Disambiguating Author Names Using Automatic Relevance Feedback

In-Su Kang\textsuperscript{1,*} and Seung-Hoon Na\textsuperscript{2}

\textsuperscript{1} Computer Science and Engineering, Kyungsung University, Pusan 608-736, South Korea
\textsuperscript{2} Department of Computer Science, National University of Singapore
dbaisk@ks.ac.kr, nash@comp.nus.edu.sg

Abstract. Author disambiguation is to associate author-name instances with author identifiers corresponding to real-world authors. Bibliographic records, which normally contain author-name occurrences to be disambiguated, are not sufficient to provide author-resolving features. This study proposed a feedback-based method for author disambiguation to tackle the insufficiency of author representation. Our method is focused on enhancing author's topical representation by adding topically-related terms obtained from feedback documents. Experiments showed the positive effect of the method on author resolution by improving the performance.

Keywords: Author Disambiguation, Automatic Relevance Feedback, Web Evidence.

1 Introduction

Author disambiguation is to map author-name occurrences into real-world individuals. Its difficulty comes from many-to-many relationships between names and persons. It can improve the quality of academic search systems by enabling author-oriented grouping of bibliographic records rather than name-oriented.

In general, a bibliographic record in which author names to be disambiguated appear consists of co-author name(s), a paper title, a publication title, a publication year, etc. However, such information is not enough to provide person-identifying features to author disambiguation systems. Researchers have thus attempted to supplement this with information from either full-text archives or personal publication-list pages on the Web. These external sources however are not always or mostly available.

This study tries to attack the above problem by expanding the basic bibliographic record with its topically-related articles obtained from a large collection of bibliographic records for which there exist many freely-accessible archives or services such as DBLP and PubMed. For this, we apply the relevance feedback technique well-developed in information retrieval (IR) communities. Experiments show that our method effectively improves the performance of author disambiguation.

* Corresponding author.
by 34.6%. From the viewpoint that author-identifying features can be classified into biographic ones (e.g., coworkers, emails and affiliations) and topical ones (e.g., academic areas and research topics), we handle the latter.

The remainder of this article is organized as follows. Section 2 reviews the related works. Section 3 describes a feedback-based method of enriching author’s topical features. Section 4 reports the experimental results, and Section 5 provides the conclusion.

2 Related Work

This section reviews the earlier uses of influential author-identifying features to enhance author representation. For an extensive survey of author disambiguation, readers can refer to [1].

Several researchers have reported on the successful application of PPLW (personal publication-list pages on the Web) features to author resolution [2-4]. PPLW indicate the web page that contains a list of bibliographic records authored by a particular researcher. Instead of directly identifying and gathering PPLW pages from the Web, previous works have attempted to indirectly find the evidence from the general search engine that the same web page includes two bibliographic records of which author names are to be disambiguated.

The full-text of an article is another rich source for improving author representation, which holds biographic features such as coauthors, email addresses and affiliations, as well as topical features such as keywords, abstracts, references, etc. Song[5] exploited the first page of the article document to train Latent Dirichlet Allocation model and observed the best disambiguation performance 93.6% for English names. Kang[6] improved the performance of Korean author disambiguation up to 91.47% in F1 by employing a combination of co-author names and emails extracted from the article full-text.

Unfortunately, both PPLW and the full-text of articles have some limitations on their accessibility and availability. In other words, author disambiguation systems cannot guarantee that all or most authors of articles would maintain their publication pages on the Web. Besides, all or more full-text archives are not freely accessible. Unlike such hardly-obtainable resources, the collections of bibliographic records that are exploited by this study are relatively easily reachable.

3 Feedback-Based Enhancement of Author Representation

Author disambiguation is normally applied to a collection of same-name author occurrences. Its general procedure is as follows. First, every name instance is converted into its author-instance representation. Next, for each pair of author-instance representations, its similarity is computed. As a result, a graph is obtained where a node indicates a name instance, and the link between two nodes corresponds to the similarity between two author-instance representations. Then, a clustering technique is applied to the graph to produce a set of author clusters of name instances.