赛道 1 论文的冷启动消歧

班级：计算机技术(专硕)

姓名：刘波

学号：201934721

摘要

本文提出了一种基于规则匹配和 DBSCAN 相结合的论文作者名消歧方法；首先对数据进行预处理，例如大小写、符号等问题；使用指定的规则对数据中相同作者的论文进行消歧；基于论文的属性信息（如姓名、机构、摘要等）提取特征，然后选取合适的聚类算法进行聚类。实验效果表明，DBSCAN 方法较适合于论文作者名的消歧任务。然后将两种方法得到的结果结合起来，提高作者名消歧的准确性。

关键字

作者名消歧，文本特征提取，聚类

1 简介

1.1 背景介绍

在许多应用中，同名消歧 (Name Disambiguation - aiming at disambiguating WhoIsWho) 一直被视为一个具有挑战性的问题，如科学文献管理、人物搜索、社交网络分析等，同时，随着科学文献的大量增长，使得该问题的解决变得愈加困难与紧迫。尽管同名消歧已经在学术界和工业界被大量研究，但由于数据的杂乱以及同名情景十分复杂，导致该问题仍未能很好解决。

1.2 问题描述

收录各种论文的线上学术搜索系统(例 Google Scholar, Dblp 和 AMiner 等)已经成为目前全球学术界重要且最受欢迎的学术交流以及论文搜索平台。然而由于论文分配算法的局限性，现有的学术系统内部存在着大量的论文分配错误；此外，每天都会有大量新论文进入系统。故如何准确快速的将论文分配到系统中已有作者档案以及维护作者档案的一致性，是现有的线上学术系统亟待解决的难题。

由于学术系统内部的数据十分巨大（AMiner 大约有 130，000，000 作者档案，以及超过 200，000，000 篇论文），导致作者同名情景十分复杂，要快速且准确的解决同名消歧问题还是有很大的障碍[1]。

竞赛希望提出一种解决问题的模型，可以根据论文的详细信息以及作者与论文之间的联系，去区分属于不同作者的同名论文，获得良好的论文消歧结果。而良好的消歧结果是确保学术系统中，专家知识搜索有效性、数字图书馆的高质量内容管理以及个性化学术服务的重要前提，也可影响到其他相关领域。

1.3 任务描述

给定一堆拥有同名作者的论文，要求返回一组论文聚类，使得一个聚类内部的论文都是一个人的，不同聚类间的论文不属于一个人。最终目的是识别出哪些同名作者的论文属于同一个人。

2 相关工作

相比于传统的人名消歧，论文作者名消歧有其特殊性。一方面数据库中的作者识别系统尚未完全开发；另一方面，论文信息一般包括作者、标题、摘要、关键字和出版物名称等内容，所包含的信息量较为有限。论文作者名的消歧也具有一些挑战。一方面，论文作者名在同一机构中可能会存在重名现象或者同音现象；另一方面，作者在论文中的署名可能存在多种形式或者简写。

本文首先利用人工构建的匹配规则对论文作者名进行匹配，根据匹配得到的作者分类提取出部分集合通过 DBSCAN 进行聚类提高准确率。

3 基于规则匹配和 DBSCAN 相结合的消歧方法

3.1 人名数据处理

如相关工作所述，论文作者名存在混淆的原因一方面是作者存在重名现象，另一方面是在论文中署名中的不同形式。对于论文中署名的不同形式可能采用了不同的姓名顺序和大小写规则，导致一个作者的中文名可能会对应多种形式的英文名字，再加上多音字的现象，会出现大量作者名混淆的情况。

针对这一现象，本文对所有的英文名制定了一个转换规则，即基于转换规则将所有的英文名统一，并且去除大小写和特殊符号（如分号）。

表 1 英文名转换规则

输入英文名	输出英文名
LI Youji	li youji
Tao Zhang	zhang tao
KONG QING-XIN	kong qingxin
Sun Xin	sun xin

3.2 论文信息预处理

数据集中的信息有题目、摘要、关键字等，如表 2 所示。

表 2 论文数据格式

域	类型	含义	举例
id	string	论文 id	53e9ab9eb7602d970354a97e
Title	string	题目	Data mining: concepts and techniques
name	string	作者姓名	Jiawei Han
org	string	作者单位	Northwest Normal University(Northwest Normal University),Lanzhou,China
venue	string	会议/期刊	Inteligencia Artificial, Revista Iberoamericana Inteligencia Artificia
year	int	发表年份	2000
keywords	list of strings	关键字	["data mining", "structured data","relational data"]
abstract	string	摘要	Our ability to generate...

本文使用到的论文信息有作者姓名、作者单位、关键字、摘要和标题。由于文本信息中存在很多噪音数据，而且没有进行分词，所有首先需要进行预处理，预处理过程依次对文论信息进行去噪处理，包括：去掉特殊字符和标点符号、去掉多余空格和换行符，去掉停顿词、字符小写话等。

3.3 基于规则匹配

在 3.1 和 3.2 中已经对人名和论文信息进行了处理。然后将处理过的信息通过本文制定的规则进行分类，具体规则如下：

1）如果两篇文章中的作者相同个数大于等于 3，认为这两篇文章属于同一个作者。

2）如果两边文章中的作者相同个数小于 3，基于下面的规则再对作者名进行分类。

①：设置权值 score=0。

②：如果两篇文章中作者姓名相同权值 score+=10。

③：如果两篇文章中关键字相同权值 score+=5。

④：对于摘要，首先用 set 集合保存，然后求两个摘要的交集，最后权值 score+=len(交集)。

⑤：如果最终的权值 score 大于 30，则认为这两篇文论属于同一作者，如果最终的权值 score 小于 30，则认为两篇文论不属于同一作者。

基于规则匹配的消歧在测试集准确度：30.0%。

3.4 DBSCAN 聚类

DBSCAN 算法步骤大致描述如下：

对于给定的邻域距离 e 和邻域最小样本个数 MinPts：

1）遍历所有样本，找出所有满足邻域距离 e 的核心对象的集合。

2）任意选择一个核心对象，找出其所有密度可达的样本并生成聚类簇。

3）从剩余的核心对象中移出 2 中找到的密度可达的样本。

4）从更新后的核心对象集合重复执行 2-3 步直到核心对象都被遍历或者移出。

经过规则匹配后，已经有多个作者聚类在一起，但有的聚类长度因为规则匹配中的不完善导致聚类长度很短(小于 10)，对于这种数据，本文将聚类长度小于 10 的数据单独提取出来，通过 DBSCAN 进行聚类。本文使用 name、org 和 abstract 作为特征，对数据进行聚类。

将 DBSCAN 得到的结果和 3.3 基于规则匹配得到的结果相结合得到最终结果。

相结合后在测试集的结果：36%。

4 结果和不足

本文设计了一种基于规则匹配和 DBSCAN 聚类相结合的作者名消歧方法。实现了英文作者名的统一化处理，并设计了合适的规则对作者名进行消歧。最终准确率 36%，排名 52。

通过这个比赛，学会了对数据进行分析处理。也有一部分思想没有实现，最初的思想是在基于规则匹配中对每一个分好的组中求加权值，判断哪个分组中加权值最大，然后再将改作者 id 加入到分组中。但可能限于提交次数，没有找到更好的参数去提高准确率。另外的不足是太依赖于调参，而不是真正的分析问题。

5 参考文献

[1] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998.

[2] Jie Tang, A.C.M. Fong, Bo Wang, and Jing Zhang. A Unified Probabilistic Framework for Name Disambiguation in Digital Library. IEEE Transaction on Knowledge and Data Engineering (TKDE), 2012, Volume 24, Issue 6, Pages 975-987.

[3] Song Y., Huang J., Councill I., Li J., & Giles C. (2007). Efficient topic-based unsupervised name disambiguation. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL '07). ACM, New York, NY, USA, 342-351.

[4] Smalheiser N R, Torvik V I, Author name disambiguation[J], Annu Rev Inf Sci Tec, 2009, 43:1.