Information Processing and Management 000 (2016) 1–12

Contents lists available at ScienceDirect

Information Processing and Management

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

A parser for authority control of author names in bibliographic records

Rafael C. Carrasco[a,*], Aureo Serrano[a], Reydi Castillo-Buergo[b]

[a] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, Spain
[b] Departamento de Computación, Universidad Agraria de La Habana, La Habana, Cuba

ARTICLE INFO

Article history:
Received 10 June 2015
Revised 8 February 2016
Accepted 16 February 2016
Available online xxx

Keywords:
Digital libraries
Cataloguing standards
Natural language processing

ABSTRACT

Bibliographic collections in traditional libraries often compile records from distributed sources where variable criteria have been applied to the normalization of the data. Furthermore, the source records often follow classical standards, such as MARC21, where a strict normalization of author names is not enforced. The identification of equivalent records in large catalogues is therefore required, for example, when migrating the data to new repositories which apply modern specifications for cataloguing, such as the FRBR and RDA standards. An open-source tool has been implemented to assist authority control in bibliographic catalogues when external features (such as the citations found in scientific articles) are not available for the disambiguation of creator names. This tool is based on similarity measures between the variants of author names combined with a parser which interprets the dates and periods associated with the creator. An efficient data structure (the unigram frequency vector trie) has been used to accelerate the identification of variants. The algorithms employed and the attribute grammar are described in detail and their implementation is distributed as an open-source resource to allow for an easier uptake.

1. Introduction

Authority control in a catalogue is defined as the maintenance of the consistency of index terms for bibliographic material. In the particular case of the author names stored in a standard catalogue format, such as the MARC21 standard (Library of Congress, Network Development & MARC Standards Office, 2009), authority control serves two main purposes:

- Distinguishing creators who have published under the same name—for example, Leopoldo Alas (1852–1901, aka Clarin) and his descendant Leopoldo Alas (1962–2008)—by adding titles or other words associated with the name, or by including information about the creator's birth, death or active dates.
- Identifying variants of the name, for instance, pen names like Fígaro (Mariano José de Larra, 1809–1837) or alternative spellings like Pedro Fernandes de Queirós and Pedro Fernández de Quirós (1565–1614).

Authority control is usually assisted by an index which helps cataloguers to identify previous occurrences of a name. However, this index cannot be guaranteed to be totally free from errors and the maintenance of catalogue entries is a routine task at libraries. Moreover, whenever open-access catalogues integrate records from different sources, it is possible that

* Corresponding author. Tel.: +34 965903978.
E-mail address: carrasco@ua.es (R.C. Carrasco).

http://dx.doi.org/10.1016/j.ipm.2016.02.002
0306-4573/

some libraries or cataloguing departments have employed different authoritative names. Remarkably, the metadata exchange protocols —such as OAI-MPH[1] (de Sompel, Nelson, Lagoze, & Warner, 2004) and SRU[2] (Network Development & MARC Standards Office, Library of Congress, 2007)—, neither address the quality control of metadata nor implement procedures to harmonize variants (Salo, 2009).

Some frequent types of inconsistency found in catalogue names include:

- Variants of the same name, like *Jan Moretus* catalogued sometimes as *Joannes Moretus* and also as *Jean Moretus*.
- Name permutations such as *Carlos Borromeo* and *Borromeo, Carlos*.
- Typos, for example, *Alfonso Díaz de Montalvo* instead of *Alonso Díez de Montalvo*.
- Stop word removal, as in *Vicente Zea* for *Vicente de Zea* or *Belén Bañas* replacing *María Belén Bañas*.
- Removal of diacritics, as in *Schoner, Johann* for *Schöner, Johann* or *Salcedo Coronel, Garcia de* for *Salzedo Coronel, García*.

However, it is not necessarily the case that two names with minor differences in their spelling correspond to a unique person. For instance, *Francisco de la Puente* and *Francisco de la Fuente* are two different persons who coexist in the catalogue of the Biblioteca Virtual Miguel de Cervantes (BVMC, 1999).

Clearly, these inconsistencies pose a challenge for the effective retrieval of bibliographic items. The application of a unique identification number for each author, such as the International Standard Name Identifier (ISO 27729), has been advocated for a long time as the solution to this problem—see, for instance, Snyman and van Rensburg (2000). Although some progress in this direction has been made (Hickey & Toves, 2014), the shared identifier approach is still far form being universally adopted and, of course, it is difficult to apply to the vast collection of former records.

A large number of techniques and tools have been developed to assist the maintenance of the creator names in catalogue records (Ferreira, Gonçalves, & Laender, 2012). For instance, Chávez-Aragón, Ramirez Cruz, Reyes-Galaviz, Ayanegui-Santiago, and Portilla (2009) use a simple Euclidean distance between feature vectors to identify variants of a name. The components in the feature vector collect semantic information such as the keywords, title words or coauthor lists. A similar approach has been followed by Lee (2007) to trace evolving names and to design a system supporting the maintenance of metadata that change over time. The cosine distance (using normalized names, titles and publication venues as features) has been used instead to measure the similarity between records by de Carvalho, Ferreira, Laender, and Gonçalves (2011) and by Cota, Ferreira, Nascimento, Gonçalves, and Laender (2010).

Some methods for name authority control are based on Bayesian classification (Warnner & Brown, 2001) and transform every text field into a vector of words and follow an incremental, supervised approach: a subset is manually disambiguated and then used as the seed for a clustering process guided by author names and their variants (for example, as found in the Library of Congress authority file), and also exploit contextual evidences such as the publication dates (which should be consistent with author's lifetime). For instance, the Levy II suite (DiLauro, Choudhury, Patton, Warner, & Brown, 2001) employs an adaptive Bayesian probability model and a threshold that triggers human intervention. Tang, Fong, Wang, and Zhang (2012) applied Hidden Markov Random Fields trained with Expectation Maximization as an alternative to traditional clustering algorithms. Machine learning techniques have been also traditionally applied to this problem (Han, Xu, Zha, & Giles, 2005; Torvik & Smalheiser, 2009). More recently, Ferreira, Veloso, Gonçalves, and Laender (2014) implemented a bootstrapping procedure for those cases where training data are missing or scarce. Following this line of work, a specific similarity function, based on terms appearing in the list of coauthors, publication and venue titles, was defined (Santana, Gonçalves, Laender, & Ferreira, 2015) and used in combination with several heuristics to improve upon previous results.

Alternatively, names can be considered as sequences of characters to which string similarity measures can be applied. For instance, Cohen, Ravikumar, and Fienberg (2003) compared different metrics to evaluate the similarity between author names and conclude that hybrid methods, such as the Monge–Elkan measure (Monge & Elkan, 1997), improve the results over the traditional Levenshtein distance (Levenshtein, 1966).

This paper presents a method which parses temporal annotations in bibliographic records in order to disambiguate author names. Due to the weak normalization of such expressions, a variety of forms must be interpreted. We have defined for this purpose an attribute grammar (Aho, Sethi, & Ullman, 1986) which parses valid dates and gives them a non-ambiguous interpretation as a temporal range or period. This grammar, described in more detail in Section 2, interprets a date expression either as a year, a century or period and associates an uncertainty to each temporal unit. The output is later used as complementary information to check the compatibility between authors.

The comparison of author names, described in Section 3, is based on simple measures for string similarity and employs a compact data structure to accelerate the search for name variants. The full method is specifically designed for those cases where additional features—such as the publication venue or the cross citations—are missing, a common case in humanistic and literary libraries such as the Miguel de Cervantes digital library (BVMC, 1999). Section 4 analyses the results obtained when the method was applied to a real collection of bibliographic records in that library and, finally, Section 5 presents the conclusions.

[1] Open Archives Initiative Protocol for Metadata Harvesting.
[2] Search/Retrieve via URL.

Please cite this article as: R.C. Carrasco et al., A parser for authority control of author names in bibliographic records, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.02.002