Computational Linguistics and Chinese Language Processing
Vol. 19, No. 1, March 2014, pp. 33-54

Linking Databases using Matched Arabic Names

Tarek El-Shishtawy[*]

Abstract

In this paper, a new hybrid algorithm that combines both token-based and character-based approaches is presented. The basic Levenshtein approach also has been extended to the token-based distance metric. The distance metric is enhanced to set the proper granularity level behavior of the algorithm. It smoothly maps a threshold of misspelling differences at the character level and the importance of token level errors in terms of token position and frequency.

Using a large Arabic dataset, the experimental results show that the proposed algorithm successfully overcomes many types of errors, such as typographical errors, omission or insertion of middle name components, omission of non-significant popular name, and different writing style character variations. When compared with other classical algorithms, using the same dataset, the proposed algorithm was found to increase the minimum success level of the best tested lower limit algorithm (Soft TFIDF) from 69% to about 80%, while achieving an upper accuracy level of 99.67%.

Keywords: Name Matching, Record Linkage, Data Integration, Arabic NLP, Information Retrieval.

1. Introduction

Information about individuals can be found in a variety of resources, such as population survey databases, national identifier databases, medical records, news articles, tax information, and educational databases. In all of these heterogeneous sources, name matching is a fundamental task for data integration that joins one or more tables to extract information referring to the same person.

Matching personal names has been used in a wide range of applications, such as record linkage or integration, database hardening, removing or cleaning up duplicated records, and searching the Web. Unfortunately, the name may not be known exactly, may be misspelled, or may have spelling variations. Therefore, in these applications, the general word matching

[*] Prof. Ass. of Computer Engineering, Faculty of Computers and Information, Benha University
E-mail: t.shishtawy@ictp.edu.eg