *in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 269–278.

[7] Wang, G. Li, and J. Fe, "Fast-join: An efficient method for fuzzy token matching based string similarity join," in Proc. IEEE 27th Int. Conf. Data Eng. ,2011, pp. 458–469

[8] C. Xiao, W. Wang, X. Lin, J. X. Yu, and G. Wang, "Efficient similarity joins for near-duplicate detection," ACM Trans. Database Syst., vol. 36, no. 3, pp. 15:1–15:41, 2011.