International Journal of Speech Technology
https://doi.org/10.1007/s10772-018-9489-6

Unsupervised learning blocking keys technique for indexing Arabic entity resolution

Marwah Alian[1,2] · **Arafat Awajan**[2] · **Bandan Ramadan**[3]

Received: 12 October 2017 / Accepted: 3 January 2018

Abstract

Attribute values in textual datasets are subjects of different types of errors due to the data entry processes such as typogr aphi-cal errors, pronun ciation errors or dialects alterations. These errors make the entity resolution process more challenging. The iterative blocking indexing technique can be used for correcting this type of errors mainly in query access where the records are stored into more than one block. Blocking indexing technique selects a subset of object pairs saved in the same block for later detailed computation for similarity discarding other pairs in other blocks considering them as irrelevant. This work aims to solving such proble ms for Arabic texts. It proposes to adapt a specific model for learning blocking keys and analyze its performance for Arabic datasets. The resulted blocking keys are passed as blocking keys to Dynamic Aware Inverted Index (DySimII) that worked efficiently with Arabic datasets. The model is tested against a telephone book dataset that contains duplicates and errors in attribute values according to phonetic and typing errors. The results reach a matching accuracy of 84% for using learned keys with small numbe r of corrupted attributes while the performance is declined with the increase of the number of corrupted attributes.

Keywords Arabic entity resolution · Learning keys · Indexing · Arabic datasets

1 Introduction

The process of finding matches for a q uery record in a dataset or multiple datasets that represent the same real world entity is called Entity resolution (ER) (Ramadan and Christen 2015). ER techniques are important for organizations to improve their business operations as well as for a data warehouse which needs integration of data from multiple sources into one consistent center. ER techniques are very useful for cleaning and standardizing data (Ramadan 2016).

Indexing is an essential step in ER process mainly for large datasets since it has the advantage of reducing the number of candidate records that are going to pass to the detailed comparison step (Ramadan 2016). There are two main approaches for indexing in ER; the first approach is blocking in which re cords in a dataset are partitioned into different blocks according to a blocking key criterion and the records that stored in the same block are compared with each other. The second approach is sorting in which records are sorted with a sorting key and thi s will bring similar records to be closed to each other to compare only records that are close to each other (Ramadan 2016).

In Entity Resolution or Record Linkage, the number of similarity computation between pairs increase quadratically as the size of t he dataset increase since the similarity func -tion will be computed for all pairs in the dataset (Bilenko et al. 2006). However, a good indexing approach should store similar records in the same block or bring them close to each other based on the key crit eria (Ramadan 2016). Thus, blocking techniques relieve the problem of compar -ing all records in the dataset by computing the similarity between pairs in the same block that are approximately simi-lar (Bilenko et al. 2006).

In real world, data always conta in errors and alterations (Hernandez and Stolfo 1998) due mainly to data entry errors.

* Marwah Alian
marwah2001@yahoo.com

Arafat Awajan
awajan@psut.edu.j

Bandan Ramadan
bramadan@psu.edu.sa

[1] Hashemite University, Zarqa, Jordan

[2] Princess Sumaya University for Technology, Amman, Jordan

[3] Prince Sultan University, Riyadh, Saudi Arabia

Published online: 15 January 2018