Information Processing and Management 47 (2011) 706–718

Contents lists available at ScienceDirect

Information Processing and Management

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

An unsupervised heuristic-based approach for bibliographic metadata deduplication

Eduardo N. Borges [a,*], Moisés G. de Carvalho [b], Renata Galante [a], Marcos André Gonçalves [b], Alberto H.F. Laender [b]

[a] Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
[b] Computer Science Dept., Federal University of Minas Gerais, Belo Horizonte, Brazil

ARTICLE INFO

Article history:
Received 22 July 2009
Received in revised form 20 November 2010
Accepted 24 January 2011
Available online 17 February 2011

Keywords:
Digital libraries
Metadata
Deduplication
Similarity

ABSTRACT

Digital libraries of scientific articles contain collections of digital objects that are usually described by bibliographic metadata records. These records can be acquired from different sources and be represented using several metadata standards. These metadata standards may be heterogeneous in both, content and structure. All of this implies that many records may be duplicated in the repository, thus affecting the quality of services, such as searching and browsing. In this article we present an approach that identifies duplicated bibliographic metadata records in an efficient and effective way. We propose similarity functions especially designed for the digital library domain and experimentally evaluate them. Our results show that the proposed functions improve the quality of metadata deduplication up to 188% compared to four different baselines. We also show that our approach achieves statistical equivalent results when compared to a state-of-the-art method for replica identification based on genetic programming, without the burden and cost of any training process.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Digital libraries (DLs) are complex information systems built to address the information needs of specific target communities (Gonçalves, Fox, Watson, & Kipp, 2004). DLs are composed of collections of rich (possibly multimedia) digital objects along with services such as searching, browsing and recommendation, that allow easy access and retrieval of these objects by the members of the target community (Fox, Akscyn, Furuta, & Leggett, 1995; Gonçalves et al., 2004).

Collections of digital objects are usually described by means of metadata records (usually organized in a metadata catalog) whose function is to describe, organize and specify how these objects can be manipulated and retrieved, including who has the rights for doing so. In order to promote interoperability among DLs and similar systems, metadata records usually conform to one or more metadata standards that specify, among others, a standardized set of metadata fields and their semantics for the description of digital objects. The Dublin Core,[1] for example, is a general descriptive metadata standard for the representation and storage of information about scientific publications and Web pages.

Although very useful, these standards do not completely solve all the interoperability problems as there is not a consensus among all existing digital libraries in terms of a unique 'de facto' standard. Moreover, even if such a consensus existed,

* Corresponding author. Tel.: +55 51 33087746; fax: +55 51 33087308.
E-mail addresses: enborges@inf.ufrgs.br (E.N. Borges), moisesgc@dcc.ufmg.br (M.G. de Carvalho), galante@inf.ufrgs.br (R. Galante), mgoncalv@dcc.ufmg.br (M.A. Gonçalves), laender@dcc.ufmg.br (A.H.F. Laender).

[1] http://dublincore.org.

0306-4573/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2011.01.009

BDBComp

```
1 <title>A Computer Vision Framework for Remote Eye Gaze Tracking</title>
2 <creator>Carlos H. Morimoto</creator>
3 <source>sibgrapi2003</source>
```

DBLP

```
4 <title>A Computer Vision Framework for Eye Gaze Tracking</title>
5 <author>Carlos Hitoshi Morimoto</author>
6 <booktitle>SIBGRAPI</booktitle>
```

IEEE Xplore

```
7 <title>A computer vision framework for eye gaze tracking</title>
8 <author>Morimoto, C.H.</author>
9 <pages>406</pages>
```

Fig. 1. Heterogeneity of metadata.

differences in practices and in the way some metadata elements are filled, not mentioning possible errors in this process (e.g., misspellings and typos), allow for the existence of several different records describing the same digital object.

Consider the example of Fig. 1 that presents excerpts of metadata records from three distinct digital libraries: BDBComp,[2] DBLP[3] and IEEE Xplore.[4] All records refer to the same digital object. The field *source* in the BDBComp metadata record (line 3) corresponds to the field *booktitle* from DBLP (line 6). The metadata structures are different, but both refer to the same information, i.e., the publication venue of a specific paper. Also, the author of the paper, which is represented by the *creator* and *author* fields, has the value "Carlos H. Morimoto" in the BDBComp record (line 2), "Carlos Hitoshi Morimoto" in the DBLP record (line 5) and "Morimoto, C.H." in the IEEE Xplore record (line 8). The value of the *title* field also differs in the word "Remote" (lines 1, 4 and 7).

Deduplication is the task of identifying in a data repository duplicated records that refer to the same real-world entity. These records may be hard to identify due to, as mentioned before, variations in spelling, writing style, metadata standard use, or even typos (Carvalho, Gonçalves, Laender, & da Silva, 2006). Deduplication is also known as record linkage, object matching or instance matching (Doan, Noy, & Halevy, 2004). In fact, this is not a new problem but a long-standing issue, for example, in the Library and Information Science field, in the context of Online Public Access Catalogs (OPACs) (Large & Beheshti, 1997), as well as in the database realm.

Several approaches to record deduplication have been proposed in recent years (Bilenko & Mooney, 2003; Dorneles et al., 2009; Carvalho et al., 2006; Carvalho, Laender, Gonçalves, & da Silva, 2008; Chaudhuri, Ganjam, Ganti, & Motwani, 2003; Cohen & Richman, 2002; Tejada, Knoblock, & Minton, 2001). Most of these works focus on the deduplication task in the context of integration of relational databases. Few automatic approaches, if any, have been specifically developed for the realm of digital libraries or in a more general sense, for bibliographic metadata records. For example, metadata fields that specify the authors of a digital object are some of the most discriminative fields of a record and this information should be used as a strong evidence for the deduplication process. In fact, there may be several objects with similar titles but there is a very small chance that they will also have authors with similar names and be a different real-world object. For instance, Baeza-Yates and Ribeiro-Neto as well as Manning have published books with similar titles (Baeza-Yates & Ribeiro-Neto, 1999; Manning, Raghavan, & Schütze, 2008). Another specific problem to deal with is the variation in the way author names are represented in bibliographic citations. Variations include abbreviations, inversions of names, different spellings and omission of suffixes as Jr. (Ley, 2002). Deduplication techniques applied to the digital libraries domain should therefore take into special consideration the fields that refer to author names to correctly identify duplicated metadata records. These techniques may even explore a number of other sources of information such as authority files to help with the task of comparing author names, although this is not the focus of this work.

This article presents an approach to identifying duplicated bibliographic metadata records. We assume that a mapping between the metadata fields in different standards is provided and we focus on the application of specially designed similarity functions for the metadata content. We are aware of the problem of schema matching (Rahm & Bernstein, 2001), however it is not the focus of our work; recent solutions in the literature for the problem could be used (Fagin et al., 2009). Here, instead, we are interested in the instance matching problem, specifically for the realm of digital libraries. In this context, the main contributions of this work are:

- an efficient and effective approach for metadata record deduplication that is based on a set of similarity functions specially designed for the digital library domain;
- the identification and analysis of the failure cases of the evaluated deduplication functions, which are valuable for the development of new approaches for automatic bibliographic metadata deduplication.

[2] http://www.lbd.dcc.ufmg.br/bdbcomp.
[3] http://www.informatik.uni-trier.de/ley/db.
[4] http://www.ieeexplore.ieee.org.

Download English Version:

https://daneshyari.com/en/article/515044

Download Persian Version:

https://daneshyari.com/article/515044

Daneshyari.com