2017 IEEE/ACS 14th International Conference on Computer Systems and Applications

Using Transliteration with Entity Resolution for Arabic Datasets

Marwah Alian
Hashemite University
Princess Sumaya University fo
Technology
marwah2001@yahoo.com

Ghazi Al-Naymat
Princess Sumaya University for
Technology
Amman, Jordan
g.naymat@psut.edu.jo

Banda Ramadan
Applied Science University
Amman,Jordan
b_ramadan@asu.edu.jo

Abstract— **Entity resolution (ER) is the operation of distinguishing records that return to the same real world entity. It is used to link records among datasets and to match query records in real-time with existing datasets. Indexing is a major step in the ER process that reduces the search space. Most existing indexing techniques that are utilized in the ER process are designed to work with English datasets. Such techniques may not be suitable for use with other languages, such as Arabic. In this paper, enhancement for indexing techniques that are designed to work with English datasets has been proposed to be used with Arabic language by applying transliteration on Arabic strings before performing the indexing step of the ER process. The proposed approach is experimented and compared with using word stems as blocking keys in the indexing step. The results show better matching accuracy for the use of transliteration over the use of words stems.**

Keywords— Entity Resolution; Transliteration; Stemming; Indexing; Arabic Dataset.

I. Introduction

The process of distinguishing and matching records that represent the same entity in datasets is called Entity Resolution (ER) [1]. An entity represents a person, an organization or a product. ER is used in a situation where records duplication exists in a dataset of an organization, since duplicates affect organization's outcomes. The role of ER is to identify duplicates in a dataset to improve the quality of decisions in an organization [2].

Most existing ER approaches are applied for static datasets using batch algorithms by comparing all records in a dataset but do not identify records related to a single query record. On the other hand, other indexing techniques, such as the Dynamic Similarity Aware Inverted Index (DySimII) which uses indexing strategies to handle query-based ER in real time for dynamic English datasets. In this approach, a small number of high-quality candidate records is generated to be compared with a query record in real-time [5].

Most of indexing techniques that are used with Arabic datasets are applied on unstructured data such as in [6][10][11][12], while rarely used for indexing structured Arabic datasets. Moreover, most available indexing techniques, which are used in the ER process, are designed to work with English datasets, but are not suitable for Arabic datasets.

Transliteration stands for an orthography that is one to one substitution of orthographical symbols or characters from one language to another. [3]. Generally, Transliteration is used in Information Retrieval (IR), especially in multilingual or cross language retrieval of proper names where the requested query in one language is searched for in another language dataset or collection [15].

The problem of using ER techniques on structured Arabic datasets can be considered as an unsolved problem, since this area of research does not get enough attention. In this research, adding a transliteration step has been proposed, which converts Arabic strings into its phonetic to the indexing step in ER process to be able to use existing indexing techniques with Arabic datasets.

This paper is organized as follows: Section II presents the related work. Section III provides a description for the proposed methodology. In section IV, the experimental evaluation is demonstrated. Finally, conclusion is presented in section V.

II. Related Work

Indexing is the main step in the ER process which aims to reduce record pairs comparison by minimizing the number of records that does not correspond to true matches and retaining record pairs that correspond to the same entity to apply more detailed comparison [7].

The dynamic similarity aware inverted indexing techniques (DySimII), proposed in [9] is an indexing technique that works with real-time ER and is designed to work with dynamic English datasets. DysimII uses encoding functions to generate blocks that aims at reducing the search space. Moreover, in this approach, the inverted index size is minimized by indexing only the most frequent attribute values and it is updated automatically with each query record. The insertion time of a new record into the DySimII is not influenced by the size growth of the index. In addition, the size of the index is minimized by applying a filtering technique based on frequency but with a minor decrease in recall.

2161-5330/17 $31.00 © 2017 IEEE
DOI 10.1109/AICCSA.2017.11