ENTITY RESOLUTION FOR HIERARCHICAL DATA USING ATTRIBUTES VALUE COMPARISON OVER DISTRIBUTED DATABASE

Asst. Prof. Vinod Ingale¹, Asst. Prof. Sonali Patil²

¹Computer Science & Engineering, AITRC Vita, District Sangli, Maharastra,India, ²Information Technology, AITRC Vita, District Sangli, Maharastra,India,

ABSTRACT

Entity resolution is the problem of identifying which records in database refer to the same entity. Records linkage is the process of matching records between two or multiple dataset that represent same real world entity. Computing similarities between all pairs of records which can be very expensive for large datasets. Blocking techniques eliminates this problem. In practice, many applications need to resolve large data sets efficiently, but do not require the ER result to be exact. For example, people data from the Web may simply be too large to completely resolve with a reasonable amount of work. As another example, real-time applications may not be able to tolerate any ER processing that takes longer than a certain amount of time. This system investigates how we can maximize the progress of ER with a limited amount of work using “hints,” which give information on records that are likely to refer to the same real-world entity.

Keywords: Application areas, Basic concept and benefits, Challenges, Structure and Methodology, Terms and definitions, System Architecture.

I. INTRODUCTION

Using Data mining, users can remotely find their data and enjoy the on-demand high quality applications and services from a distributed database of configurable computing resources, without the burden of local data storage and maintenance. Moreover, users should be able to just get require result using entity resolution algorithm with in short time without worrying about the need to verify its integrity. Thus, enabling entity resolution is the problem of identifying which records in database
refer to the same entity and also help for records linkage is the process of matching records between two or multiple dataset that represent same real world entity. Computing similarities between all pairs of records which can be very expensive for large datasets. Eg if two companies that merge may want to combine their customer records in such a case the same customer name may be represented by multiple records. Such records are compared on their matching attribute like name, postal address, date of birth, email address, mobile no. At same time it may be very important to run ER with in a limited amount of time. Entity resolution is the problem of identifying which records in database refer to the same entity. Records linkage is the process of matching records between two or multiple dataset that represent same real world entity. Computing similarities between all pairs of records which can be very expensive for large datasets. Blocking techniques eliminates this problem.

II. APPLICATION AREAS

- Online stores - To avoid multiple registrations for single user.
- University - One student may be studying in multiple courses in same University so we need to find duplicates and save memory space.
- Voter’s database - To avoid multiple registrations of voters and duplicate voting.
- Banking Database - To providing multiple account choice within single account.

<table>
<thead>
<tr>
<th>Paper Name</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier based duplicate records elimination for query results from web database.</td>
<td>Work on large database with classification.</td>
<td>Lengthy algorithm require large time of execution due to classification and then duplicates finding algorithm. Each record has to visit twice which is not efficient for large DB.</td>
</tr>
<tr>
<td>Effective incremental clustering for query results from web database.</td>
<td>Clustering reduce cost of record matching by using indexing</td>
<td>Large number of clusters is generated and this will occupy large memory space.</td>
</tr>
<tr>
<td>Pay As You Go entity resolution.</td>
<td>Works on large database and securable</td>
<td>Limitations in resources</td>
</tr>
</tbody>
</table>
III. CHALLENGES

Manual system identifies the records but its very time consuming so need of such system develop the system which require less time. Existing system requires more memory so we have to develop the system which requires less memory. In this system parsing of records by name parser and address parser is not possible. In existing system it is not possible to store the reference of record. Records are not identified accurately.

IV. BASIC CONCEPT

Developing a program which will inspect desired and dedicated distributed system is to analyze all databases from different machines. This program load record buffer from all databases then record buffer will parse by name-parser for this we get sorted record which finally will easy to find duplicate. Address–parser is an identifier sorted records according to their postal address. Finally conformation module check date of birth, mob no, and other things and send to referential module. Referential modules save one copy each duplicate pair at reference table.

V. BENEFITS

- High accuracy- As one ER module divided into three relates modules
- non-duplicates records which takes more time
- Less memory requirement-As we use sorted buffers to find duplicates so memory requirement reduced.

VI. TERMS AND DEFINITIONS

- Entity-Entity is something that exists separately from other things and has a clear identity of its own.
- Entity resolution-Entity resolution is the problem of identifying which records in database refer to the same entity.
- Hierarchical system-A hierarchical system or organization is one in which people have different ranks or positions, depending on how important they are.
- Distributed system-Distributed system is a collection of independent computer that appears to its user as a single coherent system.

VII. ANALYSIS REQUIREMENT

System analysis focuses on all system elements and not just the software. It consists of Requirement analysis phase.

VIII. IDENTIFY THE USER’S NEED

The first and foremost task was to study the user’s needs. After having overall observation on today’ traditional method of work, it was the basic requirement of District domain, that should be able to provide instant report through message that is record is duplicate or new record. As well as provide memory reduction and get correct result within sort time, which give the all information related to particular student.
IX. TECHNICAL ANALYSIS

After concluding, we found that there was need to develop Application for records checking to find duplication and reduce memory space.

X. NON-FUNCTIONAL REQUIREMENT

Specify the quality of system, is mostly related to the satisfiability of the user. In system engineering and requirement engineering, a non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of the system, rather than specific behaviors. The plan for implementing the functional requirement is detailed in the system design. The plan for implementing the non-functional requirement is detailed in the system architecture.

- Performance: This part of SRS specifies the performance constraints on the software system. All the requirements relating to the performance characteristics of the system be clear specified. There are two types of performance requirements: Static and Dynamic. Static requirements are those that do not impose constraint on the execution characteristics of the system.
- Dynamic requirements specify constraints on the execution behavior of the system.
- Reliability: The capability to maintain a specified level of performance.
- Maintainability: The capability to be modified for purposes of making corrections, improvements, or adaptation.
- Availability: Availability refers to technical innovation and procedures apply to hardware and operating system. First you have to learn about the foundation of the availability the true score theory of measurement. The pre-requirements for availability in the generalized abstraction between the application logic and system interface.
- Portability: The capability to be adapted for different specified environments without applying or means other than those provided for this purpose in the product.
- Security: The system security problem can be divided into related issues:

  Security, integrity, privacy and confidentiality the determine the file structure, data structure and access procedure. System security refers to the technical innovation and applied to the hardware and operating system to protect against accidental damage from a defined threat. System integrity refers to the proper functioning of programs, appropriate physical security and safety against external threats such as eavesdropping and wiretapping. The term confidentiality is a special status given to sensitive information in a database to minimize the possible invasion of privacy. It is an attribute of information that characterize its need for protection.

XI. STRUCTURE AND METHODOLOGY

As we started earlier, a methodology basically uses the concepts to provide guidelines and notation for the design activity. Though methodologies are useful, they do not reduce the activity of design to a sequence of steps that can be followed mechanically. Client-Server approach:
Client-Server is a program relationship in which one program (the client) requests a service or resource from another program (the server). Although the client/server model used by program within a single computer, it is a more important concept for networking. In this case, the client establishes a connection to the server over a local area network (LAN) or wide area network (WAN), such as the internet.

XII. SYSTEM ARCHITECTURE

- **Name-Parser**: This module parses the records on basis of name which gives sorted records which helps to find duplicates. Input to this module is buffer and output is sorted records as per name.
- **Address parser**: This module parses the outputted records by Name parser on the basis of their postal address to find matching records. Input to this module is sorted records on the basis of their Name and out of this module are sorted records as per their postal address.
• Conformation module:-This module includes conformation as per their date of birth, Email address and mobile number to find exact matching records to create their references into the reference table. This module is used to create references of matching records to fit them into the reference table in sorted manner.
• Reference module:-This module is used to create references of matched pairs by conformation module. Output of this module is stored in database.
• System Design: System Design enables us to design the overall structure of the system. In this phase, the system has to be designed at the finer level of details. All the aspect of the system is to be designed keeping the SRS and implementation details in mind.

**DFD (Data Flow Diagram):**

Data Flow Diagram specifies the meaning of operations and constraints. Also give the information functional relationship of the modules and flow of data. It includes the processes, actors, data flow, data stores etc.

Level 0 DFD:

![Level 0 DFD](image1)

Level 1 DFD:

![Level 1 DFD](image2)
XIII. CONCLUSION

In this paper the distributed system which allows to access database from different machine and which has high accuracy, efficiency and less memory requirement. In this way finding duplicates of entered record on different server by examining details like name, address, mobile number, Email ID, etc. to reduce memory space required to store those duplicate records. At this stage the system identifies the record by using name parsing, address parsing, mobile number, in future we will add the E-mail ID, aadhar card number to verifying the records. Which reduces the time required for record identification and also increases accuracy in the process. It is also possible to add the document verification by using the concept in this paper.

REFERENCES