IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 4, Issue 1, Feb - March, 2016
ISSN: 2320 – 8791 (Impact Factor: 2.317)
www.ijreat.org

An Effective Analysis And System Based Sampling Model For Deduplication

M.Magheswari ,E.Muthu Jeyanthi, Dr.S.Selvakumar

COMPUTER SCIENCE AND ENGINEERING
GKM COLLEGE OF ENGINNERING AND TECHNOLOGY
CHENNAI,TAMILNADU,INDIA

Abstract

Data Deduplication is a strategy for removing redundant copies of data, and has been widely used in data mining to avoid the recapitulation of data's by using T3S (Two-stage Sampling Selection strategy), Which is Used for Reducing the set of Pairs to avoid the Deduplication The T3S is mainly used to reduce the labeling effort. The Training set is used to identify where the most equivocal pairs lie and to configure the classification approach. The Sampling Selection Strategy and Redundancy Removal Stages are helped to nullify Deduplication. The Report Analysis is generated for the inputs., This will be led by the ability to eliminate duplicate and unstructured data (office files, images, secured data etc.).

Key words: Deduplication, Sampling, Fuzzy region

1. Introduction

Data Mining is the strategy for extracting hidden predictive information from large sets of data.From given databases,datamining methodology generates new opportunities in business streams by providing capabilities.

1.1 Prediction of trends and behaviors:

Data mining does the process of finding predictive information from databases automatically. A predictive problem is targeted marketing which uses data on past to identify the targets leads to increase return on investment is the best example. Other predictive problems include forecasting and identifying segments of a population. Here we present some background necessary to properly understand our proposed approach as some of the used concepts are recent and significant. First we specify the main concepts behind Sig-Dedup algorithms adopted as deduplication core by our approach in the blocking and classification steps.

2. DATA MINING TECHNIQUES

- **Artificial neural networks**: Non-linear predictive models that learn through training and common biological neural networks in structure.
- **Decision trees**: Tree-shaped structures that indicates set of conclusions. These decisions generate rules for the classification of a dataset. A specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- **Genetic algorithms**: The perfect Optimization techniques that use genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

3. Existing system

- Baseline approach is inefficient, as it will generate an enormous number of keys with the increasing number of users.
- The Existing system is unreliable.
- A semantic security for unpopular data and provides weaker security and better storage and bandwidth for popular data.
- In this approach, data deduplication is the process by which a storage provider only stores a single copy of a file owned by several of its users.
- It only focuses an analysis on scenarios where the outsourced dataset contains few instances of some data items and many instances of others.

3.1 Disadvantage:

- Deduplication method aims at reducing the number of comparisons by grouping together pairs that share common features but not the unmatched pairs.
- In this project, it is used to send only the random pairs of data and takes samples of each and every data simultaneously.
- The process is used only to avoid the labeling effort, but not the relevant data what they are going to process with help of sampling strategy.

4. Modules description

4.1 Fuzzy region

Consider X and Y are product space(X x Y) from two sets of records(X and Y) into: Z indicates the set of matching pairs, U represents the set of non-matching pairs ,and P represents the set where matching and non-matching pair co-exists. The pairs in the Z set typically share common characteristics. The pairs in the U set usually have obscure agreements of the characteristics. Set P contains pairs with the high degree of ambiguity and requires human intervention to classify them because of this, we called the set P the Fuzzy region.

4.2 T3S steps overview

In this section we proposed T3S aimed at selecting the reduced and representative sample of pairs in large scale deduplication.

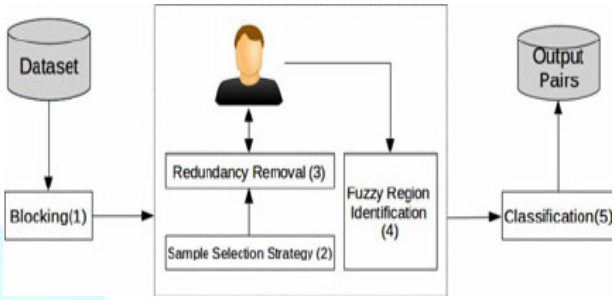

Fig 1. T3S OverView

We integrate T3S with previous FS-Dedup frame work to reduce the user effort in the main deduplication steps (e.g blocking and classification). The above figure illustrates how the T3S steps work together in a collaborative way. (Fig 1(1)) The first stage, it is used to identify the blocking threshold, and thus produce the small balanced sub samples of candidate pairs. (Fig 1(2)) Second stage, the redundant information that is selected in the sub samples is removed with the help of rule-based active sampling. (Fig 1(3)) which requires no previously labeled training set. We describe how these two steps work together to detect the boundaries of the fuzzy region. (Fig 1(4)) finally we describe our two classification approaches which are configured by using the pairs manually labeled in the two stages (Fig(1(5)).

4.3 Blocking threshold approximate identification

Blocking threshold approximation is identified using the Sig-Dedup filters that maximize recall, i.e that minimize the occurrences of pruning in actual matching pairs. This blocking threshold said as initial threshold. The initial threshold also contains the matching pairs in the set of candidate pairs. It can be also done without user intervention, if we follow generalization we can get closer to the ideal scenario. Initial threshold is used to minimize the number of “extinct” matching pairs that are outside the mean time for analysis. The set of candidate pairs produce

by using Sig-Dedup filters is worth nothing and these are configured with initial threshold. The objective of this threshold is to define number of tokens that are indexed by the sorted record(the global frequency of tokens used to resorted the records). In this step, the similarity values of each pair is not used to trim out pairs since we do not know the exact threshold value that is able to discard non-matching pairs.

In huge datasets the number of matching pairs represents a small subset of the data set. For instance two identical data sets that are matched must have less matching pairs than the total number of records in such datasets. Threshold values is incrementally increased fewer tokens in the sorted record or index that reducing the number of candidate pairs. In large datasets it may not feasible to run the Sig-Dedup filters. In light of this, we propose a stopping criterion to estimate the initial threshold. The stopping criterion which produces the threshold that avoid both : large generation of candidate pairs and recall degradation.

4.4 SAMPLE SLECTION STRATEGY (THE FIRST STAGE)

The proposed sample selection strategy is to produce the balanced sub samples of candidate pairs. The main objective of this stage is to discretize the ranking. So that small subsets of candidate pairs can be selected to decrease the computational demand of the T3S second stage.

A simplistic approach to produce samples might be to select random pairs with in the set of candidate pairs, as the set of pairs is basically formed of non-matching pairs it will result in samples having low informativness. First stage of T3S is to taken up the concept of levels to allow each sample to have a same diversity to that of the full set of pairs. The rankingness is created by blocking step, is fragmented into 10 levels.(0.0-0.1,0.1-0.2,0.2-0.3,……,and0.9-1.0).

The main two reasons for the 10 intervals: the first stage is effectiveness of the deduplication process which declines smoothly after the optimal threshold value found in 10 intervals. Second is, an increase in the number of substantially increase the number of candidate pairs will be analyzed by our method(T3S) to identify the threshold. This will force a more precise identification of the fuzzy region boundaries to minimize the error rate.

4.5 REDUNDANCY REMOVAL(SECOND STAGE)

By the random selection of pairs inside each level it will produce the samples. Sub samples are used to identify the fuzzy region boundaries. when the size of the level is quite large. The second stage is T3S is aims at incrementally removing the non-informative or redundant pairs inside each sample by using the SSAR(selective sampling using association rules) this is called as active learning method.

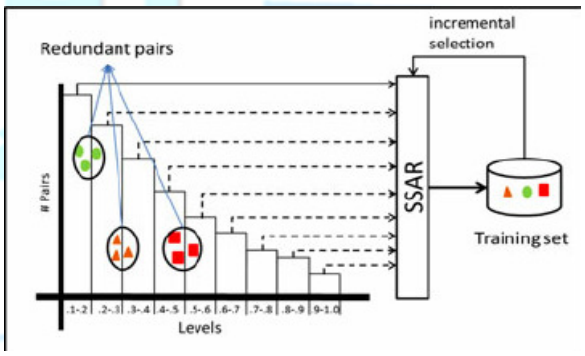

Fig 2. Illustration of the incrementally active sampling selection using fixed levels. Note that the levels[0.2-0.3] have redundant pairs. In different levels. As our SSAR active learning method is performed incrementally. Such redundancy among levels can be removed.

The objective of SSAR is to select pairs for labeling and only the most informative pairs are required to escalate the training size diversity while minimizing the labeling effort. SSAR has some advantages 1.Not requiring initial labeled set as needed by approaches based on committees. 2.Having a clear stopping criteria a property that many approaches do not possess.3.The capability of selecting very few but very informative instances on an informativeness criteria grounded on lazy

association rules. SSAR selects an unlabeled pairs u_i using inferences about the number of association rules produced with in a projected training set specific for u_i .

Algorithm 1. SSAR: Rule based Active Selective Sampling

Require: Unlabeled Set T and $\sigma_{min}(\approx 0)$

Ensure: The training set D

1. *While true do*
2. *for all* $u_i \in T$ *do*
3. $D_{ui} \leftarrow$ *D projected according to* u_i
4. $R_{ui} \leftarrow$ *extract useful rules from* D_{ui}
5. *end for*
6. *If* $D = \emptyset$ *then*
7. $\lambda_{ui} \leftarrow u_i$ *such that* u_i *is the most representative item of T*
8. *else*
9. $\lambda_{ui} \leftarrow u_i$ *such that* $\forall_{ui} : |R_{ui}| \leq |R_{ui}|$
10. *end if*
11. *if* $\lambda_{ui} \in D$ *then*
12. *Break*
13. *else*
14. *LabelPair* (λ_{ui})
15. $D \leftarrow D \cup \{\lambda_{ui}\}$
16. *end if*
17. *end while*

An unlabeled pair u_i is used as a filter to remove irrelevant features and examples from D. the objective of the procedure is to select the most dissimilar unlabeled pair by making a comparison with the current training set. *Drawback of SSAR:* high computational cost of re-projecting a large unlabeled data set at each level.

A computational complexity is $O(S^*|U|^*2^m)$ where "S" is the number of pairs selected to be labeled, "|U|" which represents the total number of candidate pairs, "m" is the one which denotes the features. (the cost of the rule generation is proportional to the number of features). Selection strategy enables SSAR to process in huge data sets reducing substantially the universe of "|U|" and the sampling selection strategy (first stage) of T3S allows SSAR to perform in huge data sets producing a reduced size of samples at each level. The final training set is created by joining together in subsamples of each level. SSAR invokes Sample selection strategy incrementally by using each level and the current training set as input.

4.6 Identifying the fuzzy region boundaries

Here we discussed, how the training set was created by the Two stages of T3S is able to detect the fuzzy region boundaries

Definition 1. Let MTP (minimum true pair) represent the matching pair with the lowest similarity value among the set of candidate pairs.

Definition 2. Let MFP(maximum false pair) represents the non-matching pair with the highest similar values among the set of non-matching pairs.

The fuzzy region is detected by using manually labeled pairs. The user is advised to manually label the pairs that are selected incrementally by the SSAR from each level. The pairs which are labeled by the users, may result in MTP and MFP pairs to reduce this problem we assume that the levels to which the MTP or MFP pairs belong are defined with in fuzzy region boundaries. For example if the MTP and MFP values are .35 and .75 expectedly all the pairs with the similarity value between .3 and .8 belong to the fuzzy region. We call the fuzzy region boundary as α and β

5. Classification step

This step aims at categorizing the candidate pairs belong to the fuzzy region as matching pairs or non-matching pairs. Two classifiers are used by us in this step are T3S-NGram and T3S-SVM. T3S-NGram is mapping the each record to a globally sorted token set then apply the Sign-Dedup filtering and defined similarity functions to the sets. The drawback of T3S-NGram is the different attributes are given the same importance lead to distortion in matching. T3S-SVM assigns different weights to different attributes by using the SVM algorithm based on

their relative discerning power[4]. There is not unique and globally similar function that can be adopted to different applications.

In our T3S frame work we use both of the classifiers even though both has drawbacks. T3S produce a set of positive and negative pairs contain a highly informative and more balanced which is used to feed the classification algorithm and to identify the fuzzy region position. T3S-NGram design to use the labeled sample pairs to detect where the matching and non-matching pairs are concentrated so that the threshold that removes the non-matching pairs can be selected. NGram tokenization is used to identify the matching pairs inside the fuzzy region called NGram threshold.

A similarity of each labeled pair will be computed. The labeled pairs are sorted incrementally by the similar value and a sliding window. The sliding window is relocated in one position until it detects the last windows with only non-matching pairs. Finally the similar value of the first matching pair encountered after the last windows with only non-matching pairs define the NGram threshold value, T3S is configured with NGram threshold value which is applied to all the fuzzy region pairs. So at last the candidate pairs that exist the filtering face and meet the NGram threshold value are considered as matching ones.

6.Conclusion

We have proposed T3S a two stage sampling strategy aimed at reducing the user labeling effort in deduplication. The first stage, T3S selects small random subsamples of candidate pairs in different fractions of data set. In second stage, subsamples are incrementally analyzed to remove redundancy. T3S is capable of reducing the user effort while keeping the same or better effectiveness. For future work, we intend to investigate genetic programming to combine similarity function and investigate whether is possible to give theoretical boundaries on how close our MTP and MFP boundary estimates are to the ideal values

7 Acknowledgement

We would like to appreciate to IJREAT for giving such wonderful platform as well as opportunity to the UG students to publish their paper. Also would like to thank to our Professor Mr. Selva Kumar, for his constant support and motivation for us. Our sincere thanks to our college, GKM COLLEGE OF ENGINEERING AND TECHNOLOGY, CHENNAI for providing a strong platform to develop our skill and capabilities

8.References

[1]. A.Arasu, M.GOtz, and R.Kaushik,"On active learning of record matching packages," in proc.ACM SIGMODInt.Conf.Manage.Data,2010,pp.783-794.

[2] P. Christen," A survey of indexing techniques for scalable record linkage and deduplication," IEEE Trans.Knowl. Data Eng., Vol.24,no.9,pp.1537-1555,Sep 2012.

[3] G.Dal Bianco,R.Galante,C.A Heuser, and M.A.Gonalves, "Tining large scale deduplication with reduced effort," in proc.25th Int.Conf.Scientific Statist.Database Manage., 2013,pp.1-12.

[5] J.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.

[4] R.M.Silva, M.A.Goncalves, and A.Veloso, "A two-stage active learning method for learning to rank," J.Assoc.Inform.Sci.Tech-nol.,vol.65,no. 1,pp.109-128,2014