IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

A Survey Study for Deduplication in Large Scale Data

Supriya Allampallewar[1], Mr. J. Ratnaraja Kumar[2]

Computer Department, G.S. Moze College of Engineering, Pune, India [1]

HOD, Computer Department, G.S. Moze College of Engineering, Pune, India [2]

Abstract: The deduplication process is nothing but finding duplicate records or duplicate data when comparing with one or more data base or data sets. The process in which we match records from several data bases is known as record linkage. The matched data (which is out- put of whole deduplication process) contains important and useable information. This information is too costly to acquire because of which deduplication process getting more attention day by day. In data cleaning process removing duplicate records in a single database is a critical step, because outcomes of subsequent data processing or data mining may get greatly influenced by duplicates. As the database size increasing day by day the matching process's complexity becoming one of the major challenges for record linkage and deduplication. To overcome this in some extent we propose a Two Stage Sampling Selection (T3S) model in this article. Basically T3S has two stages, in which, in the first stage the strategy is proposed to produce balanced subsets candidate pairs which are to be labeled. In the second stage to produce smaller and more informative training sets than in the first stage an active selection is incrementally invoked so that redundant pairs get removed which are created in the first stage. We are extending our work in classification phase by using more advanced classification approach i.e. Adaboost algorithm. Several studies said that Adaboost gives better accuracy than SVM classifier. Our experimental results on real world dataset will show the comparative analysis of both methods, which proves that proposed method, performs better as compare to SVM. This document gives formatting instructions for authors preparing papers for publication in the Proceedings of an International Journal. The authors must follow the instructions given in the document for the papers to be published. You can use this document as both an instruction set and as a template into which you can type your own text.

Keywords: Deduplication, T3S, Adaboost.

I. INTRODUCTION

In IT business DATABASE is of great importance. Many operations and decisions are carried out on the basis of outputs of databases. Therefore a quality of information depends on the quality of data, implicitly methods which are used to store and to retrieve the data from database. The system which provides comprehensive view of the linking of relational terms or joining of two or more tables can be called as error free system. But unfortunately many time data lack a unique or global identifier which permits such operations. And along with this data are neither controlled nor defined in a consistent manner in a different data sources.

In deduplication process we identify references in data records which refer to the same real world entity.

It is one of the crucial steps in data cleaning process. In collective deduplication we want to find types of real world entities in a set of records which are related. It is a generalization of deduplication. For ideal collective deduplication scenario the example can be given as , if a database of paper references is given, the system will identify all records which refer to a single paper; it will also produce a set of all conferences in which the paper was published. In this situation the output will hold a constraint about a uniqueness of paper as the same paper is not published in several conferences. So in general we can say that the output of collective deduplication contains set of several partitions of the input records that satisfy constraints in the data. Most of the existing approaches towards deduplication are designed around string similarity.

In this paper large scale deduplication, the blocking and classification phases typically rely on the user to configure or tune the process. For instance, the classification phase usually requires a manually labeled training set. However, selecting and labelling a representative training set is a very costly task which is often restricted to expert users. Active learning approaches have been proposed to alleviate this problem.

In next section II we are presenting the literature survey. In section III, the proposed the existing system. In section IV we present the proposed system. Finally conclusion is predicted in section V.

II. LITERATURE SURVEY

Many researchers have worked on deduplication process; the literatures we refer for our work are explained as follows-

On active learning of record matching packages, A. Arasu, M. Gotz, and R. Kaushik [1], - The problem of learning a record matching package or classifier comes under active learning which is attended by the author in this paper.

There is some difference between traditional learning and active learning. One of them is, in active learning the learning algorithm takes the set of records to be labelled where as in traditional learning a user selects the labelled examples. Where manually identifying suitable labels for records is difficult, here active learning comes into picture, it is important for record matching. Limitations with previous active learning algorithm for record matching was they were not guaranteed for quality & not scaled for large input, therefore new algorithms are designed to overcome these problems. These are designed differently from traditional active learning approaches to discover the problem specific to record matching.

Large-scale deduplication with constraints using dedupalog, A. Arasu, C. R_e, and D. Suciu [2], The definite framework of entity references for collective deduplication with constraints is presented by author. Constraints occur naturally and may improve the deduplication quality. An example of constraints is "each paper has a unique paper publication location"; if two paper references are duplicates, then their associated conference references must be duplicates as well. This framework supports collective deduplication, meaning that we can deduplication both conference references and paper references collectively in the example above. The above frame work is based on precise semantics with declarative Datalog-style language. Constraints are either ignored or used in ad-hoc particular domain previously in deduplication. . Author also present efficient algorithms to support the framework. Their algorithms have precise guarantees for a large subclass of our framework theoretically. They show, using a prototype implementation that our algorithms scale to very large datasets. They provide experimental results over real-world data demonstrating the utility of our framework for the ease of high-quality and scalable deduplication.

Scaling up all pairs similarity search, R. J. Bayardo, Y. Ma, and R. Srikant[3], - here a 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 . 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. 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.

Active sampling for entity matching, K. Bellare, S. Iyengar, A. G. Parameswaran, and V. Rastogi [4], 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 classifier's 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.

The author shows complexity of their algorithm is at most log n times the label complexity and also the difference is bound in the recall. The evaluation of algorithm on several real world data sets is provided which shows the effectiveness of our approaches.

Importance weighted active learning, A. Beygelzimer, S. Dasgupta, and J. Langford [5]; here the author presents a statically consistent and practical scheme for actively learning binary classifiers under a general loss function. To correct sampling bias their proposed algorithm uses importance weighting. For learning process by, controlling the variance, they are able to give rigorous label complexity bounds.

III. EXISTING SYSTEM

A typical deduplication method is divided into three main phases: Blocking, Comparison, and Classification. The Blocking phase aims at reducing the number of comparisons. The Comparison phase quantifies the degree of similarity between pairs belonging to the same block, by applying some type of similarity function (e.g. Jaccard). Finally, the Classification phase identifies which pairs are matching or non-matching. This phase can be carried out by selecting the most similar pairs by means of global thresholds, usually manually defined.

The classification phase usually requires a manually labeled training set. However, selecting and labeling a representative training set is a very costly task which is often restricted to expert users. Hence this problem may cause effect on accuracy of applied classification criteria. Also existing techniques are more time consuming in case of deduplication detection process. We will overcome all stated problem in our proposed approach.

IV. PROPOSED SYSTEM

In this paper we proposed a new advance novel approach of T3S framework for finding large scale deduplication. Our proposed method has two stages for sampling. The proposed framework is able to select a very small, non-redundant and informative set of examples with high effectiveness for large scale datasets. In more details, in the second stage a rule-based active sampling strategy, which requires no initial training set (as required in classifier committees), is incrementally applied to the selected subsamples to reduce redundancy. We are extending this framework by proposing advance classification technique known as Adaboost classifier which uses SVM as a weak classifier and performs classification based on weak classifier result hence gives

DOI 10.17148/IJARCCE.2016.5218

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

more accuracy rate as compare to SVM or other existing classification approaches.

A. Architecture

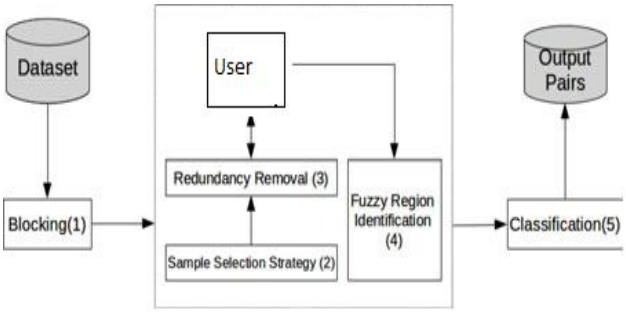

Fig1 A T3S steps overview

B. Algorithm

Adboost algorithm:

1. Start
2. Dataset load to system.
   S : {a1,a2,...,an}
   S – Dataset
   a1,a2,…,an – attributes of dataset (column names)
3. Weight assign to each attribute according their priorities.
   (Which attribute should take for consideration to find the attack? Attribute with higher priorities or weight will take first and so on.)
4. Labelling to each review by considering weight of attribute (positive or negative review)
5. Dataset will be prepared for classification with help of step 1 2 3.
6. Classification is done on basis of label of review.
7. .After classification degree of each label (positive or negative) gets calculated.
8. Compare the degree with threshold value
9. Result from step 7 show classification of dataset.
10. Stop.

C. Results of Practical Work

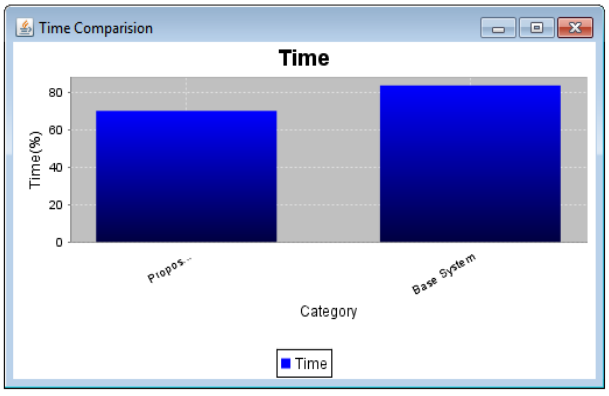

Fig. 2 Time comparison between proposed system and base system

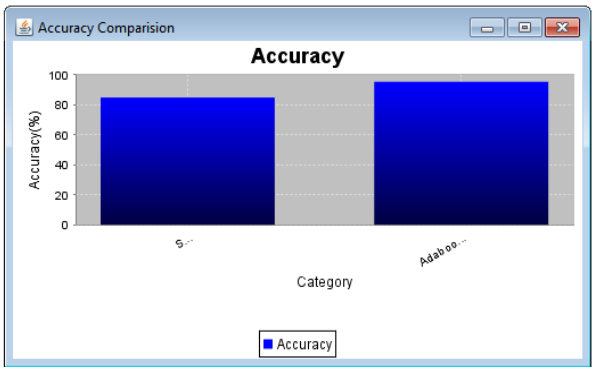

Fig. 3 Accuracy comparison between proposed system and base system

V. CONCLUSION

A two stage sampling strategy reduces the labeling effort of users in large scale deduplication tasks. The stage of T3S selects small sub samples randomly of candidate pairs where as in the second stage to remove redundancy sub samples are incrementally analyzed. In this work we have used Adaboost classifier instead of SVM classifier. The classifier which we have used gives more accuracy and less time than previous classifier.

REFERENCES

[1] A. Arasu, M. Gotz, and R. Kaushik, “On active learning of record matching packages,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2010, pp. 783–794.
[2] A. Arasu, C. R_e, and D. Suciu, “Large-scale deduplication with constraints using dedupalog,” in Proc. IEEE Int. Conf. Data Eng.,2009, pp. 952–963.
[3] R. J. Bayardo, Y. Ma, and R. Srikant, “Scaling up all pairs similarity search,” in Proc. 16th Int. Conf. World Wide Web, pp. 131–140, 2007.
[4] K. Bellare, S. Iyengar, A. G. Parameswaran, and V. Rastogi, “Active sampling for entity matching,” in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 1131–1139.
[5] A. Beygelzimer, S. Dasgupta, and J. Langford, “Importance weighted active learning,” in Proc. 26th Annu. Int. Conf. Mach. Learn., pp. 49–56, 2009.
[6] M. Bilenko and R. J. Mooney, “On evaluation and training-set construction for duplicate detection,” in Proc. Workshop KDD, 2003, pp. 7–12.
[7] S. Chaudhuri, V. Ganti, and R. Kaushik, “A primitive operator for similarity joins in data cleaning,” in Proc. 22nd Int. Conf. Data Eng., p. 5, Apr. 2006.
[8] P. Christen, “Automatic record linkage using seeded nearest neighbour and support vector machine classification,” in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008,pp. 151–159.
[9] 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.
[10] P. Christen and T. Churches, “Febrl-freely extensible biomedical record linkage,” Computer Science, Australian National University, Tech. Rep. TR-CS-02-05, 2002.
[11] D. Cohn, L. Atlas, and R. Ladner, “Improving generalization with active learning,” Mach. Learn., vol. 15, no. 2, pp. 201–221, 1994.
[12] G. Dal Bianco, R. Galante, C. A. Heuser, and M. A. Gonalves, “Tuning large scale deduplication with reduced effort,” in Proc. 25th Int. Conf. Scientific Statist. Database Manage. 2013, pp. 1–12.
[13] M. G. de Carvalho, A. H. Laender, M. A. Goncalves, and A. S. da Silva, “A genetic programming approach to record deduplication,” IEEE Trans. Knowl. Data Eng., vol. 24, no. 3, pp. 399–412, Mar. 2012.