A Survey on Data Deduplication in Large Scale Data

Saniya Sudhakaran
PG Student
Department of Computer Science & Engineering
FISAT
Angamaly, Ernakulam, India

Meera Treesa Mathews
Assistant Professor
Department of Computer Science & Engineering
FISAT
Angamaly, Ernakulam, India

ABSTRACT

This paper presents a survey on data deduplication on large scale data. deduplication is nothing but finding the duplicate records or duplicate data when compared with one or more data base or data sets.The data deduplication task has attracted a considerable amount of attention from the research community in order to provide effective and efficient solutions. Matching records from several data bases is known as record linkage. Those matched data contains important and useable information. These information is too costly to acquire because of which data deduplication process getting more attention day by day. Removing duplicate records during data cleaning process in a single database is a critical step, because the outcomes of subsequent data processing or data mining may get greatly influenced by duplicates.As database size increases day by day the matching process's complexity becoming one of the major challenges for data deduplication.
To overcome this problem we propose a Two Stage Sampling Selection (T3S) model which has two stages, in which, the strategy is proposed to produce balanced subsets candidate pairs which are to be labeled is done in the first stage and in the second stage we produced a smaller and more informative training sets than in the first stage.An active selection is incrementally invoked for removing the redundant pairs which are created in the first stage. This training set can be effectively used for identifying where the most ambiguous pairs lie and to configure the classification approaches. when compared with state-of-the-art deduplication methods in large datasets Our evaluation shows that T3S is able to reduce the labeling effort substantially while achieving a competitive or superior matching quality .

Keywords

Dedupliction,T3S,FS-Dedup

1. INTRODUCTION

Data deduplication is known as a specialized data compression technique for eliminating duplicate copies of repeating data.This technique is used to improve storage utilization and also be applied to network data transfers to reduce the number of bytes that must be sent.
A typical deduplication method is divided into three main phases: Blocking, Comparison, and Classification. The Blocking phase groups together pairs that share common features [1] which aims at reducing the number of comparisons.For avoiding quadratic generation of pairs ,Using a simplistic blocking approach then puts together all the records with the same first letter of the name and surname attributes in the same block. By applying some type of similarity function (e.g. Jaccard, Levenshtein, Jaro [2]) the Comparison phase quantifies the degree of similarity between pairs belonging to the same block.Finally, the Classification phase identifies which pairs are matching or non-matching. This phase can be carried out by means of global thresholds after selecting the most similar pairs , usually manually defined [3], [4], [5], [6] or learnt by using a classification model based on a training set.

To configure or tune the process in the case of large scale deduplication, the blocking and classification phases typically rely on the user . For instance, the classification phase requires a manually labeled training set. Thus,selecting and labeling a representative training set is a very costly task which is often restricted to expert users.The proposed method T3S is able to select a very small, non-redundant and informative set of pairs with high effectiveness for large scale datasets.

A rule-based active sampling strategy, which doesnt requires initial training set, which is incrementally applied to the selected subsamples to reduce redundancy.with mutual benefits for each other we demonstrate that the two steps of our method are complementary,While the second stage helps to remove redundancy, for the most informative pairs to be labeled the first stage allows the second one to concentrate on the most promising portions of the search space . As we noticed, just applying the rule based active sampling on the entire dataset is computationally infeasible for large datasets ,Thus to efficiently identify the position of the fuzzy region and configure suitable strategies to classify the most ambiguous pairs,T3S produces the final reduced training set which is then integrated with our previous framework .

After comparing T3S with FS-Dedup as well as with two state-of-the-art active learning methods for deduplication [7], [8] and one unsupervised approach [9]. Results obtained from synthetic and real datasets (one with about three million records) shows that when compared to FSDedup T3S reduces the training set size in about 16 times and also converges much faster to a high matching quality when compared to the baselines.

1.1 Benefits

(1) Storage-based data deduplication able to reduce the amount of storage required for a given set of files. It is more effective where many copies of very similar or even identical data are stored on a single disk.
(2) To reduce the number of bytes that must be transferred between endpoints ,Network data deduplication is used which can reduce the amount of bandwidth required.
(3) Virtual servers and virtual desktops benefit from deduplication because files for each virtual machine to be coalesced into a single storage space which allows nominally separate system .

1.2 Overview

Deduplication may occur ”in-line”, as data is flowing, or ”post-process” after it has been written.

(1) Post-process deduplication

In post-process deduplication, data is first stored on the storage device and then process at a later time will analyze the data looking for duplication. The benefit is that before storing the data there is no need to wait for the hash calculations and lookup to be completed , thereby ensuring that store performance is not degraded.

(1) In-line deduplication

Alternatively, as data enters the target device deduplication hash calculations can be done in real-time .Only a reference to the existing block is stored, rather than the whole new block if the storage system identifies a block which it has already stored .The advantage of in-line deduplication over post-process deduplication is that since duplicate data is never stored it requires only less storage . On the negative side, it is frequently argued because hash calculations and lookups take so long and data ingestion can be slower, thereby reducing the backup throughput of the device.

2. LITERATURE SURVEY

A.Arasu et al [10] Says In traditional learning a user selects the labelled examples where as in active learning algorithm takes the set of records to be labelled . active learning comes into picture when manually identifying suitable labels for records is difficult mainly important for record matching. The main objective of active learning is to maximize the recall under a precision constraint.An N-dimensional feature space composed of a set of similarity functions are created, that are manually defined, and actively selects the pairs by carrying out a binary search over the space.

limitations are as follows:

(1) They were not guaranteed for quality and not scaled for large input.
(2) Lack quality guarantees
(3) These are designed differently from traditional active learning approaches to discover the problem specific to record matching.

K.Bellare et al [7] Tells the fundamental issue in an entity matching while training a classifier to label the pairs of entities as either non duplicate or duplicate is a selecting informative example. The recent work address the issue that though active learning presents a feasible solution to problem, previous approaches minimizes the classifiers rate of misclassification, which is an unsuitable metrics for entity matching due to class imbalance. So as a solution to above problem it states to maximize recall of classification under the constraint that its precision should be greater than a specified threshold. However the proposed method also requires labelling all n input pairs in the worst case. The result of the paper is an active learning algorithm which approximately maximizes recall of the classifier with provably sub linear label complexity under a precision constraint.

P.Christen [9]Says linked data can contain information that would require time-consuming and expensive collection of specific data,or not available otherwise therefore linking databases is an important step in an increasing number of data mining projects . The aim of linking is to match and aggregate all records that refer to the same entity. while linking large databases efficient and accurate classification of record pairs challenges into matches and non-matches is one of the major Many of the more recently developed classification methods are based on supervised learning techniques when traditionally classification was based on manually-set thresholds or on statistical procedures. Therefore requires training data, which is often not available in real world situations or has to be prepared manually, an expensive, cumbersome and time-consuming process.

Author presents a novel two-step approach to automatic record pair classification. To train a support vector machine (SVM) classifier it takes training examples of high quality are automatically selected from the compared record pairs .After outperformed k-means clustering Initial experiments showed the feasibility of this approach achieving good results . Two variations of this approach are presented where first is based on a nearest-neighbour classifier.By iteratively adding more examples into the training sets the second improves a SVM classifier . Experimental results show that when compared with other unsupervised approaches this two-step approach can achieve better classification results.

G.Dal Bianco et al [12] Proposed a new framework called FS-Dedup which helps in tuning the deduplication process on large datasets with a reduced effort from the user, who is only required to label a small, automatically selected, subset of pairs. Deduplication identifies those objects which are potentially the same in a data repository. To identify some pairs representing matchings and non-matchings it usually demands user intervention in several steps of the process . These information is then used to help in identifying other potentially duplicated records. To configure the most important steps of the process (e.g., blocking and classification), the performance and matching quality depends on expert users when deduplication is applied to very large datasets.

Signature-Based Deduplication (Sig-Dedup) algorithms are exploited by FS-Dedup in its deduplication core. Sig-Dedup requires an expert user to tune several parameters while it is characterized by high efficiency and scalability in large datasets . FS-Dedup providing a framework that does not demand specialized user knowledge about the dataset or thresholds to produce high effectiveness that helps in solving drawback of sig-Dedup . Evaluation over large real and synthetic datasets (containing millions of records) shows that FS-Dedup is able to reach or even surpass the maximal matching quality obtained by Sig-Dedup techniques with a reduced manual effort from the user.

S.Sarawagi et al [8]Presents a learning-based deduplication system that uses a novel method of interactively discovering challenging training pairs using active learning. Deduplication integrates data from multiple sources. The main challenge to face is designing a function that can resolve when a pair of records refer to the same entity in spite of various data inconsistencies. As most existing systems use hand-coded functions.

To distinguish between duplicates and non-duplicates is to train a classifier thus to overcome the tedium of hand-coding . The success of this method critically hinges that able to provide a covering and challenging set of training pairs that bring out the subtlety of the deduplication function.As it requires manually searching for various data inconsistencies between any two records spread apart in large lists it is said to be non-trivial. From the experiments on real-life datasets that shows active learning reduces the number of instances required to achieve high accuracy and also investigated various design issues that arise in building a system to provide interactive response, fast convergence, and interpretable output.

3. T3S VERSUS FS-DEDUP

In This section we examined the results obtained when comparing the labeling effort and effectiveness of the proposed T3S using SVM and NGram classifiers with FS-Dedup using SVM and NGram classifier. FS-Dedup uses a random selection of pairs inside the fuzzy region to produce the training set, while in the first stage T3S uses a random selection to produce representative samples combined with an incremental active learning approach that can remove the redundant information.

In summary, the experiments show that T3S using NGram requires much fewer labeled pairs to achieve a stable F1 value in synthetic and real datasets than FS-Dedup using SVM. In the synthetic datasets, T3S-SVM depends on two similarity functions to achieve stable effectiveness. In the real datasets, both T3S-SVM and T3S-NGram (together with one similarity function and a level size of 100 pairs) achieve competitive effectiveness and use only 10 and 13 percent of the training set required by FS-Dedup in IMDBxNetFlix and DBLPxCiteseer datasets, respectively.

4. T3S VERSUS ALIAS,ALD AND CHRISTEN

A committee of classifiers are used by ALIAS[8] to actively select informative pairs in order to reduce labeling effort. ALD [7] works by looking for highly informative pairs among all unlabeled pair which corresponds to a recent active learning deduplication method, and Finally, we report experiments with the unsupervised approach proposed by Christen [9] in 2008.It can be noticed that T3S using N-Gram and SVM classifiers converge very quickly, producing good effectiveness with only a few manually labeled pairs (around 380, 103 and 31 pairs in Clean, IMDBxNetFlix and DBLPxCiteSeer datasets, respectively).In both real datasets T3S clearly outperforms ALIAS with a reduced labeling effort . also ALIAS has an unstable behavior at the beginning,while T3S starts with a high F1 value. ALIAS achieves a competitive F1 and label efforts when compared to T3S because these datasets have a pattern of dirtiness that can be easily identified by the decision tree classifier (i.e., the feature social security number is highly informative) in the synthetic dataset . Yet, in the very beginning, T3S is superior.

In case of labeling effort T3S can achieve reduction of 87 percent (765 pairs are selected by ALD compared with 103 selected by T3S), and still be around 6 percent better than ALD in terms of F1. Christen (2008) achieves 21, 68 and 70 percent in the Clean, IMDBxNetFlix and DBLPxCiteseer datasets, respectively. which shows very poor effectiveness.

In summary, T3S when compared with the baselines, it sharply reduces the training set size and, consequently, the manual labeling effort, while producing competitive or superior results in terms of effectiveness.

5. EFFICIENCY

T3S takes around 74 minutes in a single machine with about one trillion of potential comparisons.The most expensive operation represents Blocking , which consume around 70 percent (i.e., 47 minutes) of the execution time. we believe that the runtime can be reduced even further using a high performance cluster, as T3S can take advantage of the MapReduce paradigm.

6. CONCLUSION

In this paper we proposed T3S, a two-stage sampling strategy aimed at reducing the user labeling effort in large scale deduplication tasks. Initially,small random subsamples of candidate pairs are selected by T3S in different fractions of datasets. Secondly, to remove redundancy subsamples are incrementally analyzed .T3S is evaluated with synthetic and real datasets which shows comparison with four baselines, while keeping the same or a better effectiveness. T3S is able to considerably reduce user effort .

7. REFERENCES

[1] P. Christen, A survey of indexing techniques for scalable record linkage and deduplication," *IEEE Transactions on knowlwdge and data engineering*, 24, (2012)1537-1555.

[2] A. Elmagarmid, P. Ipeirotis, and V. Verykios, Duplicate record detection: A survey, *IEEE Transactions on knowlwdge and data engineering*, 19, (2007)1-16.

[3] R. J. Bayardo, Y. Ma, and R. Srikant, Scaling up all pairs similarity search, *proceedings of 16th international conference in world wide web*(2007)131-140

[4] S. Chaudhuri, V. Ganti, and R. Kaushik, A primitive operator for similarity joins in data cleaning,*proceedings in 22nd international conference in data engineering*,(2006)p.5.

[5] J. Wang, G. Li, and J. Fe, Fast-join: An efficient method for fuzzy token matching based string similarity join, *proceedings in IEEE 27th international conference in data engineering*(2011)458-469 .

[6] C. Xiao, W. Wang, X. Lin, J. X. Yu, and G. Wang, Efficient similarity joins for near-duplicate detection, *ACM transactions in database systems*,36,(2011)15:1-15:41

[7] K. Bellare, S. Iyengar, A. G. Parameswaran, and V. Rastogi, Active sampling for entity matching, *proceedings in 18th ACM SIGKDD international conference in knowledge discovery in data mining*,(2012)1131-1139.

[8] S. Sarawagi and A. Bhamidipaty, Interactive deduplication using active learning, *proceedings in 8th ACM SIGKDD international conference in knowledge discovery data mining*(2002)269-278.

[9] P. Christen, Automatic record linkage using seeded nearest neighbour and support vector machine classification, *proceedings in 14th ACM SIGKDD international conference in knowledge discovery data mining*(2008)151-159.

[10] A. Arasu, M. Gotz, and R. Kaushik, On active learning of record matching packages, *proceedings in ACM SIGMOD international conference in manage data*(2010)783-794.

[11] D. Cohn, L. Atlas, and R. Ladner, Improving generalization with active learning, *machine learning*15,(1994)201-221.

[12] G. Dal Bianco, R. Galante, C. A. Heuser, and M. A. Gonalves, Tuning large scale deduplication with reduced effort, *proceedings in international conference in scientific statist on database manage*(2013)1-12.