IJCSN International Journal of Computer Science and Network, Volume 5, Issue 3, June 2016
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 1.02

Effective Sampling Selection Strategy with Reduced Effort Implied, In Tuning Large Scale Deduplication

[1] Ashwini R. , [2] Sridevi S.

[1] PG Scholar, Department of Computer Science and Engineering,
Vels University, Chennai, India.

[2] Assistant Professor, Department of Computer Science and Engineering,
Vels University, Chennai, India

Abstract - The deduplication process is always given by a set of manually labeled pairs. But in a very large datasets, producing manually labeled pairs is a tedious process to complete. So in this article, a two-stage sampling selection procedure that reduces the set of pairs to tune the deduplication process is proposed. T3S executes in two stages. In the first stage a balanced subset of data are produced for labeling. In the next stage, the redundant and the duplicated data are removed and only the deduplicated data are produced as the output.

Keyword - *Deduplication, FS-Dedup, T3S.*

1. Introduction

There is an immense growth in the generation of information from various sources like mobile phones, gadgets, social media etc. The capability to check whether the object collected newly already exists in the database is very important to improve the data quality. The data quality can be considerably improved by detecting and removing duplicates. To remove duplicates, in this paper, the deduplication method is used in three phases

i. Blocking
ii. Comparison
iii. Classification

The Blocking phase reduces the number of comparisons by grouping the pairs together, which share the common features. The Comparison phase states the similarity level between the pairs which belong to the same block. The matching and non-matching pairs are identified in the Classification phase. The blocking and classification phase fully depend on the user, as the user has to tune the process in a large scale deduplication. A framework named FS-Dedup, which was designed to give the closely related results according to the user search criteria is used for large scale deduplication tasks with reduced effort.

In this paper, a new two-stage sampling selection strategy (T3S) for deduplication is introduced. The method which is proposed has the ability to select a small, non-duplicated and a informative site with high efficiency of a large scale database. The final reduced deduplicated set produced by T3S is then merged with FS-Dedup framework to find the fuzzy region position efficiently, so that the most unclear or ambiguous pairs are classified.

2. Methodology

Methodology is the process of analyzing the principles and procedure. The following are the four modules involved in the deduplication task efficiently.

1) Dataset and evaluation metrics
2) Identifying the Approximate Blocking Threshold
3) Two-stage sampling selection
4) Detecting the Fuzzy Region and Classification

3. Two Stage Sampling Strategy

In this section, T3S, a two-stage sampling strategy which aims at reducing the user labeling effort in large scale deduplication tasks is introduced. In the first stage, T3S selects small random subsamples of candidate pairs in different fractions of datasets. This stage produces output by performing a random selection of pairs in each level.

In the second, subsamples are incrementally analyzed to remove redundancy.

3.1 Architecture

In this paper, the system has a dataset by default while working offline or the dataset is generated online. The data are collected from the dataset. The dataset are blocked in the blocking phase. During the blocking phase the redundant date are blocked but all the redundant data are not blocked.

The data from the blocking phase enters the Sample Selection Strategy phase. In the sample selection strategy phase the search criteria of the user has to match with the strategy given. The duplicated are now not in the top of the search results. The data from this phase enters the redundancy removal phase where the whole duplicated data are not in the search results which the user needs. So the result which comes as the output after this phase is the deduplicated data.

Now the deduplicated data is sent for the fuzzy region identification. After this stage the fuzzy region is identified by knowing and classifying the ambiguous pairs. The fuzzy region being identified the data is now classified in the classification phase. After the classification phase the end data is given as the output.

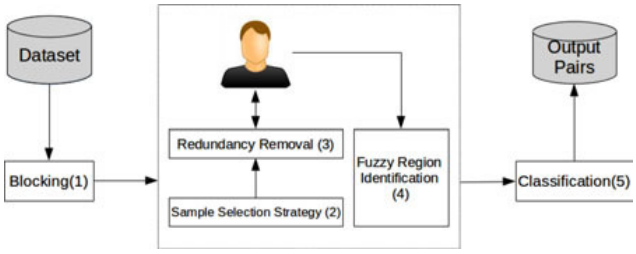


Fig 1.Architecture Diagram

4. Dataset and Evaluation Metrics

This module generates dataset. This module, it generates two types of dataset.

- **A. Synthetic**
- **B. Real**

Both synthetic and real datasets for out framework evaluation are used. The real datasets are created by combining two different real datasets from the same domain to produce a deduplication output. As the real datasets do not have proper standards, the synthetic datasets are used to create controlled output in order to evaluate the methods in a better manner. Since all solutions are compared under the same conditions, it is believed that the experimentation is balanced.

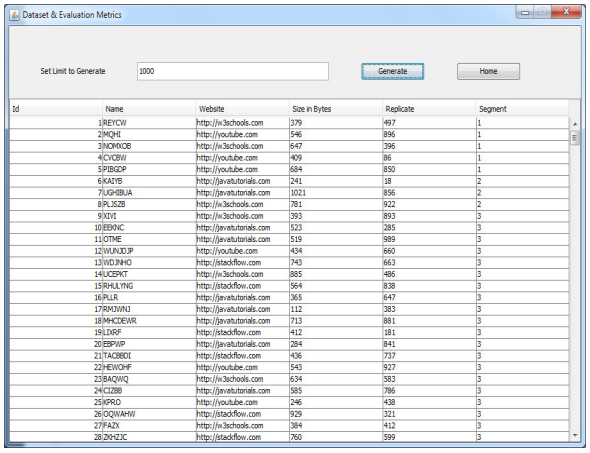


Fig 2. Dataset & Evaluation

5. Identifying The Approximate Blocking Threshold

In this step, the approximate blocking threshold is determined by using the Sig-Dedup filters that maximize recall, i.e., that minimize the chance of reducing all the actual matching pairs. This blocking threshold is called the initial threshold. Ideally, all the matching pairs are included in the set of candidate pairs produced using the initial threshold.

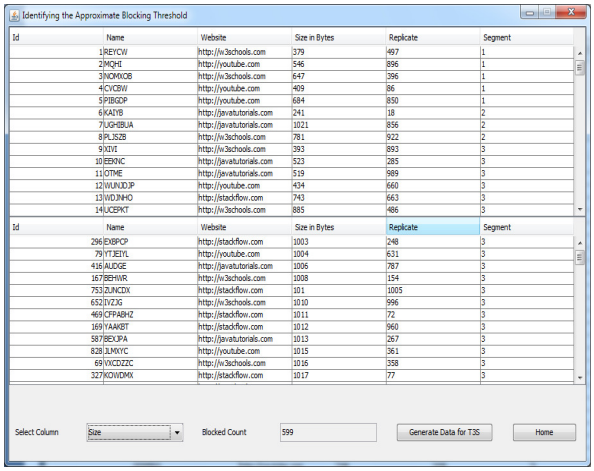
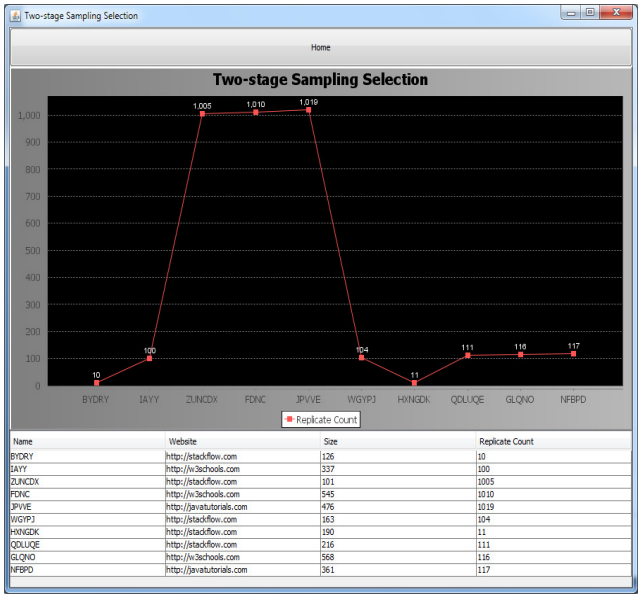


Fig 3. Identifying the Approximate Blocking Threshold

6. Two-Stage Sampling Selection

In this section, the proposed two-stage sampling selection which is aimed at selecting a reduced sample of pairs in large scale deduplication is outlined. T3S is merged with the previous FS-Dedup framework so that the user effort

in the main deduplication steps is reduced (e.g. blocking and classification).



Name	Website	Size	Replicate Count
BYDRY	http://stackflow.com	126	10
IAYY	http://w3schools.com	337	100
ZUNCDX	http://stackflow.com	101	1005
FDNC	http://w3schools.com	545	1010
JPVVE	http://javatutorials.com	476	1019
WGYPJ	http://stackflow.com	163	104
HXNGDK	http://stackflow.com	190	11
QDLUQE	http://stackflow.com	216	111
GLQNO	http://w3schools.com	568	116
NFBPD	http://javatutorials.com	361	117

Fig 4.Two Stage Sampling Selection

7. Detecting The Fuzzy Region & Classification

In this section, how the training set created by the two stages of T3S is detailed. Describing in detail the proposed approach for detecting the fuzzy region.

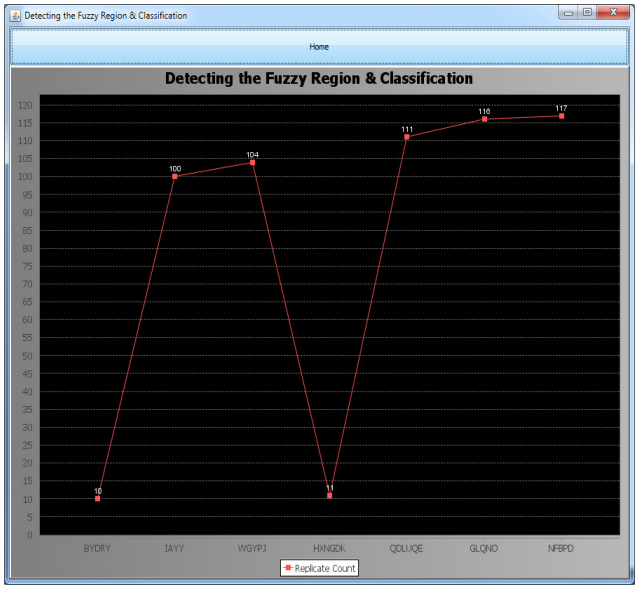


Fig 5.Detecting the Fuzzy Region & Classification

8. Conclusion

The proposed T3S, a two-stage sampling strategy aims at reducing the user labeling effort in large scale deduplication tasks. In the first stage, T3S selects small random subsamples of candidate pairs in different fractions of datasets. In the second stage, subsamples are analyzed incrementally to remove duplicated data. T3S with synthetic and real datasets is evaluated and showed that, in comparison with four modules, T3S is able to considerably reduce user effort in the large scale deduplication tasks.

References

[1] Guilherme Dal Bianco, Renata Galante, Marcos Andre Goncalves, Sergio Canuto, and Carlos A. Heuser "A Practical and Effective Sampling Selection Strategy for Large Scale Deduplication" IEEE Transactions on knowledge and data Engineering, Vol. 27, no. 9, September 2015

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] A. Beygelzimer, S. Dasgupta, and J. Langford, "Importance weighted active learning," in Proc. 26th Annu. Int. Conf. Mach. Learn., pp. 49–56, 2009.

[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.