Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 83 (2016) 488 – 495

Procedia
Computer Science

The 7th International Conference on Ambient Systems, Networks and Technologies
(ANT 2016)

Name Disambiguation Method based on Multi-step Clustering

S. Gu[a], X. Xu[a, b, *] , J. Zhu[a], L. Ji[a]

[a] *College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210023, China*
[b] *College of Computer Yancheng Big Data Research Institute, Nanjing University of Posts and Telecommunications, Yancheng 224005, China*

Abstract

Author name disambiguation is a very important and complex research topic. During the retrieval and research of literatures, the quality of the investigation results has been reduced because of the high probability of different authors sharing the same name, which lengthens the whole cycle of the scientific research. Therefore, it is necessary to find a reasonable and efficient method to distinguish the different authors who share the same name. In this paper, an author name disambiguation method based on multi-step clustering (NDMC) is proposed to disambiguate author names. First, the framework combines the brief and clear characteristics of literature system information with the comparison of co-authors' similarity to realize the initial clustering. Then, author's information is extracted from the Baidu Encyclopedia, and the semantic similarity of subordinate units is compared, as the basis of identity discrimination in the second step clustering. Finally, after extraction of two step clustering paper keywords in each class cluster, combined into corpus collection, through the characteristics of the semantic comparison, cancellation of indeterminacy results further adjustment, so as to complete the multi-step clustering. We extract literature information from the China National Knowledge Infrastructure (CNKI) to implement experiments. The experimental results show that the hybrid disambiguation framework is feasible and efficient.

Keywords: Name disambiguation; Feature extraction; semantic recognition; hierarchical clustering

1. Introduction

Authors identification online needs to be addressed. DBLP [1] (Digital Bibliography & Library Project) is first appeared in the author integrated system as the core of literature, which includes almost all computers in the major international journals published in English literature, and the meeting every quarter for a data update, the academic

* Corresponding author. Tel. : +8613813885172; fax: +862585866433.
E-mail address: xuxl@njupt.edu.cn

doi:10.1016/j.procs.2016.04.237

literature database can be a very good Computer technology, by the user to retrieve the author's name, you can find all the documents to the name of the author's record, but did not do the nuptial disambiguation.

C-DBLP [2] is developed by imitating DBLP, with the author as the core of literature integration system, and based on the co-author relationship characteristics. It has the name disambiguation function with high accuracy. However, its recall rate is relatively low.

In this paper, we focus on the Chinese literature system of the name disambiguation problem. When we retrieve objects to the author, retrieves a lot of the authors and the paper with the same information, if the name is more common, and the same information would be redundant. Often, large literature database will provide advanced retrieval function, provides the function of the restrained, but usually only from two direction constraints of units or journal, its result is just the constraints under the condition of information, can't really do the name disambiguation. Based on this, we combine the characteristics of the information system, the paper brief refining, to provide the user name repetition experts under the true character of comprehensive paper information as the goal, to automatic disambiguation methods in the literature to provide technical support to make better use in the database.

The main contributions of this work include:

- We combine the Baidu Encyclopedia classification information and the co-author information as the foundation, and then combine keyword corpus collection to obtain the higher disambiguation accuracy and recall rate.
- We utilize the retrieve full rate (RFR) according to the actual users to the author for the demand of the retrieval objects, and design the algorithm with people-oriented thoughts.
- We obtain identity recognition unit classification threshold through the experiments and improve the accuracy and recall rate of the algorithm.

The rest of the paper is organized as follows: in Section 2, a model of author and its features are presented. Section 3 presents the results and experiments and performance analysis. Section 4 concludes this paper.

2 Model of NDMC

First of all, according to the cooperator information of the information of the paper, we get different clustering by conducting the first clustering. These clusters are based on the condensed level of clustering idea, and they will only gather more and more instead of being parted again. After the first clustering, there are still many complete information remaining apart. Then, we base on the item information in the Baidu Encyclopedia and recognize multiple identity information under the name. Through the experiment, we will choose the value of the threshold to distinguish the size of clusters in order to choose clusters that are used as molding snow balls to take over other combinations of clusters. Last, basing on the first two steps of clustering, we extract keywords in the paper from all information-focused clusters, and combine them as feature corpus to compare with the rest clusters that relatively include fewer chapters in the similarity of features. So we can finish the final clustering. Detailed procedures are following:

2. 1 Similarity computation