Author Disambiguation in the YADDA2 Software Platform

Piotr Jan Dendek, Mariusz Wojewódzki, and Łukasz Bolikowski

Interdisciplinary Centre for Mathematical and Computational Modelling,
University of Warsaw, Warsaw, Poland
{p.dendek,m.wojewodzki,l.bolikowski}@icm.edu.pl

Abstract. SYNAT platform powered by the YADDA2 architecture has been extended with the Author Disambiguation Framework and the Query Framework. The former framework clusters occurrences of contributor names into identities of authors, the latter answers queries about authors and documents written by them. This paper presents an outline of the disambiguation algorithms, implementation of the query framework, integration into the platform and performance evaluation of the solution.

Keywords: author disambiguation, record deduplication, software architecture, YADDA2 software platform.

1 Introduction

1.1 Record Deduplication

A common challenge among databases is a record deduplication, which is the term describing the situation when a real-world object is described by many separate records. “Record deduplication” itself is known in different communities as “record linkage” [11], “data cleaning” [28], “data scrubbing” [3], “mirror detection” [41], “instance matching” [2], etc. This issue may occur as a result of multiple formats used for representing an attribute, for example in the case of an address or a person name. The problem is particularly acute when information is gathered over a long period of time. Even when there is only one standard of representing a record, some misspelling or diacritics handling issues may occur. Another concern is merging data from multiple heterogeneous sources, when normalization levels and record definitions may not match. Last but not least, there are cases of automatic data acquisition (e.g. from the Internet), which is an instance of combining information from many origins. Those concerns are especially problematic when coping with big datasets.

The challenge of the accurate record linkage had been addressed many times over the last five decades. Elmagarmid et al. [10] et al. coherently enumerate all mainstream approaches to the record linkage. Authors focus on specific aspects of the problem, which are:
1. Data preparation, covering a parsing and a standardization procedures.  
2. Attribute matching techniques, as approximate string matching, token based and phonetic based.  
3. Duplicate detection, covering supervised and unsupervised machine learning techniques as well as hand-crafted ones.  
4. Problem decomposition approaches as blocking, k-nearest neighbour, clustering.  

1.2 Author Disambiguation

Author disambiguation is an instance of the record linkage problem, where instances to match are authors, typically represented by first names, a surname, an affiliation and metadata of co-authored articles. It is clear that none of the mentioned attributes can single-handedly solve the entire problem. There are attributes that determine identity with a high degree of certainty, but they are frequently not present, e.g. an e-mail address appears only in 10% of articles.  

In author disambiguation all typical object deduplication obstacles arise, beginning from many standards for writing a name (“J.Smith”, “John Smith”, “J.Smith Jr.”, “Smith, J.”), misspellings (“J.Smiht”, “J.Smith”), an attribute value change over time (“Eleonore Smith”, “Eleonore Smith-Black”), diacritic handling (“José Gonçalves”, “Jose Goncalves”, “Jos? Gon?alves”), transliterations (e.g. translating “Angela Johnson” to Japanese equivalent “アンジェラ・ジョンソン” and back to English result is “anjira jyonson”) and extraction artifacts (“Smith Machine”). As Torvik and Smallheiser investigated, about 1.3% authors whose e-mail addresses match have different surnames, most likely due to inconsistencies enumerated here.

The other attributes also need further consideration, e.g. some errors may occur in an e-mail address, like “jsmith@institution.org” or “jsmith@institution.org”, hence requiring a rectification step.

Many researchers explored specifically the subject of the author disambiguation. Han et al. [12] compared Naive Bayes and Support Vector Machines (SVM) classifiers for this task, whereas in [13], they examined efficiency of k-Way Spectral Clustering. Concurrently, Dai and Storkey [8] applied hierarchical Dirichlet process and nonparametric latent Dirichlet allocation models, whereas Levin and Heuser [17] included in their solution enhancements derived from the genetic programming.

Typically, authors tried to conduct pairwise comparisons on a set of records with the same value of a major feature (e.g. surname) to determine whether two candidate author items are the same. In contrast to performing analysis in respect of all given features, Qian et al. [30] proposed to perform initial clustering with a limited number of features to obtain High Precision Clusters in the first step and then merge clusters into High Recall Clusters in the second step. They also proposed to introduce a human judgement clustering in the final step. When utilizing user feedback, it is crucial to distinguish between experts and