International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 11 (2017) pp. 2670-2674

Sampling Selection Strategy for Large Scale Deduplication for Web Data Search

R. Lavanya[1*], P. Saranya[2], D. Viji[3]
[1]*Assistant Professor, Department of Computer Science Engineering, SRM University, Chennai, India.*
[2,3] *Assistant Professor, Department of Computer Science Engineering, SRM University, Chennai, India.*
**Corresponding author mail: lavanyaconf@gmail.com*

[1]***ORCID:0000-0001-6383-6209***

Abstract

The data quality can be reduced due to the presence of duplicate pairs with misspellings, abbreviations, conflicting data, and redundant entities. Deduplication process manually labeled pairs for large data sets is a complicated process. The quality of data cannot be guaranteed. The system reduces the set of pairs required in Deduplication process of large data sets. This helps in selection of complicated pairs to provide quality data for large dataset system. This paper proposes a strategy to identify the threshold to configure step focused on recall maximization. Selection step identifies the fuzzy region boundaries and define the fuzzy region boundaries to automatically select candidate pairs to be labeled by a non-expert user with reducing effort. After defining the fuzzy region boundaries, the pairs inside are sent to the Classification step. The set below the fuzzy region is discarded while the set above is automatically sent to the output as matching pairs. Classification step classifies the candidate pairs that belong to the fuzzy region as a matching or not matching pairs.

Keywords: Sampling selection; data set; De duplication; threshold; web search

INTRODUCTION

With the advent of technology there is a large amount of increase in data. 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. Similarly, if one wishes to perform collaborative filtering on data from sites such as Amazon, the algorithms need to scale to tens of millions of the users.The ability to check whether a new collected object already exists in data repository or a close version of it is an essential is an essential task to improve data quality. As the database size increasing day by day the matching process's complexity becoming one of the major challenges for quality of a deduplication process with a redundant data.

Data quality can be degraded mostly due to the presence of duplicate pairs with misspellings, abbreviations, conflicting data, and redundant entities, among other problems. For instance, a system designed to collect scientific publications on the Web to create a central repository e.g. CiteSeer, it may suffer a lot in the quality of its provided services, e.g., search or recommendation may not produce results as expected by the end user due to the large number of replicated or near-replicated publications dispersed on the Web (e.g., a query response composed mostly by duplicates may be considered as having low informative value)one of the potential drawback is that duplicate data may be unnecessarily stored for a short time, which can be problematic if the system is nearing full capacity. The number of distinct search queries issued over a single week to any large search engine is in the tens of millions the ability check. Blocking is essential to speed up the de-duplication on large datasets. The problem is how to configure it. Usually a direct intervention is used to tune the blocking method (e.g., by setting proper similarity thresholds), implying that in most cases a combination of both direct and indirect intervention has to be performed. For instance, 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.

RELATED WORK

A. Large-Scale Deduplication

Deduplication is the process of identifying references in data records that refer to the same real-world entity. It is a crucial step in the data cleaning process. This approach creates an N dimensional binary search leading to large number of pairs to be queried. Approach have been confined to much smaller datasets..

B. Reducing the Storage Burden via Data Deduplication

Deduplication identifies and eliminates redundant information, thereby reducing volumes. Technology savvy industries such as financial services, pharmaceuticals, and telecommunications are already adopting Deduplication. The

comparison is not completely fair since this uses a manually tuned blocking threshold.

C. Record matching over query results from multiple web databases

Record matching, which identifies the records that represent the same real-world entity, is an important step for data integration. these algorithms make internal use of generalization bounds that are often loose in practice, and they can thus end uprequiring far more labels than are really necessaryese algorithms make internal use of generalizationbounds that are often loose in practice, and they can thus end up requiring far more labels than are really necessary.

D. Automatic Record Linkage using nearest neighbour

Increasingly large amounts of data are being collected by many organizations, techniques that enable efficient mining of massive databases have in recent years attracted interest from both academia and industry. Sharing of large databases between organizations is also of growing importance in many data mining projects, as data from various sources often has to be linked and aggregated in order to improve data quality.

E. Tuning Large Scale Deduplication

Record deduplication is the task of identifying which objects are an potentially the same in data repositories. Although an old problem, it still continues to receive significant attention from the database community due to its inherent difficulty, especially in the context of large datasets. Deduplication has an important role in many applications such as the data integration.

SYSTEM METHODOLOGY

A. FS Dedup - A FRAMEWORK FORSIGNATURE-BASED DEDUPLICATION

In this section, we present our proposed framework signature-based deduplication, named FS Dedup, which is able to tune most of the deduplication process in large datasets with a reduced user effort. From the point of view of the user, frame based signature dedup can be seen as a single task, avoiding an expert user intervention in specific steps (i.e. blocking and classification phases). The non-expert user intervention is requested only to label a set pairs automatically selected by our framework. In the following, we provide an overview of dedup steps as 1.sorting step: in this step, the dataset is blocked to create a sorted set of candidate pairs, without user intervention. The challenge of such step is to avoid an excessive generation of candidate pairs. We propose a strategy to identify the threshold to configure this step focused on recall maximization.2 Selection step: identifies the fuzzy region boundaries. A greedy strategy to define the fuzzy region boundaries is proposed to automatically select candidate pairs to be labeled by a non-expert user with the goal of reducing effort. After defining the fuzzy region boundaries, the pairs inside the fuzzy region are sent to the Classification step. The set below the fuzzy region is discarded while the set above is automatically sent to the output as matching pairs.3. Classification step: classifies the candidate pairs that belong to the fuzzy region as a matching or not matching pairs.

B. Two-Stage Sampling Selection

Two-stage sampling selection aimed at selecting a reduced and representative sample of pairs in large scale deduplication. We integrate three tier system with fast sign dedup framework to reduce the user effort in the main deduplication steps (e.g. blocking and classification).

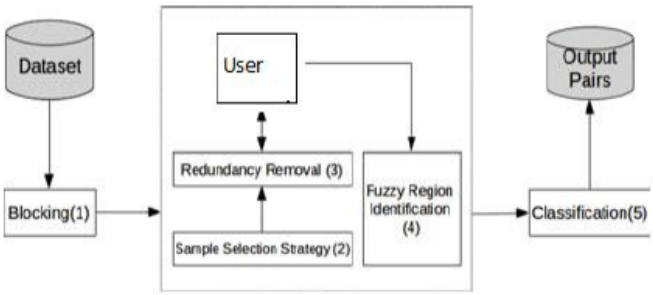

Figure 1. T3S Step Overview

First, a strategy is employed to identify the blocking threshold, and thus produce the candidate pairs. The dotted box represents the main steps of three tier system. In its first stage, produces small balanced subsamples of candidate pairs. In the second stage, the redundant information that is selected in the subsamples is removed by means of a rule-based active sampling which requires no previously labeled training set. Following this, we describe how these two steps work together to detect the boundaries of the fuzzy region. Finally, we describe our two classification approaches, also to introduced in, which configured by using the pairs manually to labeled in stages."

C. Sorting Step

The Sorting step identifies the blocking threshold using the Signature dedup filters (e.g., regarding the number of tokens to be used) that maximize recall that minimize the chance of pruning out actual matching pairs. We call this blocking threshold the initial threshold. Ideally, the set of candidate

pairs produced using the initial threshold contains all the matching pairs. As this step is performed without user intervention, we rely on generalizations as a means of becoming closer (or making an approximation) to the ideal scenario. In fact, the number of true matches and non-matches is not known a priori, but the initial thresholds are defined in order to minimize the number of "lost" matching pairs that are outside the interval for analysis. The other steps of our method are used to prune out the non-matching candidate pairs. It should be stressed that, also to avoid user intervention, the initial threshold represents a single global threshold for all the blocks. It is worth noting that the set of candidate pairs is produced using the Signature dedup filters (i.e., prefix, length, position, and suffix filtering) and that these are configured with the initial threshold. The main purpose of this threshold is to define how many tokens are indexed by the sorted record (i.e., the records are resorted using the global frequency of tokens. At the end, these candidate pairs are sorted using their similarity values to produce a ranking. In the next step, using this ranking it is possible to identify the pairs with the highest (true matching pairs) and lowest similarities (non matching pairs). This step represents a strategy to generate candidate pairs and categorize them. It makes it easier to select a specific pattern of pairs, i.e., highly positive or highly negative candidate pairs.

RESULTS AND DISCUSSION

A. Selection Step

The Selection step identifies the boundaries of the fuzzy region which, to be effectively defined, depends on two main factors: (i) the quality of the sample selection of candidate pairs to be manually labeled (ideally, the sample should be able to describe the factors to identify the fuzzy region) which should be representative of the whole dataset; and (ii) the expected manual labeling effort which should be minimized without an inaccurate boundary definition The sample selection strategy creates a balanced set of candidate pairs. We propose to discretize the ranking of candidate pairs generated in the Sorting step into fixed levels, in order to avoid that non-matching pairs dominate the sample selection.

Algorithm 1. SSAR: Rule-based Active Selective Sampling

Require: Unlabeled set T and σ_{min} (≈ 0)
Ensure: The training set D
1: **while true do**
2: **for all** $u_i \in T$ **do**
3: $D_{u_i} \leftarrow D$ projected according to u_i
4: $R_{u_i} \leftarrow$ extract useful rules from D_{u_i}
5: **end for**
6: **if** $D = \emptyset$ **then**
7: $\lambda_{u_i} \leftarrow u_i$ such that u_i is the most representative item of T.
8: **else**
9: $\lambda_{u_i} \leftarrow u_i$ such that $\forall_{u_j} : |R_{u_i}| \leq |R_{u_j}|$
10: **end if**
11: **if** $\lambda_{u_i} \in D$ **then**
12: **break**
13: **else**
14: LabelPair(λ_{u_i})
15: $D \leftarrow D \cup \{\lambda_{u_i}\}$
16: **end if**
17: **end while**

Figure 2. Rule based active selective sampling

Algorithm 2. Active Fuzzy Region selection

Require: Set of levels $L = l_1, l_2, l_3 \ldots l_9$
1: $i \leftarrow 0$; $MFP \leftarrow Null$; $MTP \leftarrow Null$; $trainingSet \leftarrow NULL$;
2: **for** $i = 0 \rightarrow 10$ **do**
3: $trainingSet \leftarrow SSAR(L_i, trainingSet)$
4: $i \leftarrow i + 1$
5: **end for**
6: **for** $i = 0 \rightarrow 10$ **do**
7: **if** L_{Pi} does not contains only *False* and $MTP = Null$ **then**
8: $MTP \leftarrow SelectLowestTruePair(L_{Pi})$;
9: continue;
10: **end if**
11: **if** L_{Pi} does not contains only *true* and $MTP! = Null$ **then**
12: $MFP \leftarrow SelectHighestFalsePair(L_{Pi})$;
13: **end if**
14: **end for**
15: **return** MTP, MFP and L_P;

Figure 3. Active Fuzzy Regionselection

The fixed levels contain a subset of candidate pairs, making easier to determine the boundaries of the fuzzy region. More specifically, the ranking, created in the Sorting step is fragmented into 9 levels (0.1-0.2, 0.2-0.3,..., and 0.9-1.0),

using the similarity value of each candidate pair. Inside each level, we randomly select candidate pairs to create the sample set to be manually labeled. approaches based on committees; (2) having a clear stopping criteria, a property that many approaches do not possess; and (3) the capability of selecting very few but very informative instances on an in formativeness criteria grounded on lazy association rules. When compared with the current training set, the unlabeled pair with less classification rules over the projected training set represents the most informative pair. A detailed example of this part of the rule based active selective sampling algorithm is shown below. Details of SSAR are shown in Algorithm 1. At each round, an unlabeled pair ui is used as a filter to remove irrelevant features and examples from D. In other words, the projected training data Dui is obtained after removing all the feature values that are not present in ui (line 3). Next, a specific classification rule-set Rui is extracted from dui. The number of rules created by each projected set represents its informativeness. The objective of this procedure is to select the most dissimilar unlabeled pair by making a comparison with the current training set. The Unlabeled pairs composed of a considerable number of common features compared with the current training set produce a large number of rules, showing that they provide the low information gain.

B. Detecting the Fuzzy Region Boundaries

We describe in detail the proposed approach for detecting the fuzzy region:

Definition 3. Let Minimum True Pair-(MTP) represent the matching pair with the lowest similarity value among the set of candidate pairs.

Definition 4. Similarly, let Maximum False Pair-(MFP) represent the non-matching pair with the highest similarity value among the set of non-matching pairs.

The fuzzy region is detected by using manually labeled pairs. The user is requested to manually label pairs that are selected incrementally by the SSAR from each level. However the pairs labeled by the user may result in MTP and MFP pairs which are far from the expected positions, as specified in Definitions 3 and 4. To minimize this problem, we assume that the levels to which the MTP or MFP pairs belong are defined within fuzzy region boundaries. For instance, if the MTP and MFP values are 0.35 and 0.75respectively, all the pairs with a similarity value between0.3 and 0.8 belong to the fuzzy region.

We call the fuzzy region boundaries a and b. Algorithm 2 identifies the fuzzy region boundaries by using the T3S strategy. First, SSAR is invoked to identify the informative pairs incrementally inside each level to produce a reduced training set (lines 2-5). The pairs labeled within a CH level are used to identify the MFP and MTP pairs. The pair labeled as true that has the lowest similarity value defines the MTP (line 8), then, the following levels are analyzed to identify the non-matching pair with the highest similarity value (line 12).It should be noted that the information that can be used at the lowest levels to identify the minimum true pairs represents the most dissimilar pairs. It should be noted that the information that can be used at the lowest levels to identify the minimum true pairs represents the most dissimilar pairs.

In this scenario, the large numbers of non-matching pairs that are present at this level are highly redundant and not informative to identify the fuzzy region boundaries. Thus, our strategy is mainly concerned with the selection of the dissimilar pairs, which are exactly the most informative means of identifying the a and b.

C. Classification Step

The Classification step aims at categorizing the candidate pairsbelonging to the fuzzy region as matching or non-matching.We use two classifiers in this step three tier ngram and three tier svm. Three tier svm maps each record toa global sorted token set and then applies both the Sig-Dedupfiltering and a defined similarity function (such as Jaccard) tothe sets. The token set does not consider the attribute positions,by allowing an exchange of attribute values. The drawback of three tier ngram is that different attributes are given thesame importance.In otherwords, an unimportant attribute value with a large lengthmay dominate the token set, and lead to distortions in thematching. On the other hand,three tier svm assigns differentweights to different attributes of the feature vector, by usingthesvm algorithm, based on their relative discriminative power four.

However, there is not an unique and globally suitablesimilarity function that can be adapted to differentapplications,and this makes it difficult to configure themethod for different situations. Moreover, long text attributescan be mapped to non-appropriated feature valuescausing a loss of information in the classification process. Asboth methods have advantages and drawbacks, we makeuse of both of them.highly informative and more balancedset of positive and negative pairs that is used forboth: to feed the classification algorithm and to identify thefuzzy region position.

CONCLUSION

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. We presented a strategy to identify the optimal configuration on large scale deduplication. In the first stage, selection little arbitrary subsamples of applicant pairs in dissimilar fractions of datasets. In the second,

subsamples are incrementally analyzed to take away redundancy. It identified the fuzzy region boundaries and define the fuzzy region boundaries to automatically select candidate pairs to be labeled by a non-expert user with reducing effort. The set below the fuzzy region is discarded while the set above is automatically sent to the output as matching pairs. For future work, genetic programming might be combined to check the similarity function to provide ideal values.

REFERENCES

[1] A. Arasu, M. Gotz, and R. Kaushik, "On active learning of recordmatching 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 withconstraints 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 similaritysearch," 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 constructionfor duplicate detection," in Proc. Workshop KDD, 2003,pp. 7–12.

[7] S. Chaudhuri, V. Ganti, and R. Kaushik, "A primitive operator forsimilarity joins in data cleaning," in Proc. 22nd Int. Conf. Data Eng.,p. 5, Apr. 2006.

[8] P. Christen, "Automatic record linkage using seeded nearestneighbour and support vector machine classification," in Proc.

[9] 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008.

[10] P. Christen, "A survey of indexing techniques for scalable record

[11] linkage and deduplication," IEEE Trans. Knowl. Data Eng., vol. 24,no. 9, pp. 1537–1555, Sep. 2012.

[12] P. Christen and T. Churches, "Febrl-freely extensible biomedicalrecord linkage," Computer Science, Australian National University,Tech. Rep. TR-CS-02-05, 2002.

[13] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization withactive learning," Mach. Learn., vol. 15, no. 2, pp. 201–221, 2010.

[14] 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, 1–12, 2013.

[15] A. Elmagarmid, P. Ipeirotis, and V. Verykios, "Duplicate record detection: A survey," IEEE Trans. Knowl. Data Eng., 19(1), 1–16, 2007.