Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210203

On the Concepts of Identity and Similarity in the Context of Biomedical Record Linkage

Murat SARIYAR[a,1] and Jürgen HOLM[a]
[a] *Bern University of Appl. Sciences, Dept. Medical Informatics, Switzerland*

Abstract. Record linkage refers to a range of methods for merging and consolidating data in a manner such that duplicates are detected and false links are avoided. It is crucial for such a task to discern between similarity and identity of entities. This paper explores the implications of the ontological concepts of identity for record linkage (RL) on biomedical data sets. In order to draw substantial conclusions, we use the differentiation between numerical identity, qualitative identity and relational identity. We will discuss the problems of using similarity measures for record pairs and quality identity for ascertaining the real status of these pairs. We conclude that relational identity should be operationalized for RL.

Keywords. Identity; record linkage; reference reconciliation; ontology matching; Levenshtein similarity

1. Introduction

Many areas of health and biomedical research require the application of record linkage (RL) methods as a preparatory step for a large number of retrospective studies using datasets from different sources [1–3]. Cross-organizational linking data of different electronic health record systems is an example for biomedical RL in the context of healthcare [4]. Two central goals for linking records are (i) increasing the size of the study population (horizontal enrichment) and (ii) enriching the information on the individuals involved (vertical enrichment). Both goals serve the overall aim of increasing the utility of available data in terms of evidence-based medicine.

RL or reference reconciliation [5] can be defined as the process of detecting records that refer to the same real-world entities but contain discrepancies, due, for example, misspelled names, changes of zip codes or different diagnoses [6]. RL is necessary when the sources do not have a common unique entity identifier. Usually, record pairs are then formed and an algorithm decides whether two records belong to the same entity (the pair is called a match: M) or to two different entities (the pair is called a non-match: U).

There are different categories of methods, especially from the fields of machine learning and classical statistics [7, 8]. In all these approaches, it is necessary to compute similarities between the entities to be linked. It is common to assume structured data, e.g., tables such as *Patient*(firstname, lastname, place, dbirth, mbirth, ybirth, sex), which

[1] Corresponding Author, Murat Sariyar, Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland; E-mail: murat.sariyar@bfh.ch.

represent the entities to be linked. In order to keep the complexity low, the similarities are usually computed in a pairwise comparison, leading to so-called comparison patterns (denoted as γ). For example, the following data pairs

('Urs', 'Schmidt', 'Bern', '18', '11', '1990', 'm')
('Urs', 'Schmitt', 'Berne', '18', '11', '1990', 'm')

leads to the following corresponding binary comparison pattern γ:

(1, 0, 0, 1, 1, 1, 1).

Continuous comparison patterns allows for higher discernibility between matches and non-matches. They can be computed by using string metrics such as the Levenshtein metric, which in our example leads to the following comparison pattern (using the levenshteinSim function in the R-package RecordLinkage [9]):

(1, 0.86, 0.8, 1, 1, 1, 1).

In probabilistic RL applications, conditional probabilities $P(\gamma|M)$ and $P(\gamma|U)$ are estimated [10], used for computing the global weight, for which a threshold for definite matches has to be determined:

$$\mathrm{w} = \log\left(\frac{(P(\gamma|M)}{(P(\gamma|U)}\right) \quad (1)$$

Relying only on similarities of such data pairs and comparison patterns in order to determine the match status (identity) has three related drawbacks. First, contextual information that could help in discerning between entities is lacking. Second, changing semantics between different data sources is not considered. Both issues are tackled by ontology-based semantic enrichments [11-13]. However, the most crucial point is the lack of consideration regarding the concept of identity (third drawback). Why should such a consideration matter? There are two main reasons: (i) there is a need to differentiate between changes in the same entity (e.g., the same virus exhibits new characteristics) and differences that relate to different entities (e.g., a new virus has emerged) and (ii) there is a lack of guidance on how to resolve synonyms (false non-matches) and homonyms (false matches), especially in the training phase of the algorithms. In the following, we provide a first step towards the translation of philosophical insights related to the concept of identity into the domain of RL.

2. Methods: Concepts of Identity

In ancient philosophy, the problem of determining identity as persistence through time is discussed by developing certain thought experiments. A prominent example is the ship of Theseus. The central question is whether the replacement of all components (materials) and changes in the shape of the ship are compatible with its persistence through time. More generally, can an enduring entity gain and/or shed arbitrary many properties without losing its identity? In the domain of ontology engineering, this question arises in the connection of data objects that change over time.

Let us return to the ship of Theseus: Since there is no definite answer to this paradox, it is common practice in philosophy to introduce distinctions in order to develop a better grasp of such problems. One distinction is between numerical and qualitative identity. Numerical identity means, that the same entity is referenced at different times, even though it might be denoted differently. For example, in astronomy, the morning star and the evening start are one and the same entity, the planet Venus. Qualitative identity is given when two entities have the same properties (they are indiscernible), but are not necessarily one and the same. An example is given by two biosamples that share the

same characteristics, e.g., genomic structure, uniformity of cell size, uniformity of cell shape, etc. Unlike the concept of similarity in RL, qualitative identity is not restricted to locally available attributes, but is related all determinable ones. The definition of numerical identity provides no guidance for the practice, since it remains unclear on what basis the different signifiers (labels or words) are defined as referring to the same entity. However, the definition of qualitative identity can be used to operationalize the concept of identity as it refers to the process of performing comparisons.

Operationalization of the identity concept by qualitative identity can be based on relational identity: Listing of all relations that an entity x has within itself (its attributes) and to other entities. In the case of a biosample, the NCBI "BioSample" attributes provide the basis for such a list. Again, the idea behind relational identity is that all entities get their identity through relations within themselves and with their environment, not just by assigning attribute values. For example, the name of a patient is not just an attribute of the patient, but a relation between the patient and those authorities that have certified and validated that name assignment. Hence, relations such as "identified by" or "certified by" are part of the relational identity of the patient.

3. Results: Consequences for the Record Linkage Practice

Using similarity measures to determine the matching status in RL applications is an application of qualitative identity in a loose sense: It is sufficient that some properties between two entities have high similarity in order to deduce numerical identity (match assignment). Due to its simplicity, such an approach cannot account for real-word changes of characteristics or for accidental similarities (i). But even if it works well in practice, the clerical reviews for homonyms and synonyms show that there are additional mechanisms for determining (numerical) identity, taking context information into consideration (ii). This reliance on implicit and contextual knowledge is unusual for a discipline that seeks to automate or to enhance decisions. One way to make such knowledge explicit is to use ontologies with their relational approach.

To enhance the practice of RL based on the qualitative identity concept that is concretized as relational identity, we propose a hierarchical approach: At the top, an ontology for the RL context is defined, which includes relations and attributes, e.g., ontology for clinical research. Based on this ontology, attributes are semantically enriched in order to be able to map different descriptions and denotations of the same concepts, e.g., mapping the attributes "target variable" and "outcome" used at different sites. At the bottom, an RL method is applied that is able to consider both attributional as well as relational similarity. Relational similarity compares the relationships of concepts (A::B) in a relation with those of other relations (C::D). Values for such relationships of concepts can be generated by word embeddings [14]. This ontology-based approach, motivated by the qualitative identity concept, allows the consideration of relational (structural) and semantic methods in addition to string-based methods. The RL methods using string metrics are often more sophisticated than their counterparts used in ontology matching, due to the focus on similarity between record pairs, reglardless of semantics. Hence, there are many extensions, for example by considering frequency of attribute values [15] or missing values [16].

In the end, this hierarchical approach realizes qualitative identity in a systematic and reflective way. In practice, the false matches and non-matches of RL methods will be reduced, as more context information is included (we ignore data protection issues here).

Regarding semantic enrichment as part of our approach, one reviewer suggested the following example related to LOINC (Logical Observation Identifiers Names and Codes): Codes 94762-2 and 94562-6 are attribute values for measuring SARS-CoV-2 antibodies. By looking up the definitions for both codes, a "94562-6 is-a 94762-2" relation can be established, which allows to match two different SARS-CoV-2 tests performed in different laboratories that use different levels of granularity when coding observations with LOINC.

Even with such an extended approach to RL, numerical identity cannot be guaranteed. As described above, numerical identity is a concept used in practice to describe our conclusion in terms of "Yes, the same biosample is represented by these two records". It has value for reflection (such as in this paper), not for direct guidance of RL practice. However, indirectly it has a huge impact, since the profundity of a theory heavily depends on the depths of the problems it sets out to tackle. Considering identity as a notion relevant for RL allows for a broader perspective, especially with respect to the self-understanding of RL as a scientific field. In addition to that, discussions with respect to master patient identifiers and electronic IDs also benefit from such a perspective.

4. Discussion

One important issue that should be investigated further is the temporal dimension of entities to be linked. In order to capture relationships throughout the history of an entity, temporal entities should be differentiated from enduring entities, e.g., an event such as an accident in contrast to the persons involved. The basic fundamental ontology (BFO), which is used as an upper-level template ontology for many biomedical fields, provides such a categorization [17]. The fundamental distinction of the BFO is between continuants (entities that persist through time while maintaining their identity) and occurrents (entities that unfold themselves in time or are the instantaneous boundaries of such entities without preserving any identity besides being registered). The additional differentiations, such as processes and temporal regions, allow a better grasp of the time dimension, when describing relationships between entities.

Even though the paper uses some philosophical notions and examples, it is intended for the field of medical informatics. Our assumption is that biomedical RL will benefit from philosophical insights into the concepts of identity, but some further work of adaptation is necessary. The next steps will be to develop an ontology-based framework for RL and to concretize identity concepts. In this connection, it will be shown that other areas, such as data protection and even virology, will benefit as well from this future work. How might virology profit from the elaboration of identity within the RL context? For example, by enriching phylogenetic analysis with the concept of relational identity. The problem of tracing back the origin of a virus through genetic analysis is similar to the paradox of the ship of Theseus: After how many mutations does a new type of virus emerge? The lack of clear decision rules for grouping non-leaf nodes in phylogenetic trees into taxonomic units reflects this uncertainty. Alignment algorithms for generating these trees represent the RL approach to sequence similarity, which is supported by the fact that the Levenshtein algorithm for strings, which is heavily used in RL, is just another name for the Needleman-Wunsch algorithm for sequences used in bioinformatics.

References

[1] Jaro MA. Probabilistic Linkage of Large Public-Health Data Files. Stat Med. 1995;14:491–498.
[2] Christen P, Churches T, Hegland M. Febrl - A parallel open-source data linkage system. Adv Knowl Discov Data Min Proc. 2004;3056:638–647.
[3] Bell GB, Sethi A. Matching records in a national medical patient index. Commun ACM. 2001;44:83–88.
[4] Hejblum BP, Weber GM, Liao KP, et al. Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. Sci Data. 2019;6:180298.
[5] Dong X, Halevy A, Madhavan J. Reference Reconciliation in Complex Information Spaces. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. New York: 2005; p. 85–96.
[6] Herzog TN, Scheuren FJ, Winkler WE. Data Quality and Record Linkage Techniques. New York: Springer 2007.
[7] Roos LL, Wajda A, Nicol JP. The Art and Science of Record Linkage - Methods That Work with Few Identifiers. Comput Biol Med. 1986;16:45–57.
[8] Sariyar M, Borg A, Pommerening K. Evaluation of record linkage methods for iterative insertions. Methods Inf Med. 2009;48:429–437.
[9] Sariyar M, Borg A. The RecordLinkage Package: Detecting Errors in Data. R Journal. 2010;2:61–67.
[10] Fellegi IP, Sunter AB. A Theory for Record Linkage. J Am Stat Assoc. 1969;64:1183–1210.
[11] Castano S, Ferrara A, Montanelli S, et al. Ontology and Instance Matching. In: Paliouras G, Spyropoulos CD, Tsatsaronis G (eds). Knowledge-Driven Multimedia Information Extraction and Ontology Evolution. Springer Berlin Heidelberg. 2011. p. 167–195.
[12] Batet M, Sánchez D, Valls A, et al. Semantic similarity estimation from multiple ontologies. Appl Intell. 2013;38:29–44.
[13] Gagnon M. Ontology-based integration of data sources. In: 2007 10th International Conference on Information Fusion. 2007; p. 1–8.
[14] Zhu VJ, Overhage MJ, Egg J, et al. An Empiric Modification to the Probabilistic Record Linkage Algorithm Using Frequency-Based Weight Scaling. J Am Med Inform Assoc. 2009;16:738–745.
[15] Sariyar M, Borg A, Pommerening K. Missing values in deduplication of electronic patient data. J Am Med Inform Assoc. 2012;19:e76-82.
[16] Fathiamini S, Johnson AM, Zeng J, et al. Rapamycin - mTOR + BRAF =? Using Relational Similarity to Find Therapeutically Relevant Drug-Gene Relationships in Unstructured Text. J Biomed Inform. 2019;90:103094.
[17] Arp R, Smith B, Spear AD. Building Ontologies with Basic Formal Ontology. The MIT Press 2015.