DATA STREAMING IN PAGE RANKING FOR ONTOLOGY BASED SEARCH ENGINE

L. Mohan¹, Dr. D. Gunaseelan², P.K. Kumaresan³

¹Test Lead, STAG Software Private Ltd., Adyar, Chennai-20,
²Professor, College of Computer Science, JAZAN University, JAZAN, Kingdom of Saudi Arabia,
³Professor, Department of Information Technology, VMKV Engineering College, Vinayaka Missions University, Periyaseeragapadi, Salem, India.

Abstract - The majority of algorithms proposed in recommender systems literature has focused on improving recommendation accuracy, other important aspects of recommendation quality, the diversity of recommendations, have often been overlooked. Number of item ranking techniques that can generate substantially more diverse recommendations across all users while maintaining comparable levels of recommendation accuracy. Comprehensive empirical evaluation consistently shows the diversity gains of the proposed techniques using several real-world rating data sets and different rating prediction algorithms. We present an unsupervised, online record matching method, UDD, which can effectively identify duplicates from the query result records of multiple Web databases.

Keywords - Unsupervised Duplicate Detection (UDD), Comprehensive empirical evaluation, Data de-duplication, Bayesian Multiple imputation, Regression imputation, WCSS

I. INTRODUCTION

The Web contains a vast amount of non-crawlable content. This hidden part of the Web is comprised of a large number of online Web databases consisting of a searchable interface (usually an HTML form) and a backend database, which dynamically provides information in response to user queries. In the hidden Web, it is usually difficult or even impossible to directly obtain the schemas of the Web databases without cooperation from the web sites. Instead, the web sites present two other distinct schemas, interface and result schema, to users.

With many businesses, government organizations and research projects collecting massive amounts of data, the techniques collectively known as data mining have in recent years attracted interest both from academia and industry. While there is much ongoing research in data mining algorithms and techniques, it is well known that a large proportion of the time and effort in real-world data mining projects is spent understanding the data to be analyzed, as well as in the data preparation and preprocessing steps (which may dominate the actual data mining activity). It is generally accepted that about 20% to 30% of the time and effort in a data mining project is used for data understanding, and about 50% to 70% for data preparation.

DATA DE-DUPLICATION:

Data de-duplication or Single Instancing essentially refers to the elimination of redundant data. In the de-duplication process, duplicate data is deleted, leaving only one copy (single instance) of the data to be stored. However, indexing of all data is still retained should that data ever be required. Example: A typical email system might contain 100 instances of the same 1 MB file attachment. If the email platform is backed up or archived, all 100 instances are saved, requiring 100 MB storage space. With data de-duplication, only one instance of the attachment is actually stored; each subsequent instance is just referenced back to the one saved copy reducing storage and bandwidth demand to only 1 MB.

EXISTING SYSTEM

Duplicate elimination problem of detecting multiple tuples. Real world entity- important data cleaning problem. Previous domain independent solutions to this problem relied on standard textual similarity functions between multi-attribute tuples. Duplicate detection is an important step in data integration. Existing concept prefix filtering principle to avoid computing similarity values for all possible pairs of records.

PROPOSED SYSTEM
New filtering techniques are proposed by exploiting the ordering information that are integrated into the existing methods and drastically reduce the user sizes and hence improve the efficiency. A new record matching method named Unsupervised Duplicate Detection (UDD) was proposed for the specific record matching problem of identifying duplicates among records in query results from multiple Web databases. A new exact similarity join algorithms is introduced with application to near duplicate detection. The techniques are used for adjusting the weights of the record fields in calculating the similarity between two records. Two records are considered as duplicates if they are “similar enough” on their fields. Different fields may need to be assigned different importance weights in an adaptive and dynamic manner. Finally an efficient algorithm is proposed and presented several optimizations that significantly reduce the overall computation time using real data set and synthetic data set.

COMPREHENSIVE EMPIRICAL EVOLUTION METHOD:

The handling of missing values is a topic of growing interest in the software quality modeling domain. Data values may be absent from a dataset for numerous reasons, for example, the inability to measure certain attributes. As software engineering datasets are sometimes small in size, discarding observations (or program modules) with incomplete data is usually not desirable. Deleting data from a dataset can result in a significant loss of potentially valuable information. This is especially true when the missing data is located in an attribute that measures the quality of the program module, such as the number of faults observed in the program module during testing and after release. Comprehensive experimental analysis of five commonly used imputation techniques. This work also considers three different mechanisms governing the distribution of missing values in a dataset, and examines the impact of noise on the imputation process. To our knowledge, this is the first study to thoroughly evaluate the relationship between data quality and imputation. Further, our work is unique in that it employs a software engineering expert to oversee the evaluation of all of the procedures and to ensure that the results are not inadvertently influenced by poor quality data. Based on a comprehensive set of carefully controlled experiments, we conclude that Bayesian multiple imputation and regression imputation are the most effective techniques, while mean imputation performs extremely poorly. Although a preliminary evaluation has been conducted using Bayesian multiple imputation in the empirical software engineering domain, this is the first work to provide a thorough and detailed analysis of this technique. Our studies also demonstrate conclusively that the presence of noisy data has a dramatic impact on the effectiveness of imputation techniques.

WEIGHTED COMPONENT SIMILARITY SUMMING (WCSS) CLASSIFIER IS USED

Duplication methods in UDD are used which focuses on Web databases from the same domain, i.e., Web databases that provide the same type of records in response to user queries. Suppose there are s records in data source A and there are t records in data source B, with each record having a set of fields/attributes. Each of the t records in data source B can potentially be a duplicate of each of these records in data source A. The goal of duplicate detection is to determine the matching status, i.e., duplicate or non-duplicate, of these s - t record pairs. Record linkage algorithms depend on string similarity functions for record fields as well as on record similarity functions for string fields. Similarity computation functions depend on the data type. Therefore the user must choose the function according to the attribute’s data type, for example numerical, string and so on. Token has been created for the selected attributes. Each function measures the similarity of selected attributes with other record fields and assigns a similarity value for each field. The clustering techniques have been selected to group the fields based on the similarity values. This matching and non-matching pairs is used for clustering and to eliminate the duplicates. The rule based duplicate detection and elimination approach is used for detecting and eliminating the records. The elimination process is very important to produce a cleaned data. The above steps are used to identify the duplicate records. Before the elimination process, the similarity threshold values for all the records in the dataset are calculated. The similarity threshold values are important for the elimination process. The threshold criteria and certainty factors are used to detect and eliminate the duplicate records. Finally one record is retained as prime representative and maintained this value in the log file. This approach can substantially reduce the probability of false mismatches, with a relatively small increase in the running time.

![Image of Weighted Component similarity summing](image_url)

**Fig.1.2 Weighted Component similarity summing**

UNSUPERVISED DUPLICATION DETECTION:

The problem of identifying duplicates, that is, two (or more) records describing the same entity, has been performed. Most previous work is based on predefined matching rules hand-coded by domain experts or matching rules learned offline by some learning method from a set of training examples. Such approaches work well in a traditional database environment, where all instances of the target databases can be readily accessed, as long as a set of high-quality representative records can be examined by experts or selected for the user to label. UDD (Unsupervised Duplication Detection) is an algorithm newly implemented for detection of duplicates.
The UDD provides techniques for adjusting the weights of the record fields in calculating the similarity between two records. Two records are considered as duplicates if they are “similar enough” on their fields.

![Architecture of Web ontology](image)

Fig:1.3 Architecture of Web ontology

Due to the absence of labeled training examples, a sample of universal data consisting of record pairs from different data sources\(^5\) as an approximation for a negative training set as well as the record pairs from the same data source are taken. The proportion of duplicate records in the universal set is usually much smaller than the proportion of non-duplicates.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Number of websites</th>
<th>Number of user specified fields</th>
<th>Duplicate ratio</th>
<th>Duplicate pair reduction ratio</th>
<th>Residue duplicate ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book</td>
<td>10</td>
<td>2</td>
<td>7.3%</td>
<td>96%</td>
<td>0.64%</td>
</tr>
<tr>
<td>Email</td>
<td>10</td>
<td>3</td>
<td>3.2%</td>
<td>76%</td>
<td>0.07%</td>
</tr>
<tr>
<td>Movie</td>
<td>10</td>
<td>3</td>
<td>6.4%</td>
<td>96%</td>
<td>0.26%</td>
</tr>
<tr>
<td>Movie Record</td>
<td>10</td>
<td>3</td>
<td>2.1%</td>
<td>88%</td>
<td>0.05%</td>
</tr>
<tr>
<td>Average</td>
<td>10</td>
<td></td>
<td>4.8%</td>
<td>89%</td>
<td>0.05%</td>
</tr>
</tbody>
</table>

Table.1 Duplicate Detection in UDD

CONCLUSION
Recent work in information retrieval, federated database systems and data mining have proposed alternatives to key components of UDD. Record linkage is the task of quickly and accurately identifying records corresponding to the same entity from one or more data sources. Record linkage is also known as data cleaning, entity reconciliation or identification and the merge/purge problem. Standard file compression tools identify short repeated substrings inside single files, the focus of data de-duplication is to take a very large volume of data and identify large sections - such as entire files or large sections of files - that are identical, and store only one copy of it. Duplicate detection is state-of-the-art methods are based on offline learning techniques, which require training data. The huge reduction of storage achieved by using data de-duplication technology makes disk an attractive, viable, and less-expensive alternative. Disk libraries can now provide faster and more reliable recovery of data at a total cost of ownership much less than tape.

REFERENCE:
2. Gediminas Adomavicius, Member, IEEE, and YoungOk Kwon “Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 5, MAY 2012.