

# Two Supervised Learning Approaches for Name Disambiguation in Author Citations

Hui Han  
Department of Computer  
Science and Engineering  
The Pennsylvania State  
University  
University Park, PA, 16802  
hhan@cse.psu.edu

Lee Giles  
School of Information  
Sciences and Technology  
The Pennsylvania State  
University  
University Park, PA, 16802  
giles@ist.psu.edu

Hongyuan Zha  
Department of Computer  
Science and Engineering  
The Pennsylvania State  
University  
University Park, PA, 16802  
zha@cse.psu.edu

Cheng Li  
Department of Biostatistics  
Harvard School of Public  
Health  
Boston, MA, 02115  
cli@hsph.harvard.edu

Kostas Tsioutsoulis  
NEC Laboratories America,  
Inc.  
4 Independence Way,  
Princeton, NJ 08540  
kt@nec-labs.com

## ABSTRACT

Due to name abbreviations, identical names, name misspellings, and pseudonyms in publications or bibliographies (citations), an author may have multiple names and multiple authors may share the same name. Such name ambiguity affects the performance of document retrieval, web search, database integration, and may cause improper attribution to authors. This paper investigates two supervised learning approaches to disambiguate authors in the citations<sup>1</sup>. One approach uses the naive Bayes probability model, a generative model; the other uses Support Vector Machines(SVMs) [39] and the vector space representation of citations, a discriminative model. Both approaches utilize three types of citation attributes: co-author names, the title of the paper, and the title of the journal or proceeding. We illustrate these two approaches on two types of data, one collected from the web, mainly publication lists from homepages, the other collected from the DBLP citation databases.

## Categories and Subject Descriptors

H.3.3 [Information Systems]: Information Search and Retrieval

## General Terms

Algorithms

<sup>1</sup>“Citations” refer to an author’s publication list in the citation format.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL’04, June 7–11, 2004, Tucson, Arizona, USA.

Copyright 2004 ACM 1-58113-832-6/04/0006 ...\$5.00.

## Keywords

Naive Bayes, Name Disambiguation, Support Vector Machine

## 1. INTRODUCTION

Due to name variation, identical names, name misspellings, and pseudonyms, we observe two types of name ambiguities in research papers or bibliographies (citations). The first type is that an author has multiple name labels. For example, the author “David S. Johnson” may appear in multiple publications under different name abbreviations such as “David Johnson”, “D. Johnson”, or “D. S. Johnson”, or a misspelled name such as “Davav Johnson”. The second type is that multiple authors may share the same name label. For example, “D. Johnson” may refer to “David B. Johnson” from Rice University, “David S. Johnson” from AT&T research lab, or “David E. Johnson” from Utah University (assuming the authors still have these affiliations).

Name ambiguity can affect the quality of scientific data gathering, can decrease the performance of information retrieval and web search, and can cause the incorrect identification of and credit attribution to authors. For example, identical names cause the ambiguity of the “author page” in the web DBLP (Digital Bibliography & Library Project)<sup>2</sup>. The author page of “Yu Chen” in the DBLP contains citations from three different people with the same name: Yu Chen from University of California, Los Angeles; Yu Chen from Microsoft Beijing; Yu Chen as the senior professor from Renmin University of China. Such name ambiguity causes the incorrect identification of authors. For example, the author page of “Jia Li” in the DBLP refers to the “Jia Li” from the Department of Statistics at the Pennsylvania State University. However, the “Home Page” link in her author page directs to the professor with the identical name in the Department of Mathematical Sciences at the University of Alabama in Huntsville. We observe from CiteSeer [18] the incorrect attribution to the authors due to similar ambiguity. “D. Johnson” is the most cited author in Computer Science accord-

<sup>2</sup><http://WWW.Informatik.Uni-Trier.DE/~ley/db/index.html>

















