DETECTING THE COGNATE DATA USING DEDUPLICATION FOR LARGE SCALE

Avni S Galathiya
Lecturer in Computer Engineering
R C Technical Institute, Ahmedabad

ABSTRACT: Databases and database technology have a major impact on the growing use of computers. A good database system would bring rapid advancement in the organization. Database plays a essential role approximately all sections where computers are used .Database quality is demeaned due to the presence of replicate information. Large speculations are made by the organizations to clean the clone created from the repository .Normal search is a time taking process. So, we can replace selection search instead of Normal search so that we can degrade user effort and time. We can use this technique in number of applications. By this technique we can also collect most user wanted data and compress the storage of database. Finally, we investigate the performance analysis through representation of graphs.

I. INTRODUCTION

We have witnessed a dramatic growth in the generation of information from a wide range of sources such as mobile devices, streaming media, and social networks. This has opened opportunities for the emergence of several new applications such as comparison shopping websites digital libraries and media streaming. These applications presuppose high quality data to provide reliable services. However, 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 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). The ability to check whether a new collected object already exists in the data repository (or a close version of it) is an essential task to improve data quality.

II. EXISTING SYSTEM

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 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. A typical de-duplication method is divided into three main phases: Blocking, Comparison, and Classification. The Blocking phase aims at reducing the number of comparisons by grouping together pairs that Share common features.

Disadvantages:
1. Training set is very costly.
2. Learning pairs are not informative.

III. PROPOSED SYSTEM

We have successfully proposed the FS-Dedup framework, designed to select the “closeto-optimum configuration” for large scale de-duplication tasks with reduced user effort. A heuristic was proposed to select a balanced and informative set of candidate pairs to be labeled by the user to accurately identify the boundaries of the fuzzy region. The labeling effort is basically concentrated in a random selection of pairs inside the fuzzy region (the “challenging” pairs for the classification process). The proposed sampling within fixed similarity levels creates balanced subsamples, thus avoiding a sample selection bias while reducing the potential size of the training set. This allows more useful information to be obtained for the classification process in a faster pace. FS-Dedup was demonstrated to be more effective than manually tuned methods, while still reducing labeling efforts. However, the resulting subsamples may still be composed of redundant pairs, with negative impacts in the labeling effort.
Advantages:
1. Reducing the redundancy in the subsamples.
2. Reducing training set cost using informative pairs.
It gives faster results compare to previous approaches.
MODULES:
1. Identifying The Approximate Blocking Threshold
2. Sample Selection Strategy
3. Redudancy Removal
4. Detecting Fuzzy Region Boundaries

Modules Description:
Identify 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 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.

Sample Selection Strategy

Our proposed sample selection strategy to produce balanced subsamples of candidate pairs. The main idea of the first stage is to discretize the ranking (produced in the previous step) so that small subsets of candidate pairs can be selected to reduce the computational demand of the T3S second stage. A simplistic approach to produce samples might be to select random pairs within the set of candidate pairs. However, as the set of pairs is basically formed of non-matching pairs, this kind of approach will result in samples having low informativeness. In this way, the second stage of T3S will hardly be able to select representative samples, because of the lack of in formativeness in the matching pairs.

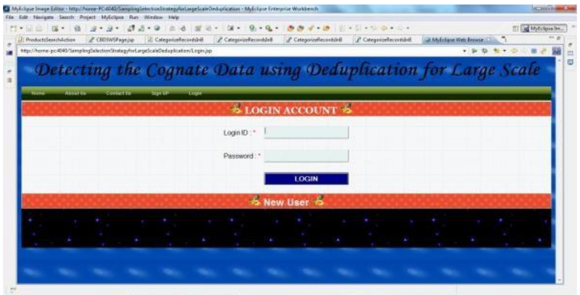
Redudancy Removal

The first stage produces samples by carrying out a random selection of pairs inside each level. As we observed in, the subsamples are an effective means of detecting the fuzzy region boundaries, especially when the size of the level is quite large. However, several pairs selected inside each level are composed of redundant information which does not help to increase the training set diversity. The second stage of T3S aims at incrementally removing the noninformative or redundant pairs inside each sample level by using the SSAR (Selective Sampling using Association Rules) active learning method. By redundant, we mean pairs carrying very similar information; the inclusion of a redundant pair in the training set for the classification step does not contribute with useful information for the learning process.

Detecting Fuzzy Region Boundaries

We detail how the training set created by the two stages of T3S is able to detect the fuzzy region boundaries. We describe in detail the proposed approach for detecting the fuzzy region: 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.

SCREEN SHOTS


MyEclipse Image Editor - http://home-PC4040/SamplingSelectionStrategyforLargeScaleDeduplication - MyEclipse Enterprise Workbench
File Edit Navigate Search Project MyEclipse Run Window Help
ProductsSearchAction
CBDSWSPage.jsp
CategorizeRecordsInB
MyEclipse Web Browse
Home
About Us
Contact Us
Sign UP
Login
LOGIN
New User

http://home-pc4040/SamplingSelectionStrategyforLargeScaleDeduplication/Login.jsp

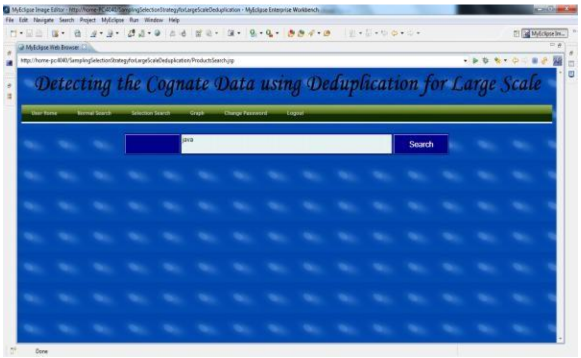
Detecting the Cognate Data using Deduplication for Large Scale

LOGIN ACCOUNT

Login ID : *

Password : *

Figure: User Login

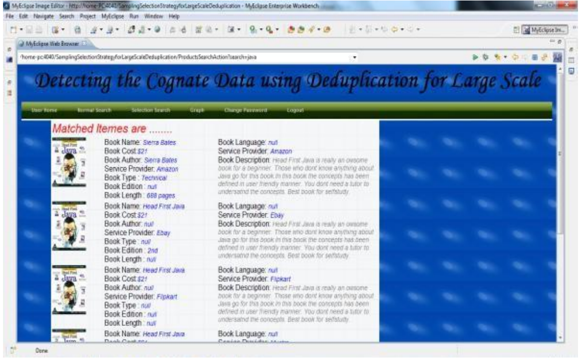

MyEclipse Image Editor - http://home-PC4040/SamplingSelectionStrategyforLargeScaleDeduplication - MyEclipse Enterprise Workbench
File Edit Navigate Search Project MyEclipse Run Window Help
MyEclipse Web Browser
User Home
Normal Search
Selection Search
Graph
Change Password
Logout
Search
Done

http://home-pc4040/SamplingSelectionStrategyforLargeScaleDeduplication/ProductsSearch.jsp

Detecting the Cognate Data using Deduplication for Large Scale

java

Figure : Normal Search


MyEclipse Image Editor - http://home-PC4040/SamplingSelectionStrategyforLargeScaleDeduplication - MyEclipse Enterprise Workbench
File Edit Navigate Search Project MyEclipse Run Window Help
MyEclipse Web Browser
User Home
Normal Search
Selection Search
Graph
Change Password
Logout
Done

home-pc4040/SamplingSelectionStrategyforLargeScaleDeduplication/ProductsSearchAction?search=java

Detecting the Cognate Data using Deduplication for Large Scale

Matched Itemes are ....

Book Name: Sierra Bates
Book Cost $21
Book Author: Sierra Bates
Service Provider: Amazon
Book Type : Technical
Book Edition : null
Book Length : 688 pages

Book Language: null
Service Provider: Amazon
Book Description: Head First Java is realy an owsome book for a beginner. Those who dont know anything about Java go for this book.In this book the concepts has been defined in user friendly manner. You dont need a tutor to undersatnd the concepts. Best book for selfstudy.

Book Name: Head First Java
Book Cost $21
Book Author: null
Service Provider: Ebay
Book Type : null
Book Edition : 2nd
Book Length : null

Book Language: null
Service Provider: Ebay
Book Description: Head First Java is realy an owsome book for a beginner. Those who dont know anything about Java go for this book.In this book the concepts has been defined in user friendly manner. You dont need a tutor to undersatnd the concepts. Best book for selfstudy.

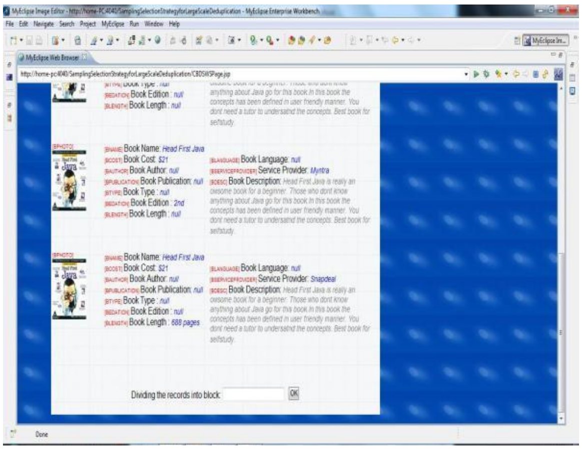
Book Name: Head First Java
Book Cost $21
Book Author: null
Service Provider: Flipkart
Book Type : null
Book Edition : null
Book Length : null

Book Language: null
Service Provider: Flipkart
Book Description: Head First Java is realy an owsome book for a beginner. Those who dont know anything about Java go for this book.In this book the concepts has been defined in user friendly manner. You dont need a tutor to undersatnd the concepts. Best book for selfstudy.

Book Name: Head First Java

Book Language: null

Figure: Normal Search Results


MyEclipse Image Editor - http://home-PC4040/SamplingSelectionStrategyforLargeScaleDeduplication - MyEclipse Enterprise Workbench
File Edit Navigate Search Project MyEclipse Run Window Help
MyEclipse Web Browser
OK
Done

http://home-pc4040/SamplingSelectionStrategyforLargeScaleDeduplication/CBDSWSPage.jsp

[BEDATION] Book Edition : null
[BLENGTH] Book Length : null

anything about Java go for this book.In this book the concepts has been defined in user friendly manner. You dont need a tutor to undersatnd the concepts. Best book for selfstudy.

[BPHOTO]
[BNAME] Book Name: Head First Java
[BCOST] Book Cost: $21
[BAUTHOR] Book Author: null
[BPUBLICATION] Book Publication: null
[BTYPE] Book Type : null
[BEDATION] Book Edition : 2nd
[BLENGTH] Book Length : null

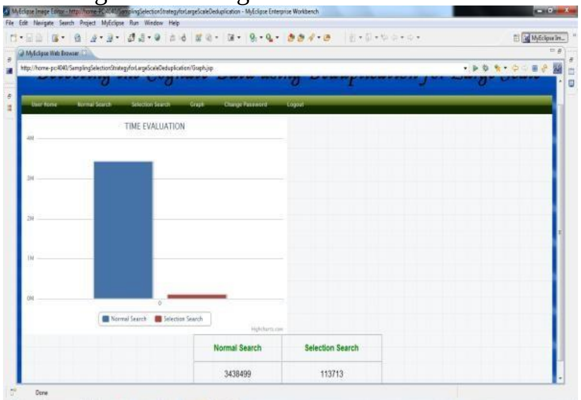
[BLANGUAGE] Book Language: null
[BSERVICEPROVIDER] Service Provider: Myntra
[BDESC] Book Description: Head First Java is realy an owsome book for a beginner. Those who dont know anything about Java go for this book.In this book the concepts has been defined in user friendly manner. You dont need a tutor to undersatnd the concepts. Best book for selfstudy.

[BPHOTO]
[BNAME] Book Name: Head First Java
[BCOST] Book Cost: $21
[BAUTHOR] Book Author: null
[BPUBLICATION] Book Publication: null
[BTYPE] Book Type : null
[BEDATION] Book Edition : null
[BLENGTH] Book Length : 688 pages

[BLANGUAGE] Book Language: null
[BSERVICEPROVIDER] Service Provider: Snapdeal
[BDESC] Book Description: Head First Java is realy an owsome book for a beginner. Those who dont know anything about Java go for this book.In this book the concepts has been defined in user friendly manner. You dont need a tutor to undersatnd the concepts. Best book for selfstudy.

Dividing the records into block:

Figure: Dividing the Records into Blocks


MyEclipse Image Editor - http://home-PC4040/SamplingSelectionStrategyforLargeScaleDeduplication - MyEclipse Enterprise Workbench
File Edit Navigate Search Project MyEclipse Run Window Help
MyEclipse Web Browser
User Home
Normal Search
Selection Search
Graph
Change Password
Logout
TIME EVALUATION
4M
3M
2M
1M
0M
0
Normal Search
Selection Search
Highcharts.com
Done

http://home-pc4040/SamplingSelectionStrategyforLargeScaleDeduplication/Graph.jsp

Normal Search	Selection Search
3438499	113713

Figure: Graph

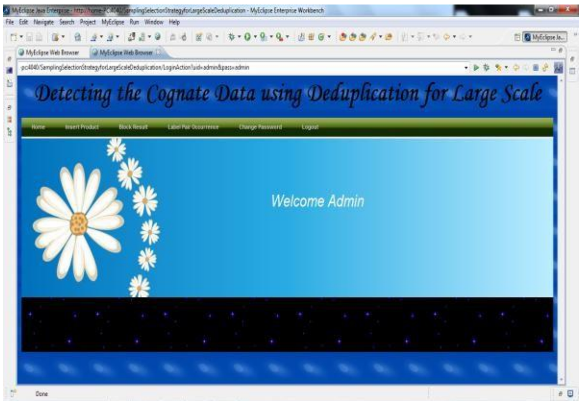
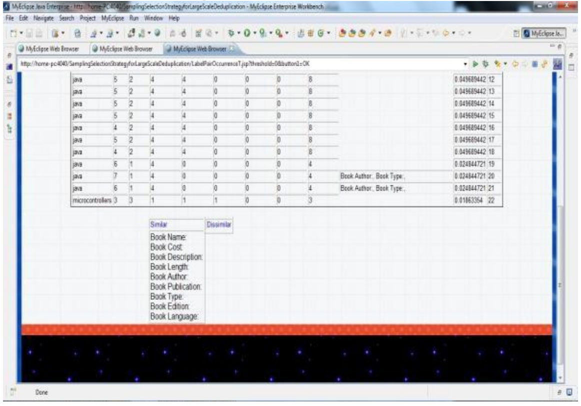


Figure: Admin Home



http://home-pc4040/SamplingSelectionStrategyforLargeScaleDeduplication/LabelPairOccurrenceT.jsp?threshold=0&button1=OK

java	5	2	4	4	0	0	0	8		0.049689442	12
java	5	2	4	4	0	0	0	8		0.049689442	13
java	5	2	4	4	0	0	0	8		0.049689442	14
java	5	2	4	4	0	0	0	8		0.049689442	15
java	4	2	4	4	0	0	0	8		0.049689442	16
java	5	2	4	4	0	0	0	8		0.049689442	17
java	4	2	4	4	0	0	0	8		0.049689442	18
java	6	1	4	0	0	0	0	4		0.024844721	19
java	7	1	4	0	0	0	0	4	Book Author:, Book Type:,	0.024844721	20
java	6	1	4	0	0	0	0	4	Book Author:, Book Type:,	0.024844721	21
microcontrollers	3	3	1	1	1	0	0	3		0.01863354	22

Similar	Dissimilar
Book Name: Book Cost: Book Description: Book Length: Book Author: Book Publication: Book Type: Book Edition: Book Language:	

Figure: Classification of similar and dissimilar items

IV. CONCLUSION AND FUTURE ENHANCEMENTS

We have proposed T3S, a two-stage sampling strategy aimed 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, subsamples are incrementally analyzed to remove redundancy. We evaluated T3S with synthetic and real datasets and empirically showed that, in comparison with four baselines, T3S is able to considerably reduce user effort while keeping the same or a better effectiveness. For future work, we intend to investigate genetic programming to combine similarity functions and investigate whether is possible to provide theoretical boundaries on how close our MTP and MFP boundary estimates are to the ideal values.

REFERENCES

[1] “Tuning large scale deduplication with reduced effort” G. Dal Bianco, R. Galante, C. A. Heuser, and M. A. Gonalves ,in Proc. 25th Int. Conf. Scientific Statist. Database Manage., 2013, pp. 1–12.

[2] “Active Sampling for entity matching” K.Bellare, S. Iyengar ,A. andV. Rastogi,in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 1131–1139

[3] ”Agenetic programming approach torecordde duplication,”M. G. de Carvalho, A. H. Laender, A. Goncalves, and A. S. da Silva IEEETrans. Knowl. DataEng., vol.24,no. 3, pp. 399–412, Mar.2012.

[4] “On active learning of record matching packages,” A. Arasu, M. Gotz, and R. Kaushik, in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2010, pp. 783–794.

[5] ”Large-scale deduplication with constraints using dedupalog,” A. Arasu, C. Re, and D. Suciu in Proc. IEEE Int. Conf. Data Eng., 2009, pp. 952–963.

[6] “A Survey on Data Deduplication in Large Scale Data” by Saniya Sudhakaran, Meera Treesa Mathews, IJCA vol 65 pp 1 , 2017