IOSR Journal of Engineering (IOSRJEN)
e-ISSN: 2250-3021, **p**-ISSN: 2278-8719
Vol. 3, Issue 8 (August. 2013), ||V2 || PP 45-51

Data Cleaning by Genetic Programming Technique

Rajnish Kumar, Pradeep Bhaskar Salve, Pritam Desale
Sir Visvesvaraya Institute of Technology, Nashik, University of Greenwich, United Kingdom

Abstract: - Today Generally all the websites having à search box from which data can be effectively extracted. When any query is submitted to databases then it retrieves the information from that database and get extracted but as the number of data is increasing rapidly in database, technique to extract the clean data i.e. non duplicate data has not been simultaneously updated so it becomes very hard to detect duplicate data and extract with clean data in effective manner. When data is uploaded in the database server from different location then it having more chances about duplicate data. User may use same link on same page. Links having some information. When link which having some information is duplicated more than one time then due to this data duplication, memory is wasted, space is wasted and hence performance and computational cost increases. In this paper, we propose a technique as a genetic programming technique which contains the three major operations. Those operations are selection, crossover and mutation. Execution of operations applies the de-duplication function. After removing the duplicate records apply the suggested function. Compare to all previous approaches present approach provides less burden, efficient and accurate results display here. It can provide good evidence based results.

Keywords: - *Evolutionary Programming, Precision and Recall, XML, XHTML.*

I. INTRODUCTION

Every day in data repositories many number of knowledgeable people update the data. Due to this, Data increases on World Wide Web. Already existing data may be added in two or more number of databases. These kinds of data repositories come under dirty repositories. Any user who forward the query, search engine extract and display the results. Extraction results may contain useless data. Query shows much number of problems like high response amount of time, availability, quality assurance and security. Websites do not provide any useful services in extraction. These services show the problems in performance, quality and operational cost. Actually, the main problem in the existing system is:

- **Data Complexity**

As the online databases is increasing day by day so a huge amount of data exist in www so from this data complexity, previous approaches (such as in vision based approach, page level extraction, TSIMMIS, Web OQL) are not effective because they are low efficient and time consuming approach.

- **Web page programming language and version Dependency**

Generally, all previous solution related to web page extraction was html dependent. Let's take an example of any website such as Pune University. Two years ago, there was some one administrator who was maintaining that website and he was using html and version was 3.0. After one year, new administrator was appointed and he maintained that website using html version 4.0 and now a time, some different administrator is maintaining that website by using some other web page programming language such as XHTML and XML. Will data be extracted effectively? Of course no. so this is a problem of web page programming language and version related dependency.

- **Problem with Record Duplication**

The most important thing is that when data is uploaded from different location then it may having chance of data duplication. If we consider any digital library website such as google, paper publication journal website then there exist so many unwanted data. One data repeats so many time hence so many space is wasted. Due this, processing time is very high when we submit any query in the database.

Let's take an example of Wikipedia website

In Wikipedia search box, suppose we type java then related page is extracted. What we see on this page?

Fig. Wikipedia Search Box for Java

We see that all the information has been indexed in unique manner that means no duplication is occurred. There are so many links available on this page but all the links are unique. No any links comes two time hence data has been indexed in effective manner. It means Wikipedia uses such type of approach in which all the duplicate value is automatically removed and no two links comes on a single page because when two links comes on a same page then there is wastage of space and memory and time because of links having same information.

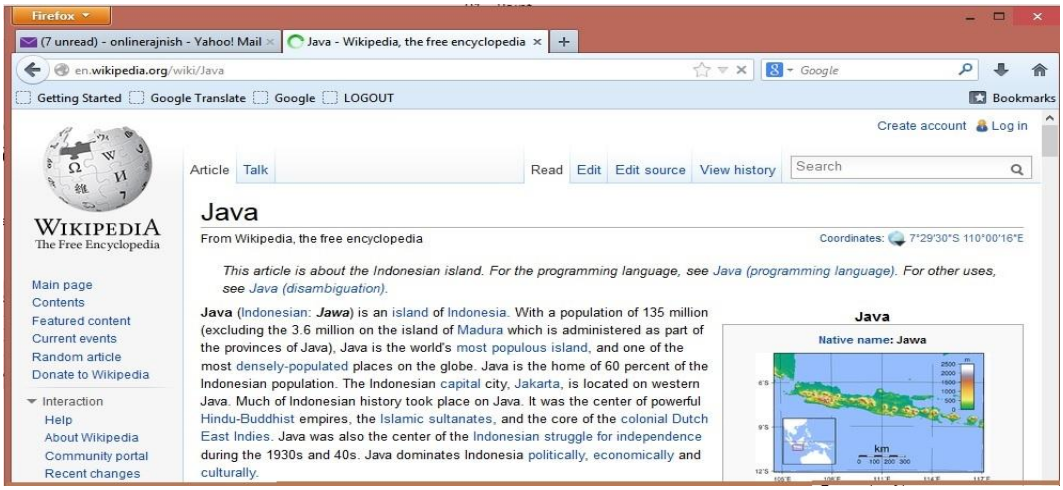

Fig. Wikipedia Search Result for Java

Like Wikipedia extraction approach, we have developed that type of project in which duplicate data will be automatically detected and after then dirty data will be removed. All the web page will contain unique link. No same links will repeat again that mean will be duplicated again hence this automatically detecting feature can be termed as genetic programming technique so it is clear that in our project, we will extract the data from any search engine or any digital libraries website and duplicate data will be removed hence we will be able to extract the effective data

So, in our proposed system, we will remove the duplicate data, all dependency whatever we explained above. Here, we will perform Data Cleaning. Data Cleaning is the process of identifying the records in a data repository that refers to the same real world entity or object in spite of misspelling words, types, different writing styles or different schema representation or data types hence clean and replica free repositories not only allow the retrieval of higher quality information but also lead to a more concise data representation and potential saving in time and resources. Our approach combines different pieces of evidence extracted from the data content to produce a deduplication function that is able to identify whether two or more entities in a repository are replicas or not. The function used for record deduplication will accomplish distinct but conflicting objectives. It will maximize the identification of record replicas.

II. LITERATURE SURVEY

2.1 Introduction

Duplicate record detection is the process of identifying different or multiple records that refer to one unique real-world entity or object. Typically, the process of duplicate detection is preceded by a data preparation stage, during which data entries are stored in a uniform manner in the database, resolving (at least partially) the structural heterogeneity problem. The data preparation stage includes a parsing, a data transformation, and a standardization step. The approaches that deal with data preparation are also described under the using the term ETL (Extraction, Transformation, Loading). These steps improve the quality of the in-flow data and the data comparable and more usable. While data preparation is not the focus of this survey, for completeness we describe briefly the tasks performed in that stage. A comprehensive collection of papers related to various data transformation approaches can be found in. Parsing is the first critical component in the data preparation stage. Parsing locates, identifies and isolates individual data elements. Parsing makes it easier to correct, standardize, and match data because it allows the comparison of individual components, rather than of long complex strings of data. For example, the appropriate parsing of name and address components into consistent packets of information is a crucial part in the data cleaning process. Multiple parsing methods have been proposed recently in the literature and the area continues to be an active of research. Data transformation refers to simple conversions that can be applied to the data in order for them to conform to the data types of their corresponding domains. In other words, this type of conversion focuses on manipulating one field at a time, without taking into account the values in related fields. The most common form of a simple transformation is the conversion of a data element from one data type to another. Such a data type conversion is usually required when a legacy or parent application stored data in a data type that makes sense within the context of the original application, but not in a newly developed or subsequent system.

Renaming of a field from one name to another is considered data transformation. Encoded values in operational systems and in external data is another problem that is addressed at this stage. These values should be converted to their decoded equivalents, so records from different sources can be compared in a uniform manner. Range checking is yet another kind of data transformation which involves examining data in a field to ensure that it falls within the expected range, usually a numeric or date range. Lastly, dependency checking is slightly more involved since it requires comparing the value in a particular field to the values in another field, to ensure a minimal level of consistency in the data.

2.2 Related Work

As Record Deduplication is a growing topic in Data mining so three types of works has been employed:

- **Domain Knowledge Based Work**

In this type of work, work has been done depending on specific domain knowledge or specific string distance metric. In this type of work, matching algorithm has been proposed in which given a record in a file or repository, looks for another record in a reference file that matches the first record according to the given similarity function and it is selected based upon threshold. For this, weight is calculated and hence duplicate data is removed.

- **Probabilistic Based Work**

In this category, work has been done on probability based. This method relies on the definition of two boundary value that are used to classify a pair of record as being replicas or not.

- **Machine Learning Based Work**

Our technique is more related to this approach. This approach apply machine learning technique for deriving record level similarities function that combine field level similarity function including the proper assignment of weight. This proposal use small portion of available data for their purpose. The extracted evidence is encoded as a feature vector that are used to train a support vector machine classifier to better combine in order to identify replicas/web page dependency and all other dependencies and hence cleaned data is extracted.

2.3 Information Retrieval

Information Retrieval is the scientific method of searching information either in documents or searching for documents themselves. It also includes searching within databases which could either be relational standalone databases or hyper textually-networked databases like the World Wide Web. It is the science of locating, from a large document collection; those documents that fulfill a specified information need. Information overload are reduced through the use of automated Information Retrieval System (IRS). IRS is used in places like universities and some other tertiary institutions. Access to books, journals, and other documents

are easily achieved in libraries by the use of IRS. Programmers have succeeded in their effort at creating applications that have very well functioned in information retrieval. In our project, we have used such applications which can be found within our search engines.

III. ARCHITECTURE

Fig. Architecture of Data Detection & Cleaning

IV. GENETIC PROGRAMMING TECHNIQUE

We have used genetic programming technique. As the name from genetic; in this approach data will be automatically selected and duplicated data will be detected with the help of relevant programming.

Genetic Programming is one of the evolutionary programming technique which having the properties of natural selection. The main aspect that distinguish genetic programming from other evolutionary technique is that it represents the concept and interpretation of a problem as a computer program and even the data are viewed and manipulated in this way. This genetic programming is able to discover the variable and relationship with each other and find the correct functional form. It having mainly three operation such as selection, crossover and mutation. All the operation has been included in the algorithm.

V. ALGORITHM USED

Algorithm used in this project is nothing but Genetic Detection Algorithm.
Following Operations are performed in this algorithm:

- First populated data is initialized that all the data is discovered.
- After then all the individual data are evaluated and a numeric fitness value is assigned to each one.
- Selection process is performed that means all the n individual are selected into next generation population without modifying the data.
- After then Crossover operation is performed in which m individual that will compose the next generation with the best parent is selected and replace the existing generation i.e. in this process two parent tree are selected according to matching policy and then a random sub tree is selected in each parent.
- And finally Mutation operation is performed in which the best individual are produced in the population.

VI. MODULES DESCRIPTION

We have five modules for this project:

1. Evolutionary Programming or Supervised learning procedure
2. Genetic operations
3. Genetic Detection algorithm
4. Record De-duplication with Genetic programming
5. Precision and Recall operations

1. Evolutionary Programming or Supervised learning procedure

User forwards the query and extracts results from the database. Under extraction of results apply the operation is selection operation. This selection performs in different databases and extracts the results with interactive query processing. It is not provides any optimal solution. These results contains some duplicates. It can display the nearly optimal solution results only. I have submitted the screen shot of my project containing the duplicate data. Following figure shows the same.

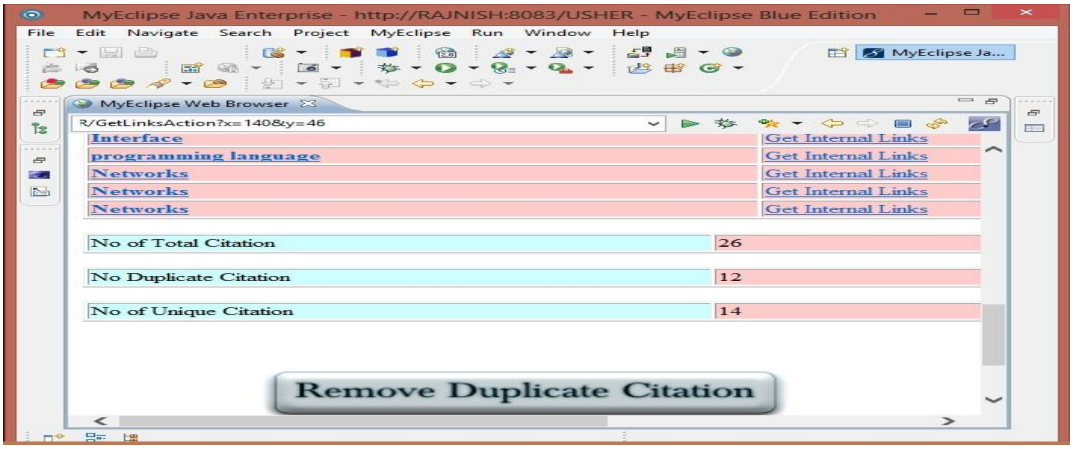

Fig. Screenshot of our project containg Duplicate Citation

2. Genetic Operations

This module try to provide the structure based results. Here first selects root terminals. This is zero level of results. Next we are finding out next level of children's. This procedure applies till reaches to leaf nodes for extraction of results. In this procedure all internal nodes we are find out here in implementation part. These internal nodes identification and create the structure possible with three operations here. Those are called selection, crossover and mutation. Following Screenshot show all the process.

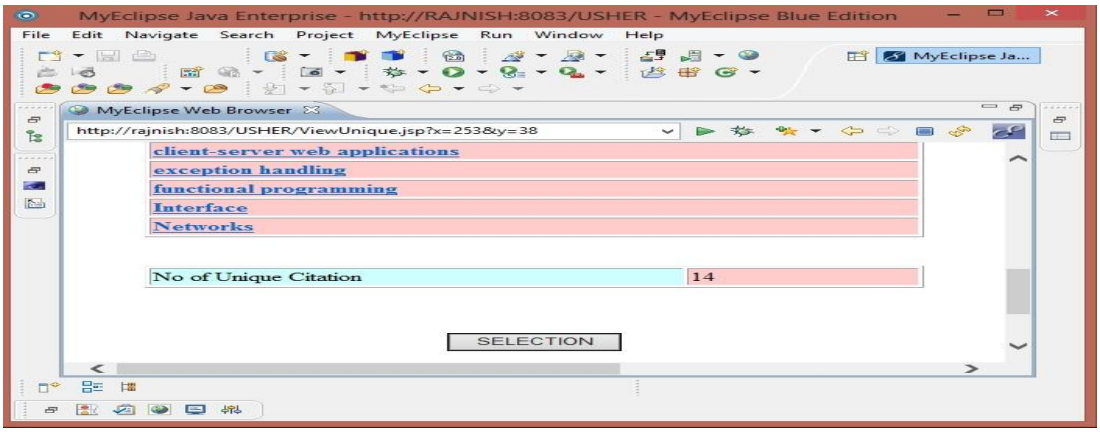

Fig. Selection Operation

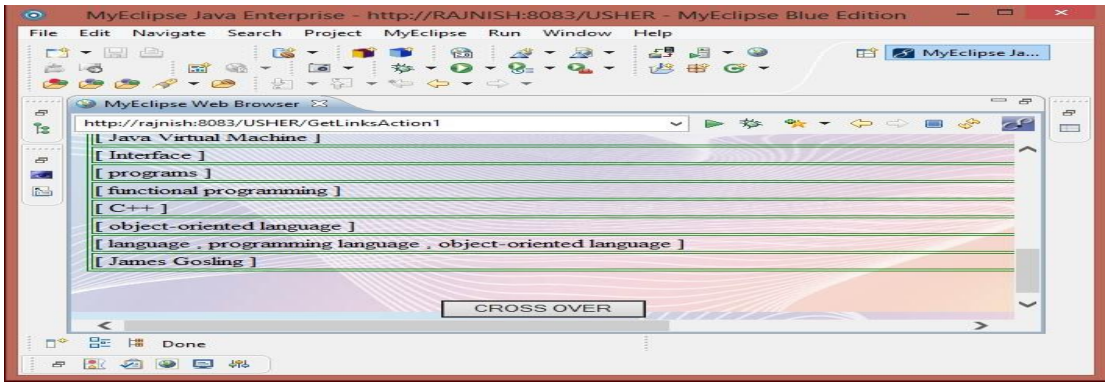

Fig. Crossover Operation

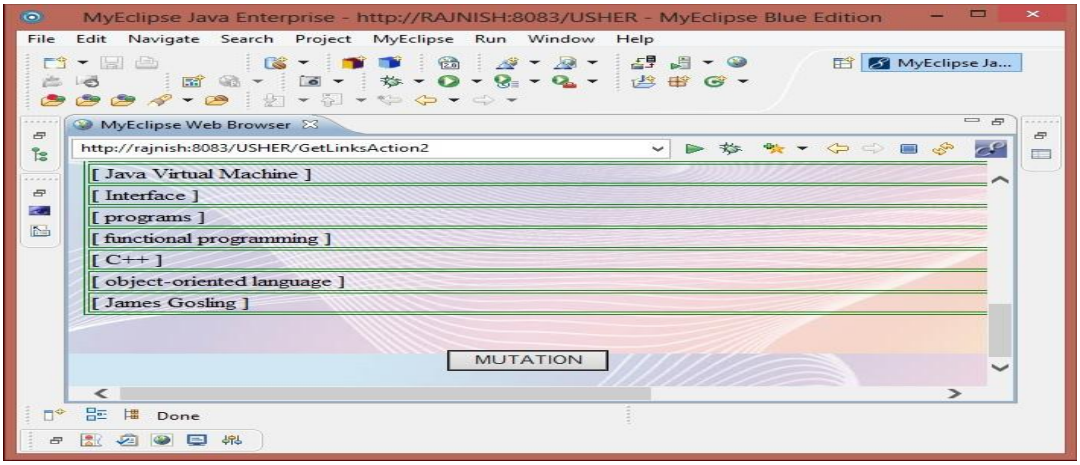

Fig. Mutation Operation

3. Genetic Detection Algorithm

Initialize all the results of nodes. Each and every node of rating we are calculates here. According to rating value calculates fitness. After finding the fitness node it's possible for creates the reproduced tree in implementation. It can contains all nodes are best. This same process applies till for finding the optimal tree identification. This same procedure repeatedly performs here.

4. Record Cleaning with Genetic programming

According to requirement automatically it can changes here in implementation. It can show the efficient results in tree data structure in implementation. It is the good evidence based results display. All those nodes are display with the help of similarity function in implementation process.

5. Precision and Recall operations

Two addition operation i.e. Precision and Recall operation is performed for accurate extraction calculation. For this, following formulae has been used:

P = Number of Correctly Identified Duplicated P airs / Number of Identified Duplicated P airs

R = Number of Correctly Identified Duplicated P airs / Number of True Duplicated pairs

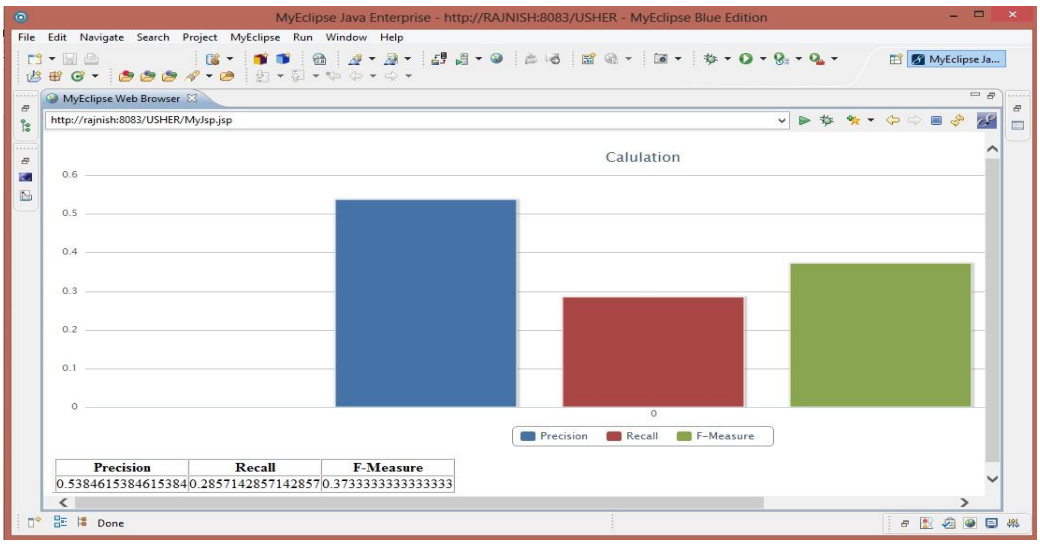

Fig. Graph Generated Based on Precision and Recall Value

VII. FUTURE WORK

As Data Mining is a vast topic and Data Cleaning is one of them. In this topic, many future work to this system can be done such as:

- Security can be enhanced.
- Testing can be done for different option where this system fails.
- Adding of More Modules so that user can access this system easily.

VIII. CONCLUSION

Following are the conclusion of this project:

- It perform an existing state of the art machine learning based method.
- It provides solution less intensive since it suggest deduplication function.
- It frees the user from the burden of choosing how to combine similarity functions and repository attributes.
- It frees from the user from the burden of choosing the replica identification boundary value since it is able to automatically select the deduplication function.
- Independent of Web Page Programming language, version and scripting related extraction.
-

REFERENCES

[1]. H. Zhao, W. Meng, Z. Wu, and C. Yu, "Automatic Extraction of Dynamic Record Sections from Search Engine Result Pages," Proc. 32nd Int'1 Conf. Very Large data Bases (VLDB), 2006.

[2]. V. Crescenzi, P. Merialdo, and P. Missier, "Clustering Web Pages Based on Their Structure," Data and Knowledge Eng., vol.54, pp. 279-299, 2005.

[3]. B. Liu, R.L. Grossman, and Y. Zhai, "Mining Data Records in Web Pages," Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 601-606, 2003.

[4]. K. Simon and G. Lausen, "ViPER: Augmenting Automatic Information Extraction with Visual Perceptions," Proc. Conf. Information and Knowledge Management (CIKM), pp. 381-388, 2005.

[5]. M. Wheatley, "Operation Clean Data", CIO Asia Magazine.

[6]. N. Koudas, S. Sarawagi and D. Srivastava, "Record Linkage: Similarity Measures and Algorithms", Proc. ACM SIGMOD Int'1 Conf. Management of Data, pp. 802-803, 2006.

[7]. R. Bell and F. Dravis, "Is You Data Dirty? and Does that Matter?," Accenture Whiter Paper, http://www.accenture.com, 2006.

[8]. J.R. Koza, Gentic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, 1992.

[9]. W. Banzhaf, P. Nordin, R.E. Keller, and F.D. Francone, Genetic Programming - An Introduction: On the Automatic Evolution of Computer Programs and Its Applications. Morgan Kaufmann Publishers, 1998.

[10]. H.M. de Almeida, M.A. Gonc¸alves, M. Cristo, and P. Calado, "A Combined Component Approach for Finding Collection-Adapted Ranking Functions Based on Genetic Programming," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 399-406, 2007.