A Review of Similarity Measurement for Record Duplication Detection

Saleh Rehiel Alenazi, Kamsuriah Ahmad
Research Center for Software Technology and Management, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, 43600, Malaysia
E-mail: sala207@hotmail.com, kamsuriah@ukm.edu.my

Akeem Olowolayemo
Department of Cognitive Science, Faculty of Cognitive Science & Human Development, University Malaysia Sarawak
oakeem@unimas.my

***Abstract*— Similarity measurement is a significant process to determine the degree of similarity between two records. This paper presents a comparative analysis of important similarity measurements which are utilised for the detection of duplicated records in databases. The work evaluates their strengths based on the efficiency of prevailing algorithms, the time required to process and identify duplications as well as performance accuracy. The analysis conducted found that among the most common similarity measurements, those based on the Jaro-Winkler algorithm significantly outperformed the other algorithms. This paper presents an enhanced strategy based on the Jaro-Winkler algorithm to improve the detection of similarity among database records. The ability to provide solutions to this problem will greatly enhance the quality of data used in decision-making.**

Keywords: record duplication deduction; similarity measurement; character-based; Jaro-Winkler.

I. INTRODUCTION

The similarity measure plays a vital role in nearly every field of science and engineering. A similarity measure can be described as a process to determine the degree of similarity that exists between two objects [1], [2]. The identification of similar database records is an important entity matching application. The term 'duplicate record detection' is used to describe the process of recognising records that represent the same real-world entity in a specific database. The difficulty associated with duplication is that duplicated records may not share the same record key. Various methods to resolve this issue have been employed to locate and cleanse erroneous duplicated records in a typical dataset. The duplicated or erroneous data can result from several factors which include, data entry errors, such as typing the name of a person like "John" as "Jon", etc. Moreover, there could be a missing validation check or restriction issue such as an age value of 320, or of multiple conventions such as 22 E, 7th St vs. 22 East Seventh Street). An additional problem may also result from structural differences between database sources [3].

An essential phase associated with data cleansing is referred to as the preprocessing stage which aims to detect and remove duplicate records relating to identical entities in a database. The purpose of this stage is to match records belonging to the same entities within one or more databases. This is because information is often acquired from multiple sources, often different, and merged (combined) to enhance the data or used for data mining analysis [3].

Often when combining data that is integrated or used from several sources, heterogeneity will result in two forms, namely structural and lexical forms. The term structural heterogeneity is commonly used to describe the condition where two databases having different field structures match. For instance, matching a database where the address has been stored in a single field, namely "address", while in another database the address is stored in three columns, namely "street", "city", and "state". Conversely, lexical heterogeneity results in the condition where the structure of both databases are the same such as using "address" or "street", "city", and "state" in both databases. However, there are cases of quite diverse representations of the data, for instance, "M. Harry" and "Harry Marshal" [3].

There are also similarities that exist between all databases and the approaches to compare the values of important fields of records. Vatsalan and Christen [4] employed a similar patient matching (SPM) technique comparing field values. For example, in the 'Patient' records database, patient fields such as age, gender, body, mass index, blood pressure and fasting blood sugar, etc., are used. Probabilistic-based techniques have often been employed in duplication detection applications such as FEBRL [5], TAILOR [6] and BigMatch [7], [8]. The probabilistic-based techniques depend on training the datasets to determine the maximum likelihood that can be employed to verify whether pairs of records match (or not). Figure 1 illustrates a typical data cleansing process. The initial step involves collecting data from various sources followed by the extraction procedure which collects the data from a relational database, XML, JSON or from any other source. Next, the data runs through a number of transformation procedures. The pre-processing procedures ensure that all the data is in a consistent format. This is followed by the record duplication process which involves key generation, preparation of records, similarity measurements and the existence and detection of duplication. Similarity measurements will sometimes utilise a blocking approach or window approach to track duplicated records. The results of the duplicate detection process are lastly, loaded into the data warehouse.

* This work is partially supported by the Universiti Kebangsaan Malaysia for funding the research under the Fundamental Research Grant Scheme FRGS/1/2014/ICT07/UKM/02/3).

978-1-5386-0475-5/17/$31.00 ©2017 IEEE