Chapter XXIII
Duplicate Journal Title Detection in References

Ana Kovacevic
University of Belgrade, Serbia

Vladan Devedzic
University of Belgrade, Serbia

ABSTRACT

Our research efforts are oriented towards applying text mining techniques in order to help librarians make more informative decisions when selecting learning resources to be included in the library's offer. The proper selection of learning resources to be included in the library's offer is one of the key factors determining the overall usefulness of the library. Our task was to match abbreviated journal titles from citations with journals in existing digital libraries. The main problem is that for one journal there is often a number of different abbreviated forms in the citation report, hence the matching depends on the detection of duplicate records. We used character-based and token-based metrics together with a generated thesaurus for detecting duplicate records.

INTRODUCTION

Digital libraries need to continuously improve their collections. Knowing how a digital library and its collection are used is inextricably tied to the library's ability to sustain itself, improve its services, and meet its users' needs (McMartin, Iverson, Manduca, Wolf, & Morgan, 2006).

In Serbia, the major provider of digital learning resources is KOBSON[1] (Consortium of Serbian Libraries), which provides Serbian students, teachers, and researchers with access to foreign journals and other learning resources (Kosanović, 2002). Since the available funds are rather modest, the appropriate selection of journals to be made available through KOBSON is highly important and poses a challenge for their staff. Accordingly, our research efforts are aimed at helping librarians in general and KOBSON staff in particular to identify the journals that would be of interest

to their users in order to include those journals in the library collection. In addition, we aim to help to improve the services offered to KOBSON's users so that the users can find the resources they are interested in more easily.

This chapter presents part of our current work on the realization of the illustrated research challenge. Specifically, we are working on the identification of those journals that are frequently used by domestic researchers and are therefore relevant for inclusion in KOBSON's offer. An indication that a certain journal is read and considered important by a researcher is the appearance of the journal's title in the citations of the researcher's published papers. However, manual analysis of citations is impossible due to the large volume of data that need to be processed. Likewise, automatic analysis is impeded by the fact that authors tend to use different kinds of abbreviations when writing citations. Accordingly, we are currently working on matching the journal title abbreviations found in citations with the journals (i.e., their full titles) in the KOBSON digital libraries.

In the following section we present the problem of data heterogeneity that impedes the matching process and offer solutions for resolving this problem in digital libraries.

DATA HETEROGENITY

In the real world, data are not perfectly clean and there are various reasons for that, such as data entry errors, missing check constraints, lack of standardization in recording data in different sources, and so forth. In general, data originating from different sources can vary in value, structure, semantics and the underlying assumptions (Elmagarmid et al, 2007). This is the problem of data heterogeneity. There are two basic types of data heterogeneity: structural (differently structured data in different databases) and lexical (diverse representations of the same word entity) (Elmagarmid, Ipeirotis, & Verykios, 2007). The task of lexical heterogeneity has been explored in different research areas, such as statistics, databases, data mining, digital libraries, and natural language processing. Researchers in different areas have proposed various techniques and refer to the problem differently: record linkage (Newcomb & Kennedy, 1962), data duplication (Sarawagi & Bhamidipaty, 2002), database hardening and name matching (Bilenko, Mooney, Cohen, Ravikumar, & Fienber, 2003), data cleaning (McCallum & Wellner, 2003) or object identification (Tejada, Knoblock, & Minton, 2002), approximate matching (Guha, Koudas, Marathe, & Srivastava, 2004), fuzzy matching (Ananthakrishna, Chaudhuri, & Ganti, 2002), and entity resolution (Benjelloum, Garcia-Molina, Su, & Widom, 2005).

Data heterogeneity can have a negative impact on many common data library services. In our work we are addressing the problem of lexical heterogeneity in general and duplicate record detection in particular. The technique for matching fields depends on the particular problem, and there is no absolute solution. Basically, these techniques may be classified into the following categories:

- Character-based similarity metrics, which consider distance as the difference between characters, and is useful in the case of typographical errors (i.e., Levenshtein distance) (Levenshtein, 1966), Jaro-Winkler metrics (Winkler, 1995), and Q-Grams (Ukkonen, 1992).
- Token-based similarity metrics, which is based on statistics for common words, is useful when word order is not important (i.e., atomic strings [Monge & Elkan, 1996] and WHIRL [Cohen, 1998]).
- Phonetic similarity metrics, based on the fact that strings may be phonetically similar even if they are not similar in character or token level (i.e., double metaphone) (Philips, 2000).