Advances in Intelligent Systems and Computing (AISC)
Volume 1247, 2021, Pages 518–533

Deduplication Method for Ukrainian Last Names, Medicinal Names, and Toponyms Based on Metaphone Phonetic Algorithm (Conference Paper)

Hu, Z.[a], Buriachok, V.[b], Sokolov, V.[b]

[a]Central China Normal University, Wuhan, China
[b]Borys Grinchenko Kyiv University, Kiev, Ukraine

Abstract

This paper attempts to optimize the phonetic search processes for fuzzy matching tasks, such as deduplication of data in various databases and registers to reduce the number of errors in personal data entry (for instance, last names). The analysis of the most common last names in the territory of Ukraine shows that the majority of these last names are of Ukrainian and Russian origin (which are also reduced to phonetic rules of the Ukrainian language). The rules for pronouncing and writing last names in Ukrainian are fundamentally different from the basic algorithms for English and quite different for the Russian language, so the phonetic algorithm should take into account the peculiarities of the formation of Ukrainian last names. The use of the phonetic algorithm gives significant advantages in search and deduplication in comparison with already known algorithms: calculation of Levenshtein, Damerau-Levenshtein, Hamming, Jaro or Jaro-Winkler distance, Q-gram index, etc. The task of searching by last name was previously formalized in English, Russian and some other languages, but for the Ukrainian language such an attempt was made for the first time. The paper presents the results of the experiment on the formation of phonetic indices, as well as the results of increasing productivity when using the generated indices. A method of tailoring the search to other domains and several related languages is presented separately, for example, the search for medicines. Also, search optimization by place names in Ukrainian and Russian was separately worked out. Since in Ukraine there is an abrupt change in the names of cities and streets, the latest relevant data was collected to obtain an up-to-date list of names. Among the existing phonetic search algorithms for the Cyrillic language group, the Metaphone has proven itself in the best way.

Author keywords

Deduplication, Drug, Fuzzy coincidence, International nonproprietary name, Medication, Medicine, Metaphone, Phonetic algorithm, Phonetic rule, Toponym, Ukrainian last name, Ukrainian surname

Funding details

Funding sponsor	Funding number	Acronym
Ministry of Education - Singapore	CCNU19TS022	MOE

Funding text

This scientific work was partially supported by RAMECS and self-determined research funds of CCNU from the colleges' primary research and operation of MOE (CCNU19TS022). In addition, the authors of the paper thank the management of the medical information system of Helsi LLC for access to depersonalized medical data, a database of medical preparations and information resources for analysis and creation of a phonetic algorithm.

ISSN: 2194-5357
Print ISBN: 978-303055505-4
DOI: 10.1007/978-3-030-55506-1_47
EID: 2-s2.0-85089719737

First Online: 06 August 2020
Source Type: Book Series
Document Type: Conference Paper
Publisher: Springer, Cham