A Survey of Indexing Techniques for

Scalable Record Linkage and Deduplication

1. **ABSTRACT**

Record linkage is the process of matching records from several databases that refer to the same entities. When applied on a single database, this process is known as deduplication. Increasingly, matched data are becoming important in many application areas, because they can contain information that is not available otherwise, or that is too costly to acquire. Removing duplicate records in a single database is a crucial step in the data cleaning process, because duplicates can severely influence the outcomes of any subsequent data processing or data mining. With the increasing size of today’s databases, the complexity of the matching process becomes one of the major challenges for record linkage and deduplication. In recent years, various indexing techniques have been developed for record linkage and deduplication. They are aimed at reducing the number of record pairs to be compared in the matching process by removing obvious non-matching pairs, while at the same time maintaining high matching quality. This paper presents a survey of twelve variations of six indexing techniques. Their complexity is analyzed, and their performance and scalability is evaluated within an experimental framework using both synthetic and real data sets.

**2. INTRODUCTION**

**INTRODUCTION:**

As many businesses, government agencies and research projects collect increasingly large amounts of data, techniques that allow efficient rocessing, analyzing and mining of such massive databases have in recent years attracted interest from both academia and industry. One task that has been recognized to be of increasing importance in many application domains is the matching of records that relate to the same entities from several databases. Often, information from multiple sources needs to be integrated and combined in order to improve data quality, or to enrich data to facilitate more detailed data analysis. The records to be matched frequently correspond to entities that refer to people, such as clients or customers, patients, employees, tax payers, students, or travelers.

**DESIGN PRINCIPLES & EXPLANATION**

**MODULES**

The Project has the following modules

* Admin
* Users
* Key Generation
* Deduplication

1. The **administrator** should be able to
   * Admin is owner of this sit, he can login to the site with his specific username and password.
   * After login he accept the new users who newly registered with the site i.e He can give the login permissions to the users.
   * Admin Generate the key based on users first name. This Key will be generate for two databases(Ration Card and Bank A/c).Based on this key he can compare two databases and finally find the matching database and unmatching database.
2. **Users** should be able to
   * login to the system through the initial screen of the system
   * change the password after logging into the system
   * After login based on his requirement he can application for ration card and bank a/c.
   * He can search based on first name or last name.admin will compare both databases give matching database to the users.
3. The **Key Generation** should be able to

* Key will be generate for users fisrt name based on soudex functions of user names. This key used for comapring two databases.

1. The **Deduplication** should be able to

In this process we are creating a new database fron existing two databases

#### Abstract…….

# REQUIREMENT SPECIFICATION

**Specification Principles:**

Software Requirements Specification plays an important role in creating quality software solutions. Specification is basically a representation process. Requirements are represented in a manner that ultimately leads to successful software implementation.

Requirements may be specified in a variety of ways. However there are some guidelines worth following: -

* Representation format and content should be relevant to the problem
* Information contained within the specification should be nested
* Diagrams and other notational forms should be restricted in number and consistent in use.
* Representations should be revisable.

**Software Requirements Specifications:**

The software requirements specification is produced at the culmination of the analysis task. The function and performance allocated to the software as a part of system engineering are refined by establishing a complete information description, a detailed functional and behavioral description, and indication of performance requirements and design constraints, appropriate validation criteria and other data pertinent to requirements.

An outline of the Software Requirements Specification:

A simplified outline can be given for the framework of the specifications. This is according to the IEEE Standards.

**SOFTWARE AND HARDwARE REQUIREMENTS**

**Hardware Requirements:**

Processor Pentium IV

Speed 800 MHz

Hard Disk 20 GB

RAM 512 M B

**Software Requirements:**

Operating System Windows NT/XP.

Languages C#.NET

Web Technologies ASP.NET,

Back End SQL SERVER2000

Web/Application server Internet Information services (IIS)