



An Automatic Blocking Keys Selection For Efficient Record Linkage

Hamid Naceur Benkhlaed, EEDIS Laboratory, University of Djillali Liabes, Sidi Bel Abbes, Algeria

 <https://orcid.org/0000-0001-7354-4920>

Djamal Berrabah, EEDIS Laboratory, University of Djillali Liabes, Sidi Bel Abbes, Algeria

Nassima Dif, EEDIS Laboratory, University of Djillali Liabes, Sidi Bel Abbes, Algeria

 <https://orcid.org/0000-0002-8683-3163>

Fauouzi Boufares, LIPN Laboratory, Paris 13 University, France

ABSTRACT

One of the important processes in the data quality field is record linkage (RL). RL (also known as entity resolution) is the process of detecting duplicates that refer to the same real-world entity in one or more datasets. The most critical step during the RL process is blocking, which reduces the quadratic complexity of the process by dividing the data into a set of blocks. By that way, matching is done only between the records in the same block. However, selecting the best blocking keys to divide the data is a hard task, and in most cases, it's done by a domain expert. In this paper, a novel unsupervised approach for an automatic blocking key selection is proposed. This approach is based on the recently proposed meta-heuristic bald eagles search (bes) optimization algorithm, where the problem is treated as a feature selection case. The obtained results from experiments on real-world datasets showed the efficiency of the proposition where the BES for feature selection outperformed existed approaches in the literature and returned the best blocking keys.

KEYWORDS

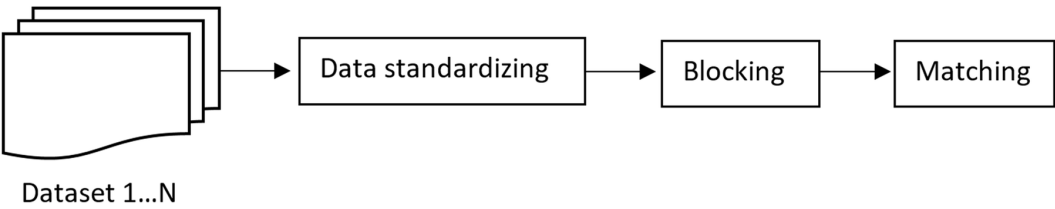
Bald Eagle Search Optimization Algorithm, Blocking Keys, Feature Selection, Meta-Heuristic, Record Linkage

1. INTRODUCTION

In recent years, the world is witnessing a massive explosion in the volume of data. Specifically, after the adoption of smartphones and social Media which generates a huge amount of data in a daily manner. Organizations around the world found themselves in the need of integrating their own data coming from various sources in different formats. This data have to be integrated in order to facilitate the process of data analyzing and extracting useful information out of it. However, data integration can become a very time-consuming process due to data quality problems, such as duplicates values, missing values, and referential integrity issues. The stakeholders are now more aware of the importance of data quality. A lot of money is invested in order to improve the quality of the stored data.

DOI: 10.4018/IJOI.2021010104

Figure 1. Record Linkage's phases



Record Linkage (RL) is one of the most important tasks in the data quality field. RL is defined as the process of identifying the records that represent the same real-world entity during the merge of different data sources. When the RL process is executed on a single database, it can be referred to as the deduplication process (Sarawagi and Bhamidipaty, 2002). Recently, the RL process was exploited in several domains for multiple goals such as privacy-preserving, removing duplicates from bibliographic citations, prices comparison and fraud detection.

The best way to detect all the tuples that refer to the same real-world entity is to compare each one in the dataset to all the others. However, with the case of a very large dataset, the Cartesian product could end-up with an unacceptable number of comparisons. For example, applying the RL process on database A and B, each one contains 2 million records will end-up by doing 4 billion matching operations, which is not reasonable.

To overcome this problem, the RL community proposed the blocking technique. Blocking consists of dividing the data into a set of blocks. In a way that all the records in the same block share a similar value called “The Blocking Key Value” (BKV). With this technique, matching between the records is done only on the records that are in the same block.

Generally, Record Linkage based on blocking consists of three main steps: Data standardization, Blocking and Matching (Christen, 2012) (Figure 1). Data standardization is the process of standardizing the dataset attributes that may represent the same information but with different identification in each one. For example, the sex attribute can be found in one dataset as a binary value (0/1) and in another one as (M/F). So standardizing the data is a very important preprocessing step. The next step is blocking, where one of the blocking techniques that exist in the literature (will be discussed in the next sections) is used to divide the data into a set of blocks. For example, using the standard blocking approach, all the records that have the same address are grouped in one block. After that, matching the records of the same block is done.

Two important parameters control the performance of a good blocking technique. The first one is the blocking key value. A blocking key can be formed using one field (attribute) or a concatenation of several parts from a set of fields. For example, a BKV can be formed using the First-Name value or it can be formed by the concatenation of the first three characters from the First-Name field and the ZIP code from the address field. Table 1 shows an example of blocking keys generating from the restaurant dataset. Two blocking keys were generated. The first one (BK1) is the Soundex phonetic

Table 1. A Blocking keys example from the restaurant dataset

BK1	BK2	Name	Address	City	Phone	Type
A6553102461501	LASANG435	arnie morton's of Chicago	435 s. la cienega blv.	Los Angeles	310/246-1501	American
H3413104721211	STADAC12224	art's deli	12224 ventura bold	Studio city	818-762-1221	delis

16 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/an-automatic-blocking-keys-selection-for-efficient-record-linkage/267171

Related Content

The Use of Evolutionary Algorithm-Based Methods in EEG Based BCI Systems

Adham Atyabi, Martin Luerssen, Sean P. Fitzgibbon and David M W Powers (2013). *Swarm Intelligence for Electric and Electronic Engineering* (pp. 326-344). www.irma-international.org/chapter/use-evolutionary-algorithm-based-methods/72835

Insurance-Based Business Web Services Composition

An Liu, Liu Wenyin, Liusheng Huang, Qing Li and Mingjun Xiao (2012). *Intelligent and Knowledge-Based Computing for Business and Organizational Advancements* (pp. 157-173). www.irma-international.org/chapter/insurance-based-business-web-services/65792

Improvement of Restaurant Operation by Sharing Order and Customer Information

Takeshi Shimmura, Takeshi Takenaka and Motoyuki Akamatsu (2012). *Intelligent and Knowledge-Based Computing for Business and Organizational Advancements* (pp. 241-257). www.irma-international.org/chapter/improvement-restaurant-operation-sharing-order/65797

The Need for Traffic Based Virtualisation Management for Sustainable Clouds

Kiran Voderhobli (2015). *International Journal of Organizational and Collective Intelligence* (pp. 8-19). www.irma-international.org/article/the-need-for-traffic-based-virtualisation-management-for-sustainable-clouds/137717

Principled Reference Data Management for Big Data and Business Intelligence

Sushain Pandit, Ivan Milman, Martin Oberhofer and Yinle Zhou (2017). *International Journal of Organizational and Collective Intelligence* (pp. 47-66). www.irma-international.org/article/principled-reference-data-management-for-big-data-and-business-intelligence/165392