Information Systems 77 (2018) 151–166

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

A new technique of selecting an optimal blocking method for better record linkage

Kevin O'Hare[a,*], Anna Jurek[a], Cassio de Campos[a,b]

[a] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Computer Science Building, 18 Malone Road, BT9 5BN Belfast, United Kingdom
[b] Buys Ballotgebouw, Utrecht University, Utrecht 3584 CC, The Netherlands

ARTICLE INFO

Article history:
Received 13 March 2018
Revised 15 June 2018
Accepted 16 June 2018
Available online 22 June 2018

Keywords:
Record linkage
Blocking
Entity resolution

ABSTRACT

Record linkage, referred to also as entity resolution, is the process of identifying pairs of records representing the same real world entity (e.g. a person) within a dataset or across multiple datasets. In order to reduce the number of record comparisons, record linkage frameworks initially perform a process referred to as blocking, which involves splitting records into a set of blocks using a partition (or blocking) scheme. This restricts comparisons among records that belong to the same block during the linkage process. Existing blocking methods are often evaluated using different metrics and independently of the choice of the subsequent linkage method, which makes the choice of an optimal approach very subjective. In this paper we demonstrate that existing evaluation metrics fail to provide strong evidence to support the selection of an optimal blocking method. We conduct an extensive evaluation of different blocking methods using multiple datasets and some commonly applied linkage techniques to show that evaluation of a blocking method must take into consideration the subsequent linkage phase. We propose a novel evaluation technique that takes into consideration multiple factors including the end-to-end running time of the combined blocking and linkage phases as well as the linkage technique used. We empirically demonstrate using multiple datasets that according to this novel evaluation technique some blocking methods can be fairly considered superior to others, while some should be deemed incomparable according to those factors. Finally, we propose a novel blocking method selection procedure that takes into consideration the linkage proficiency and end-to-end time of different blocking methods combined with a given linkage technique. We show that this technique is able to select the best or near best blocking method for unseen data.

1. Introduction

Record Linkage (RL) is a process of identifying and linking pairs of records representing the same real world entity. An overview of a general RL process is demonstrated in Fig. 1. As the number of record pairs that require comparison during linkage grows exponentially with dataset sizes, linkage often incurs great computational expense even for moderately sized datasets. For this reason, a blocking phase is implemented prior to linkage to reduce the otherwise high computational cost of exhaustively comparing all record pairs.

Blocking is a process of dividing records into groups (blocks) in such a way that records within each group hold a high chance of being linked in the subsequent linkage process. Following the blocking process, linkage is performed exclusively upon the record pairs within each of the generated blocks.

During a blocking process a set of blocking keys is used to determine which records should be placed in the same block. Consider a dataset of records $R = r_1, \ldots, r_n$, where each record comprises values it takes for attributes from a scheme $A = a_1, \ldots, a_m$. Accordingly, we can represent a record r_i as $[r_{i1}, \ldots, r_{im}]$, where r_{ij} is the value that the i^{th} record takes for the j^{th} attribute. A blocking key is defined as follows.

Definition 1.1. (Blocking key) A blocking key is an $\langle a_j, h \rangle$ combination where $a_j \in A$ is an attribute and h is an indexing function. For each $r_i \in R$, h takes r_{ij} as an input and provides a set of values, referred to as blocking key values (BKVs), as an output.

For example, the blocking key $\langle$*Name, Contain common tokens*$\rangle$ applied to a record containing "Information Systems Journal" in the

* Corresponding author.
E-mail addresses: kohare08@qub.ac.uk (K. O'Hare), a.jurek@qub.ac.uk (A. Jurek), c.decampos@uu.nl (C. de Campos).

https://doi.org/10.1016/j.is.2018.06.006
0306-4379/

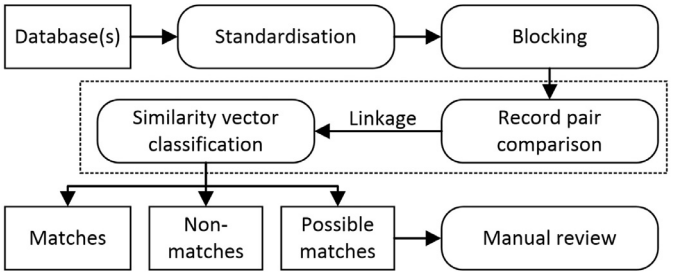

Fig. 1. General overview of record linkage process.

Name attribute field would generate a BKV set containing three BKVs {"Information", "Systems", "Journal"}. BKVs determine into which block(s) records are placed, with each unique BKV referring to a specific block. Our example record would therefore be placed in three different blocks, each associated with one of the three aforementioned BKVs.

A good blocking method places many matching record pairs and few non-matching record pairs into the generated blocks thus allowing for an efficient subsequent linkage phase. A number of different linkage methods exist which classify each record pair within each block as either match or non-match based on the similarity between them [12,15,16,21,24]. Due to the complexity of datasets (i.e. missing values, typographical errors, acronyms, initialisations, etc.) a single blocking key is rarely likely to capture all matching record pairs efficiently, therefore multiple blocking keys may be needed in the form of a blocking scheme.

Definition 1.2. (Blocking Schemes) Given a set of individual blocking keys, $K = k_1, \ldots, k_{k'}$, a blocking scheme is a combination of blocking keys, which can be disjunctive i.e. $\langle k_i \rangle \cup \cdots \cup \langle k_j \rangle$, conjunctive i.e. $\langle k_i \rangle \cap \cdots \cap \langle k_j \rangle$ or of disjunctive normal form i.e. $\langle \langle k_i \rangle \cap \cdots \cap \langle k_j \rangle \rangle \cup \cdots \cup \langle \langle k_{i'} \rangle \cap \cdots \cap \langle k_{j'} \rangle \rangle$

Blocking schemes may be created manually [13,15] or automatically learned [2,20,27] using a blocking scheme learning algorithm and labelled data.

Blocking methods are commonly evaluated with labelled data (with known matching status of each record pair) using evaluation metrics such as *reduction ratio (RR), pairs completeness (PC)* and/or a harmonic mean $F_{\mathrm{RR,PC}}$ of *RR* and *PC* [18].

Definition 1.3. (Reduction Ratio) For two datasets, A and B, reduction ratio is defined as:

$$RR = 1 - \frac{N}{|A| \times |B|}, \tag{1}$$

where $|A|$ and $|B|$ are the sizes of respective datasets and $N \leq (|A| \times |B|)$ is the number of record pairs formed by a blocking method.

RR indicates how much the comparison space is reduced after the blocking phase. For example, if a potential comparison space of 1,000,000 record pairs was reduced by blocking to 5000 record pairs, that would equate to $RR = 1 - (5,000/1,000,000) = 0.995$.

Definition 1.4. (Pairs completeness) Pairs completeness is defined as:

$$PC = \frac{N_m}{|M|}, \tag{2}$$

with $N_m \leq |M|$ being the number of matching record pairs contained within the reduced comparison space after blocking and $|M|$ being the number of matches within the entire dataset.

PC is the ratio of matching record pairs found within the formed blocks. One can notice that there is a trade-off between RR and PC. Comparing all record pairs (placing all the records in the same block) minimises RR but maximises PC, whereas performing no comparisons at all (placing each record in an individual block) maximises RR and minimises PC. Ideally one looks for a blocking scheme that maximises both RR and PC. A commonly applied evaluation metric, which balances the trade-off between RR and PC, is the harmonic mean of RR and PC.

Definition 1.5. (Harmonic mean of RR and PC) For a given RR and PC, the harmonic mean is defined as:

$$F_{\mathrm{RR,PC}} = \frac{2 * \mathrm{RR} * \mathrm{PC}}{\mathrm{RR} + \mathrm{PC}}. \tag{3}$$

In this paper we make the following contributions: (1) We compare existing blocking scheme learning methods using results from the respective papers to show that current blocking evaluation metrics are insufficient when performed independently of a subsequent linkage phase. Analysis of these results show that no blocking method is superior to the others in every instance, and that the choice highly depends on which evaluation metric is prioritised. (2) We propose a novel technique that evaluates blocking methods as part of an RL framework (i.e. takes the quality and runtime of the subsequent linkage into consideration) and visualises the results graphically. This allows an optimal blocking method to be easily identified according to multiple factors, including resources (in particular running time), datasets and linkage method. (3) We propose a new selection technique that uses results obtain by blocking methods on known labelled datasets to select an optimal blocking method for a new unlabelled dataset for a given RL method. We perform a number of experiments using different blocking methods and some of the commonly used RL methods. We compare the results of our selected methods against all others on multiple datasets to show that an optimal or near optimal blocking method is selected in every case.

2. Relevant work

Automatic blocking scheme learning approaches [2,20,27] commonly evaluate an initial set of individual blocking keys against a set of labelled data. The best individual keys, according to a predetermined criterion, continue to iteratively form blocking schemes with remaining individual keys often re-ranked between iterations. These schemes are then evaluated against labelled data using evaluation metrics. A blocking key or a blocking scheme is commonly evaluated with reduction ratio (RR), pairs completeness (PC) and harmonic mean of RR and PC ($F_{\mathrm{RR,PC}}$), following the blocking phase.

The supervised approach in [27] ranks individual keys with PC above a predetermined threshold by RR. Each top key is extended by other keys as conjunctions so RR improves while maintaining PC above the threshold. This continues until RR no longer improves for each conjunction. The idea is that although each individual conjunction may only cover a certain proportion of the matches, their disjunction will collectively detect most if not all matches. The proficiency of learned blocking schemes against different datasets are presented using RR and PC. While this paper presents good results, they are presented independently of computational run-time.

Another supervised approach [2] ranks keys by their ratio of detected matches to non-matches. Top keys are then iteratively applied to the labelled record pairs as a disjunctive blocking scheme until a predetermined proportion of labelled positives are detected. Disjunctive normal form schemes may also be learned by iteratively extending each top key by others so that the ratio is maximally improved. This continues until a conjunction of desired length is generated. The individual keys are supplemented by the conjunctions formed at each iteration. The supplemented set is then ranked and iteratively applied to the