密级:（涉密论文填写密级，公开论文不填写）

中国科学院研究生院

硕士学位论文

面向自动化学科中文期刊论文的文本挖掘系统

作者姓名：刘禹

指导教师：杨一平　研究员

中国科学院自动化研究所

学位类别：工程硕士

学科专业：计算机技术

培养单位：中国科学院自动化研究所

2012 年　05 月

The Text Mining System Subjected to Automatic Journal Articles in Chinese

By

LiuYu

A Dissertation Submitted to
Graduate University of Chinese Academy of Sciences
In partial fulfillment of the requirement
For the degree of
Master of Engineering

Insititute of Automation, Chinese Academy of Sciences

05, 2012

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确地说明并表示了谢意。

签名：刘禹 日期：2012.5.31

关于论文使用授权的说明

本人完全了解中国科学院自动化研究所有关保留、使用学位论文的规定，即：中国科学院自动化研究所有权保留送交论文的复印件，允许论文被查阅和借阅；可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签名：刘禹 导师签名：[illegible] 日期：2012.5.31

摘 要

“自动化学科创新思想与方法研究”课题对影响国内自动化学科发展的因素进行系统分析，并利用各因素之间的相互联系构建自动化学科知识体系，通过对已有思想与方法的形成和发展规律进行总结，对学科发展方向进行前瞻性预测。该课题的最终目标是在学科知识体系的基础上开发学科知识服务网络平台，为相关领域的研究人员和技术人员提供知识服务，进而推动知识创新。

知识要素（包括研究对象、研究方法、研究工具、研究人员、研究机构等）是建设学科知识体系的基本要素，因此知识要素获取是该课题的首要环节。本文以课题中的知识要素获取需求为研究课题，在大量文献调研和实验的基础上设计和实现了用于知识要素抽取的文本挖掘系统，并在项目中得到很好的应用。论文的主要工作和贡献如下：

①文本分类和特征词选择技术在数据清洗中的应用。本文实现了文本分类的文档向量模型(VSM)，将其用于区分自动化学科和非自动化学科的文献；提出了基于卡方拟合优度的特征词选择方法(chifit)，该方法能够使用较低的特征维度达到较好的分类效果。

②提出了基于编辑距离二次计算的关键词语义聚类算法。项目数据中有大量文献关键词在形态上相似且语义上相同，该算法充分利用这一特性将语义聚类问题转换成形态聚类问题。

③提出并实现了知识族谱构建方案。该方案把与被查询知识点在时间上可能存在继承、发展、演变关系的知识点以亲疏程度和时间切片为依据展现出来，用以辅助用户进行文献检索和知识理解。

④提出了基于距离属性的二叉分裂算法。该算法属于分裂式层次聚类算法，算法的执行过程即是层序建立二叉树的过程，叶子结点就是最终的聚类。该算法有效解决了人物名称与机构名称对齐问题。

⑤提出了基于图聚类的人名消歧算法。汉语中存在大量人名重复现象，给准确统计学者的学术成果带来困难。该算法将名字视为图上的结点，根据两个结点之间的属性相似情况，决定是否加边，最后根据图的连通特性，将每一个连通分量视为指向同一人物实体的聚类。

⑥提出了一种无监督的机构名称归一化算法，该算法充分利用同一个人物实体所涉及的机构名称之间的关系，提取一级机构名称，不需要事先准备规范化的机构名称列表，也不需要定义复杂的机构名称结构规则。

关键词：文本挖掘，文本分类，文本聚类，命名实体识别与消歧，知识服务

Abstract

The research of innovative ideology and methodology in automation discipline aims to give a systematic analysis of the factors which play important roles in the development of domestic automation discipline. It also aims to explore the relationship among those factors to build a knowledge system, whose ultimate goal is to develop a network platform offering knowledge services to potential users.

Factors that include research objects, researchers, institutions, methods, theories, tools, etc. are so vital to the knowledge system that it is of great prominence to retrieve them precisely. This paper designs and implements a text mining system focusing on information extraction. The main contributions are summarized as follows:

① The application of text categorization and feature word selection technique in data cleaning. The vector space model approach is implemented to predict articles' categories. A feature selection method named chifit is proposed, which can achieve higher precision with lower feature dimension.

② A method that reduces the problem of semantic clustering to morphological similarity computation is proposed to resolve keywords clustering.

③ A novel scheme "knowledge pedigree" is proposed and implemented to facilitate users in literature research and knowledge understanding.

④ A divisive clustering approach is used for person-institution alignment. This method is very similar to constructing a binary tree in a level-order traverse.

⑤ In order to evaluate the scholars' academic influence precisely, a clustering approach based on graph is presented for person name disambiguation.

⑥ An unsupervised institution name normalization method is proposed, fully exploring the institution data within each person entity.

Key Words: text mining, text classification, text clustering, name entity recognition and disambiguation, knowledge service

目 录

插图目录

表格目录

第一章 绪论

1.1 研究背景和意义

2007 年，王大珩、刘东生、叶笃正三位老科学家向温家宝总理提交了《关于加强创新方法工作的建议》，提出了“自主创新，方法先行”的观点。温总理对此高度重视并做出重要批示[2]。作为国家科技部创新方法工作的一部分，中国科学院自动化研究所于 2009 年 10 月至 2011 年 10 月，承担了编号为 2009IM020300 的课题“自动化学科创新思想与科学方法研究”。该课题拟对影响国内自动化学科发展的因素进行系统调研，并利用各因素之间的相互联系构建自动化学科的知识体系，通过对已有思想与方法形成和发展的规律进行总结，对学科发展方向进行前瞻性预测。简而言之，该课题的两个基本目标是：首先，建设包含人物、机构、学术研究方向等因素在内的学科知识体系；其次，在学科知识体系的基础上开发自动化学科知识服务网络平台。课题的最终目标是知识服务平台不仅可以满足潜在使用者（相关学科的科研人员和技术人员）的多层次和个性化的知识需求，而且可以在推动使用者进行知识发现和知识创新方面有所贡献。

1.2 相关工作

1.2.1 知识、知识管理与知识服务

何谓知识？自古希腊时代哲学家们就开始试图给知识下定义。关于如何定义知识，认识论上有很多分歧。文献[147]中对知识的几种定义，及其对应的知识系统的涵义进行了综述。本文仅采用通过数据、信息、知识之间的区别和转换关系进行知识定义的观点[149][150]。这种观点认为数据、信息、知识之间存在转换关系，即：数据是信息和知识的物理载体和来源；信息是有意义的数据；知识是准确、有效，且被人吸收了的信息如图 1-1 所示。

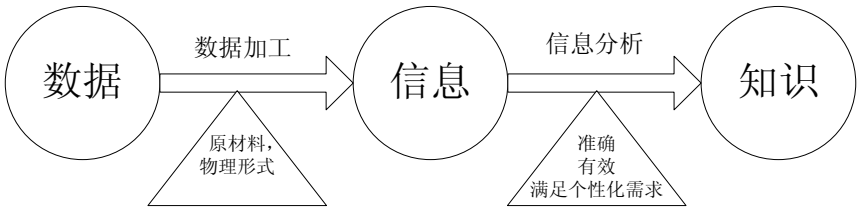


图 1-1 数据、信息、知识之间的演变关系

本文以图 1-2 为例，进一步阐明数据、信息、知识之间的关系。图 1-2 中的四个数据源中的文本都是数据；但是数据源 2 中的数据由于不具有可读性，所以不是信息，其余的文本都是信息；数据源 4 中阐述清雍正帝是清康熙帝的父亲，与历史事实不符，因此不是知识。知识是经过不同个体进行倾向性选择的正确的信息[147]，比如程序员通过阅读上述文本中的信息，能够获得“c、java、python 都是编程语言”的知识，一位刚刚开始从事自然语言处理相关研究的青年学生获得是“python 在自然语言处理领域内广泛应用”的知识。图 1-3 从个体角度阐释数据、信息、知识的转换过程。

Data source1:

C : the third letter of the English alphabet;

java: a kind of coffee

python: a large tropical snake that kills animals for food.

Data source2:

Dldlg=age\khg'ag\skhg\heh\ajh'rkh'klh\whk

Data source 3:

C: a general-purpose computer programming language developed between 1969 and 1973 by Dennis Ritchie at the Bell Telephone Laboratories for use with the Unix operating system.

java: Java is a programming language originally developed by James Gosling at Sun Microsystems (which has since merged into Oracle Corporation) and released in 1995 as a core component of Sun Microsystems' Java platform.

python: python is a general-purpose, high-level programming language[6] whose design philosophy emphasizes code readability. Nowadays, python is widely used in Natural Language Processing.

Data source 4:

清雍正帝是清康熙帝的父亲。

图 1-2 数据、信息、知识之间关系的举例阐释

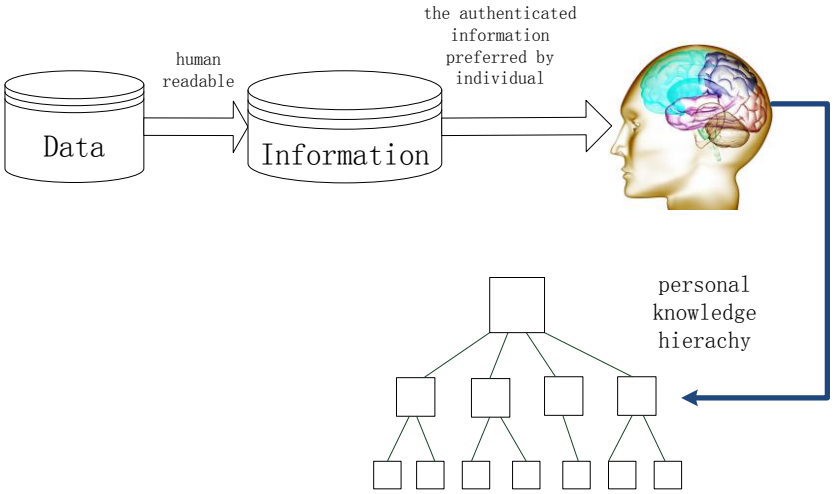


图 1-3 从个体角度理解数据、信息、知识的演变过程

从数据、信息、知识之间的演化关系定义知识的观点常被 IT 领域的学者和工程师所接受。他们认为知识系统是信息系统的高级阶段。信息系统的目的是提供信息，并且让用户能够检索到信息，但并不保证用户检索到的信息是否对用户有用；知识系统在信息系统的基础上进行功能扩展，其目的在于：通过合理地组织、管理和展现知识，一方面帮助用户吸收新知识，一方面帮助用户找到对其存在潜在价值的信息[147]。

学术情报研究（academic informatics study）通过收集、整理、加工和分析学术信息，对已有学术信息梳理出其传承演变，对未来的学术趋势进行预测推断。从根本上来讲，学术情报研究旨在推进知识的传承与创新，所以受到了国内外各大研究机构的广泛重视。数字图书馆技术（Digital Bibliography & Library Project，DBLP）的发展使学术信息不再被经院学府束之高阁，从源头上得到解放。这给学术情报研究带来了极大的方便。随着我国知识工程及数字图书馆建设的展开，知识服务已逐渐成为国内图书情报学的重要研究领域[152]。关于知识服务概念内涵，国内图书情报界存在一定争论。如文献[153]认为知识服务是立足于知识层次上的服务。本文关注知识服务系统的实践层面，不对理论纠纷做过多探讨，采纳张晓林等人的观点，认为知识服务是信息服务的高级阶段，是帮助用户获取知识和解决方案的服务。在知识服务网络平台设计上，秉承张晓林老师有关知识服务的相关论点，即认为：知识服务是用户目标驱动的服务，是面向知识内容的服务，是面向解决方案的服务，是面向增值服务的服务[151]。

1.2.2 相关学术知识服务网络平台

学术情报研究和学术知识服务受到了国内外研究者的广泛重视。于 1986 年成立的日本学术情报中心是全日本学术信息收集、加工、提供以及学术信息系统的研究与开发中枢[3]。中国科学院国家文献情报中心至今已经成立 50 余年，旨在进行学术情报方面的专门研究。

学术情报研究与学术知识服务也受到了各大互联网公司的青睐。谷歌学术搜索（Google Scholar）[1]是检索学术资源的免费搜索工具，专门用于帮助用户查找期刊论文、学位论文、专业图书、预印本、文摘和技术报告等学术文献。它的搜索范围涉及诸多学科领域，覆盖面广、权威性强。目前已成为科技人员和教师、学生查找专业文献资料的首选工具[4]。微软学术搜索（Microsoft Academic Search）[2]不仅提供论文检索服务，而且提供学者相关的检索服务。

知网[3]、万方[4]、维普[5]等以中文期刊论文检索为主要服务的网络平台也开始利用自身数据优势，提供其他类型的知识服务。如作者科研关系检索[6][7]等。清华大学计算机系软件研究所知识工程研究室的唐杰老师等人研制的 ArnetMiner 系统[8]旨在抽取和挖掘学术社交网络，其数据来源是计算机学科的英文论文文献[5]。人民大学网络与移动数据管理实验室的孟晓峰老师等人研制的以学者为中心的 C-DBLP 系统[9]旨在从中文计算机领域论文中抽取和挖掘学术社交网络。

图 1-4 至图 1-6 分别为知网、万方、C-DBLP 提供的学者检索服务。从图中可以发现，知网的学者检索服务，没有进行同名消歧，返回的检索结果精确度不高；万方的学者检索服务，返回结果未进行机构名称消歧，错把“北京万方数据股份有限公司”与“万方数据股份有限公司”当成不同的机构；C-DBLP 的

[1]http://scholar.google.com.hk/

[2]http://academic.research.microsoft.com/

[3]http://www.cnki.net/

[4]http://www.wanfangdata.com.cn/

[5]http://www.cqvip.com/

[6]http://stads.wanfangdata.com.cn/zz/

[7]http://scholar.cnki.net/beta/Result.aspx

[8]http://arnetminer.org/

[9]http://www.cdblp.cn/

学者检索服务，较之前两者有很大改善，这些说明提高学术知识检索结果的精度，需要在命名实体识别与消歧等技术上有所突破。

图 1-4 CNKI 学者检索服务

图 1-5 万方学者检索服务

图 1-6 C-DBLP 学者检索服务

1.3 自动化学科知识服务网络平台

从满足用户多层次、个性化学术知识需求的角度出发，自动化学科知识服务网络平台设计了如图 1-7 所示的各个功能模块。各个检索模块之间具有深度的耦合关联关系，即从任何一个检索界面，都可以通过超链接的方式进入其他检索页面，进而极大方便用户使用，使用户通过一次检索即能获得相关学术活动的立体全景。图 1-8 到 1-12 给出了自动化学科知识服务网络平台的各检索页面视图。检索页面之间通过超链接进行关联，实际上是不同实体之间通过实体属性建立关联关系。本平台共涉及文章、学者、机构、知识点、期刊五种实体，五种实体的各种属性，以及关联关系见附录 5。

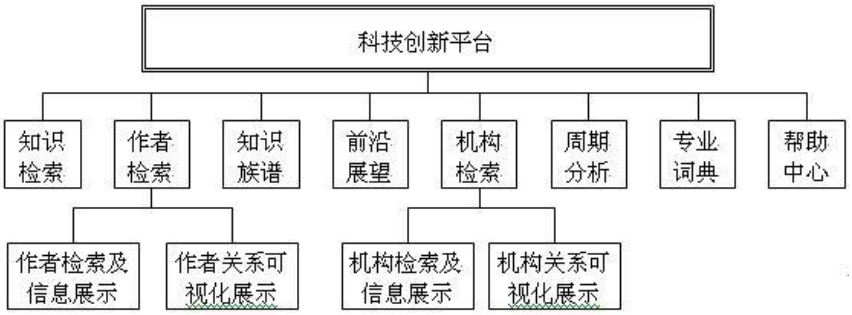


图 1-7 自动化学科知识服务网络平台框架

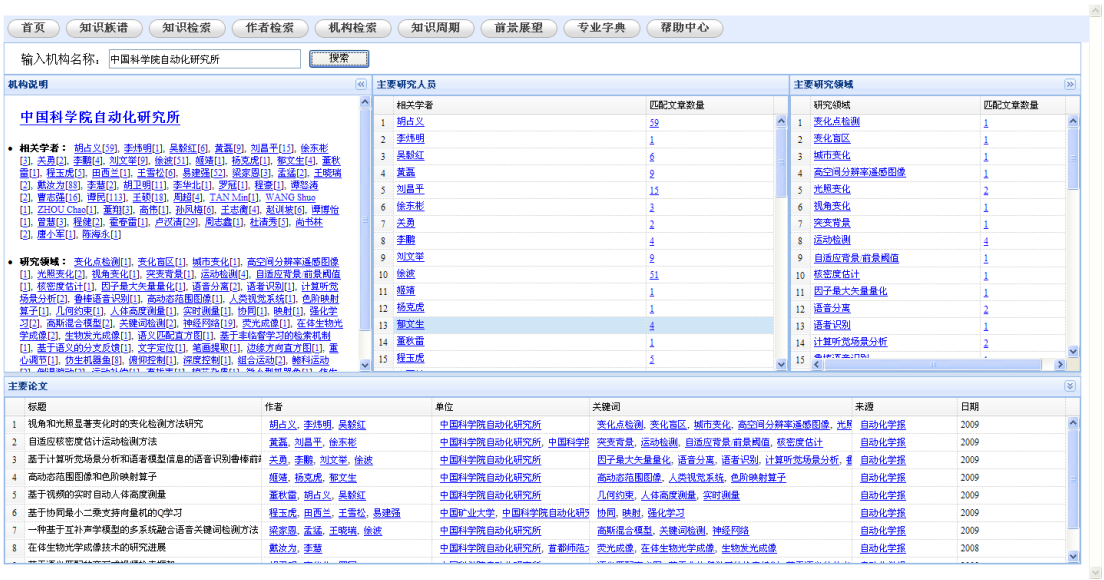


图 1-8 机构检索页面

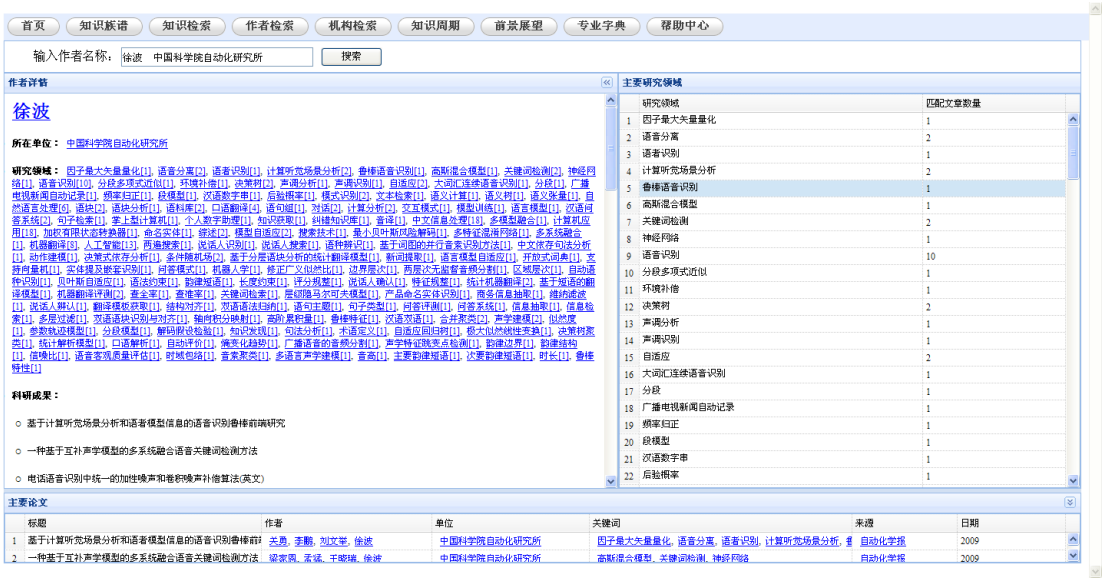


图 1-9 学者检索页面

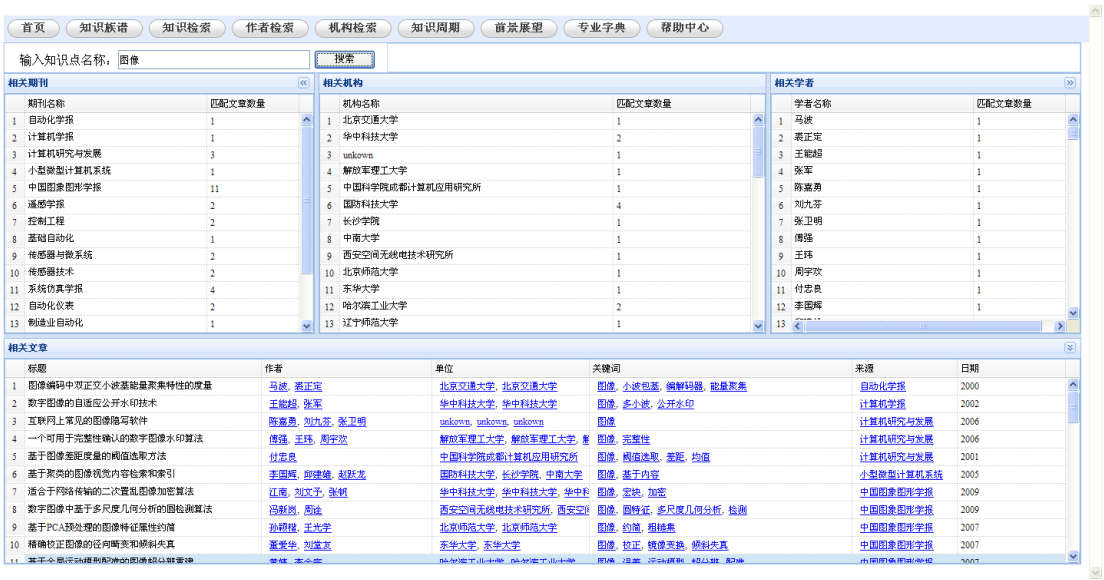


图 1-10 知识检索页面

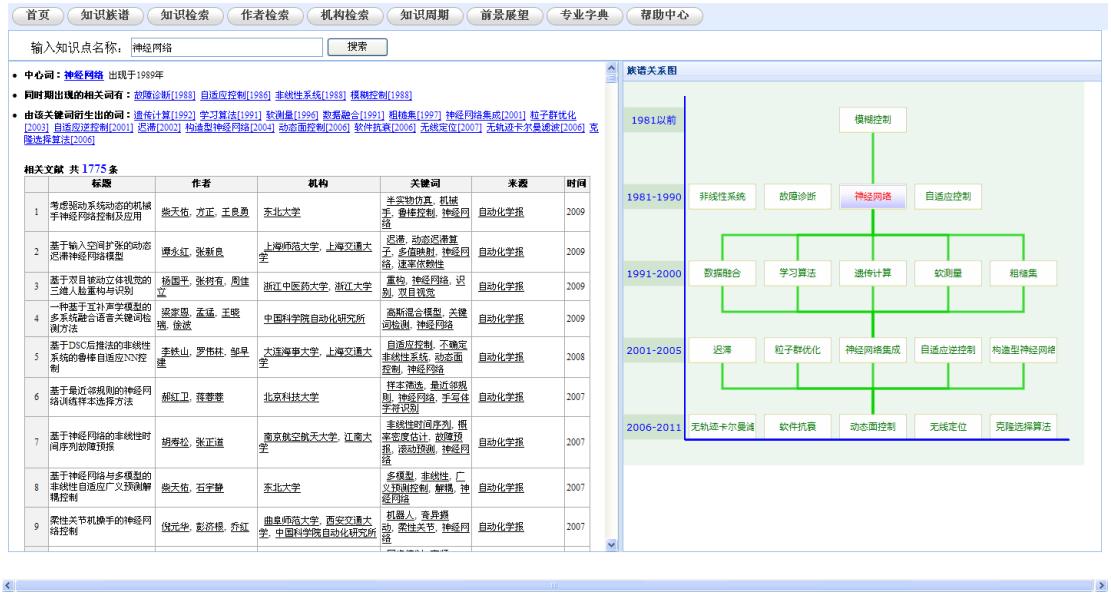


图 1-11“知识族谱”检索页面


首页 知识族谱 知识检索 作者检索 机构检索 知识周期 前景展望 专业字典 帮助中心

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

术语列表

抽象	Abstract	交-交变频	AC-AC frequency converter	自主规划	autonomous planning
抽象代数	abstract algebra	交-交变频器	AC-AC frequency converter	自主计算	AOC
抽象体系	abstract framework	交-直-交变频	AC-DC-AC frequency conversion	自主调度执行	autonomous scheduling execution
抽象分析	abstract analysis	交-直-交型变频器	AC-DC-AC converter	自主越障	autonomous negotiation
抽象化	abstraction	交-直流变换	AC-DC Conversion	自主车辆	autonomous vehicle
抽象图归约机	abstract reduction machine	交交变频器	A.c.—a.c.cycloconverter	自主轮式机器人	autonomous wheeled robot
抽象实现结构图	Abstract implement struture diagrams	交叉耦合系数	across couple coefficient	自主运算	automatic computing
抽象层APII	abstract hierarchy APII	交变控制	alternate control	自主运行与管理	autonomous running and governing
抽象工作流	Abstract workflow	交变磁场	alternating magnetic field	自主运行分布式卫星	autonomous distributed satellites
抽象工厂模式	abstract factory pattern	交变磁场测量	Alternating Current Field Measurement	自主运行编队	autonomous satellite formation
抽象归纳	abstract generalization	交换深度	alternation depth	自主通信	autonomic communication
抽象技术	Abstraction technique	交替互补逻辑	alternating-complementary logic	自主配置	autonomous configuration
抽象指令集	abstract instruction set	交替活跃	alternate activity	自主陆地车辆	autonomous land vehicle
抽象描述模型	Abstract description model	交替游程码	alternating run-length code	自主预测	Autonomic prediction
抽象数据切片	abstract data slice	交替的ω－有穷自动机	Alternating ω-finite automata	自主飞艇	Autonomous Flight
抽象数据类型	ADT	交替类估计	Alternating cluster estimation	自主飞行器	autonomous flight vehicle
抽象数据类型（ADT）编译器	abstract data type compiler	交替迭代算法	alternative iteration algorithm	自主驾驶	autonomous driving
抽象数据类形。	abstract data type.	交替顺序重建滤波	alternating sequential filtering by reconstruction	自准直	auto-collimating
抽象文法	abstract grammar	交汇天幕	across sky screen	自制体	agent
抽象服务	Abstract service	交流传动	AC Drive	自动摘要	Automatic Summarization
抽象机	abstract machine	交流伺服	AC serve	自动	Automation

图 1-12 汉英术语词典检索页面

1.4 研究内容以及结构安排

长期以来，图书情报服务的主流是劳动密集性和资源依赖性工作，即通过简单重复的人工劳动和资源的垄断提供情报服务[1]。可接触的学术信息量骤然增多，给过去的主要依赖于人工手段进行信息收集、整理和加工的工作方式带来了新的挑战。“自动化学科创新思想与方法学研究”课题在信息收集、整理和加工上的具体实现方式是：①经过领域内专家多次商讨，选择国内 23 本中文期刊（见附录 1）作为自动化学科论文知识仓库数据源，基本覆盖了自动化学科的相关论文，并收集它们自创刊以来到 2010 年期间全部期刊论文题录信息。②为了能够获得关于同一篇论文的详实题录信息，我们从知网、万方等多个数据源采集数据，进行属性扩充。③从题录信息中提取自动化学科知识要素（论文标题、关键词、作者、机构等），并对知识要素进行本体建模和关联分析。尽管文献题录属于半规则文本，但是从中提取有价值的知识要素，并不是一件容易的事情。以关键词数据为例，表达同一语义的关键词可能存在多种形式，如“李雅普夫定理”和“李亚普夫定理”。另外数据中也普遍存在学者同名现象。因此设计和开发用于知识要素抽取的文本挖掘系统是本论文的核心内容。

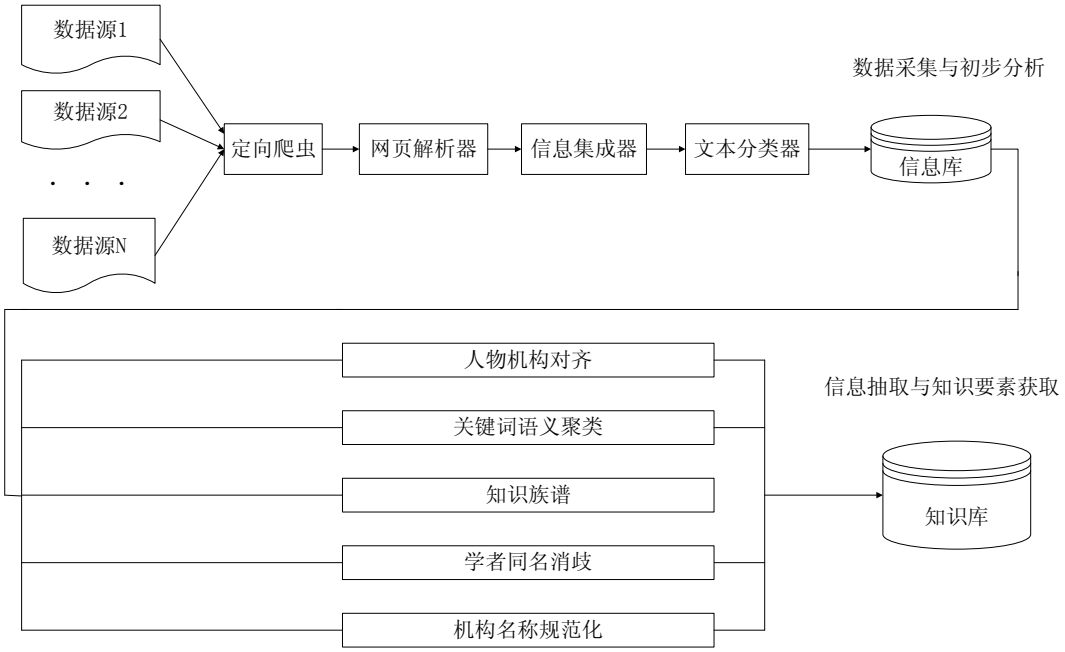


图 1-13 文本挖掘系统框架图

图 1-13 为本文设计的文本挖掘系统的结构框架图。经过调研分析，学者同名消歧、人物机构对齐、关键词的语义聚类等问题是实现高精细粒度的知识要素抽取的关键因素，因此也是本文的核心研究内容。文本分类的相关技术，在数据初步分析中有重要应用，因此也是本文的重点内容。本文按如下结构组织：第一章介绍研究背景和相关工作；第二章介绍文本分类和特征词选择技术及其在数据清洗中的应用；第三章介绍关键词语义聚类算法以及知识族谱；第四章介绍信息抽取技术及其在知识要素获取中的应用；第五章对全文进行了总结，并展望了未来的研究工作。

第二章 文本分类和特征词选择技术及其在数据清洗中的应用

文本分类技术是信息检索、自然语言处理等领域的基本技术，有许多的应用场景，如网页分类、垃圾邮件过滤等。很多研究问题如词义消歧（Word Sense Disambiguation,WSD）、词性标注（Part-Of-Speech, POS），实体链接消解（Entity Linking）等都可以最终归结为分类问题。自 1961 年 Maron 提出了概率文本分类[12]到至今为止，文本分类技术已经有逾 50 年的发展历史，经历了由知识工程方法(knowledge engineering approach)到机器学习方法(machine learning approach)的转变。知识工程方法依靠知识工程师或领域专家人工总结分类规则，代表系统有 CONSTRUE[13]-[16]等。在文档管理领域，该类方法的效果通常要优于机器学习的方法。由于“知识获取瓶颈”，也即该类方法的实施过程中需要大量人工劳动和专家知识来创造和维护知识编码规则，所以基于知识工程的文本分类方法在 80 年代末期经历短暂的流行之后，迅速被基于机器学习的文本分类方法所取代。另一方面由于互联网的发展、海量文本数据的产生，加速了基于知识工程的文本分类方法的淘汰，带动了基于机器学习的文本分类方法的研究和发展。

本章主要介绍基于机器学习的文本分类方法的一般性知识及其在“自动化学科创新思想与科学方法研究”项目中的工程应用，除此之外本文还对文本分类问题中的特征词选择算法做了深入探讨，并提出了 chifit 特征词选择算法。实验结果表明该算法的效果和传统卡方特征词选择算法、信息增益特征词选择算法的效果基本持平，并可以在较低的特征维度上能够取得较好的效果。

本章论文将按如下组织，2.1 节是文本分类技术概述，2.2 节介绍实验中用到的基本分类器 kNN、决策树、贝叶斯、SVM 等，2.3 节介绍文本分类器评价，2.4 节介绍常用特征词选择算法，2.5 节给提出 chifit 特征词选择算法，2.6 节给出课题任务的具体需求以及相应的工程实践，2.7 节为本章小结。

2.1 文本分类技术概述

文本分类是有监督学习的一种，可以形式化定义为对具体的文档类别组$\langle d_j, c_i\rangle \in D \times C$赋予真值,其中 D 是文档集合，$C=\{c_1, c_2, \ldots, c_{|c|}\}$为预先定义好的

类别集合。如果文档d_j隶属于类别c_i则应该被赋予真值，反之应赋予假值。目标函数$\emptyset: D \times C \rightarrow \{T, F\}$被称为分类函数。通常情况下，$\emptyset$是未知的，需要我们进行估计和假设。经过估计得出的函数$\tilde{\phi}$与目标函数$\emptyset$之间的覆盖程度，被定义为分类算法的有效性(effectiveness)[17]。

从类别数目来看，分类问题可以划分为二类分类问题和多类分类问题。在二类（binary case）分类问题中，类别体系由两个互补的类别组成，一篇文档属于或者不属于该类别；在多类问题中，类别体系由三个或者以上的类别组成，如果类别之间具有统计独立性，那么多类问题可以通过转变为二类分类问题解决。文本分类过程中涉及的很多度量指标的计算，比如特征词选择算法的权重计算、分类效果的评定指标计算等，都是以二类分类问题为基础的。从文档是否可以兼类来看，文本分类又可以分为单标签(single label)模式和多标签(multi-label)模式。在单标签模式中，每篇文档只允许隶属于一个类别；在多标签模式中，每篇文章可以隶属于一个以上的类别。

上文中提到的类别标签未必要一定符合人类的语义。分类体系是人为构造的，只要满足一定的逻辑标准即可，比如本人将在附录 2 提供的中文新闻语料共分为包含 Reading、Entertainment、History、Education、Society & Law、Culture、It、Military 等八个类别。因为对于计算机而言，它只需要获得训练样本和类别之间的对应关系即可，而不需要考虑类别标签的语义。

2.1.1 文本分类的一般流程

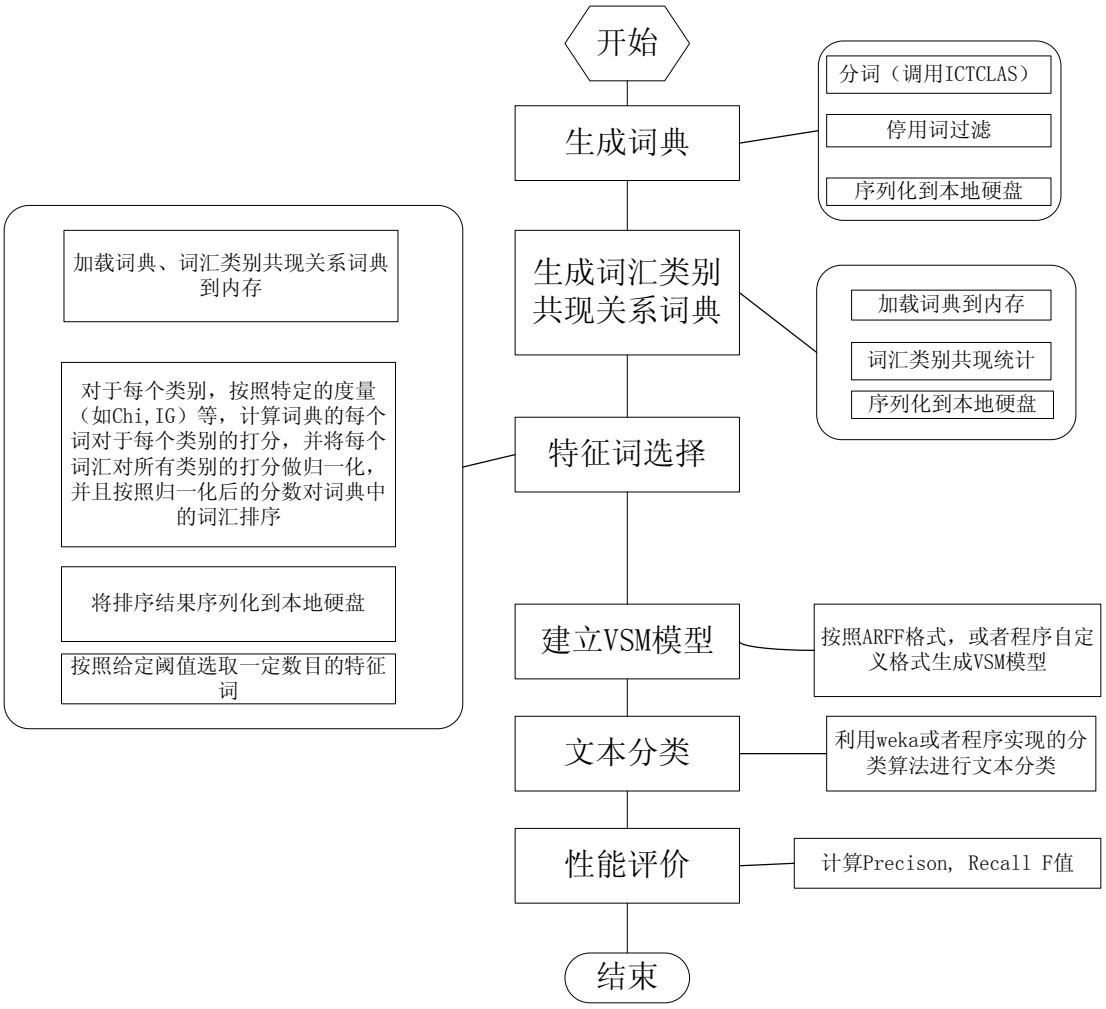


图 2-1 文本分类流程图

图 2-1 给出文本分类过程的基本流程图，文本分类过程涉及到预处理、分类器选择、性能评价三个部分。关于分类器，以及分类器评价将在 2.2，2.3 小节已经给出相应介绍。预处理部分包括生成词典、生成词汇类别共现关系词典、特征选择和建立文档向量模型（Vector Space Model,VSM）四部分，本节将从工程实现的角度介绍文本分类的预处理环节，本节相关代码等请参见附录 3。

①生成词典：文本分类的第一步是根据训练样本集生成一个公共词典，此词典将作为后续步骤的处理对象。通常情况下，在这一步也可以做一些简单的特征选择，比如根据停用词表去除停用词，根据词性过滤一部分虚词等等。如

果训练样本集是中文素材，还需要先对训练样本集进行分词；如果训练样本集是英文素材，必要时还可以进行词根还原。一些开源分词组件诸如 ictclas 等，内部已经实现并封装了针对中英素材的不同分词处理方式，因此可以直接调用进行分词和词性标注，必要时还可以加入短语识别等开源组件。本文在分类程序实现中为词典定义了如下数据结构：

typedef map<string,vector<pair<int,int>.> DICTIONARY

其中 string 为词的文本，pair<int,int>中第一个 int 表示文章的 ID 号，第二个 int 表示该词在文章中出现的次数。所以，整个词典储存的是词，以及词在训练样本集合中的出现情况，如图 2-2 所示。

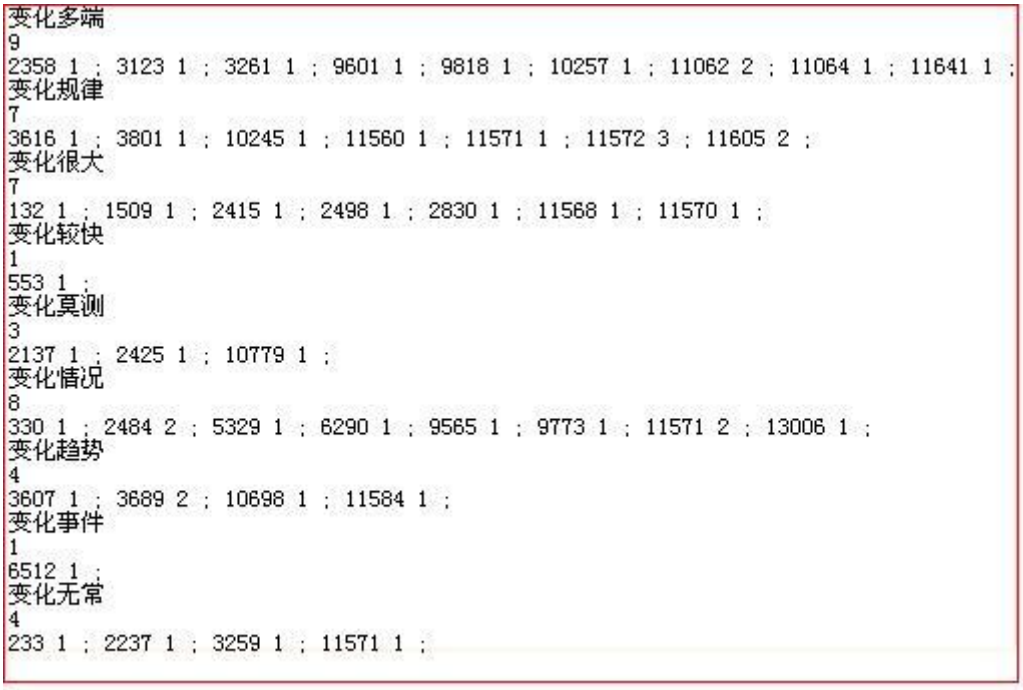


图 2-2 词典数据结构及其内容样例

②生成词汇类别共现关系词典：词汇类别对应关系词典与上文提到的词典一同为特征词选择模块提供统计数据，它主要储存词汇和类别之间的共现次数等内容。本文在分类程序中如下定义此数据结构：

typedef map<pair<string,string>,pair<int,int>> CONTIGENCY

其中 pair<string,string>中第一个 string 表示词，第二个 string 表示类别；pair<int,int>

中第一个 int 表示该类别中含有该词的训练集文档数目，第二个 int 表示该类别中不含有该词的训练文档数目，如图 2-3 所示。

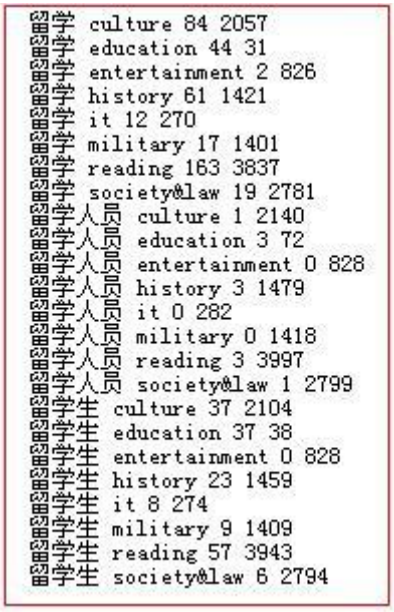


图 2-3 词汇类别对应关系词典数据结构以及内容样例

③特征词选择：特征选择一般分为局部特征词选择和全局特征词选择两种方式。局部特征词选择方式针对不同类别分别选择特征词集合，在计算文档类别隶属关系时，针对不同的类别，同一篇文档要根据不同的特征词集合进行向量化；全局特征词选择方式首先针对不同类别分别选择特征词集合，然后按照某种测度进行归一化，选取针对全局类别的特征词集合。归一化的方式有按照词汇会对类别的最大贡献度[10]归一化或者词汇对于类别的平均贡献度归一化等方式[23]。本文在程序实现上采用的是全局特征词选择方式，设计如下数据结构保存词汇对类别的贡献度：

```
typedef    map<string,vector<pair<string,double>>> FeatureWeight
```

其中 string 为词汇的文本表示，pair<string,double>中第一个 string 为类别标识，double 为词汇对类别的贡献度。如图 2-4 所示，为以卡方为贡献度计算方式时的词汇对类别贡献度情况。如图中的词汇“华为”、“华硕”对 it 类的贡献度最大，分别为 49.1629 和 90.3969；对 education 类的贡献度最小，分别为 0.162499

[10]所谓贡献度，是指按照卡方、信息增益等公式计算出的、能够表征词汇与类别相关度的量化数值。

和 0.0115839。

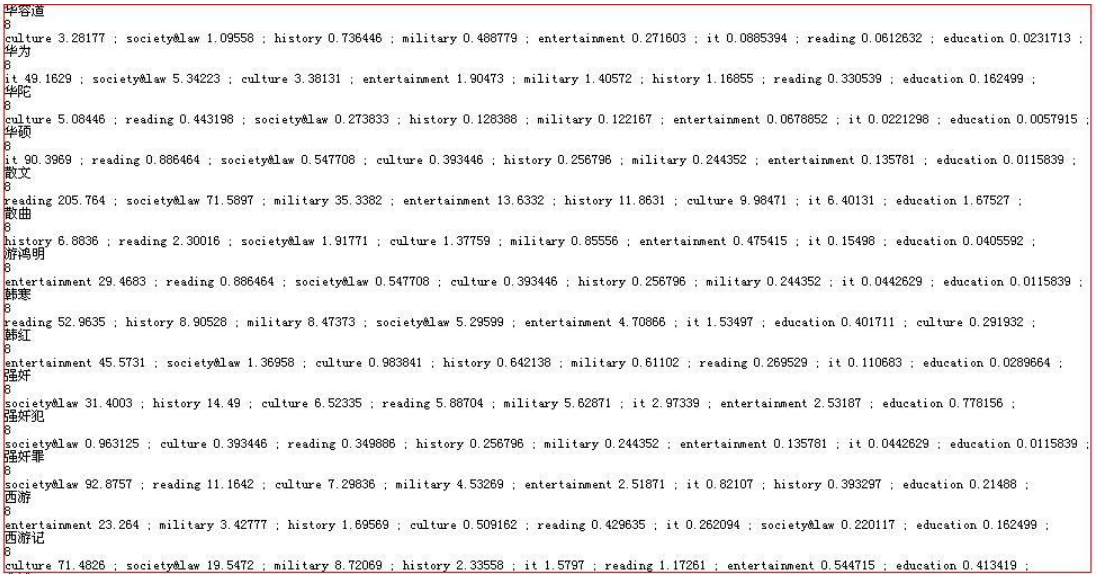


图 2-4 词汇对类别的贡献度存储数据结构以及内容样例

归一化公式如公式 2-1、2-2 所示，其中 m 为类别数目：

$$tc_{avg}(t)=\sum_{i=1}^{m}p(c_i)tc(t,c_i) \qquad \text{公式 2-1}$$

$$tc_{max}(t)=\underset{i=1}{\overset{m}{Max}}\{tc(t,c_i)\} \qquad \text{公式 2-2}$$

将词汇对类别贡献度归一化后，按贡献度从高到低排列，即可按照特征维度要求选取不同数目的特征词。

④建立文档向量模型（VSM）：这一步骤是为特征词集合和文档集合建立如图 2-5 所示的文档—词汇标引矩阵，其中 m 为文档数目，n 为特征词数目，a_{ij} 表示特征权重。特征权重一般为词频(Term Frequency,TF)和倒文档频率（Inverse Document Frequency,IDF）的某种加权计算[24][25][26]。SMART 系统的 TFIDF 权重计算方式见图 2-6 所示，本文程序中采用的是 *l-t-c* 模式。

$$A = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ \\ \\ t_n \end{array} \begin{array}{c} \begin{array}{cccccc} d_1 & d_2 & d_3 & \cdots & d_{m-1} & d_m \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & & & \vdots \\ \\ \\ a_{n1} & a_{n2} & a_{n3} & \cdots & \cdots & a_{nm} \end{bmatrix} \end{array}$$

图 2-5 文档——词汇标引矩阵

term frequency		document frequency		normalization	
n (natural)	$tf_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1+\log(tf_{t,d})$	t (idf)	$\log \frac{N}{df_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\cdots+w_M^2}}$
a (augmented)	$0.5+\frac{0.5\times tf_{t,d}}{\max_t(tf_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-df_t}{df_t}\}$	u (pivoted unique)	$1/u$ (Section 17.4.4)
b (boolean)	$\begin{cases}1 \text{ if } tf_{t,d}>0 \\ 0 \text{ otherwise}\end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha<1$
L (log ave)	$\frac{1+\log(tf_{t,d})}{1+\log(ave_{t\in d}(tf_{t,d}))}$				

图 2-6 不同 tf-idf 方法的 SMART 系统记号[11]

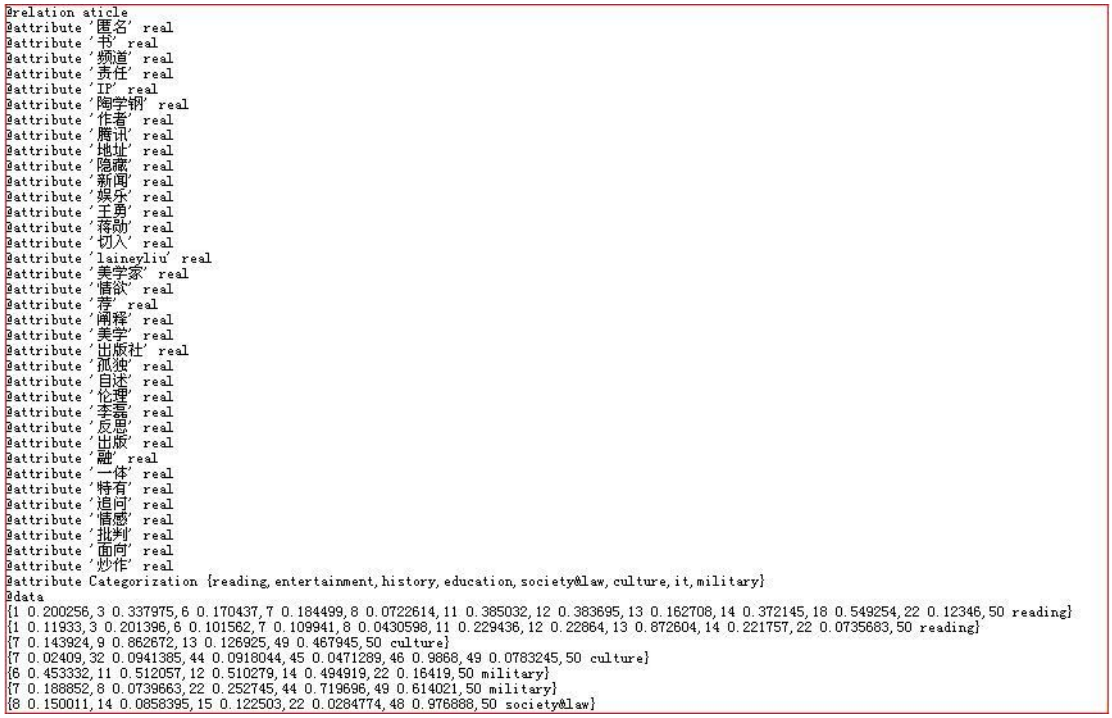
程序中定义文档向量模型的存储数据结构如下

typedef map<int,vector<pair<int,double>>> DOCMATRIX_1

其中 int 为文档 ID 号，pair<int,double>中的 int 为特征词序号，double 为 tf-idf 权重值。程序最终形成的 ARFF[12]文本格式的文档向量模型，如图 2-7 所示。

[11]本图摘自文献[27]，其中 CharLength 指的是以字节为单位的文档长度

[12]ARFF 为 Weka 软件要求的数据格式



```
@relation aticle
@attribute '匿名' real
@attribute '书' real
@attribute '频道' real
@attribute '责任' real
@attribute 'IP' real
@attribute '陶学钢' real
@attribute '作者' real
@attribute '腾讯' real
@attribute '地址' real
@attribute '隐藏' real
@attribute '新闻' real
@attribute '娱乐' real
@attribute '王勇' real
@attribute '蒋勋' real
@attribute '切入' real
@attribute 'laineyliu' real
@attribute '美学家' real
@attribute '情欲' real
@attribute '荐' real
@attribute '阐释' real
@attribute '美学' real
@attribute '出版社' real
@attribute '孤独' real
@attribute '自述' real
@attribute '伦理' real
@attribute '李磊' real
@attribute '反思' real
@attribute '出版' real
@attribute '融' real
@attribute '一体' real
@attribute '特有' real
@attribute '追问' real
@attribute '情感' real
@attribute '批判' real
@attribute '面向' real
@attribute '炒作' real
@attribute Categorization {reading,entertainment,history,education,society&law,culture,it,military}
@data
{1 0.200256,3 0.337975,6 0.170437,7 0.184499,8 0.0722614,11 0.385032,12 0.383695,13 0.162708,14 0.372145,18 0.549254,22 0.12346,50 reading}
{1 0.11933,3 0.201396,6 0.101562,7 0.109941,8 0.0430598,11 0.229436,12 0.22864,13 0.872604,14 0.221757,22 0.0735683,50 reading}
{7 0.143924,9 0.862672,13 0.126925,49 0.467945,50 culture}
{7 0.02409,32 0.0941385,44 0.0918044,45 0.0471289,46 0.9868,49 0.0783245,50 culture}
{6 0.453332,11 0.512057,12 0.510279,14 0.494919,22 0.16419,50 military}
{7 0.188852,8 0.0739663,22 0.252745,44 0.719696,49 0.614021,50 military}
{8 0.150011,14 0.0858395,15 0.122503,22 0.0284774,48 0.976888,50 society&law}
```

图 2-7 ARFF 格式的文档向量模型

2.1.2 基于规则与基于统计的文本分类方法

基于机器学习的文本分类方法，又可以细分为基于规则(rule-based)的分类方法和基于统计(probability-based)的分类方法。基于规则的分类方法如决策树等，容易被人所理解，但是分类性能不是很稳定，而且容易出现过拟合(overfitting)问题。例如用决策树进行分类时，可能出现预测训练集数据时其精度非常高，但是预测测试集数据时，其精度又非常低的情况。所以在不需要人来理解分类过程的应用场景中，很少使用决策树分类器。目前基于统计的文本分类方法占据主流地位。

2.2 常用文本分类器介绍

基于统计的文本分类方法又可以根据是对类别和观察数据进行联合概率建模，还是对类别和观察数据进行条件概率建模，具体分为产生式分类和鉴别式分类。产生式模型对输入 x 和标签 y 的联合概率 $p(x,y)$进行建模，并通过贝叶斯准则计算后验概率 $p(y|x)$，然后选出最有可能的类别标签 y；鉴别式模型直接对后验概率 $p(y|x)$进行建模。一般来说，鉴别式分类器的分类效果要好于产生式分

类器的分类效果。根据 Vapnik 的统计学习理论[18]“应该试图直接解决问题，而不是试图解决一个比待解决问题更具一般性的问题”，以及 Occam 剃刀原理[19]“除非必要，‘实体’（或‘解释’）不应该随便增加”，研究者更看好鉴别式分类器。Andrew Y.Ng 等学者指出：随着训练样本集的不断增大，鉴别式分类器能够达到更低的渐近错误率，但是产生式分类器能够更快地达到它的渐近错误率[20]。

2.2.1 kNN 分类器

kNN 分类算法属于鉴别式分类算法，又被称为基于实例的分类算法，或者懒惰学习方法。该算法没有训练过程，给定一个待分类的文档 d，在训练文档集合 $\boldsymbol{D}$ 中选取离 d 最近的 k 个样本，并统计这 k 个样本的类别标签，将出现次数最多的类别标签赋予文档 d，作为 d 的类别标签，因此它是一种基于局部信息的方法。在度量文档之间的距离属性时，通常选取余弦相似度。kNN 算法具有很好的分类效果，但是当特征维度增高时，会出现维度灾难（the curse of dimensionality），导致分类性能下降[154]。图 2-17 和图 2-19 中也可以观察到 kNN 分类算法在高维特征空间的分类效果下降现象。

2.2.2 朴素贝叶斯与多项式贝叶斯分类器

从概率论的角度来看，分类过程实际上是一个求最大后验概率的过程 $c = arg\,max_{c\epsilon C}\ p(c_j|d)$。假设文档由词汇组成，即：$d = [(t_1, n_1),\ (t_2, n_2), \dots,\ (t_i, n_i), \dots, (t_v, n_v)]$，其中 v 为词汇表大小，n_i 表示词汇 t_i 在文档 d 中出现的次数。若 n_i 的取值只能为布尔值，也就是只考虑词汇在文档中是否出现，而不考虑词汇在文档中出现的次数，称相应的分类算法为朴素贝叶斯或者多变量伯努利贝叶斯；若 n_i 可取非 0 或 1 的其他正整数值，则表示文档是多项式分布的一个具体实现，称相应的分类算法为多项式贝叶斯。

朴素贝叶斯

根据类别条件独立性假设：当类别给定时，词汇之间具有独立性，朴素贝叶斯的后验概率可以进一步推导如公式 2-3 所示。

$$P(c_j \mid d) = p(c_j \mid t_1 t_2 ... t_i ... t_v) = \frac{p(c_j t_1 t_2 ... t_i ... t_v)}{p(d)}$$
$$= \frac{p(c_j)\prod_i^v p(t_i \mid c_j)}{p(d)} \quad \text{公式 2-3}$$

由于对同一篇文档进行分类，p(d)相同，所以计算时可以省去，最终的类别计算公式如公式 2-4 所示。

$$c = \arg\max_{c_j \in C} p(c_j)\prod_i^v p(t_i \mid c_j) \quad \text{公式 2-4}$$

此时关于概率 $p(c_j)$和 $p(t_i|c_j)$的估计成为核心问题。朴素贝叶斯和多项式贝叶斯在这两个概率值的估计上有所不同。朴素贝叶斯做如下估计：

$$p(c_j) = \frac{\#\{\text{属于}c_j\text{类的训练样本集数目}\}}{\#\{\text{训练样本集总数目}\}} \quad \text{公式 2-5}$$

$$p(t_i \mid c_j) = \frac{\#\{\text{属于}c_j\text{类且含有}t_i\text{的训练样本数目}\}}{\#\{\text{属于}c_j\text{类的训练样本数目}\}} \quad \text{公式 2-6}$$

多项式贝叶斯

在多项式贝叶斯建模中，认为文档由多项式分步生成[155]，即：

$$p(d_i \mid c_j) = p(|d_i|)(|d_i|!)\prod_{t=1}^{|V|}\frac{p(w_t \mid c_j)^{N_{ti}}}{N_{ti}!}$$
$$\text{其中}\sum_{t=1}^{|V|} N_{ti} = |d_i|, \sum_{t=1}^{|V|} p(w_t \mid c_j) = 1 \quad \text{公式 2-7}$$

后验概率为：

$$p(c_j \mid d_i) = \frac{p(c_j)p(d_i \mid c_j)}{p(d_i)} = \frac{p(c_j)\prod_{k=1}^{|d_i|} p(w_{d_i,k} \mid c_j)}{p(d_i)} \quad \text{公式 2-8}$$

其中

$$p(c_j)=\frac{\sum_{i=1}^{|D|}p(c_j\mid d_i)}{|D|},\ p(w_t\mid c_j)=\frac{\sum_{i=1}^{|D|}N_{ti}p(c_j\mid d_i)}{\sum_{s=1}^{|V|}\sum_{i=1}^{|D|}N_{si}p(c_j\mid d_i)} \qquad \text{公式 2-9}$$

如果d_i在c_j中，则$p(c_j|d_i)$取值为 1，否则为 0。

2.2.3 决策树

决策树学习算法基于分治策略（Divide-and-Conquer），算法根据混杂度函数每次选取最佳分类属性作为分隔当前数据实例集的属性，直到所有的当前结点中的数据都属于同一类时算法终止。其中，常用的混杂度函数包括信息增益、信息增益率等。

2.2.4 支持向量机

支持向量机是一个线性学习系统，适用于二分类问题。支持向量机寻找最优分隔平面，使得该平面到两类数据点的最小距离之和最大，也即和两类的支持向量之间的距离之和最大。根据凸优化问题的拉格朗日定理以及对偶定理，寻找最优分隔平面的问题归结为求α使得公式 2-10 成立。

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\alpha^TQ\alpha\text{-}e^T\alpha\\&subject\,to\ y^T\alpha=0,\quad 0\le\alpha_i\le C,i=1,...,l.\end{aligned} \qquad \text{公式 2-10}$$

其中 e 是全 1 向量，C 是上界，Q 是$l\times l$的半正定矩阵，$Q_{ij}=y_iy_jK(x_i,x_j)$，$K(x_i,x_j)=\varphi(x_i)^T\varphi(x_j)$是核函数。类别判定函数为$\mathrm{sgn}\left(\sum_{i=1}^{l}y_i\alpha_iK(x_i,x)+b\right)$.

2.3 分类器的有效性评价

2.3.1 常用评价指标

在文本分类中，常用查准确(Precision，p)、查全率(Recall,r)和 F 值(F-score,f)作为衡量分类器性能的标准。为了定义以上三个指标，首先定义分类器的混合矩阵，如表 2-1 所示。查准率 p、查全率 r 和 F 值 f 分别由公式 2-11、2-12、2-13 定义。其中查准率的含义是测试集中被正确分类的正例数量除以被分为正例的

数据数量；查全率的含义是测试集中被正确分类的正例的数量除以测试集中实际正例的数量；F 值为查准率与查全率的调和平均数。

表格 2-1 分类器的混合矩阵

	分类为正例	分类为负例
实际上为正例	TP	FN
实际上为负例	FP	TN

$$p = \frac{TP}{TP + FP} \quad \text{公式 2-11}$$

$$r = \frac{TP}{TP + FN} \quad \text{公式 2-12}$$

$$f = \frac{2pr}{p + r} \quad \text{公式 2-13}$$

对于多分类问题，如附录 2 中的中、英新闻分类语料，则需要先分别计算各个类别的查准率、查全率和 F 值，然后按照某种测度求平均。一般有宏平均(macroaveraging)和微平均(microaveraging)两种做法，宏平均将各个类别同等对待是在类别之间求算术平均值，而微平均则将每篇文档的分类判定结果等同对待，它将各个类别的分类判定级联起来，视为一个分类判定。宏平均指标在一定程度上反映稀有类别上的效果，而微平均指标主要反映普通类别上的效果。

本文对文本分类器的研究在于评价某种分类器是否适合在某个具体的实际数据场景中应用，其侧重点在于考证分类器的总体精准性能，因此使用兰德指数(Rand Index,RI)作为评估指标。假定所有文档只允许有一个类别标签，RI 衡量的是所有被正确分类的文档数量与测试文档集总数量的比率，定义见公式 2-14。

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{公式 2-14}$$

2.3.2 分类效果的一般性评价

分类器性能的评估与文档集合情况（包括类别数目，是否为均衡语料等）以及实验设置（包括评估方法等）有关，因此在不同研究者的研究成果之间做关于分类器性能的比较是很难得到一般性结论的。根据大多数研究者的研究结论，有如下三条总结[21]：

①一般来说 SVM、AdaBoost、kNN 以及 regression 分类器性能较好，没有充足的统计证据能够证明其中哪一种分类器性能最好。在针对具体问题的实际应用中，选择分类器时应考虑效率、实现复杂度等多重因素。

②Rocchio 和 Naïve Bayes 在所有基于机器学习的分类器中效果最差，经常被用作基准分类器；

③神经网络和决策树的分类效果不稳定，跟具体实验相关。在一些实验中他们能够取得和 SVM 相当的分类效果，而在另一些实验中效果又相当差。

2.3.3 分类器评价实验与分析

本文针对 Naïve Bayes, Multinomial Bayes, 决策树、kNN,SMO 等基本分类器在中英两种分类语料，采用卡方、chifit 两种特征词选择算法生成的 VSM 模型上，分别选取 50,100,150…500 等 10 个特征维度的 RI 值作为评价标准，得出了大致相同的结论。实验结果如图 2-8、2-9、2-10、2-11 所示。

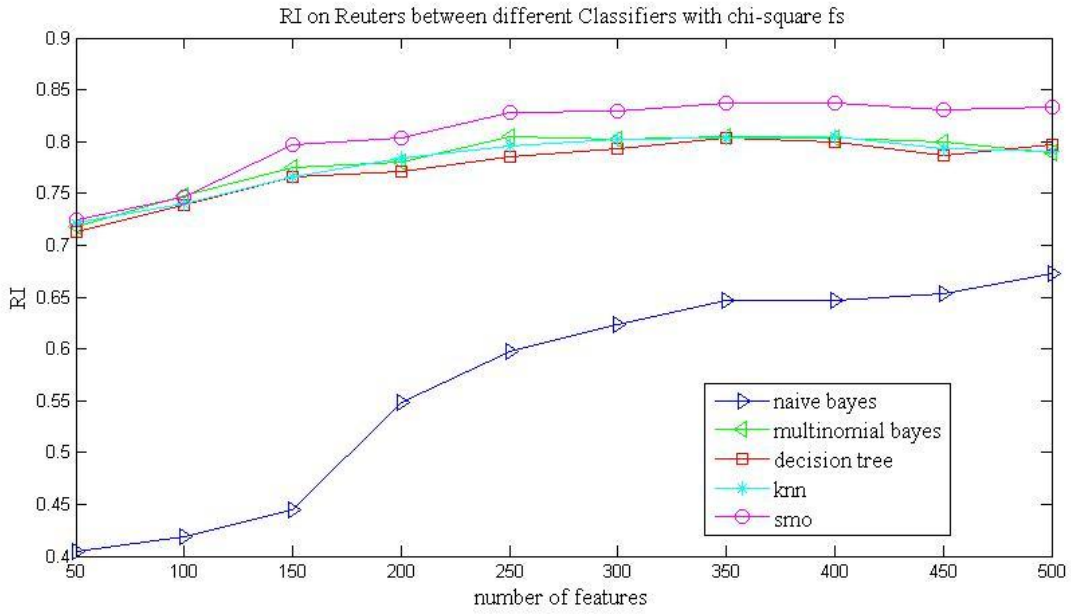


图 2-8 Reuters 语料上卡方特征选择算法的分类效果评估

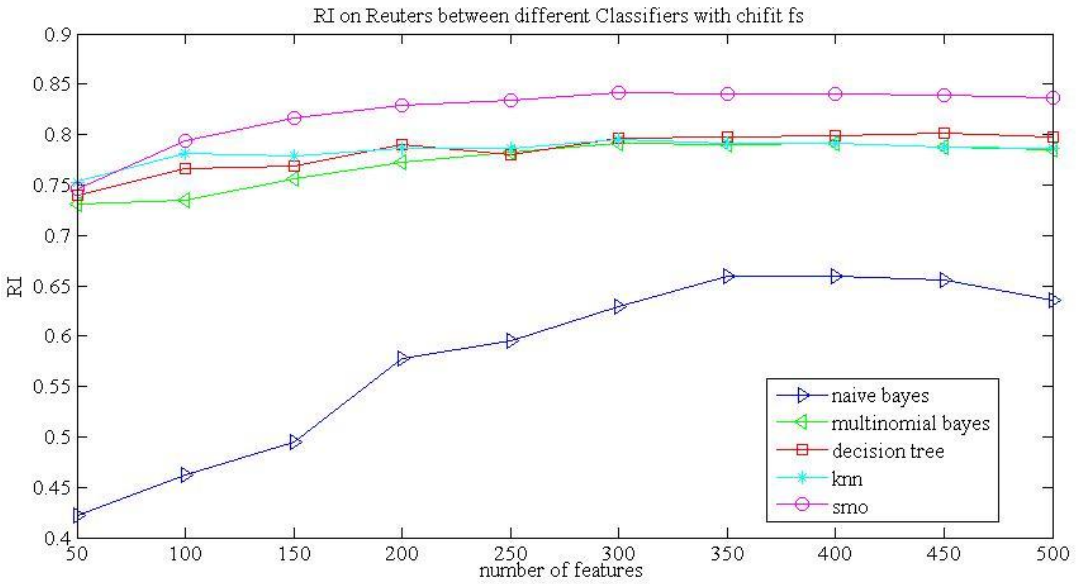


图 2-9 Reuters 语料上 chifit 特征选择算法的分类效果评估

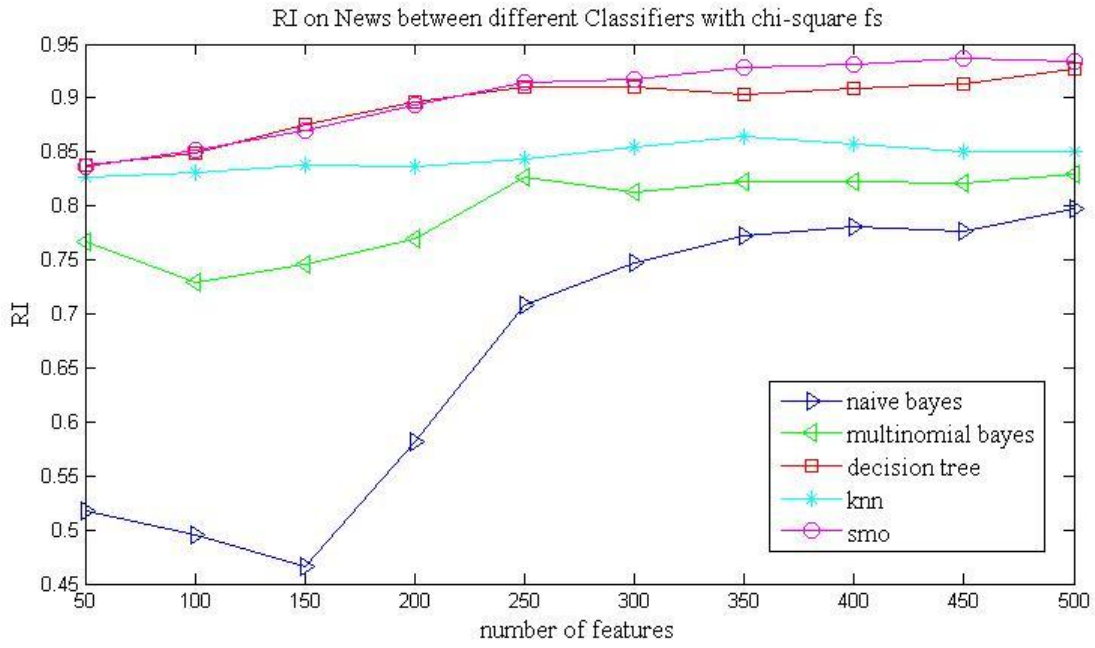


图 2-10 中文新闻分类语料上卡方特征选择算法的分类效果评估

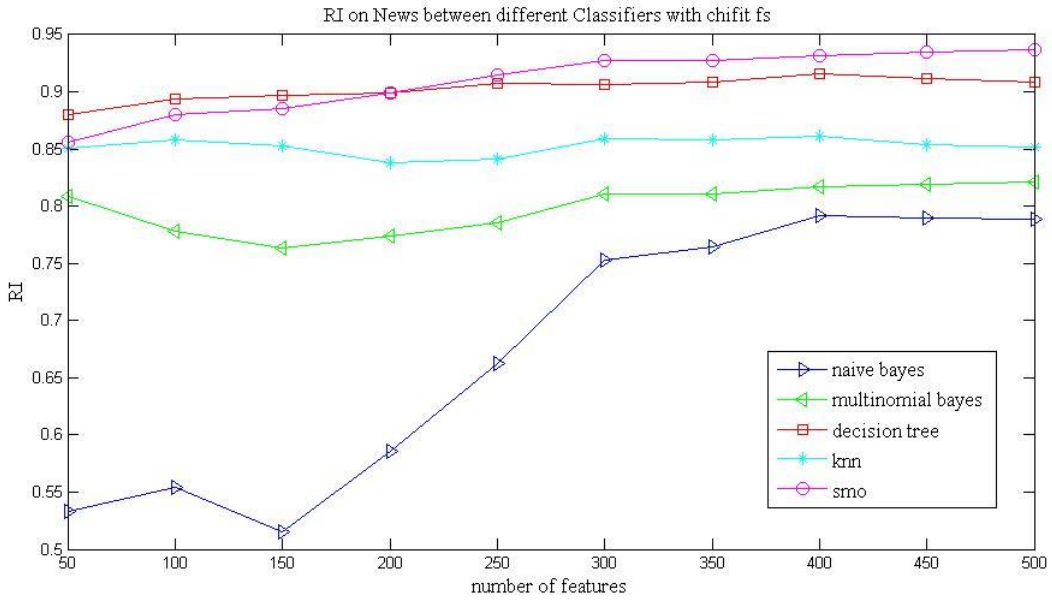


图 2-11 中文新闻分类语料上 chifit 特征选择算法的分类效果评估

关于文本分类语料和实验中所使用的 VSM 模型的具体介绍见附录 2、附录 4。分类器为 Weka 自带的基本分类器，分别为 weka.classifiers.bayes.NaiveBayes（朴素贝叶斯分类器）、weka.classifiers.bayes.NaiveBayesMultinomial（多项式贝叶斯分类器）、weka.classifiers.trees.J48（决策树分类器）、weka.classifier.lazy.IBk（kNN 分类器）、weka.classifier.functions.SMO（SVM 分类器）[22]。其中除了 kNN 分类器中参数 k 取 5 之外，其他分类器参数采用 Weka 默认值。图 2-8、2-9 说明在 Reuters 语料上 SMO 分类器的分类效果要好于其他分类器；图 2-10、2-11 说明在中文新闻分类语料库上 kNN 的效果要恒定优于 multinomial bayes，决策树和 smo 效果相差不大；以上四幅图都说明 multinomial bayes 算法效果要优于 naïve bayes 算法的效果；Reuters 语料共有单词 66，363 个，中文新闻分类语料共有单词 147，907 个，以上四幅图中 RI 值分布在(0.6,0.95)之间，可以说明特征词选择对特征降维有明显的促进作用。

2.4 现有特征词选择算法分析与实现

2.4.1 特征词选择算法应具备的特点

特征词选择算法是否有效，主要看利用该算法生成的特征词集合是否具有以下四个特征：①类别区分性，即如果文章中出现某些词就可以基本断定文章

的类别；②所选取的特征词应该在它所“标识”的类别中频繁出现；③所选取的特征词应该在它所“标识”的类别的文章中均匀出现，能够索引该类别的多数文章；④特征词集合能够尽可能多地索引测试样本集的文章。

除此之外，算法在实现上是否简单、易用，算法和其他算法相比能否在更小的特征集上达到相同的效果等也是工业中选择特征词选择算法的重要标准。

如何寻找最优特征，以何种标准和评估手段来评价所选取的特征是否是最优特征一直是特征选择研究面临的两个首要问题[29]。在文本分类中一般采用评价分类器有效性的评价指标来评估特征词选择算法的性能。具体来说是在多个语料上运行待测试的特征词选择算法，并将生成的文档向量模型分别通过多种分类器，测试分类效果。一些算法在一些论文中获得良好的效果，但实际上可能只是实验不充分的“杰作”（例如仅在一种语料上做对比实验，仅用一种分类器做测试，仅和一种特征词选择算法做对比等等）。

本人在文献调研中发现，一些特征词选择算法是对常用特征词选择算法的改进或者和常用特征词选择算法近似等价。如文献[30]中给出的特征词选择算法计算公式是只保留卡方特征词选择算法计算公式的分子，并开平方的结果。文献[31]给出的特征词选择算法实际上是仅对卡方特征词选择算法计算公式的分子开平方的结果。文献[32]中基于贝叶斯理论提出了以词汇类别后验概率$p(c_i|t)$为基础，并加入阈值限制的特征词选择算法，不难证明当训练语料为均衡语料，即每个类别所含有的文档数目相同时，该算法等价于在点互信息特征词选择算法中加入阈值限制，从而克服点互信息特征词选择算法在选词上具有稀疏性的缺点。文献[33]中提出的 MI 特征词选择算法在一些语料集合上同样克服了点互信息特征词选择算法的选词具有稀疏性特点。MI 算法实际上是求词汇事件集合 A={t 出现，t 不出现}与类别事件集合 B={c 出现}之间的集合平均互信息。

另外一些特征词选择算法在理论上创新很大，但是实现起来又过于复杂。文献[34]认为特征词选择主要依赖于频度、集中度和分散度指标。并提出了基于对数似然比测试的特征词选择算法，将以上三种指标一体化。在算实现上将词汇在类别上的条件概率用 e 指数家族的概率密度函数进行参数化建模，增加了算法复杂度，导致算法的可用性降低。

2.4.2 常用特征词选择算法

特征词选择算法按照是否需要统计词汇和类别关联信息，可以分为有监督的特征词选择算法和无监督的特征词选择算法两种。其中无监督的特征词选择算法不需要依赖类别信息，有监督的特征词选择算法需要依赖类别信息。有监督特征词选择算法一般对词汇相关的事件集合、类别相关的事件集合、词汇——类别相关的事件集合共同进行概率估计，上述事件集合的定义如下：

i 词汇相关的事件集合$E_{t=}\{e_t = 0, e_t = 1\}$；

ii 类别相关的事件集合$E_{c=}\{e_c = 0, e_c = 1\}$；

iii 词汇—类别相关的事件集合$E_{tc} = \{e_{tc} = 00, e_{tc} = 01, e_{tc} = 10, e_{tc} = 11\}$；

其中用 0 表示“不出现”，1 表示“出现”；

借助以上形式化描述体系，N_{11}表示词汇 t 和类别 c 共现的文档数目；N_{10}表示含有词汇 t 但是不属于类别 c 的文档数目；N_{01}表示属于类别 c 但是不包含词汇 t 的文档数目；N_{00}表示既不包含词汇 t，又不属于类别 c 的文档数目；$N_{11} + N_{10}$表示含有词汇 t 的文档数目；$N_{11} + N_{01}$表示属于类别 c 的文档数目；$N = N_{11} + N_{01} + N_{11} + N_{10}$表示训练样本集合总体的文档数目。

相应概率的 MLE 估计为：

$$p(e_t = 1) = \frac{N_{11} + N_{10}}{N} \quad \text{公式 2-15}$$

$$p(e_t = 0) = \frac{N_{00} + N_{01}}{N} \quad \text{公式 2-16}$$

$$p(e_c = 0) = \frac{N_{00} + N_{10}}{N} \quad \text{公式 2-17}$$

$$p(e_c = 1) = \frac{N_{11} + N_{01}}{N} \quad \text{公式 2-18}$$

$$p(e_t = 1, e_c = 1) = \frac{N_{11}}{N} \quad \text{公式 2-19}$$

$$p(e_t = 1, e_c = 0) = \frac{N_{10}}{N} \quad \text{公式 2-20}$$

$$p(e_t = 0, e_c = 0) = \frac{N_{00}}{N} \quad \text{公式 2-21}$$

$$p(e_t=0,e_c=1)=\frac{N_{01}}{N} \qquad \text{公式 2-22}$$

①文档频率（Document Frequency,DF）：该方法将文档频率小于一定阈值的词汇过滤掉，只保留在训练文档集合中高频出现的词汇，属于无监督的特征词选择算法，因此也可以应用在聚类问题中。文献[23]认为 DF 特征词选择方法能够在保持分类效果的前提下将特征词维度降到训练文档集合全部词汇的 10%。

②卡方（卡方,CHI）：卡方是统计学范畴的度量。在统计学中常用卡方检测两个事件 A、B 之间的统计独立性。即$P(AB)=P(A)P(B)$。在特征词选择算法这一特定应用场景下，先假设词汇 t 与类别 c 之间统计独立，则在词汇—类别相关的事件集合上有：

$$\chi^2(\mathrm{D},\mathrm{t},\mathrm{c})=\sum_{e_t\in\{0,1\}}\sum_{e_c\in\{0,1\}}\frac{(N_{e_te_c}-E_{e_te_c})^2}{E_{e_te_c}} \qquad \text{公式 2-23}$$

根据统计独立性假设有

$$E_{01}=N\times p(e_t=0)\times p(e_c=1)) \qquad \text{公式 2-24}$$

$$E_{01}=N\times p(e_t=0)\times p(e_c=1)) \qquad \text{公式 2-25}$$

$$E_{00}=N\times p(e_t=0)\times p(e_c=0)) \qquad \text{公式 2-26}$$

$$E_{11}=N\times p(e_t=1)\times p(e_c=1)) \qquad \text{公式 2-27}$$

最终得到卡方的 MLE 计算公式为：

$$\chi^2(D,t,c)=\frac{N\times(N_{11}N_{00}-N_{10}N_{01})^2}{(N_{11}+N_{01})(N_{11}+N_{10})(N_{10}+N_{00})(N_{01}+N_{00})} \qquad \text{公式 2-28}$$

卡方公式的结果为 0，表示词汇与类别无关，卡方公式的结果越大，则表示词汇与类别之间的相关性越大。

③信息增益(InformationGain,IG)：信息增益属于信息理论范畴的度量，它实际上是词汇相关的事件集合与类别相关的的事件集合的集合平均互信息，因此其取值范围在区间$[0,+\infty)$，当其取值为 0 时，表示两个事件集合相互独立。信

息论中对集合平均互信息的定义为：

$$I(X,Y)=\sum_{XY}p(xy)\log\frac{p(xy)}{p(x)p(y)} \quad \text{公式 2-29}$$

在特征词选择应用背景之下信息增益的公式具体为：

$$I(U,C)=\sum_{e_t\in\{0,1\}}\sum_{e_c\in\{0,1\}}p(U=e_t,C=e_c)\log\frac{p(U=e_t,C=e_c)}{p(U=e_t)p(C=e_c)} \quad \text{公式 2-30}$$

最终得到信息增益的 MLE 计算公式为：

$$I(U,C)=\frac{N_{11}}{N}\log\frac{NN_{11}}{(N_{11}+N_{10})(N_{11}+N_{01})}+\frac{N_{01}}{N}\log\frac{NN_{01}}{(N_{00}+N_{01})(N_{11}+N_{01})}$$
$$+\frac{N_{10}}{N}\log\frac{NN_{10}}{(N_{00}+N_{10})(N_{11}+N_{10})}+\frac{N_{00}}{N}\log\frac{NN_{00}}{(N_{00}+N_{01})(N_{00}+N_{10})} \quad \text{公式 2-31}$$

④点互信息（point-wise MI,MI）：点互信息也属于信息理论范畴的度量，它实际上事件“词汇出现”与事件“类别出现”之间的事件互信息。信息论中关于事件互信息的定义为：

$$I(x_i,y_j)=\log\frac{p(x_i\mid y_j)}{p(x_i)}=\log\frac{p(x_iy_j)}{p(x_i)p(y_j)}=\log\frac{p(y_j\mid x_i)}{p(y_j)}=I(y_j,x_i) \quad \text{公式 2-32}$$

当事件x_i、y_j统计独立时，点互信息量为 0。点互信息量可正可负。在给定事件y_j的条件下，事件x_i出现的概率$p(x_i|y_j)$称为后验概率。当后验概率$p(x_i|y_j)$大于先验概率$p(x_i)$时，点互信息量$I(x_i,y_j)$大于 0，含义是事件y_j的出现有利于事件x_i的出现；当后验概率$p(x_i|y_j)$小于先验概率$p(x_i)$时，点互信息量$I(x_i,y_j)$小于 0，含义是是事件y_j的出现不利于事件x_i的出现[36]。

在特征词选择应用背景之下点互信息的公式为：

$$I(t,c)=\log\frac{p(e_t=1,e_c=1)}{p(e_t=1)p(e_c=1)} \quad \text{公式 2-33}$$

最终得到信息增益的 MLE 计算公式为：

$$I(t,c)=\log\frac{NN_{11}}{(N_{11}+N_{10})(N_{11}+N_{01})} \qquad \text{公式 2-34}$$

2.4.3 算法实验与分析

本文对常用特征词选择算法 CHI,IG,DF,MI 的性能分别采用 KNN 分类器、多项式贝叶斯分类器在 Reuters 语料和中文新闻语料上做验证，实验结果如图 2-12 至 2-15 所示。图 2-12、图 2-13 分别为在 Reuters 语料采用多项式贝叶斯分类器和 KNN 分类器的实验结果；图 2-14、图 2-15 分别为在中文新闻语料上采用多项式贝叶斯和 KNN 分类器的实验结果。

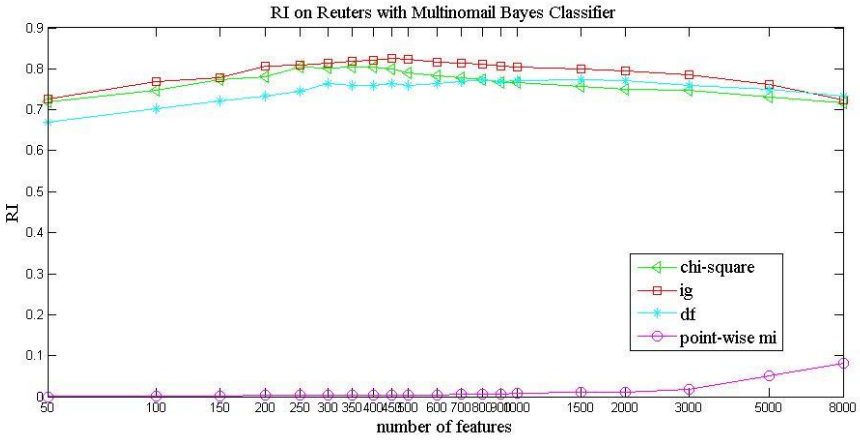


图 2-12 四种常用特征词选择算法在 Reuters 语料上的效果（一）

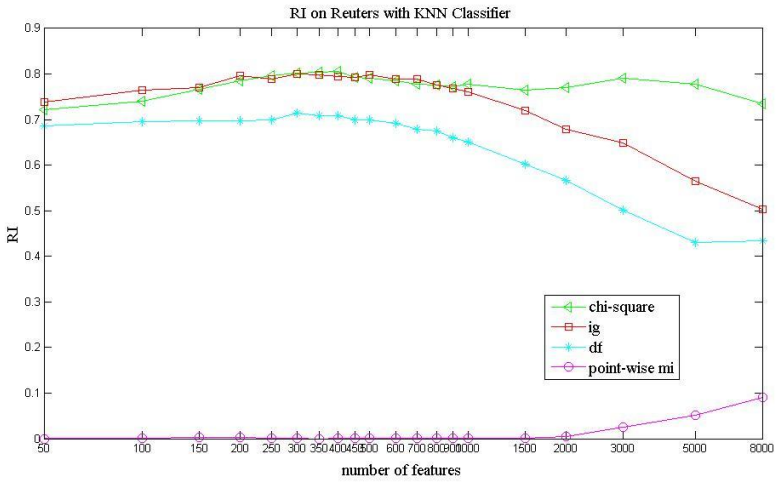


图 2-13 四种常用特征词选择算法在 Reuters 语料上的效果（二）

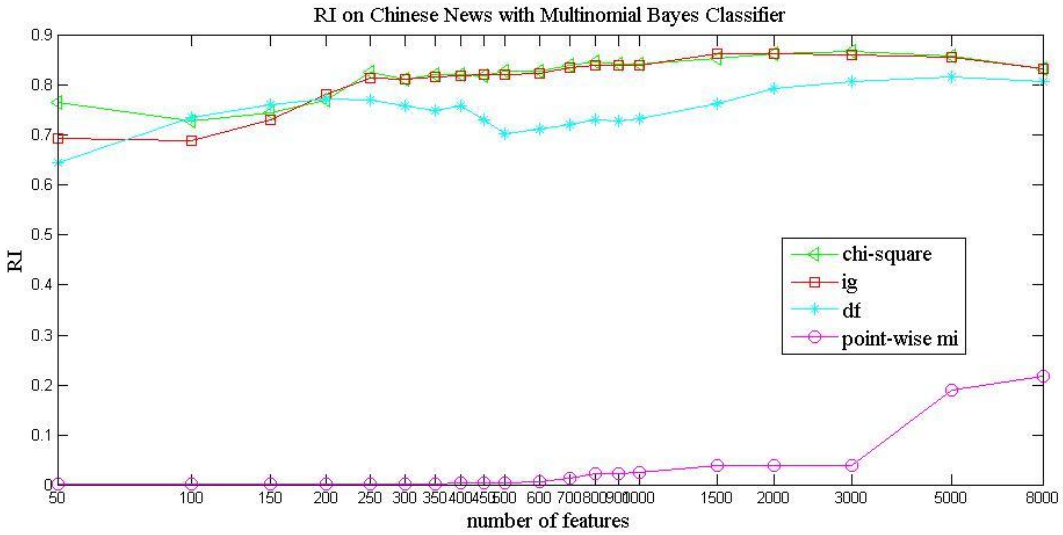


图 2-14 四种常用特征词选择算法在中文新闻语料上的效果（一）

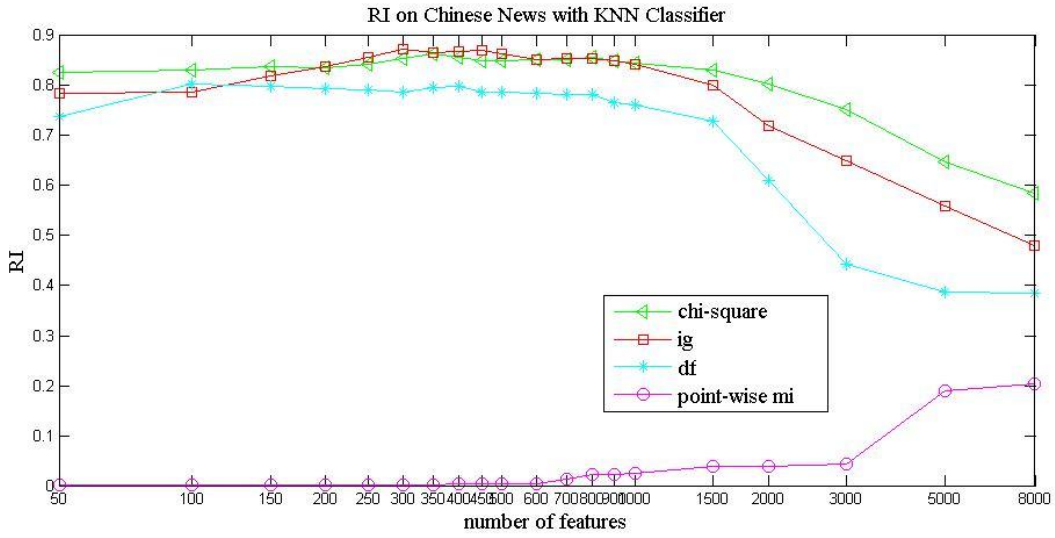


图 2-15 四种常用特征词选择算法在中文新闻语料上的效果（二）

综合以上四幅图，可以得出和文献[23]类似的结论，即：卡方、信息增益以及文档频率特征词选择算法的效果要远远优于点互信息特征词选择算法。观察点互信息公式 2-32 可以发现，点互信息特征词选择算法的效果不好的原因在于该方法倾向于选择稀有词汇作为特征词。本文分别选取了卡方、信息增益、点互信息特征词选择算法在 Reuters 和中文新闻语料上生成的前 50 位特征词（见附录 20），并分别与文档频率特征词选择算法在同一份语料上生成的前 50 位和

8000 位特征词做比对，结果如表 2-2、2-3 所示。表 2-2、2-3 中的实验结果表明点互信息特征词选择算法确实具有倾向于选择稀有词汇作为特征词的特点。

表格 2-2 常用特征词选择算法产生的前 50 位特征词中含有常见词的情况（一）

	in df_top_50	in df_top_8000
卡方	7	49
ig	27	50
mi	0	0

表格 2-3 常用特征词选择算法产生的前 50 位特征词中含有常见词的情况（二）

	in df_top_50	in df_top_8000
卡方	10	47
ig	12	50
mi	0	0

本文认为，点互信息特征词选择算法的效果较之其他特征词选择算法要差很多的原因在于：点互信息度量公式根本不适合应用于文本分类中的特征词选择问题中。阅读 2.4.2 节中关于常用特征词选择算法相关理论的介绍，不难发现卡方和信息增益这两种特征词选择算法在问题建模时针对的是词汇出现情况和类别出现情况的两个事件集合，所以测度更鲁棒；而点互信息特征词选择算法在问题建模时仅仅是针对“词汇出现”和“类别出现”两个事件，所以测度不够鲁棒。其实很多和点互信息在问题建模上有相同特点，且在其他场合又非常有用的度量也不适于直接用于特征词选择算法中，比如后验概率 *p(c|t)*，还有按照公式 2-35 进行问题建模的 Kullback-Leibler Distance。总之，数学公式和测度本身是无罪的，关键是看解决问题的工程师从何种视角看待问题，如何对问题进行数学建模。

$$KL(t,C)=\sum_{i=1}^{k}p(e_{c_i}=1\,|\,e_t=1)\log\frac{p(e_{c_i}=1\,|\,e_t=1)}{p(e_{c_i}=1)} \qquad \text{公式 2-35}$$

2.5 基于卡方拟合优度(chifit)的特征词选择算法

2.5.1 相关工作

在总体的分布函数完全未知或只知其形式但不知其参数的情况下，为了推断总体的某些未知特性，提出关于总体的假设，然后根据样本对所提出的假设作出是接受还是拒绝的决策，**假设检验**是作出这一决策的过程[37]。在假设检验问题中，有两个互补的假设，分别被叫做**原假设**（null hypothesis）和**备择假设**（alternative hypothesis），用H_0和H_1表示。假设检验的过程是决定哪些样本值使得H_0被接受，哪些样本值使得H_0被拒绝，H_1被接受，即确定**接受域**和**拒绝域**的过程。在假设检验过程中，拒绝域通常表示为检验统计量的不等式形式，**检验统计量**是随机样本的函数。假设检验的有效性用**显著性水平**来衡量。假定α为某假设检验的显著性水平，则表示该假设检验在应该接受H_0时犯下拒绝接受H_0的第一类错误的概率不大于α[38]。

皮尔逊卡方检验（pearson 卡方 test）是假设检验中的重要理论，主要应用在检测两个随机变量是否独立（如卡方特征词选择算法）和检验分布的拟合当中。检验分布的拟合是评价抽样样本拟合后的分布与某个理论上的分布之间的差异程度。

使用卡方检验法对总体分布进行检验时，首先提出原假设：
H_0：总体 X 的分布函数为 $F(X)$
然后根据样本的经验分布与所假设的理论分布之间的吻合程度来决定是否接受原假设。这种假设通常称作**拟合优度检验(test of goodness of fit)**，它是一种非参数检验。分布拟合的卡方检验法的基本原理和步骤如下：
①将总体 X 的取值范围分成 k 个互不重叠的小区间，记作A_1，A_2，…，A_k.
②把落入第 i 个区间A_i的样本值的个数记作f_i，称为实测频数。所有实测频数之和$f_1+f_2+\cdots+f_k$等于样本容量 n.
③根据假设的理论分布，可以算出总体 X 的值落入每个A_i的概率p_i，于是 np_i就是落入A_i的样本值的理论频数.

皮尔逊引入如公式 2-36 所示的统计量表示经验分布与理论分布之间的差异，并证明在理论分布 $F(X)$完全给定的情况下，每个p_i都是确定的常数。由棣莫佛—

拉普拉斯中心极限定理，当 n 充分大时，实测频数f_i的分布渐近正态分布，因此公式2-36是k个近似正态分布的变量的平方和，且满足公式2-37中的约束条件，因此是渐近 k-1 个自由度的卡方分布[156]。

$$\chi^2=\sum_{i=1}^{k}\frac{\left(f_i-np_i\right)^2}{np_i} \qquad \text{公式 2-36}$$

$$\sum_{i=1}^{k}\frac{\sqrt{p_i}\left(f_i-np_i\right)}{\sqrt{np_i}}=0 \qquad \text{公式 2-37}$$

根据皮尔逊卡方定理，对给定的显著性水平α，查卡方分布表获得临界值χ_a^2，得到拒绝域范围为$\left\{\chi^2>\chi_a^2\left(k-1\right)\right\}$，如果根据样本算得的卡方统计量的实测值在拒绝域范围内，则拒绝原假设，否则则接受原假设。

2.5.2 问题建模与推导

正如卡方特征词选择算法是皮尔逊卡方检验在检测事件独立性方面的应用特例一样，本文所设计的 chifit 特征词选择算法实际上是皮尔逊卡方检验在拟合优度检验方面的应用特例。针对特征词选择问题，我们做如下问题建模：

H_0：t 与 c 统计独立，形式化的表示为$p(e_c|e_t=0)=\ p(e_c|e_t=1)$

H_1：t 与 c 并非统计独立，即$p(e_c|e_t=0)$与$p(e_c|e_t=1)$之间存在较大的差异。

设 t 不出现时类别 c 的分布为总体的理论分布；t 出现时类别 c 的分布为样本的经验分布。根据$e_c=0$ 或者$e_c=1$可以把样本空间分为两个部分。对应的实测频数分别为N_{10}和N_{11}，样本总数为$N_{10}+N_{11}$。对应的理论分布概率分别如公式 2-38、公式 2-39 所示，最后得到卡方检验统计量的 MLE 形式，如公 2-40 所示。

$$p(e_c=1|e_t=0)=N_{01}/\left(N_{01}+N_{00}\right) \qquad \text{公式 2-38}$$

$$p\left(e_c=0|e_t=0\right)=N_{00}/\left(N_{00}+N_{01}\right) \qquad \text{公式 2-39}$$

$$\chi^2(D,t,c)=\frac{\left(N_{11}-(N_{11}+N_{10})\frac{N_{01}}{N_{01}+N_{00}}\right)^2}{(N_{11}+N_{10})\frac{N_{01}}{N_{01}+N_{00}}}+\frac{\left(N_{10}-(N_{11}+N_{10})\frac{N_{00}}{N_{01}+N_{00}}\right)^2}{(N_{11}+N_{10})\frac{N_{00}}{N_{01}+N_{00}}} \quad \text{公式}$$

$$=\frac{(N_{11}*N_{00}-N_{10}*N_{01})^2}{(N_{11}+N_{10})*N_{00}*N_{01}}$$

2-40

从公式2-40可以看出，本文设计的chifit特征词选择算法依然是对N_{11}、N_{10}、N_{01}、N_{00}进行加权计算，其中N_{00}表示既不包含词汇 t 又不属于类别 c 的文档数目。本文在文献调研时，发现一些有关特征词选择的论文认为N_{00}是无用的，因此直接将其从卡方特征词选择算法或者信息增益特征词选择算法中剔除。尽管这么做可能会在某些语料集上产生效果，但本文认为这种做法过于启发式，不具有合理性。它一方面破坏了原有特征词选择算法的理论基础，另一方面没有考虑到在训练数据和测试数据满足独立同分布时，N_{00}其实也是一个可以说明一定问题的重要指标。

仔细观察公式 2-40 可以发现，该公式的分母可以为 0，因此实际的 chifit 特征词选择算法计算公式是在公式 2-40 的基础之上对边界条件情况下做了修正的结果。在给出 chifit 特征词选择算法公式之前，我们先对边界条件的语义作如下理解和分析：

①N_{11}+N_{10}表示含有词汇 t 的训练文档集合文档数目。特征词选择算法的处理对象是训练文档集合生成的词汇，因此N_{11}+N_{10}不可能为 0。

②N_{00}=0 的含义是训练文档集合中的文章要么包含词汇 t，要么属于类别 c，暗含的语义是词汇 t 是高频词或者类别 c 是大类。因此根据 2.4.1 小节关于特征词选择算法的特点论述，此时公式的值应该被放大。

③N_{01}=0 的含义是训练文档集合中凡是属于类别 c 的文章都含有词汇 t，这说明词汇 t 对于类别 c 具有显著性，同上原因此时公式的值应该被放大。

④N_{00}=0 且N_{01}=0 表示训练文档集合中的文章要么包含词汇 t，要么属于类别 c，并且凡是属于类别 c 的文章都含有词汇 t。更近一步说，此时训练集合中的所有文档都含有词汇 t，词汇 t 对于类别 c 是否具有显著性，应该看N_{11}与N_{10}之间的对比度。假设N_{00}与N_{01}为同阶无穷小，则当 $N_{00}\to 0$且$N_{01}\to 0$时 公式 2-40

的极限值如公式 2-41 所示，考虑在此边界条件下应该将公式的值进行放大，所以去掉分母 N，得到最终的 chifit 计算公式如公式 2-42 所示。

$$\lim_{N_{00}\to 0,N_{01}\to 0}\frac{(N_{11}*N_{00}-N_{10}*N_{01})^2}{(N_{11}+N_{10})*N_{00}*N_{01}}$$
$$=\lim_{N_{00}\to 0,N_{01}\to 0}\frac{(N_{11}-N_{10}\frac{N_{01}}{N_{00}})\left(N_{11}\frac{N_{00}}{N_{01}}-N_{10}\right)}{\left(N_{11}+N_{10}\right)} \qquad \text{公式 2-41}$$
$$=\frac{\left(N_{11}-N_{10}\right)^2}{\left(N_{11}+N_{10}\right)}=\frac{\left(N_{11}-N_{10}\right)^2}{N}$$

$$chifit_{value}=\begin{cases}\infty,\text{if }\left(N_{00}=0\ or\ N_{01}=0\right)\text{and not }\left(N_{00}=0\ and\ N_{01}=0\right)\\ \left(N_{11}-N_{10}\right)^2,\text{if }\left(N_{00}=0\text{ and }N_{01}=0\right)\\ \left(N_{11}N_{00}-N_{10}N_{01}\right)^2/\left(\left(N_{11}+N_{10}\right)N_{00}N_{01}\right),\text{otherwise}\end{cases} \qquad \text{公式 2-42}$$

值得指出的：如果在问题建模的时候，假设 t 不出现时类别 c 的分布为样本的经验分布；t 出现时类别 c 的分布为总体的理论分布，即把原来的假设倒过来，则可以得到如公式 2-43 所示的结果。尽管此公式和公式 2-40 在形式上相近，但是此时$N_{00}+N_{01}$的结果可以等于 0，表示训练文档集中的所有文档都含有词汇 t，且N_{11}和N_{10}不能同时为 0，情况比较复杂，也不适合对边界条件的取值进行进一步建模，所以本文不做这样的假设。

$$\chi^2=\frac{\left(N_{00}N_{11}-N_{01}N_{10}\right)^2}{N_{10}N_{11}\left(N_{00}+N_{01}\right)} \qquad \text{公式 2-43}$$

2.5.3 算法实验与分析

本文对 chifit 特征词选择算法和 2.4.2 节介绍的常用特征词选择算法，在 Reuters 和中文新闻分类语料库上，分别用多项式贝叶斯以及 KNN 分类算器用 RI 指标进行效果验证（数据见附录 21），得到的结论是 chifit 特征词选择算法和卡方、信息增益等特征词选择算法效果相当，要好于点互信息特征词选择算法，如图 2-16 至图 2-19 所示。

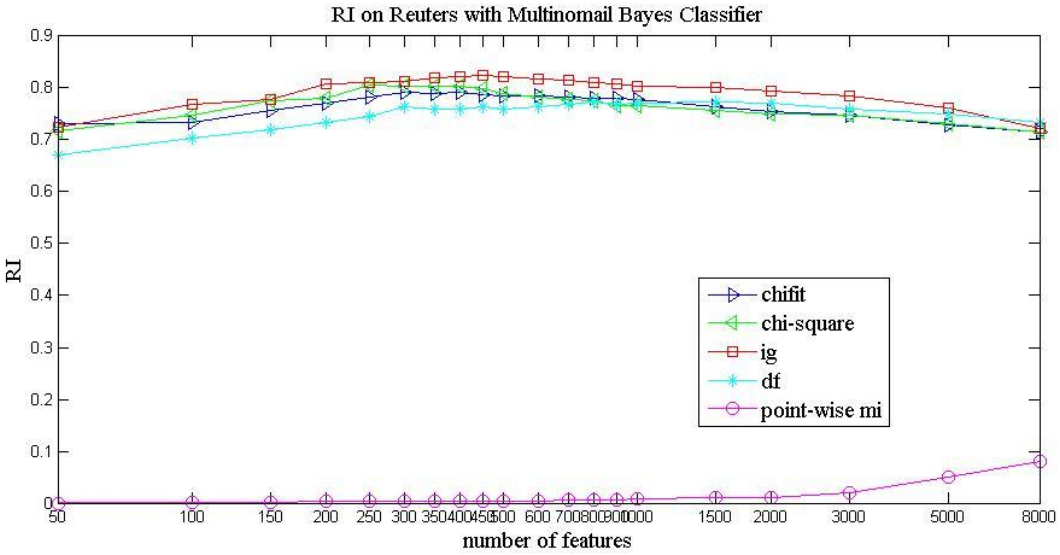


图 2-16 几种特征词选择算法在 Reuters 语料多项式贝叶斯分类器下的效果

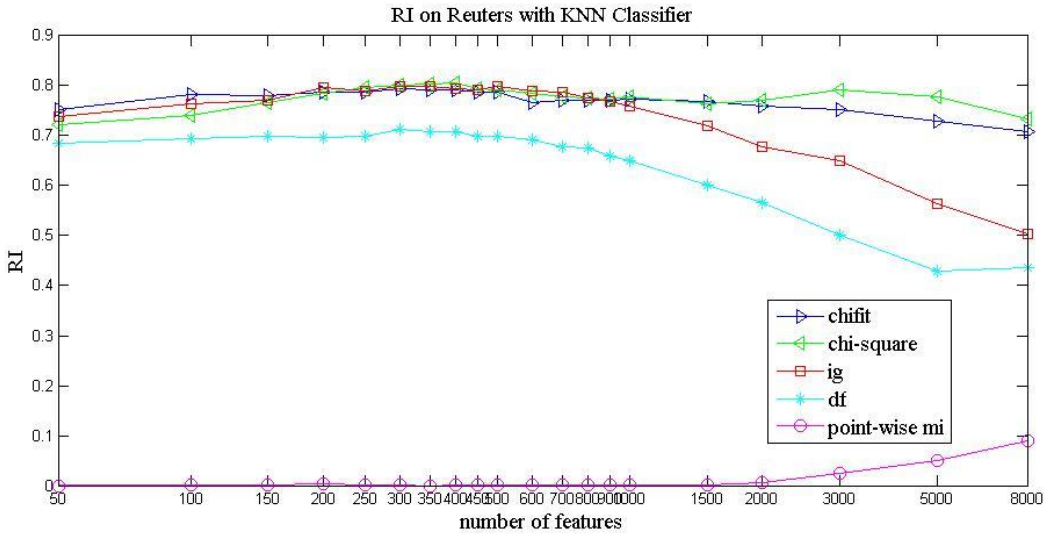


图 2-17 几种特征词选择算法在 Reuters 语料 KNN 分类器下的效果

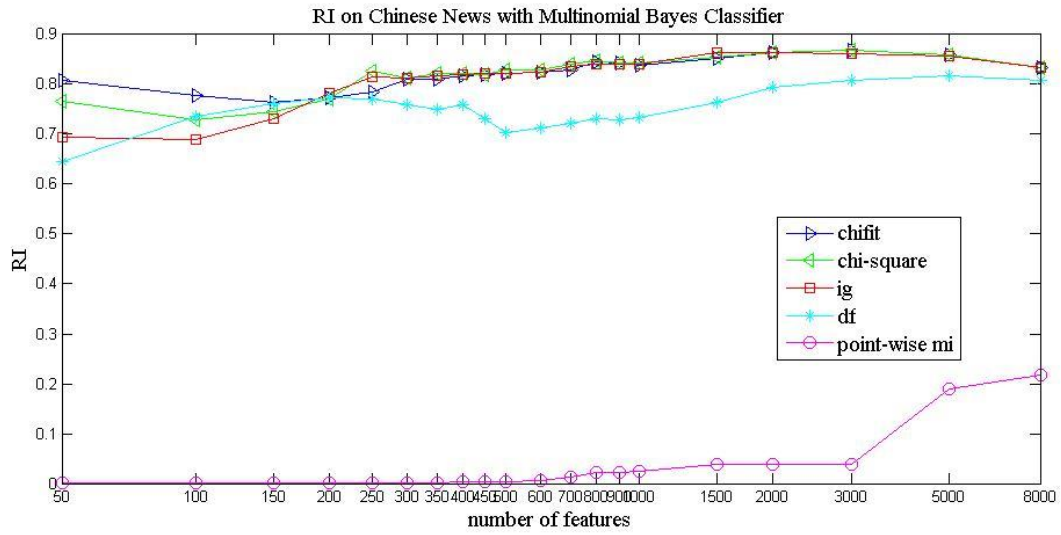


图 2-18 几种特征词选择算法在中文新闻语料多项式贝叶斯分类器下的效果

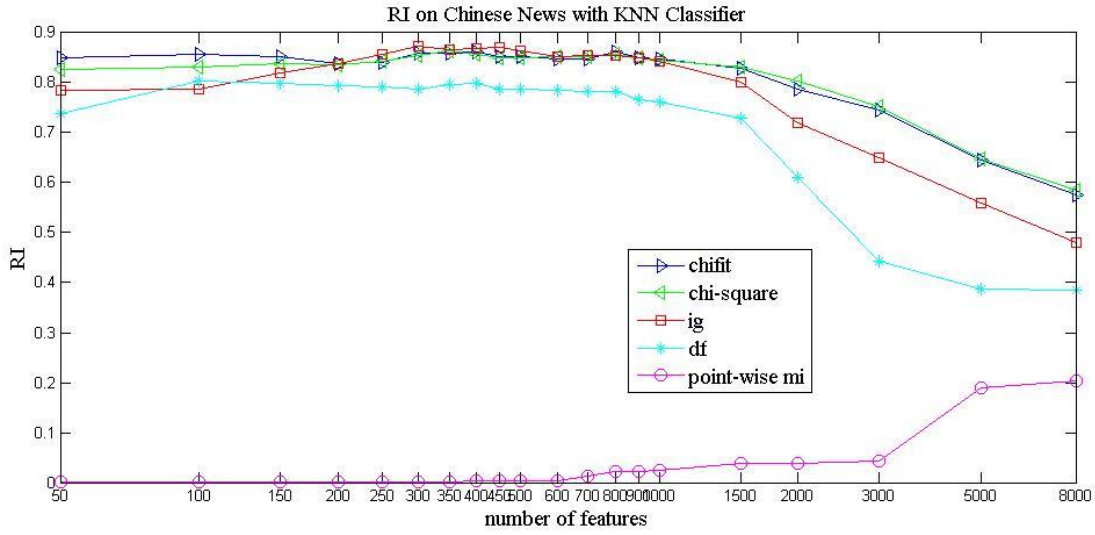


图 2-19 几种特征词选择算法在中文新闻语料 KNN 分类器下的效果

由于 chifit 特征词选择算法和卡方特征词选择算法的理论基础都是皮尔逊卡方检验，因此本文又对 chifit 特征词选择算法和卡方特征词选择算法做了详细的对比实验。

①首先两种算法在不同维度上标引测试集样本的能力上有所区别。如图 2-20、2-21 所示，chifit 特征词选择算法选取的特征词标引测试样本集的能力要好于卡方特征词选择算法选取的特征词的标引能力。

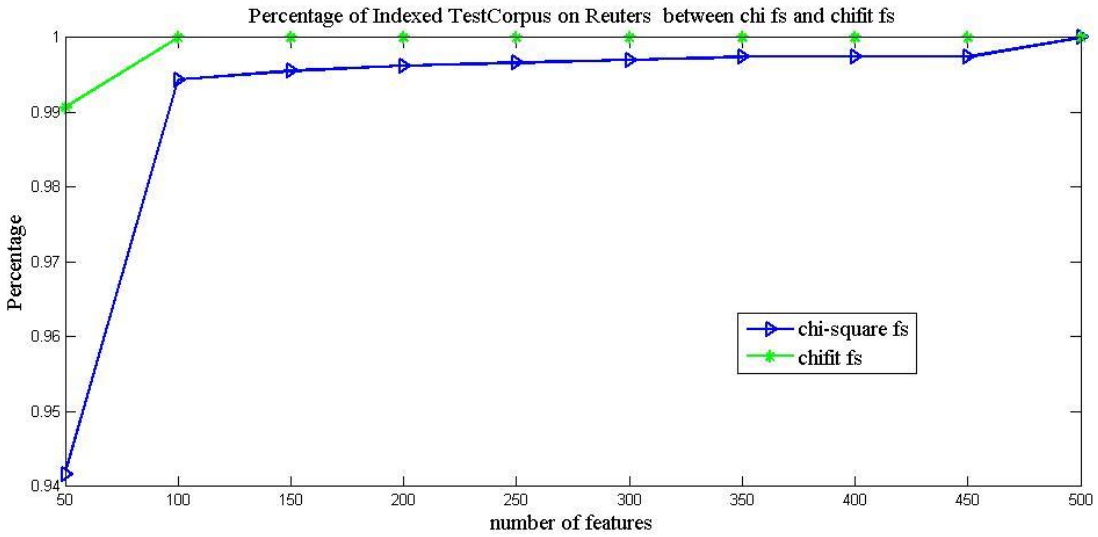


图 2-20 chifit 和卡方在 Reuters 上标引测试集样本的能力

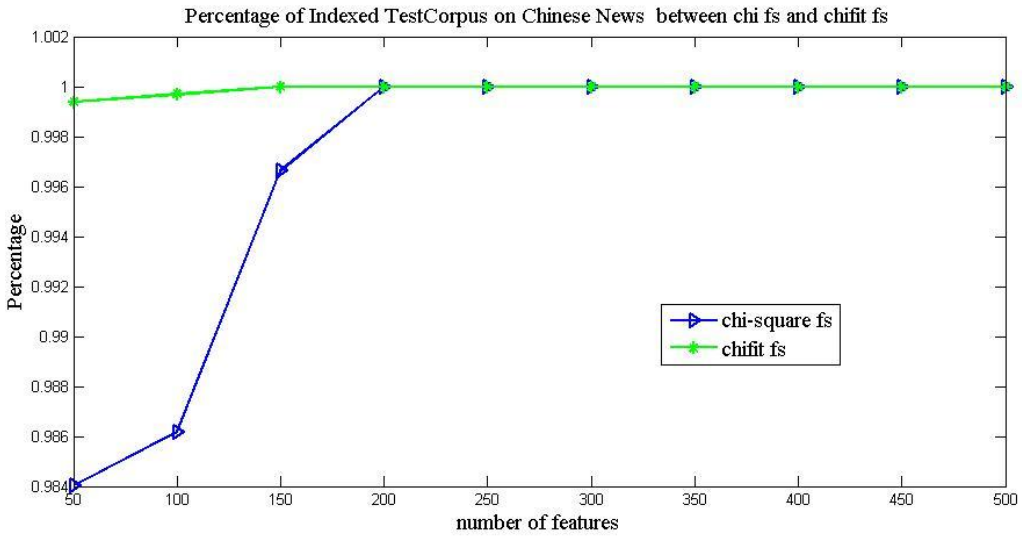


图 2-21 chifit 和卡方在中文新闻上标引测试集样本的能力

②两个集合的 *A,B* 的对称差（symmetric difference）的定义如公式 2-44 所示

$$A \oplus B = (A - B) \cup (B - A)$$
$$\text{其中} A - B = A \setminus A \cap B,\ \ B - A = \text{B\A} \cap \text{B} \qquad \text{公式 2-44}$$

借助于集合的对称差定义，我们可以定义卡方和 chifit 特征词选择算法在同一特征维度的差异度，如公式 2-45 所示，其中 *size_n* 表示特征维度

$$diversity = (featureset_{chi-square} \oplus featureset_{chifit}) / size\_n \quad \text{公式 2-45}$$

图 2-22 显示了卡方和 chifit 特征词选择算法在 Reuters 和中文新闻语料库上不同维度上生成的特征词集合的差异程度（具体数据将参见附录 22），仔细观察不难发现，随着特征维度的升高，卡方和 chifit 特征词选择算法的差异程度也就越来越低，其效果也就越来越来越相似。

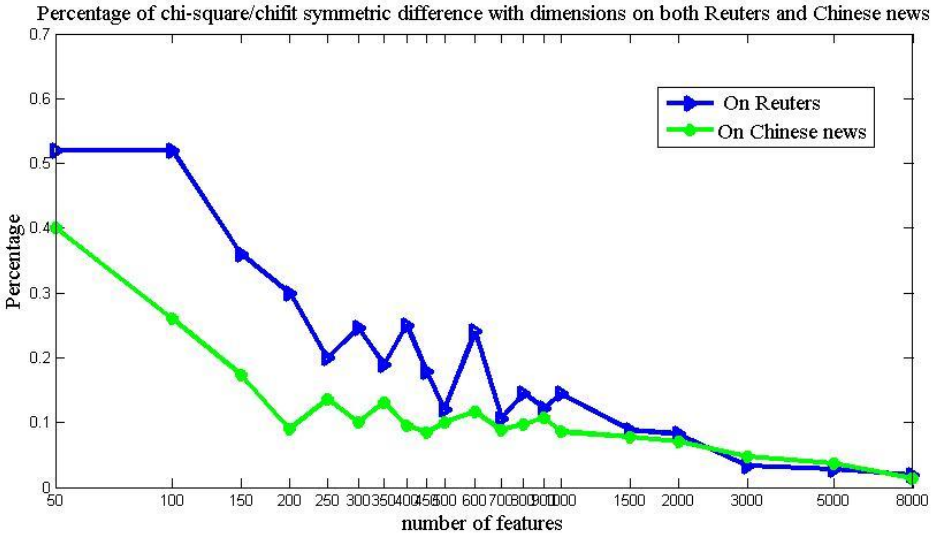


图 2-22 卡方和 chifit 特征词选择算法在中英语料上的差异度曲线

③从图 2-22 我们发现在低特征维度时，卡方和 chifit 两种特征词选择算法存在较大差异度，因此本文在 Reuters 语料和中文新闻语料上分别采用 KNN、朴素贝叶斯、多项式贝叶斯、SMO、决策树五种分类器，对卡方和 chifit 特征词选择算法生成的 50 到 500 特征的 VSM 模型进行测试实验，并采用 RI 作为评测指标，结果如图 2-23 至图 2-32 所示。图 2-23 至图 2-32 分别为在 Reuters 上和中文新闻语料上依次采用朴素贝叶斯、多项式贝叶斯、kNN，决策树，SMO 分类算法的实验结果。从图中不难看出，尽管两种特征词选择算法生成的 VSM 模型在不同分类器上的分类效果有所区别，但是当特征维度较低时，chifit 特征词选择算法生成的 VSM 模型无论在那种分类器上，那个语料集上进行测试，其效果都优于卡方特征词选择算法生成的 VSM 模型的分类效果。

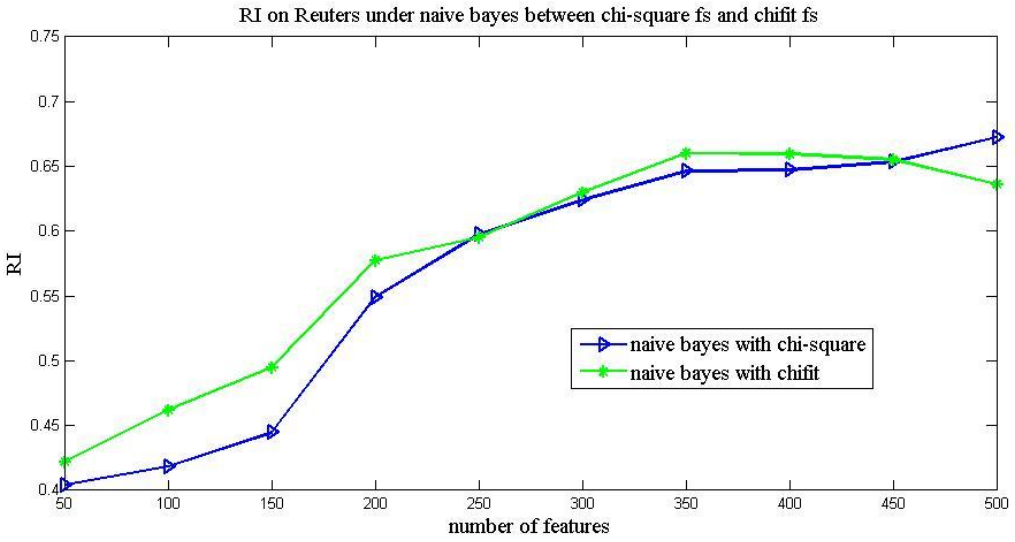


图 2-23 卡方与 chifit 算法在 Reuters 语料上的实验结果（一）

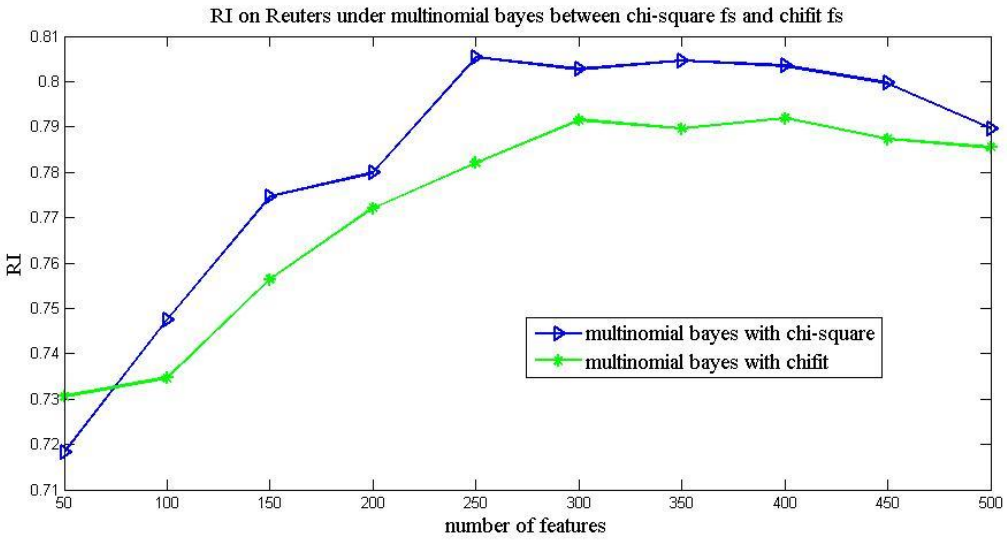


图 2-24 卡方与 chifit 算法在 Reuters 语料上的实验结果（二）

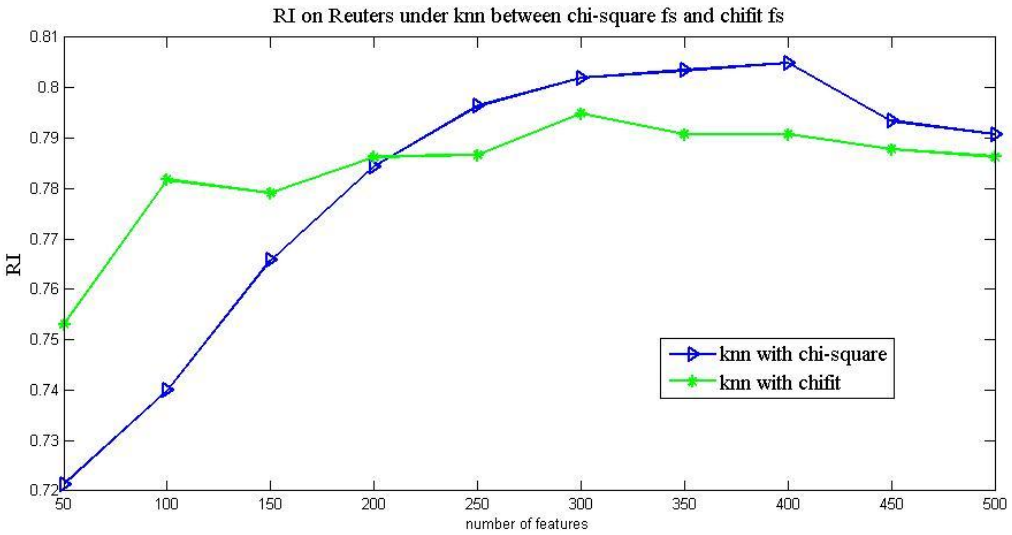


图 2-25 卡方与 chifit 算法在 Reuters 语料上的实验结果（三）

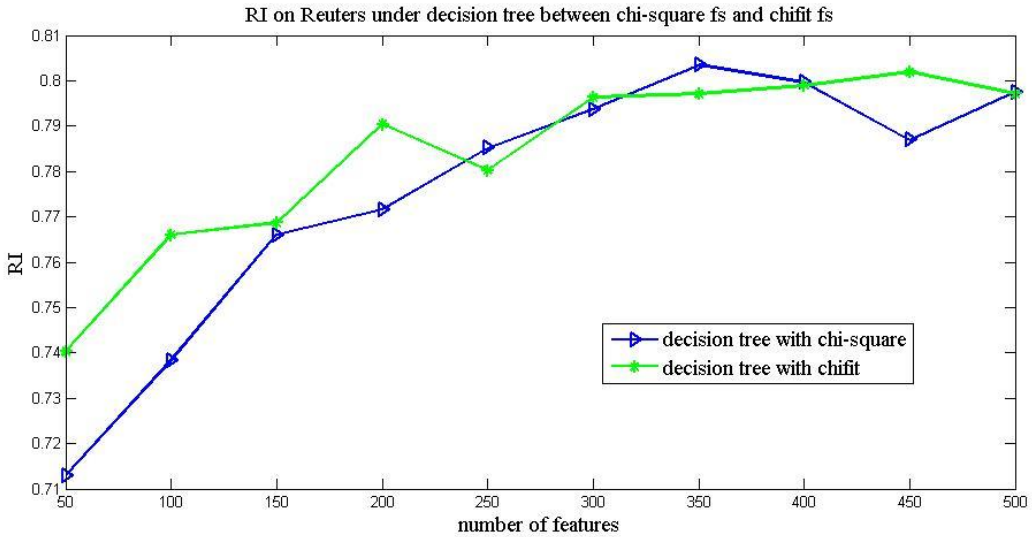


图 2-26 卡方与 chifit 算法在 Reuters 语料上的实验结果（四）

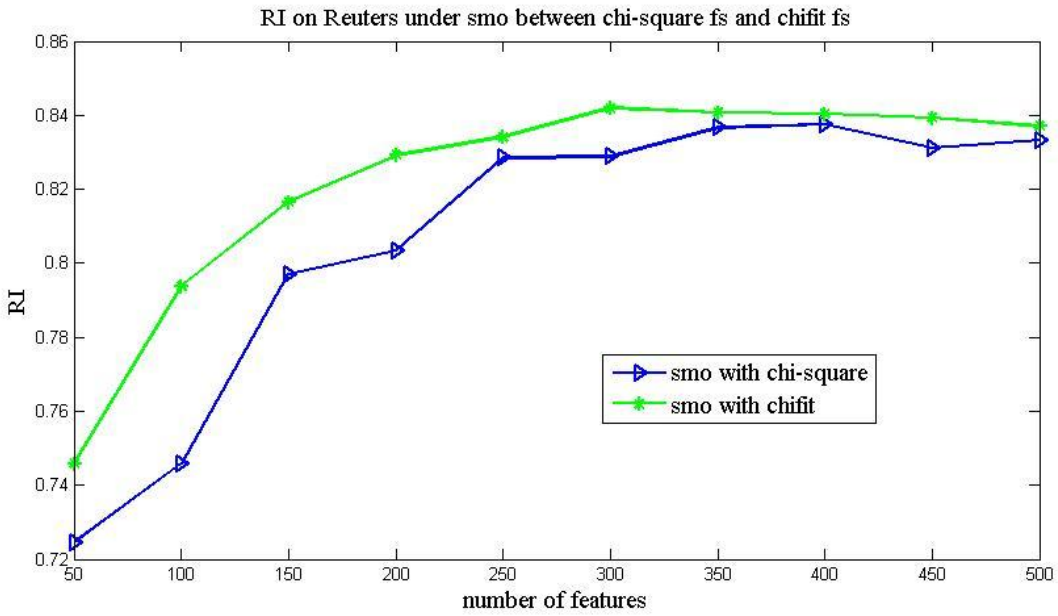


图 2-27 卡方与 chifit 算法在 Reuters 语料上的实验结果（五）

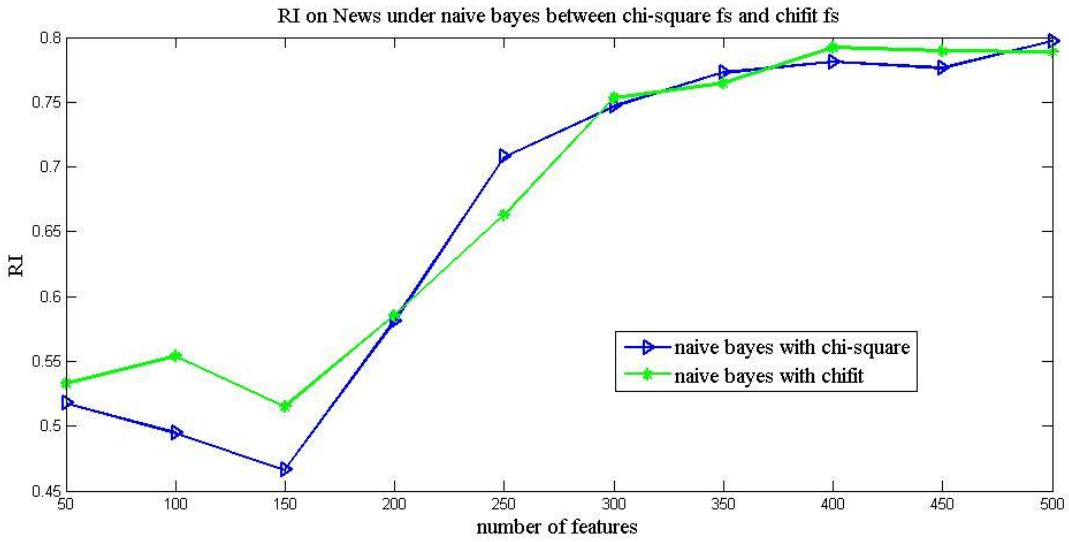


图 2-28 卡方与 chifit 算法在中文新闻分类语料上的实验结果（一）

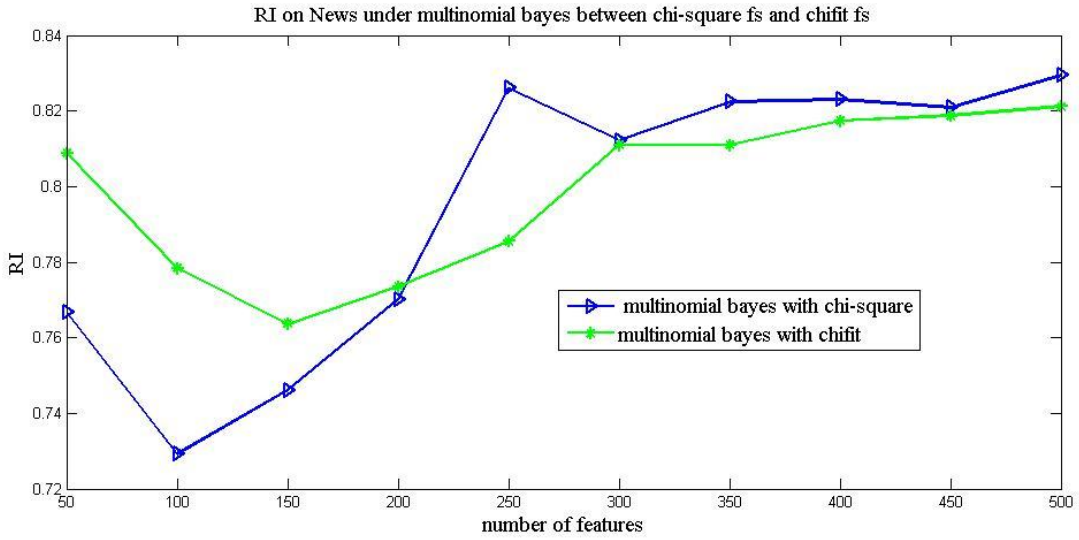


图 2-29 卡方与 chifit 算法在中文新闻分类语料上的实验结果（二）

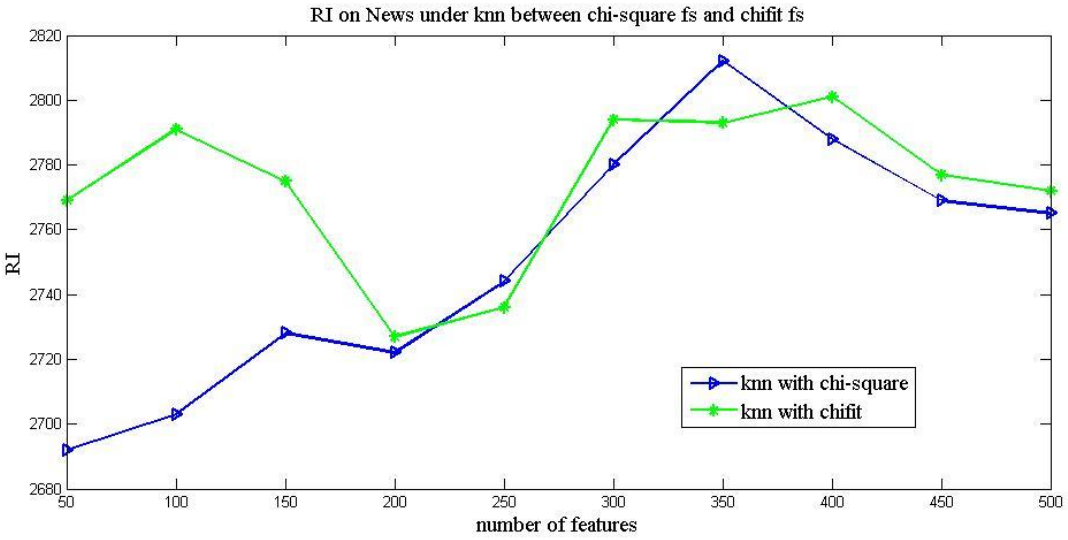


图 2-30 卡方与 chifit 算法在中文新闻分类语料上的实验结果（三）

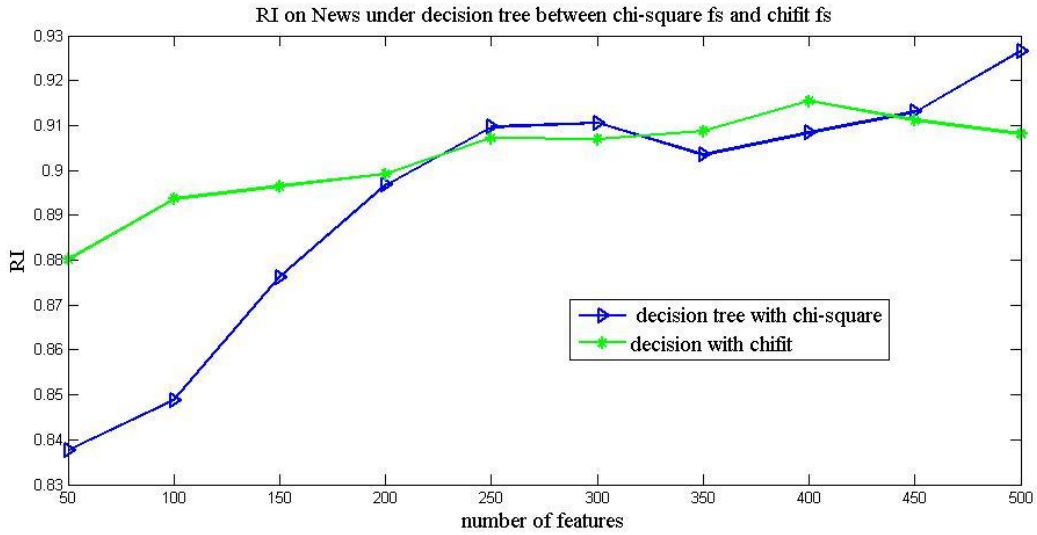


图 2-31 卡方与 chifit 算法在中文新闻分类语料上的实验结果（四）

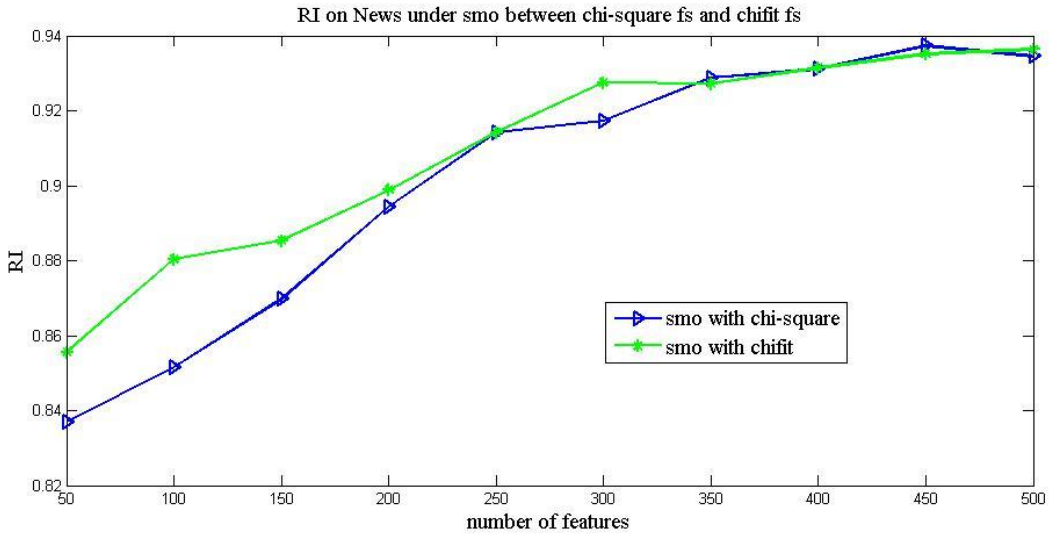


图 2-32 卡方与 chifit 算法在中文新闻分类语料上的实验结果（五）

④表 2-4 给出了两种算法针对 5 种分类器，在两种语料，特征维度[50,500]上的最佳分类效果及其出现的维度。观察表 2-4 可以看出 chifit 特征词选择算法在 Reuters 语料上，使用 SMO 分类器评测时的最优效果略高于卡方特征词选择算法，其他情况下都略微低于卡方特征词选择算法的最优效果，但是两者的效果差距并不是很大。

表格 2-4 卡方与 chifit 特征词选择算法在多个语料、多个分类器上的最优效果与出现维度

		朴素贝叶斯	多项式贝叶斯	KNN	决策树	SMO
Reuters	**卡方**	(0.6723,500)	(0.8053,250)	(0.8049,400)	(0.8034,350)	(0.8374,400)
	chifit	(0.6599,350)	(0.7919,400)	(0.7948,300)	(0.8019,450)	(0.8419,300)
中文新闻	**卡方**	(0.7972,500)	(0.8294,500)	(0.8642,350)	(0.9266,500)	(0.9373,450)
	chifit	(0.7923,400)	(0.8211,500)	(0.8608,400)	(0.9155,400)	(0.9364,500)

⑤表格 2-5 给出了卡方与 chifit 特征词选择算法在 50 维度特征下、在多个语料、多个分类器上的效果，观察表格 2-5 不难发现，在选取 50 维度特征词时，无论在哪种语料、采用哪种分类器做验证实验，chifit 算法的效果都要优于卡方算法。

表格 2-5 卡方与 chifit 特征词选择算法在 50 维度特征下、在多个语料、多个分类器上的效果

		朴素贝叶斯	多项式贝叶斯	KNN	决策树	SMO
Reuters	**卡方**	0.4040	0.7182	0.7212	0.7130	0.7246
	chifit	0.4219	0.7306	0.7530	0.7403	0.7459
中文新闻	**卡方**	0.5175	0.7667	0.8273	0.8377	0.8368
	chifit	0.5329	0.8089	0.8510	0.8801	0.8556

$$set(chifit-chi) = featureset_{chifit} - featureset_{chi} = featureset_{chifit} \setminus \left(featureset_{chi} \cap featureset_{chifit} \right) \quad \text{公式 2-46}$$

$$set(chi-chifit) = featureset_{chi} - featureset_{chifit} = featureset_{chi} \setminus \left(featureset_{chi} \cap featureset_{chifit} \right) \quad \text{公式 2-47}$$

公式 2-46 和 2-47 中的 *set(chifit-chi)*和 *set(chi-chifit)*分别表示仅在 chifit 特征词选择算法生成的特征集合中存在的特征词和仅在卡方特征词选择算法生成的特征集合中存在的特征词。附录 23 分别给出了在 Reuters 和中文新闻语料下，选取 50 个特征词时 *set(chifit-chi)*和 *set(chi-chifit)*中的词汇、含有这些词汇的训练集合文档数目、这些词汇对应的最大贡献度类别、以及该类别中含有词汇的文档数目和类别的文档总数目。为了进一步探讨 chifit 和卡方特征词选择算法在分别选取 50 维特征词时，chifit 特征词选择算法的效果要好的原因，并参考 2.4.1

节中给出的特征词算法应具备的特点，本文定义如 2-48、2-49、2-50 所示的三个指标。这三个指标分别作用于 set(chifit-chi)和 set(chi-chifit)上。其中索引比重 $Ratio_{index}$ 体现的是词汇是在训练文档集合上是频繁出现还是稀疏出现；最大贡献类别浓度$Ratio_{contribution}$ 体现的是该词汇被选为特征词的有效性，显然浓度越高，越能表明该词汇在该类别中频繁出现，从类别局部来讲，就越能表明该词对该类别越具有标识性；索引集中度$Ratio_{concentration}$ 体现词汇在类别分布上是否均衡，显然集中度越大，从训练样本集合全局上来讲，越能表明该词汇越具有类别标识性。

i 索引比重：

$$Ratio_{index} = \frac{\#\left(\text{词汇所引文档的总数}\right)}{\#\left(\text{训练集合文档总数}\right)} \quad \text{公式 2-48}$$

ii 最大贡献类别浓度：

$$Ratio_{contribution} = \frac{\#\left(\text{最大贡献度类别中含有词汇的文档数目}\right)}{\#\left(\text{最大贡献度类别文档总数}\right)} \quad \text{公式 2-49}$$

iii 索引集中度：

$$Ratio_{concentration} = \frac{\#\left(\text{最大贡献度类别中含有词汇的文档数目}\right)}{\#\left(\text{词汇所引文档的总数}\right)} \quad \text{公式 2-50}$$

图 2-33、2-34 分别为两个语料上的$Ratio_{index}$ 对比直方图；图 2-35、2-36 分别为两个语料上的$Ratio_{contribution}$ 对比直方图；图 2-37、2-38 分别为两个语料上的$Ratio_{concentration}$ 对比直方图。观察图 2-33、2-34 可以得出结论：在两个语料集上 set(chifit-chi)中的词汇都要比 set(chi-chifit)中的词汇出现更频繁；观察图 2-35、2-36 可以得出结论：在两个语料集上 set(chifit-chi)中的词汇都要比 set(chi-chifit)中的词汇更具有局部类别标识性；观察图 2-37、2-38 可以得出结论：在两个语料集上 set(chi-chifit)中的词汇都要比 set(chifit-chi)中的词汇更具有全局类别标识性。尽管 set(chi-chifit)中的词汇都要比 set(chifit-chi)中的词汇更具有全局类别标识性，由于 set(chi-chifit)中的词汇相对稀疏，所以会出现特征维度低时，chifit 特征词选择算法的效果要好于卡方特征词选择算法的结果。

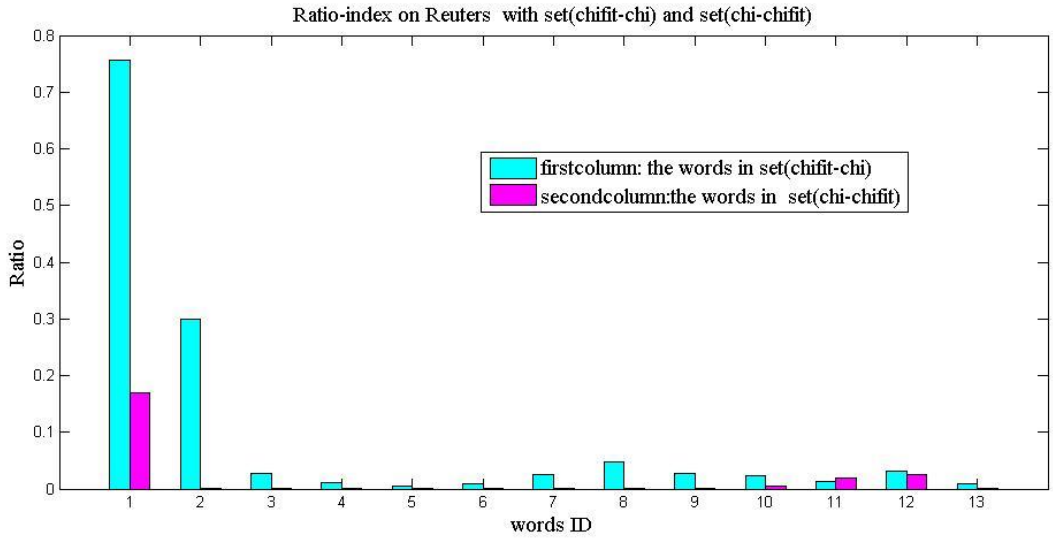


图 2-33 Reuters 语料上 set(chifit-chi)和 set(chi-chifit)索引比重

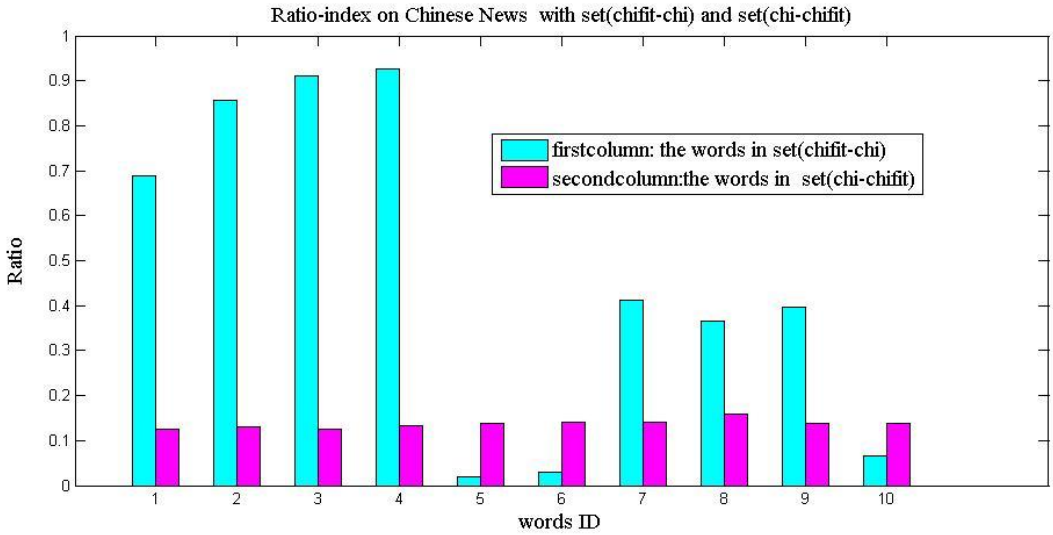


图 2-34 中文新闻语料上 set(chifit-chi)和 set(chi-chifit)的索引比重

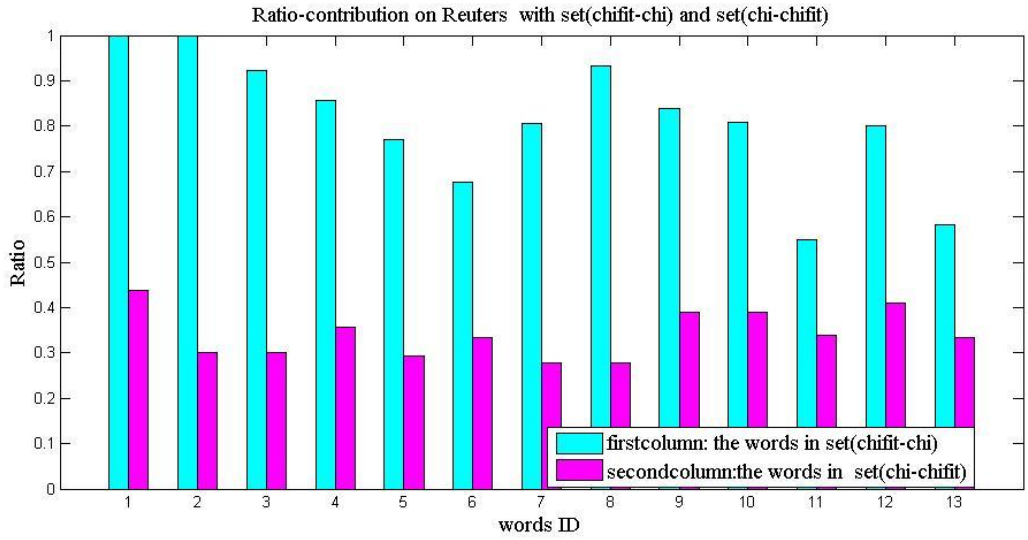


图 2-35 Reuters 语料上 set(chifit-chi)和 set(chi-chifit)的最大类别贡献度

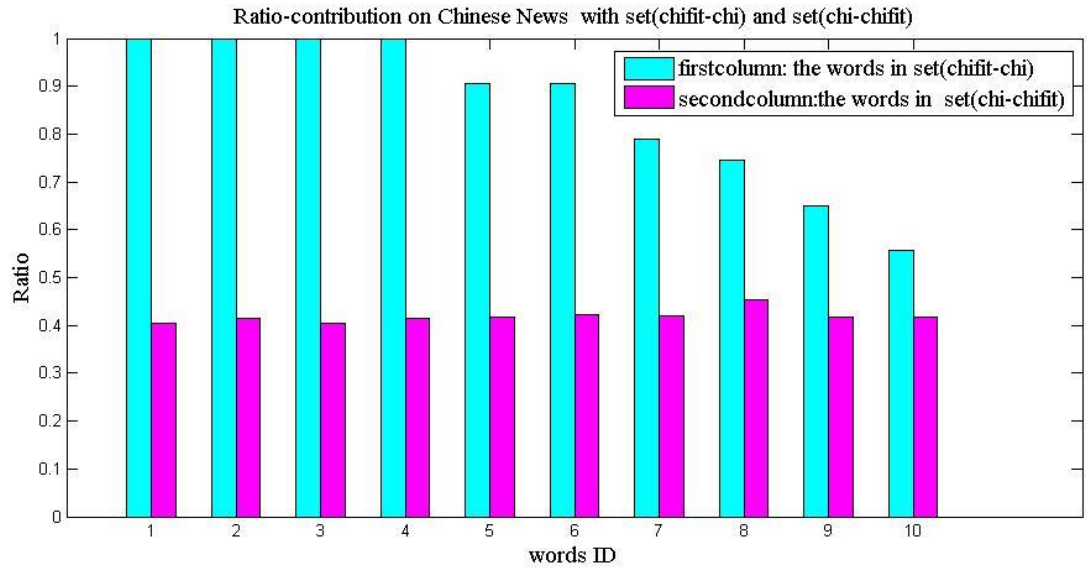


图 2-36 中文新闻语料上 set(chifit-chi)和 set(chi-chifit)的最大类别贡献度

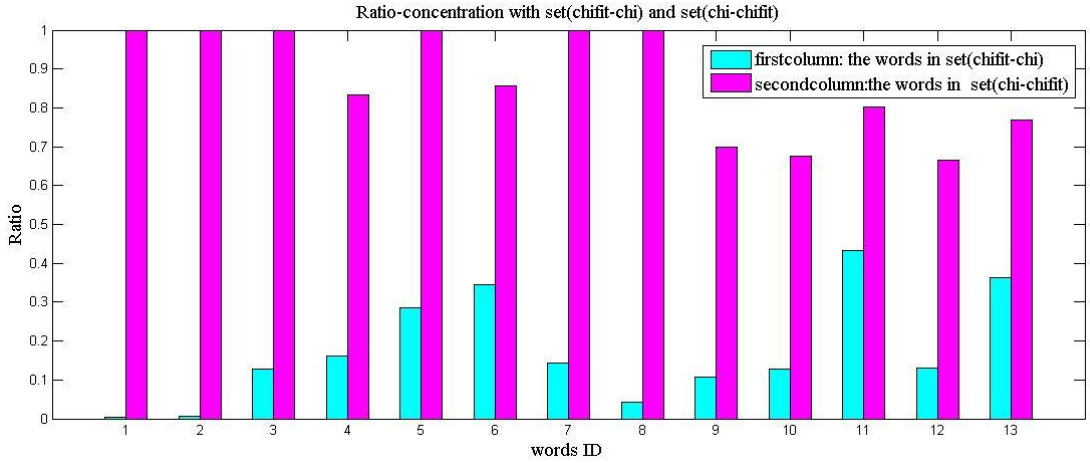


图 2-37 Reuters 语料上的 set(chifit-chi)和 set(chi-chifit)的索引集中度

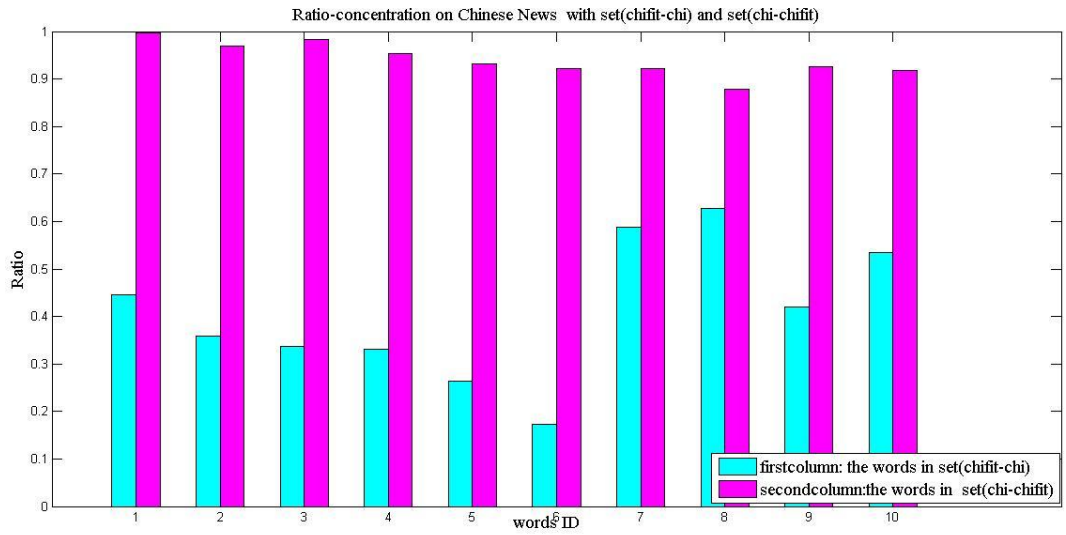


图 2-38 中文新闻语料上的 set(chifit-chi)和 set(chi-chifit)的索引集中度

2.6 针对课题任务的算法方案设计与实现

2.6.1 任务需求说明

（一）“自动化学科创新思想与方法研究”课题爬取了附录 1 中所示的期刊列表中的论文题录信息，这些期刊涉及计算机、自动化、通信、电子等交叉学科，因此需要将自动化类的论文题录信息从中剥离出来。考虑到来自万方数据源的 58,235 条题录信息（参见附录 14）中包含中图分类号，因此可以通过中图

分类号将自动化类文章和非自动化类文章直接区分开来；由于来自知网数据源的 116,642 条题录信息（参见附录 14）中不含有中图分类号，因此无法直接确定是否为自动化类文章，所以可以采用分类算法，以来自万方数据源的题录信息作为训练集，对来自知网数据源的题录信息的类别（是否属于自动化学科）进行预测。

（二）“自动化学科创新思想与方法研究”课题结合专家知识从数据源中抽取知识要素，确立自动化学科的知识点，使用杨一平研究员开发的概念管理中心软件平台（Concept Management Center,CMC）构建自动化学科知识体系。由于自动化学科类的关键词数量较多，仅靠人工肉眼进行筛选容易出现错误和疏漏，因此需要用机器学习的方法先从自动化学科类的关键词中选出“代表性”子集，之后交给编辑人员筛选定夺。

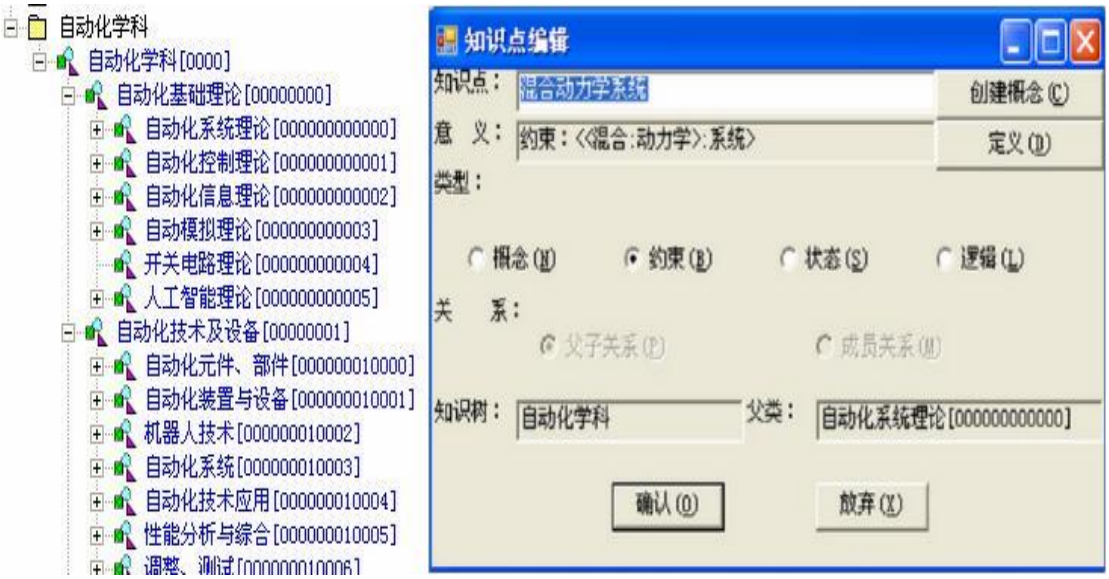


图 2-39 采用 CMC 平台构建自动化学科知识体系[13]

2.6.2 实验设计与分析

（一）考虑来自知网数据源的题录信息条目与来自万方数据源的题录信息条目数量之比近似 2:1，我们在模拟实验中将来自万方数据源的题录信息条目随机分成三份，其中一份作为训练集合，两份作为测试集合，训练样本集规模为

[13]此图来自刘贤达师兄的硕士毕设论文《基于语义的文本关联分析》

17,050，测试样本集规模为 34,102[14]。特征词选择算法采用本文提出的 chifit 特征词选择算法，分类器采用 KNN 和 libsvm[157]，在多种特征维度进行实验，分别得到的最高分类准确率为 41.2787%与 63.7646%。

由此可见，在现有数据条件下，无法将自动化类别的论文题录信息与其他类别的论文题录信息进行有效区分。所以“自动化学科知识服务网络平台”实际上是“信息学科知识服务网络平台”。我们也可以为统计分类算法的失效找到现实依据：首先，分类的文本是针对题录信息的摘要。期刊论文摘要具有一定的模板规范和字数限制，一般不超过 500 字，所以信息量有限。其次，学科融合与学科借鉴是学科发展的基本趋势之一，所以一些传统的自动化学科概念如“反馈”等也逐渐渗透到了通信、计算机等领域。

尽管数据清洗模块的工作以失败告终，但是它对于整个文本挖掘系统的后续模块在算法设计上有重要的借鉴意义，如本论文第五章中的同名作者消歧算法，就没有将题录摘要信息融入算法建模中。

（二）采用 chifit 特征词选择算法对万方数据源中属于自动化学科的关键词进行权重计算，按权值由高到低排列，交给编辑人员。

2.7 本章小结

本章论文首先对文本分类的基本概念和基本分类器原理进行介绍。然后以实践和验证的方式对分本分类流程、基本分类器性能进行分析和探讨，与传统课本中对文本分类的“黑盒模式”讲解方式有所不同，本文力求深究文本分类流程中的每个技术细节，并对流程中的重要数据结构进行透视分析。随后本文介绍了业内常用的特征词选择算法及其理论基础，并重点介绍了本文提出的 chifit 特征词选择算法，此部分内容包括 chifit 特征词选择算法的理论基础、问题建模与推导、以及分类实验验证。尤其重点比对分析了 chifit 特征词选择算法和卡方特征词选择算法的性能差异。经过一系列实验表明，chifit 特征词选择算法可以在较低的特征维度上获得较好的分类效果，因此适于用计算资源有限的应用场景中，具有一定的可实践性。

[14]以往课本、论文中的分类实验设置中一般会让训练样本集合的数目大于测试样本集合的数目，但这与实际应用情境不符。实际应用中往往需要用“有限”的训练样本去预测“无限”、“未知”的测试样本，为了更贴近实际，本文这里故意反其道而为之。

第三章 关键词语义聚类与知识族谱

本章节将按如下组织：3.1 节给出实际工程需求的描述；3.2 节对自动化领域汉语科技术语的特点进行总结，进而论证由形态相似度度量衡量语义相似度的可行性；3.3 节详细讲解基于编辑距离二次计算方法的相似度计算；3.4 节给出实验以及分析；3.5 节介绍知识族谱的构建方案；3.6 节为本章小结。

3.1 需求分析

本文从万方、知网论文题录信息网页中抽取了近 15 万余条自动化相关学科的汉英术语对照组（见附录 14），计划制作一部汉英对照互译电子词典。由于原始数据中含有大量冗余和噪声，即：很多词条语义上相同，但是词形上有所区别，所以无法直接形成优质的词典资源。因此需要将其中的汉英术语对照组按照语义是否相同，采用聚类等技术进行去重，形成初步可实用的汉英术语对照词典。

双语词典在计算机辅助写作、机器翻译、跨语言检索、查询扩展等领域有重要的应用价值。除此之外，领域内术语双语词典在学术检索、图书馆管理学、领域学科建设中有重大意义。一直以来双语词典的获取和编撰都要依靠大量手工工作，因此需要投入大量的资金和人力。为此本文提出了一种基于编辑距离二次计算的术语相似度计算方法，将该方法应用于对网络爬虫获得的双语术语素材的相似度计算和聚类问题中，进而完成词典的自动构建，减轻人工工作量。

3.2 数据分析

(1)Monte Carlo法;Monte Carlo方法;Monte-Carlo法;Monte-Carlo方法;Monte-carlo法
(2)Ad Hoc 网络;Ad Hoc网;Ad Hoc网络;Ad hoc 网络;Ad hoc网络;AdHoc网;AdHoc网络;Adhoc网;
(3)词汇相似度;词相似度;
(4)词序相似度;
(5)词义相似度;词语语义相似度;
(6)单纯形法;单纯形方法;
(7)虚拟专网;虚拟专用网络;
(8)李雅普诺夫稳定理论;李雅普诺夫稳定性理论;李亚普诺夫稳定理论;李亚普诺夫稳定性理论;
(9)语法分析
(10)语义分析
(11)2-d系统;2—d系统;2－d系统;

图 3-1 关键词数数据情况

图 3-1 给出了原始文献题录数据中关键词数据的一些特性。通过观察可以发现，很多关键词语义相同，但是形态上却有若干差别。经过人工对汉语关键词（或称为科技术语）进行分析，总结出如下几条比较稳定的规律：

⑴术语翻译的规则性强，具体表现为：短术语意译如“time delay”翻译成“时延”或“时滞”，或音译如“bayes”翻译成“贝叶斯”。长术语的翻译由短术语和普通汉语词汇构成，长术语的翻译一般要借助短术语的翻译，或者其中部分词汇不翻译如“BP neural network”翻译成“BP 神经网络”。

⑵语义上同指或相近的术语，形态上也相近。如“BP 神经网”和“bp 神经网络”，极少会出现如“小苏打”和“碳酸氢钠”、“勾股定理”和“毕达哥拉斯定理”这样术语语义相同，但是形态上差别很大的情况。

⑶将一个汉字、数字、英文字母、拉丁字母的长度视为单位长度，本文对从互联网上获取的约 15 万个自动化领域汉英术语对的长度分别进行统计分析，得到如下结果：汉语术语的平均长度为 5.10869，英语术语的平均长度为 19.6761。

考虑到自动化专业汉语术语在形式和语义上都较英语术语都更简洁、紧致和凝练。本文通过对汉英术语组中的汉语术语进行相似度计算来达到同义术语组聚类的目的。考虑到语义相似的汉语术语在形态上相近，我们将语义相似度

计算转换为形态相似度计算，从而降低问题的复杂度，并借助于融入形态约束规则的编辑距离二次计算方法计算形态相似度。

3.3 基于编辑距离二次计算的术语相似度计算方法

基于编辑距离二次计算的术语相似度计算方法的核心思想是：首先根据传统编辑距离算法计算出两个术语之间的初始编辑距离$ed^{initial}$以及最优编辑路径集合，然后将启发式规则集合作用于最优路径集合上得到最终距离ed^{final}，如图 3-2 所示。

如果计算得到的最终距离ed^{final}为 0，或者显著小于初始距离$ed^{initial}$，则说明两个术语同义，如算法 3-1 所示。其中阈值 *threshold1* 和 *threshold2* 可以根据实际情况自行设置。本论文中取 *threshold1=min(s.len,t.len)*， *threshold2*=0.5。

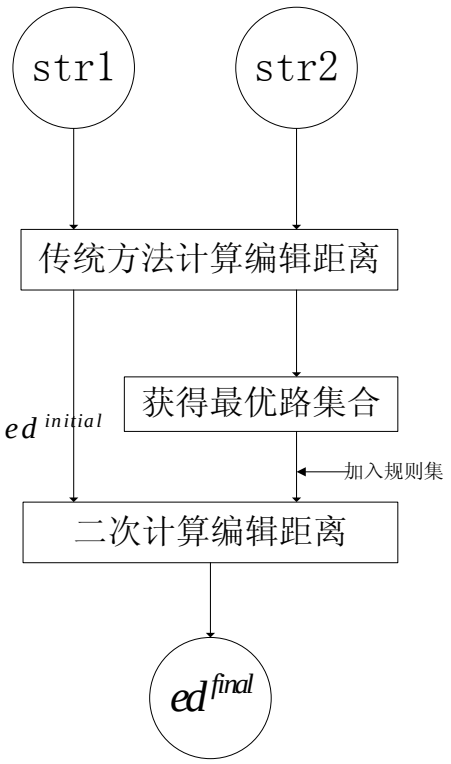


图 3-2 编辑距离二次计算框架

算法 3-1 术语相似度算法

```
Input:term1,term2
Output:是否相似
Begin:
OptimalPaths  optimals;
1.  传统方式计算编辑距离，并保存最优路径集
    edinitial=CalTraditionEditDistance(optimals);
2.  如果edintial >threshold1则返回不相似，如果ediintial=0则返回相似，否则进入3
3.  将规则集合作用于最优路径集上，第二次计算编辑距离
    edfinal=CalEditDistanceByRule(optimals);
4.  如果edfinal=0返回相似，否则进入5
5.  如果(edinitial-edfinal)/edinitial<threshold2则返回相似，否则不相似
```

3.3.1 相关工作

编辑距离又被称为 Levenshtein 距离[39]，是指一个字符串经过增加字符、删除字符、替换字符变成另一个字符串的代价，常被用来衡量两个字符串之间的相似程度。原始的编辑距离定义中只包含添加、删除、替换三种操作作为原子操作。在计算机应用中，编辑距离的概念又有进一步推广，如允许不同的编辑操作具有不同的权重[40]；加入其它操作作为原子操作[41][42]；根据应用需要在编辑距离计算过程中加入启发式规则[43]等。无论是改变操作权重、增加原子操作类型，还是计算过程中加入规则，都属于“事前方法”。这种工作模式无法有效利用全局信息，当规则集合过于庞大时，运算效率也会降低。因此我们提出了属于事后模式的编辑距离二次计算方法。该方法将规则集合作用于传统方法求得的最优编辑路径之上，从而避免了事前方法必须面临的以上问题。此外该方法可以解决多重错误，并能更好地反映人脑中的规则。我们知道在传统编辑距离框架中加入转置操作之后，得到“OT”与“TO”之间的距离为 1，而“OST”与“TO”之间的距离为 3 的情况[44]，这时计算机和人在理解规则的时候出现了“二义性”，采用编辑距离二次计算框架则可以避免以上问题出现。

3.3.2 编辑距离二次计算方法

3.3.2.1 计算最小编辑距离

设字母 i、d 、r、e 分别代表插入（insert）、删除(delete)、替换(replace)、空操作(empty)。insert、delete、replace 的操作权重均设为 1，空操作的权重设为 0。本文把 insert、delete、replace 称为带权操作，把 empty 定义为不带权操作。求

最小编辑距离，也就是求所有操作权重累加值最小。

本文在原有的编辑距离计算算法中加入存储结构，保存每个步骤上的操作，以及当前的权重累加值。如图 3-3 所示，其中要改变的字符串 s=“abc”，目标字符串 t=“ca”。为了便利算法的实现，给字符串分别加入了开始符号“#”和终止符号“$”。图中的箭头表示操作路径，方格里面的数字表示操作的总权重，上角标表示当前执行的操作。如 02 号方格的内容为2^{i}，它代表的意思是从空字符串得到字符串“ca”要经过两步带权操作，且最后一步操作为插入操作。

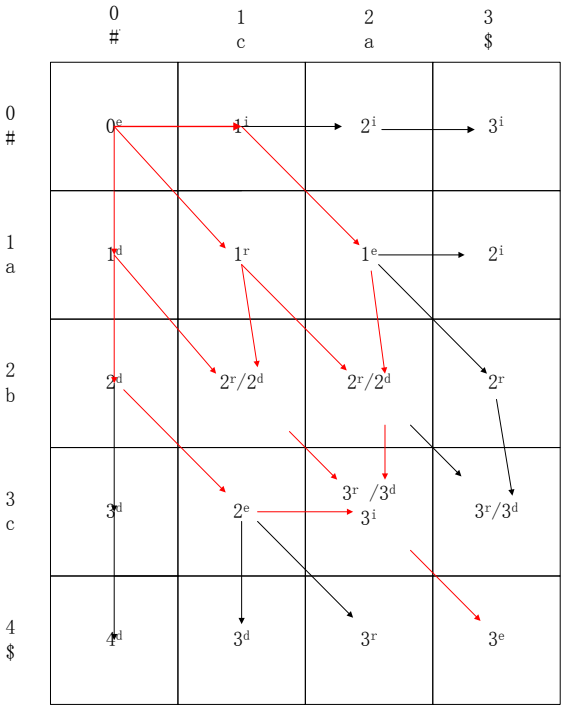


图 3-3 传统方法求最小编辑距离

3.3.2.2 获取最优编辑路径集合

将图 3-3 中标注成红颜色的最优路径信息用图的数据结构表示出来，如图 3-4 所示。该路径图是起点唯一、终点唯一、且没有回边，所以是有向无环图(DAG)。由此求两个字符串间的最小编辑距离的问题转换为求图上两定点之间的所有路径的问题。为解决该问题，本文设计了求图上两点之间所有路径的算法 3-2。该算法本质上是枚举的，因此有指数级复杂度，不适合应用在大规模稠密图上。但该算法不仅仅可以应用于有向无环图，也适用带环有向图，以及无向图。

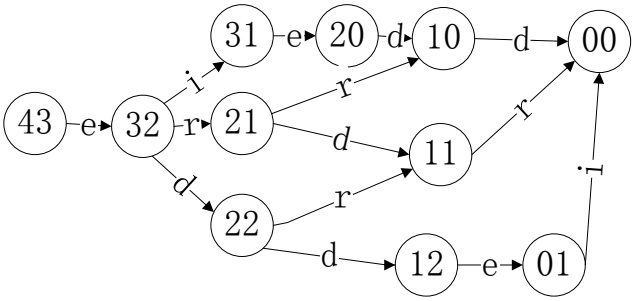


图 3-4 最优路径集合图[15]

算法 3-2 求两点之间所有路径的算法

```
ArcMatrix//保存图结构的邻接矩阵
VertexStatus[VertexNum]//结点访问标志数组，如果结点进栈，则对应的位置置1，否则为0
ArcStatus[VertexNum][VertexNum]//边访问标志数组，如果边至少有一个顶点在栈内，则置1，否则为0
void GetAllPaths(Paths&pathcollection, int beginVertex, int endVertex)
{
    Stack s;
    s.pushback(beginVertex);
    VertexStatus[beginVertex]=1
    while(!s.empty())
    {
        int topelem=s.top();
        if(topelem==endVertex)//找到了一条路径
        {
            Path p=TraverseStack(s);
            s.pop();
            vertexsStatus[topelem]=0;
            UpdateArcStatus(topelem); 更新ArcStatus[][]，使得所有两个端点都不在栈内的边的状态为0
        }
        else
        {
            i=0;
            for(;i<VertexNum;i++)//找到栈顶结点的第一个孩子结点
            {
                if(VertexStatus[i]=0&&ArcStatus[elem][i]=0&&ArcMatrix.contain(elem,i))
                {
                    VertexStatus[i]=1;
                    ArcStatus[elem][i]=1;
                    Mystack.push(i);//入栈
                    break;
                }
            }
            if(i==VertexNum)//该节点没有符合要求的后续结点
            {
                VertexStatus[elem]=0;
                UpdateArcStatus();//更新ArcStatus，使得所有两个端点都不在栈内的边的状态为0
                Mystack.pop();//出栈
                vertexsStatus[elem]=0;

            }
        }
    }
}
```

[15] 圆圈内的标号对应于图 3-3 中方格的标号

3.3.2.3 获取和更新最优编辑路径上的操作数

在获得了最优编辑路径集合以及路径上的操作之后，需要用迭代的方法获取路径上每个点的操作数，即操作前的状态和操作后的状态。假设当前位置对应的字符为 a，如果当前操作是“替换”，那么操作后的当前字符就是替换后的字符；如果当前操作是“插入”，那么操作后的当前字符则是被插入的字符；如果当前操作为“删除”,那么操作后的当前字符则为原当前字符的前一个字符。

3.3.2.4 用于二次计算的启发式规则集

基于数据分析，本文制定了如下规则集：

①初次计算编辑距离时，插入、删除、替换操作的权重均为 1。

②在编辑距离再计算中加入如下启发式规则：

i 插入/删除一个非汉字、非数字、非英文字母、非希腊字母的字符，其操作权重调整为 0；ii 两个非汉字、非数字、非英文字母、非希腊字母的字符替换，替换权重为零；iii 大小写英文字母替换，替换权重为 0； iv 将一个汉字替换成另一个汉字，如果这两个汉字读音相同，则替换权重为 0；v 如果插入或者删除一个汉字，且该汉字是一个不承担实际意义的前缀或后缀词素，则插入或者删除权重为 0。

这里对 v 中关于如何判定汉字是否是不承担实际意义的前缀或者后缀词素做一下特别说明：首先对源字符串和目的字符串进行分词[16]，若源字符串插入或者删除该汉字后的分词序列更趋近目的字符串的分词序列，则认为此汉字为附着词素。

例如源字符串为 s=“ADHOC 网”，目标字符串为 t=“Adhoc 网络”，采用传统编辑距离算法，计算得到的原始最优编辑距离权重为 6。图 3-5 为原始最优路径集合，此处用元组（operator,operandbefore,operandafter）表示路径中的每个节点的操作情况。将上述规则集合作用于图 3-5 中的最优路径集合上，得到最终的编辑权重为 0。

[16]笔者采用 ictlas50 作为分词工具，下载地址为 http://ictclas.org/Down_OpenSrc.asp

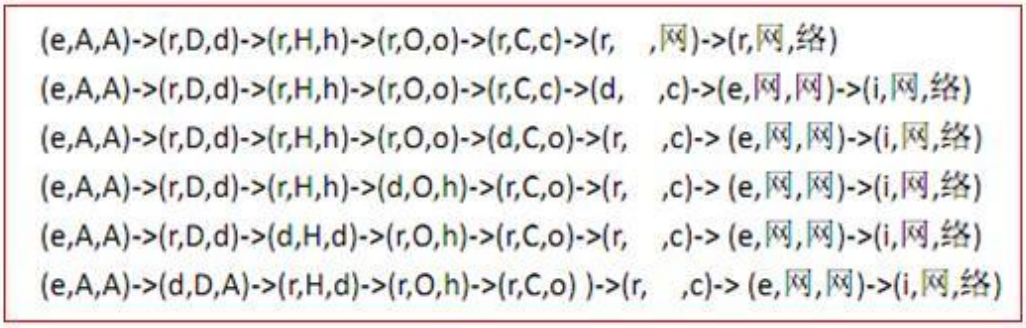


图 3-5 原始最优路径集合以及路径结点操作

3.4 算法实验与分析

3.4.1 评价指标

本文中借用分类性能评价指标正确率 P、召回率 R、F 值评价同义词汇聚类的效果：

把意义相同的词汇看做是一个组，则

$$P=NUM(\text{算法结果组} \cap \text{标准答案组})/NUM(\text{算法结果组}) \quad \text{公式 3-1}$$

$$R=NUM(\text{算法结果组} \cap \text{标准大案组})/NUM(\text{标准答案组}) \quad \text{公式 3-2}$$

$$F=2*P*R/P+R \quad \text{公式 3-3}$$

3.4.2 数据集

将从网络上获取的 15 万条英汉对照组进行初步处理，去掉重复元组，随机抽取 224 个能够反映数据总体情况的元组进行试验。为了借用分类性能评价指标，将这 224 个元组根据语义上是否同义进行人工聚类，作为“标准答案”。人工标注的依据是语义相似程度，原则为“词汇在上下文中的可替换性”[45]。标准答案中共有 138 个元组类。测试样本、标准答案以及编辑距离二次计算方法

和传统编辑距离算法的聚类结果见附录 24。其中采用传统编辑距离算法计算相似度时，如果两个术语之间的编辑距离小于等于 1，则认为两个术语相似。

3.4.3 实验结果与分析

表格 3-1 中所示的各项指标表明：采用融入规则的编辑距离二次计算方法作为术语相似度计算方法，其聚类效果要好于以传统编辑距离计算方法作为相似度计算方法的聚类结果。表格 3-2 所示为通过网络爬虫获得的原始关键词数目以及经过本章提出的关键词语义聚类算法处理后的关键词数目。通过两者对比可知，该算法在工程应用中具有一定的实用价值。

表格 3-1 编辑距离二次计算方法 VS 传统编辑距离计算方法

	准确率（P）	召回率（R）	F 值
编辑距离二次计算方法	**0.9214**	**0.9348**	**0.9281**
传统编辑距离计算方法	0.4833	0.4202	0.4495

表格 3-2 处理前后的关键词数据对比

原始数据条目数	算法处理后数据条目数
148,825	83,602

3.5 知识族谱

文献调研是科研工作的先导和向导，对于科研工作有重要的价值和意义。充实的文献调研是优秀的科研成果不可或缺的基本条件。有调查统计，科研人员做文献调研的时间往往占其全部工作时间的三分之一。互联网的发展导致网上文献资源越来越丰富；科学研究的日益专门化、学科之间渗透交叉等现象日趋强烈，导致各学科的文献越来越分散[131]。基于以上两点原因科研工作者耗

费在文献检索上的时间成了提高文献调研效率的瓶颈。为了提高文献调研效率，本章提出了知识族谱模型以及实现方案，该方案已经在自动化学科知识服务网络平台[17]上实现。

3.5.1 需求分析

创新来自联想，联想源于博学广识和集体智慧[133]。知识创新源于知识积累，它们是创新和继承的关系。对于一个学科、一个专业、一个岗位，都存在大量的已有知识。只有迅速掌握这些已有知识，在头脑中建立该学科、该专业的知识框架，才能够在已有知识基础上进行知识检索、知识关联和知识创新。特别是对于初学者，是否能够快速掌握知识框架直接关系到知识创新的效率和结果[132]。

科学研究活动立足于对已有知识的搜集、梳理和积累，着眼于对未知世界的探索和钻研。文献调研是按照科学研究活动的需要，有目的、有计划地收集和阅读文献资料，其目的是为了充分了解具体的科研方向。文献检索有助于使用者掌握本课题研究的进展动态，开拓思路、避免重复劳动，把研究水平提到新的高度[131]。

现有的检索系统需要根据使用者输入的检索关键词进行基于布尔逻辑的匹配查询。很多专业领域的初学者由于不了解该专业领域的知识体系，在确定和调整检索关键词时缺乏参考和指导，往往会造成检索效果不佳。此外现有的文献检索系统的检索特点导致了“由点及面”的调研方式。即研究者在调研过程中，需要根据一篇文献的信息（作者、关键词、参考文献等）二次检索找到其他的文献。这种由点及面的调研方式对于海量的文献信息来讲，不仅调研效率低，而且不能保证调研活动的查全率与查准率。即：不能保证文献调研活动在有限的时间内获得最有价值的信息。由此可见，将用户所要了解的研究领域的发展态势纲举目张地展现出来，使得背景知识前景化，从而解决用户缺乏背景知识的问题是提高文献检索效率的关键所在。本章设计的知识族谱即是通过数据挖掘算法发掘关键词之间的共现关系，并按照一定层次可视化展现出来，将背景知识前景化。

[17]http://autoinnovation.ia.ac.cn

3.5.2 相关工作

共现分析(cooccurrence)是指将研究对象的共现频率作为基本研究素材的一种分析手段。这种方法假设研究对象之间的亲疏程度由他们在同一个场景中共同出现的频率或概率决定。信息检索领域常用的查询扩展方法(关联簇、度量簇、标量簇等[134])和隐式语义索引技术（LSI Latent Indexing）[135]，以及数据挖掘领域的 Apriori 算法等都属于共现分析范畴。

将共现分析用于文献情报研究已有先例[138]-[141]，然而，这些研究和应用中，均没有考虑时间因素，忽略了研究对象间可能存在继承和发展关系，而这种关系对于研究者了解学科领域发展整体状况是非常有价值的。对于一个专业领域而言，学科知识框架至少应该包含该学科的研究主题关联和发展脉络两个因素，其中研究主题关联通过论文中的关键词共现关系体现，而发展脉络则可以通过年代来划分层级。由此形成可视化的学科知识谱系，可以为使用者的知识创新提供更多的辅助参考[132]。

3.5.3 知识族谱构建方案

知识族谱常见于人文社科类文章中，本文首次将“知识族谱”这一概念由社会学领域引入到数字图书馆领域，并给出了采用计算机技术自动建立“知识族谱”的方案[18]。本文对“知识点”以及知识族谱做如下定义：

1.关于什么是知识，认识论范畴一直存在争论。不同的研究者根据其方法论立场不同，采纳不同的观点[142]。本文采用知识是物化的、过程化的观点[143][144][145]。本文认为有四种类型的知识点：（1）理论：如“贝叶斯”、“卡方检验”等；（2）方法或算法：如“贝叶斯分类算法”、“卡方特征词选择算法”等；（3）工具：如“贝叶斯分类器”、“网络采集器”等；（4）总结性知识点，它的内容涉及理论、方法或算法、工具，一般为代表研究方向的短语，如“自然语言处理”、“模糊控制”等。知识点是知识族谱上的结点。

2.知识族谱由知识点构成，是一种知识组织和表达的形式。它以知识点产生的年代为主线，将知识点表示的理论、方法或算法、工具以及研究方向按照时

[18]“知识族谱”这一设计理念及其基本形态由中科院自动化所刘禹老师在“自动化学科创新思想与科学方法研究”课题中提出，本文的工作为立足于数据分析，设计了工程上可实现的构建方案和可视化方案。

序顺序组织起来，着重体现知识点之间的演变、继承和发展关系。

本文所述知识族谱的详细构建步骤如下：

⑴年代划分：参考自然的年代划分方法，比如以 10 年或者 5 年为一个时间段；在此基础之上，要兼顾到现有语料素材的分布情况，尽量让落在各个时间段内的文献数量均衡。

⑵知识点的选取：知识点的内涵是理论、方法或算法、工具以及研究方向，是有价值的信息。考虑到文献的关键词由论文作者给出，而论文的著作者一般是熟悉某个领域的专家，因此可以将文献的关键词看作是作者知识的体现。本文从关键词中按如下标准提炼知识点：①关键词总是和其他关键词共同出现，出现频率高的关键词，和其他关键词关系就会越紧密，越具有代表性；由此本文认为知识点是在全体语料素材中出现频率（Document Frequency）高于一定阈值的关键词。②根据事物是普遍联系的哲学原则，要求所选出的知识点能够形成连通图。③知识点的时间标签为其首现年代。④综合以上三条标准，为每个年代选取代表性知识点。

⑶知识点关联强度排序策略：①将与被查询知识点有共现关系的知识点按照时间标签分类；②对于每个年代，按如下规则进行关联强度排序 a.按照和被查询知识点共现的次数由高到低排序；b.若两个知识点与被查询知识点共现次数相同，则出现次数（Document Frequency）高的知识点排在前面。

3.5.4 知识族谱可视化方案

知识族谱表示方案如图 3-6 所示，红色标注的知识点代表被查询知识点。知识族谱首先将与被查询知识点相关联的知识点按照产生年代进行层次划分，每个层级上的知识点又根据与其与被查询知识点之间的关联强度由高到低，由中心到两端排列。

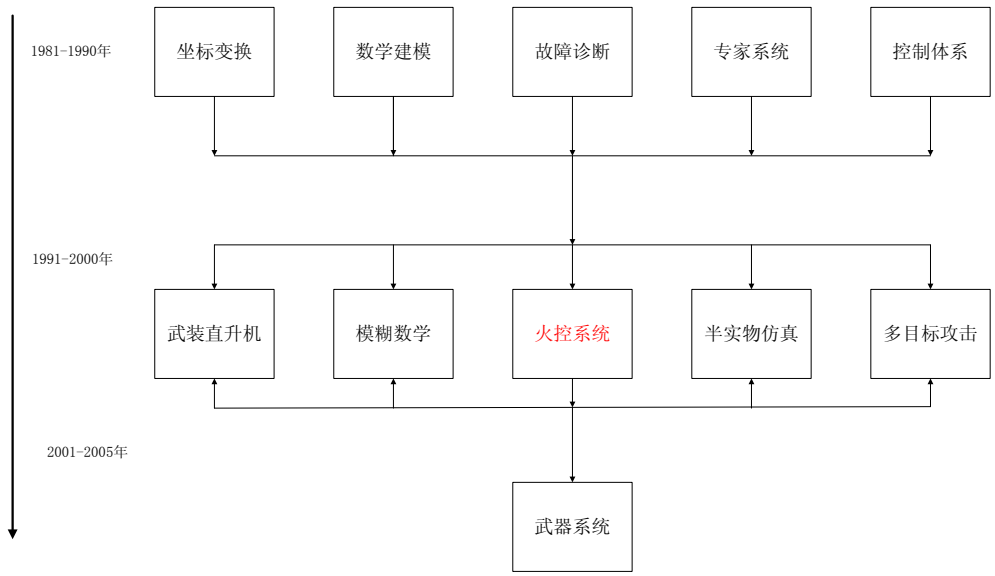


图 3-6 被查询知识点为“火控系统”的知识族谱

3.5.5 有益效果

知识不是孤立于理解者而独立存在的，它受理解者的知识需求和原始知识体系结构的影响，是外界刺激引发认知过程活动的结果[146]。本论文所阐述的知识族谱构建方案以及可视化方案在背景知识前景化，帮助使用者形成专业领域知识体系，或者在文献检索查询词扩展方面有一定的有益效果。如图 3-7 表示在“知识发现”这一概念提出之前，研究者主要从事“机器学习”、“专家系统”、“分类”、“聚类”、“时间序列”相关的理论研究。“知识发现”是在上述理论基础上继承发展来的。与“知识发现”同一时期出现的其它知识点，如“数据挖掘”、“关联规则”、“粗糙集”、“数据仓库”等方面的研究，为“知识发现”的研究提供辅助和补充。随着人们对这一领域的研究深入，“案例推理”、“客户关系管理”等概念相继出现，进一步丰富和扩展了“知识发现”这一研究领域的深度和广度。

图 3-7 检索知识点为“知识发现”的知识族谱

3.6 本章小结

关键词语义聚类算法的结果，将作为知识族谱构建算法中的输入，因此关键词语义聚类算法的效果和粒度关乎知识族谱的效果。用于关键词语义聚类算法的启发式规则集还有一些不完善的地方，比如没有包含同义词和近义词相关的规则。下一步可以考虑加入同义词林来描述构成术语的基本词语的词义相似程度。本章提出的编辑距离二次计算框架也可以应用于其他计算字符串形态相似度的场合；本章设计的求两点之间所有路径的算法，不局限于 DAG 图，对其他有向图以及无向图同样适用。

本章中设计的“知识族谱”仅仅利用了知识点之间的共现关系，由于没有融入常识知识，更没有对知识点进行细粒度的属性刻画，因此止步于浅层知识表达层面。本人在设计过程中试图采用人工编辑知识点的方法对知识点进行属性扩充，但是这又似乎违背了通过计算机自动完成知识族谱构建的设计初衷。对于如何采用计算机技术自动完成富属性注入，本人目前尚没有成熟且可实践的思路。

第四章 信息抽取技术与知识要素获取

本章 4.1 节论述信息抽取技术在数字图书馆以及文献情报学研究中应用的必要性；4.2 节给出信息抽取技术的定义、应用前景及其发展历史和现状；在 4.2 节论述的基础上，4.3 节给出本论文所涉及的技术方面以及设计知识要素获取算法的指导原则；4.4 节至 4.6 节分别给出人物机构对齐、人名消歧、机构名称识别算法的设计与评价。

4.1 信息抽取技术在数字图书馆应用中的意义

数字图书馆的长远目标是从信息检索服务转向知识提供服务。知识服务的前提是知识的获取，通过信息抽取，能够从自由文本抽取出数值数据和结构化信息，建立起可供研究分析的联机分析系统，进而进行大规模的数据挖掘和信息分析[46]。

文献情报学研究立足于对特定领域内的文献、学者、机构、以及研究方向等（以下简称知识要素）做全方位的关联分析和计量分析。如统计学者之间的合作情况、机构之间的合作情况、学者的研究方向、学者发文数量、机构发文数量、关键词与学者之间的对应情况、关键词与文章之间的对应情况等等。

4.1.1 人名消歧的必要性

我国新华字典共收录 231 个姓氏[47]；现代汉语词典第五版共收录姓氏条目 1818 个[48]；中国科学院遗传与发育生物学研究所于 2006 年 1 月 10 日发布了用两年时间调查完成的《中国姓氏统计》，该调查表明在全国目前使用的约 4100 个姓氏里，除去双姓、三姓，绝大多数是单字的姓，其中李、王、张分别以 9600 万人次、9400 万人次、8800 万人次位列三甲[49]。文献[50]对 2001 年-2003 年福建省高考录取信息资料的 71,797 个名字进行了统计分析，得出中国人名具有文化传承性、性别区别性、时代性、悦耳性等特点。该文指出大约每 25 个男性名字中的第一个字具有一个相同的字种；约每 30 个女性名字中的第一个字会出现一个相同的字种。本文将自动化学科知识服务网络平台后台数据库中出现次数在 50 次以上（也即关联的文章数目大于等于 50）的名字按照出现次数从高到低排序，依据常识知识选取了 42 个常用名字。这 42 个名字在数据库中出现的次

数以及经过同名消歧算法处理后对应的人物实体数目参见附录 15。本文进一步对这 42 个名字的姓名用字进行分析，如表格 4-1、4-2 所示。通过观察可以得出与文献[49][50]类似的结论，即：汉语人名倾向于姓名常用字种重叠。

表格 4-1 自动化学科知识服务网络平台数据库后台常用姓字种统计

排序（按出现次数）	姓用字种	出现次数（次）
1	王	10
2	李	8
3	张	8
4	刘	6
5	吴	2
6	陈	2
7	杨	2
8	谢	1
9	高	1
10	赵	1
11	徐	1

表格 4-2 自动化学科知识服务网络平台数据库后台常用名字种统计

排序（按出现次数）	名用字种	出现次数（次）
1	刚	5
2	军	4
3	伟	4
4	斌	3
5	强	2
6	波	2
7	勇	2
8	平	2
9	涛	2
10	明	2
11	磊	1
12	立	1
13	辉	1
14	文	1
15	飞	1
16	宏	1
17	健	1
18	华	1
19	芳	1
20	俊	1
21	浩	1
22	超	1
23	鹏	1
24	敏	1

据附录 15 中的统计数据，1,369 人次累计发表了 4,423 篇文章。作者和文章之间的数据对应比较稀少，这也是自动化学科知识服务网络平台数据库的总体情况。即：少数作者比如高校教师由于长期处于教育研究第一线，所以数据库中会有大量论文记录；大多数作者比如高校学生仅在学习期间有论文记录，毕业之后也就离开学术科研岗位了。自动化学科知识服务网络平台数据库中论文数目最多的作者为南京理工大学的杨靖宇老师，他的最早著作可以追溯到 1979 年，截至 2010 年他共参与发表了 274 篇论文。这个统计结果与杨静宇老师个人主页上[19]的介绍基本一致。据附录 15，名字“王伟”共参与到了 245 篇论文的发表工作中。通过图 4-1 中关于“王伟”的单位字符串可以进一步鉴定“王伟”这个名字对应很多不同的实体。因此若不进行人名消歧而直接对学者的发文量进行统计，很可能把不同“王伟”发表过的文章归结于同一个“王伟”名下。

AuthorId	AuthorName	AuthorWorkPlace
175175	王伟	大连理工大学信息与控制研究中心;
175340	王伟	东北大学自动化研究中心
175769	王伟	东北工学院自控系沈阳
176459	王伟	东北大学自动化研究中心
178188	王伟	西北工业大学
185026	王伟	西安交通大学ＣＩＭＳ研究中心
185090	王伟	北京冶金管理干部学院;
191146	王伟	哈尔滨工业大学自动化测试与控制系
191749	王伟	哈尔滨工业大学自动化测试与控制系
194179	王伟	北卡罗来纳大学;
195433	王伟	中国科学院计算技术研究所智能开放实验室
198487	王伟	上海交通大学图像处理与模式识别研究所
200929	王伟	河南师范大学计算机与信息技术学院
201850	王伟	上海交通大学电气工程系;
206618	王伟	中国地质大学研究生院武汉
207086	王伟	中国海洋大学青岛海关;中国电子口岸数据中...
209278	王伟	武警工程学院研究生队;
209354	王伟	上海大学计算机工程与科学学院
213573	王伟	东北工学院自动控制系;东北工学院自动控制...
213959	王伟	东北大学自动化研究中心
214804	王伟	西北工业大学航空自动控制系
216605	王伟	北京科技大学智能及计算机科学研究所;
216661	王伟	哈尔滨工业大学自动化测试与控制系

图 4-1 有关名字“王伟”的记录

[19]http://cs.njust.edu.cn/szdw/ShowArticle.asp?ArticleID=27

4.1.2 机构名称抽取与消歧的必要性

自动化学科知识服务网络平台数据库中收录了附录 1 所示的 29 本相关领域中文期刊自创刊以来截至 2010 年的论文题录数据。其中最早的期刊如《计算机研究与发展》可以追溯到 1960 年，年代跨度很大。在近 50 年的新中国自动化相关领域的学术发展历程中，有不少学术机构涌现，有不少学术机构消亡或者合并，也有不少学术机构虽然依然存在，但是机构名称历经更迭。其中最常见的一种机构名称变化方式为院校内部的院系调整，如“北京邮电大学信息工程学院”与“北京邮电大学电信工程学院”在 2008 年合并为“北京邮电大学信息与通信工程学院”。因此在统计学术机构的学术贡献时，我们不能简单地依据文献题录信息中所记录的机构名称进行统计，需要从中识别和抽取出一级机构名称，如从“第二炮兵工程学院 302 教研室；清华大学自动化系陕西西安”中抽取“第二炮兵工程学院”和“清华大学”[20]。

4.2 信息抽取技术的发展和现状

4.2.1 信息抽取技术的定义

信息抽取技术在信息检索技术与信息过滤之后发展起来[51]，这三种技术的根本宗旨相同，即满足用户的信息需求，但是在具体内容上有各有侧重。信息检索技术的基本内容为是从海量数据中获取相关文档列表，信息过滤技术侧重对用户建模和个性化信息的定制与推送[52]；信息抽取技术是指从特定的文档或者文档集合中抽取事实性信息[51]，如图 4-2 所示。文献[53]中具体总结了信息抽取系统与信息检索系统的三点区别：（1）功能不同，信息检索系统主要是找到用户需求相关的文档列表；信息抽取系统主要从文本中抽取用户可能感兴趣的事实信息。（2）处理技术不同，信息检索系统通常采用关键词匹配等技术，把文档看成词的集合，不需要对文本进行深入分析与理解；信息抽取系统往往要借助自然语言理解技术，通过对文档中的句子进行分析后才能完成，可参考图 4-3 所示的信息抽取系统典型架构；（3）适用领域不同，信息检索系统通常是领域无关的；信息抽取系统则是领域相关的，只能抽取预先设定好的、有限的

[20]由于中国科学院的各个研究所地域分布广阔且学术成就众多，在一级机构名称抽取中，本文将抽取粒度细化到中国科学院的各个院所，这点与普通的大学以及学院的机构名称抽取粒度不同。

几类事实信息。

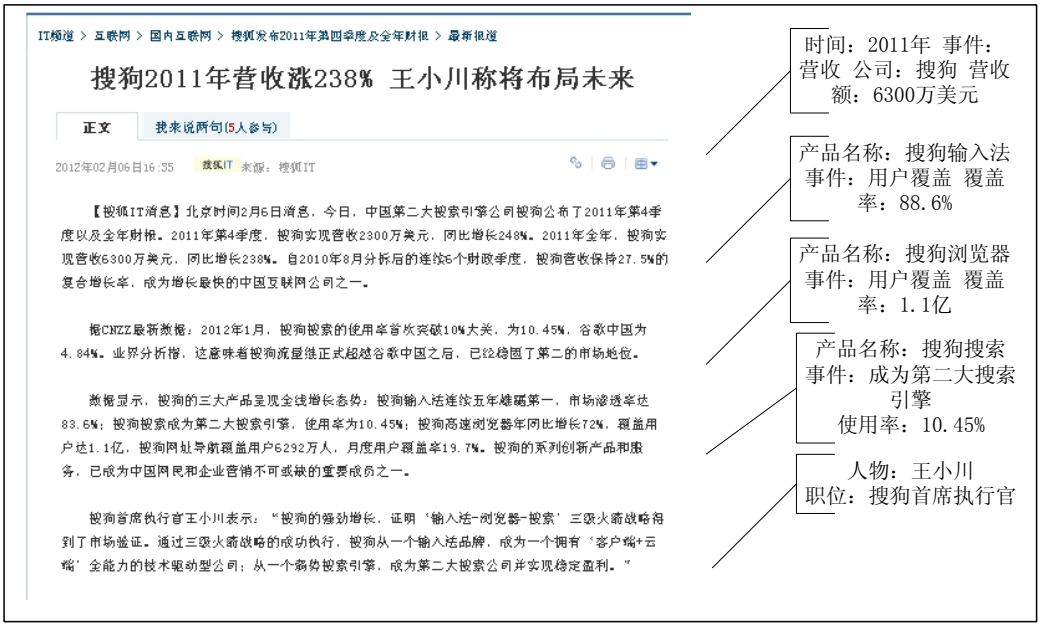


图 4-2 信息抽取技术说明图例

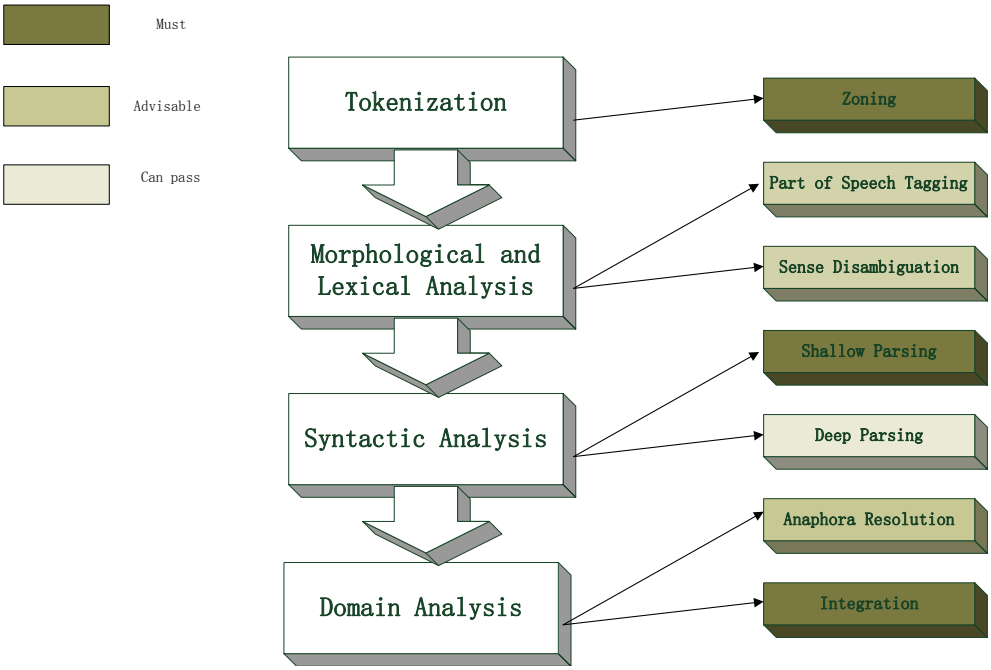


图 4-3 信息抽取系统典型架构[21]

[21]本图来自文献[54]

4.2.2 信息抽取技术的发展历史和现状

FRUMP 是最早被记录在案的信息抽取系统，该系统于 1979 至 1982 年间由 Gerald deJong 设计实现。该系统是一个融合关键词和句子语义分析，将期望驱动(top down)与数据驱动(bottom up)处理方法相结合的新闻抽取系统[58][59]，该系统的设计方案被后来的很多其他系统借鉴。自 20 世纪 80 年代以来国际学术界举办的一系列评测会议 MUC[60]、ACE[61]、TAC[62]大力推进了信息抽取技术的发展。MUC（1987-1997）包括命名实体识别（Name Entity Recognition,NE）、模板元素填充（Templates Element Task,TE）、模板关系确定（Templates Relationship Task,TR）、同指关系消解（CoreferenceTask,CO）、场景模板填充（Scenario Template,ST）五类任务，数据来源是限定领域语料，如军事情报等；ACE（1999-2008）的数据来源主要为书面新闻语料，涉及实体检测与跟踪（Entity DectionandTracking,EDT）、数值检测与识别（Value Detection and Recognition,VDR）、时间识别和规范化（Time Expression Recognition and Normalization,TERN）、关系检测与描述（Relation Detection and Characterization,RDC）、事件检测与描述（Event Detection and Characterization,EDC）、实体翻译（Entity Translation,ET）等评测任务；TAC-KBP（2009-2012）包括实体链接（Entity Linking）和实体属性抽取（Slot Filling）两项评测任务，数据来源为新闻和网络数据[57]。

回顾信息抽取技术 20 余年的发展历程，从方法论上大致可将信息抽取技术分为基于规则的和基于统计的两类。在互联网高速发展之前，由于数据的结构相对简单，规模较小，所以基于规则的方法相对稳定和流行如 WHISK[63]，BWI[64]，[65]等。互联网的飞速发展促使了大量异质的、冗余的、不规范且含有噪声的数据的产生，进而推进了基于统计的信息抽取方法的发展。一般来讲，基于规则的方法一般比较适用于半结构化数据的应用场景，在自由文本场景下，基于统计的方法效果会更好。

4.2.3 信息抽取系统的相关评测结果以及可用性分析

信息抽取系统的评价与比较必须要考虑其内部每一个信息抽取算法的具体实验设置。包括训练集与测试集的划分方式、特征集合是否一致、如何定义完全匹配等[66]。文献[55]中表 2 给出了 BAKEOFF-3[68]命名实体识别六个评测任

务中性能最好的系统的测试指标；表 4 中给出了 2004 年“863”命名实体识别两个评测任务中性能最好的系统的测试指标。文献[55]同时指出表 2 中评测结果好于表 4 的评测结果，表 2 中 MSRC 语料和 CITYU 语料上的评测结果好于 LDC 语料上的评测结果。文献[55]中指出出现以上结果的原因可能在于：863 评测不提供训练集，MSRC 和 CITYU 语料拥有相当规模的训练集，LDC 只提供了小规模的训练集。由此可见训练集规模、训练集与测试集的相似程度等对于基于语料的信息抽取算法效果的影响。在实际应用中由于无法获得或者根本不存在人工标注的大规模训练集，所以语料依赖型的信息抽取算法的效果可能会显著下降。

4.3 知识要素获取算法设计的指导原则

命名实体相关的研究工作主要包括：命名识别识别、命名实体消歧、实体跨语言关联、实体属性抽取、实体关系检测等[55]。如图 4-4 所示，本文中的知识要素获取任务共包括人物机构对齐、人名消歧、顶级机构名称抽取等三部分工作。就本论文而言，人物机构对齐属于实体跨语言关联范畴，人名消歧属于命名实体消歧范畴，顶级机构名称抽取属于命名实体识别范畴。

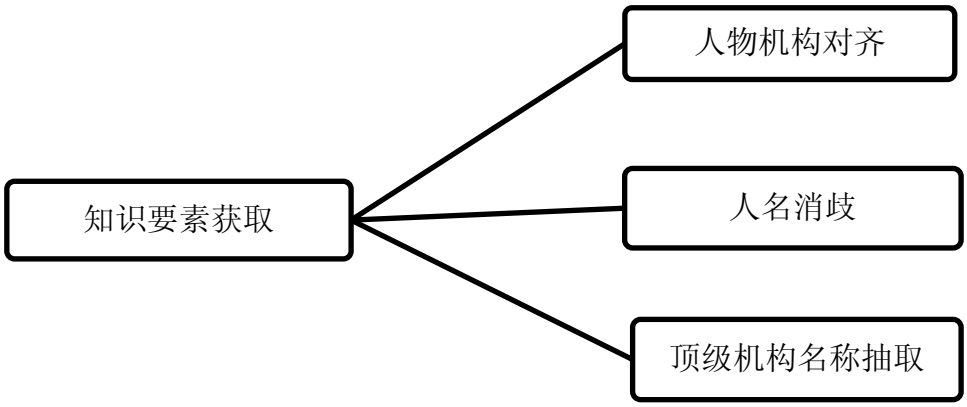


图 4-4 知识要素获取工作的组成部分

在知识获取阶段，为了达到充分利用原始数据资料、精准定位和抽取各项知识要素的目的，本文制定了如下三条知识要素抽取规则：①最大限度地利用

原始数据源，从原始数据源提取信息，进而从信息中提炼出知识；②提取的信息要尽量保证精确；③在保证以上两条工作准则的基础上，利用自动化算法实现知识要素提取，尽量减少人工参与和手工劳动。

本文用信息抽取技术解决实际问题，属于技术应用范畴。因此必须分清理论研究与技术应用之间的区别：①理论研究立足于前沿和新颖，而技术应用立足于成熟和稳定。②理论研究者通常试图融入更多的特征，力求用一个复杂的模型来达到数学上的形式完备性，而技术应用则更侧重考虑效率和可实现性[22]。③理论研究面向一般性问题，而技术应用一般有具体的数据场景，因此基于统计机器学习的方法尤其是基于指数家族参数化模型的统计机器学习方法，未必一定会在效果上优于基于规则和背景知识的方法。④面向理论研究的评测任务中，评测方往往提供大规模的训练集，而且训练集规模往往远远大于测试集规模，在实际应用中，高质量的训练集需要依赖大量的人工劳动，因此通常是无法获得的；此外，实际应用面向的是开放式问题，即使有标注语料，那么相对于预测文本“无限性”来说也是“有限”的。⑤理论研究中评测指标的高低与测试集规模以及测试集与训练集的相似程度有关，所以不能因为某个算法在某篇论文中达到了某种效果就认为该算法可以照搬套用，采用算法方案之前必须针对实际数据进行验证性实验；⑥理论研究者通常会借助开源的算法工具包如CRF[68]，MaxEnt[69]，LDA[70][71]，LibSVM[72]等，并在原来的模型中添加特征。所以研究领域内很多算法的广泛使用或许很大程度上是开源工作推动的结果，应用者不能因为算法被广泛使用，就简单认为该算法可以应用在某个具体问题中；⑦出于经济成本等方面的考虑，应用者在设计算法时应充分利用手头已有的或者可以廉价获得的数据资源，避免使用高成本数据资源。

综上，本文在对研究领域内的算法进行调研的基础上，结合实际情况，在算法设计上力求简单、可行和有效。

[22]在“自动化学科创新思想与科学方法研究”课题中，由于结题期限的限制，从信息抽取算法的调研，知识要素提取算法的设计、实现、实验和评测，以及最后算法在工程中的正式使用，只有 5 个月的时间。

4.4 人物机构对齐

4.4.1 需求分析

在如图 4-5 所示的数据源上，无法直接获得作者与作者单位之间的对应信息，因此只能先获取作者的拼音形式姓名与英文机构名称之间的对应关系，然后进一步获得汉字形式的姓名与中文机构名称之间的对应关系。

图 4-5 人物、机构字符串数据情况

4.4.2 相关工作

先获取作者的拼音形式姓名与英文机构名称之间的对应关系，然后进一步获得汉字形式的姓名与中文机构名称之间的对应关系，一般来说需要获得中文机构名称对应的英语翻译、汉字形式的姓名对应的拼音形式姓名。其中汉字形式的姓名转换成拼音形式姓名，主要依赖于汉字拼音对应表，不存在技术难度。机构名称的翻译问题包含音译、意译、音译和意译区分、虚词对应、习惯用法等问题。学术界通常从统计机器翻译[73][74]、双语语料抽取[75][76]、借助搜索引擎等三个角度解决机构名称翻译的问题[77][78]。机构名称翻译的一般过程是对中文机构名称进行分词，获得对每个分词后的单元获得对应的英文翻译，最后对翻译进行调序，形成整个中文机构名称的翻译。这里涉及到两个问题，一

是由于机构名称中可能含有地名、人名等其他专有名词，而目前的分词算法依赖于词典，对未登录词的切分结果不尽人意，所以直接对中文机构名称字符串进行分词可能会因为切分不当导致翻译结果不佳；二是在获取每个词汇单元的英文翻译的过程中，需要对中文、英文字符串都进行命名实体识别，然后再找出对应关系，这又会受到命名实体识别算法的有效性的限制。杨帆等人提出了一种基于启发式网络挖掘和非对称英汉对齐的中英文机构名称翻译方法[79]。该方法定义了地点组块、名称组块、修饰组块、关键词组块。首先对中文机构名称进行组块分析[80]，对可能含有未登录词的地点组块和名称组块不再进行分词，仅对关键词组块和修饰组块内部进行二次分词，从而避免了直接对机构名称分词所造成的切分不当问题。之后将切分结果放入统计翻译系统，并从翻译结果中挑出部分词汇和中文机构名称一起构成启发式查询，从返回的双语网页中提取英文句子；最后将问题转换为带权二部图求最大权匹配的问题，并用 KM 算法求解[81][82]。这种问题转换方式有效规避了传统机构名称翻译中要对源语言和目标语言字符串分别进行命名实体识别的需求。

4.4.3 算法设计

本文认为文献[79]中提出的方法可以在此处使用，但是该算法的实现需要有 LDC 语料[83]的作为训练集，而本文无法获得，只能另想办法。本文在算法实现上，继承了该文献作者在算法设计上规避不成熟算法干扰的思路。

4.4.3.1 数据分析与规则总结

观察图 4-5 中所示的数据源情况，可以总结出如下几条稳定规律。

I 英文字段中机构名称出现的次序与中文字段中机构名称出现的次序一致；

II 英文字段中标识的作者机构所属关系有如下两种情况：

情况①：当同一作者隶属的单位不止一个时，采用作者拼音姓名之后用数字注明所属单位的形式，本文将这种描述作者机构所属关系的方式成为模式 1；

情况②：若干作者拼音姓名之后加上机构名称，表明这些作者同时隶属于某个单位，本文称之为模式 2；

基于以上数据分析，本文设计了如图 4-6 所示的人物机构对齐算法流程。以下着重介绍汉字形式名字字符串到拼音形式名字字符串的转换算法，以及基于距离属性的二叉分裂算法。

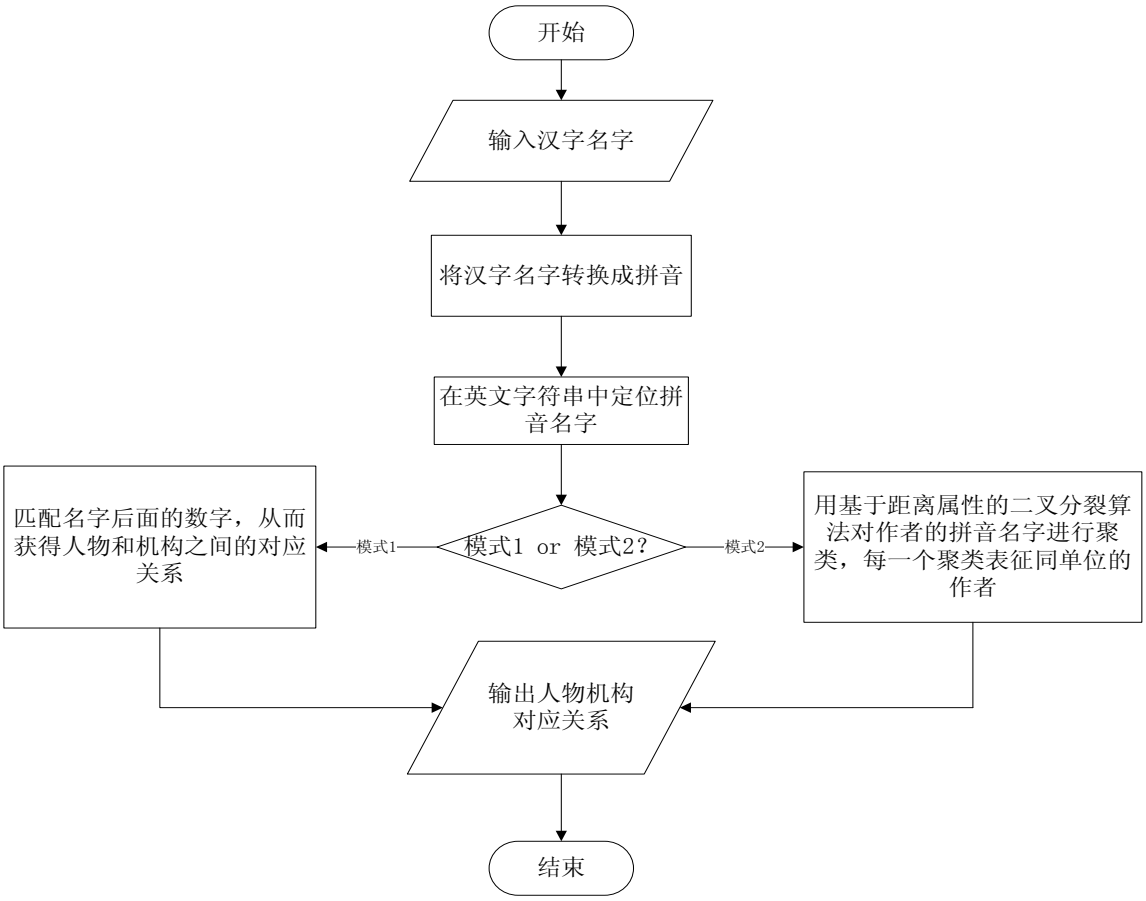


图 4-6 人物机构对齐算法流程图

4.4.3.2 汉字字符串到拼音字符串的转换算法

将汉字形式的名字字符串转换成拼音形式的名字字符串，主要通过汉字拼音转换码表实现。考虑到名字中包含的汉字可能是多音字，以及拼音之间的连接符类型可能有多种形式，我们对每一个汉字名字组合形成多个候选的拼音名字正则表达式字符串。可以将生成多候选的拼音名字正则表达式字符串的问题转换为求有向图上两点之间所有路径的问题，如图 4-7 所示，并用算法 3-2 求解。

输入汉字形式名字为“曾校靓”

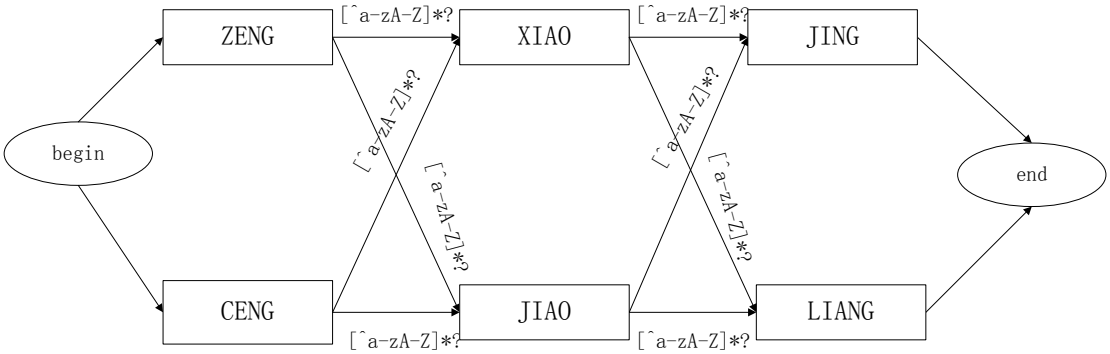


图 4-7 汉字字符串到拼音字符串的转换

4.4.3.3 基于距离属性的二叉分裂算法

基于距离属性的二叉分裂算法主要解决的是拼音形式名字和英文形式的机构名称之间的对应关系问题。基于 4.4.3.1 节对数据情况所做的分析，这里设计的距离属性的二叉分裂算法是一种保持原有数据位序属性的分裂型层次聚类方法如图 4-8 所示。图中的结点 a、b、c、d、e、f 等分别表示图 4-5 中所示的情况 2 中的拼音形式名字，边表示两个拼音形式名字之间的距离也即是机构英文名称字符串和其他字符的长度和。当两个名字之间的距离大于一定阈值时，就可以断定这两个名字不属于同一个机构。该分裂型聚类算法实际上是一个以层序方式建立二叉树的过程，叶子节点即是最终的聚类，分别顺序对应于中文机构名称字符串中的机构名称。

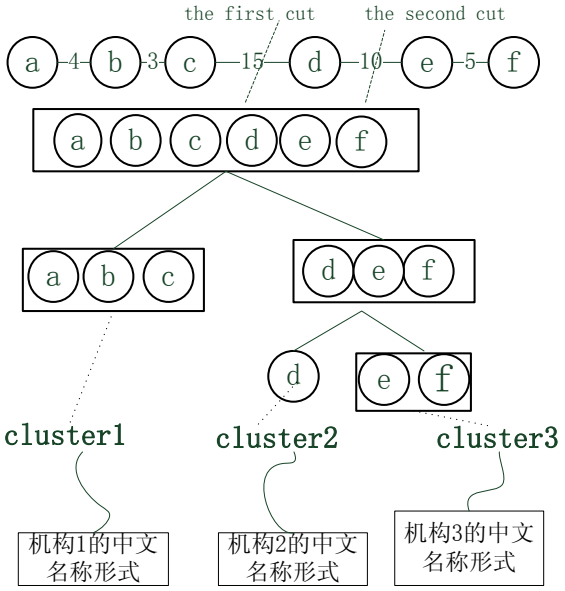


图 4-8 基于距离属性的二叉分裂算法示意图

4.4.4 算法有效性与局限性分析

4.4.3 小节设计的人物机构对齐算法，存在以下结点不足：

i. 如果数据源中没有任何能够提示人物机构对齐的信息，则算法失效。

ii.本文设计的汉字拼音转换算法考虑到了汉语名字中含有字音转换码表中不存在的汉字情况，如果有些汉字不能被正确翻译，则用正则表达式[A-Za-z]*?代替。但是此法不能完全避免潜在错误的产生。

iii.二叉分裂算法在决定是否将一个聚类分裂成两个聚类时，要依赖启发式阈值，阈值的设定会对算法的准确性产生影响；

本文在算法运行中记录下可能出现潜在错误的实例，最终计算得到人物机构对齐算法的准确率为 262,896/299,823=87.6837%。其中 299,823 为总的人物条目，262,896 为被正确进行人物机构对齐的条目。

4.5 人名消歧

4.5.1 需求分析

4.1.1 节指出了人名消歧的必要性与重要性，在数字图书馆领域，目前有两种方式进行人名消歧，即人工消歧[85]和计算机实现消歧算法两种策略[84]。前者的准确率高，但是不适用于大规模数据的应用场景；后者利用程序实现消歧算法，可以适用于大规模数据的应用场景，但是准确率会有所下降。本文采用计算机实现消歧算法的策略。

4.5.2 相关工作

4.5.2.1 关于聚类的背景知识

聚类是无监督学习（unsupervised learning）最普遍的一种形式，它一般应用于事先不知道类别标签的场合。聚类算法将文档集合按照“类内紧密、类间分散”的原则聚团为若干文档簇。聚类问题一般涉及特征选择与变换、相似度计算以及聚类算法选取三个阶段。

①常用的特征选择算法，一般以词袋子模型（Bag of Words,BoW）为出发点，按照某种测度，选取一定量的特征词，将文档表示成文档向量模型（Vector Space Model,VSM）模型。由于缺少类别标签，分类中常用到的特征选择算法如卡方（卡方）、信息增益（IG）等都无法直接应用到聚类问题中，一般只有文档频率（DF）方法可以直接在聚类问题中使用。值得指出的是特征选择与变换往往和实际问题密切关联，因此需要根据实际应用的具体场景并结合处理者的背景知识来确定特征选择和变换的方法。

②由于在计算文档相似度之前，文档已经按照所选特征用可计算的模型进行表示，所以相似度计算实际上是一个与实际问题无关的几何量层面的问题。常用的相似度计算方式有余弦相似度、欧式距离、曼哈顿距离、闵科夫斯基距离、切比雪夫距离等。

③聚类算法依据相似度计算的结果将文档聚团成簇，一般来讲聚类算法及其优化问题属于数学范畴，与实际问题关联不大，但也并不是完全无关。聚类算法按照结果簇是否具有层次组织形式可以分为扁平聚类算法和层次聚类算法

两种。扁平聚类算法包括 Kmeans，EM 等，一般来讲扁平聚类算法需要事先对聚类数目进行初始估计。层次聚类算法包括凝聚式层次聚类和分裂式层次聚类两种。凝聚式层次聚类首先将每个数据点本身看做是一个聚类，然后根据相似度算法的结果对原有聚类进行合并，直到全部合成一类或者满足某个停止条件为止，用于凝聚式层次聚类的相似度计算方法一般包括单连接、全连接、组平均以及质心相似度方法。分裂式层次聚类算法首先将所有数据点看做一个聚类，然后根据分裂条件不断地将大聚类进行分割，直到每个聚类只有一个数据点或者满足某个停止条件为止。4.4.3.3 中设计的基于距离属性的二叉分裂算法就属于分裂式层次聚类算法的一种。一般来讲，层次聚类算法不需要对聚类类别数目事先进行估计。综上，扁平聚类算法适用于能够对聚类类别数目进行大致估计的应用场合，而层次聚类算法试用于不能够对聚类类别数目进行大致估计的应用场合。

以上①②③是一般文本聚类算法的必经步骤。近年来，由于受社交网络的兴起，以及概率图模型逐渐成为研究热点的影响，基于图论的聚类方法也有了越来越多的应用场景。该类聚类算法的特点是不需要对聚类数目进行初始估计如 Affinity Propagation[89][90][91]，Rosvall's information-theoretic clustering algorithm[92]，spectral clustering[93]，Betweenness-Based Clustering[94]，Min-Cut Clustering[95][96]，Minmum Spanning Tree Based Clustering Algorithms[97][98]等。

4.5.2.2 命名实体消歧的定义和工作

命名实体消歧工作可通过如下三元组来定义：

$$M = \{N, E, \sigma\}$$

其中$N = \{N_1, N_2, N_3, \dots, N_l\}$是待消歧的实体名称或称实体指称项集合；$E = \{E_1, E_2, \dots, E_k\}$是现实世界的实体概念集合；$\sigma{:}N \to E$是消歧函数。

按照实体概念列表是否给出，命名实体消歧有可以分为：实体链接型实体消歧和聚类型实体消歧；根据应用领域不同，也可分为数据库实体消歧和文本实体消歧。本论文工作所涉及到的人名消歧属于数据库实体消歧范畴。

4.5.2.3 研究现状

UNED 大学针对 Web 检索结果中的人名歧义问题，专门组织了 web people search evaluation(WePS)[87]，截止至目前为止已经举办了三届。WePS 主要针对英文人名进行歧义消解。WePS 任务给定包含歧义人名的网页集合，要求参与评测的系统按照网页中的实体指称项所指向的人物概念对网页进行聚类，可以通过抽取关于网页中某个人的特定属性信息来辅助人名消歧。2010 年 CIPS-sighan[88]的评测任务 3 是专门针对中文人名消歧的评测，

现有的命名实体消歧算法可以分为四类：①基于表层特征的指称项相似度计算，即对指称项的上下文词语用向量空间模型进行建模如[99][100]；②基于扩展特征的实体指称项相似度计算，如借助人物传记属性等[101]和 Wikipedia[107]；③基于社会网络的实体指称项相似度计算，如[102][103]用 Random Walk 来计算两个实体指称项在社交网络之间距离，如果距离低于一定阈值，那么则认为这两个实体指称项指向同一个实体。④系统集成法，即采用多方法融合、多阶段处理的策略进行命名实体消歧如[104]-[106]，其中[104]先用凝聚式层次聚类（HAC）对人名进行聚类消歧，然后用 boostrap 算法对 document-cluster matrix 和 feature-cluster matrix 进行迭代更新，从而对聚类结果进行修正。韩先培认为现有的命名实体消歧系统在关键技术上主要存在以下两方面的问题（1）过分依赖于统计自然语言处理技术，忽略语义知识；（2）缺少能支撑自然语言处理的语义知识库，虽然这些知识广泛存在于互联网的知识源中，却难于为自然语言处理技术所使用。所以，传统的命名实体消歧研究由于不能有效挖掘和利用语义知识，难以取得令人满意的消歧性能[86]。

面向数字图书馆的作者消歧问题与网页中人名消歧相比，具有鲜明的属性信息，如作者工作单位、作者 email 等，以及鲜明的社会网络信息，如作者合作关系等，但是缺少上下文信息。在消歧方法上，有只利用作者合作关系的，如[108]-[111]；有综合利用作者合作关系、论文标题、发表日期、论文内容、发表期刊、论文摘要、关键词等特征的如[112]-[115]；还有借助于网络信息的如[116]-[118]。

4.5.3 算法设计

4.5.3.1 算法设计思路

由于本文只是针对自动化学科论文作者进行同名消歧处理，所以论文发表期刊、论文摘要、关键词等特征的作用不是很大。考虑到题录信息中的作者单位在命名上比较规范且很少出现简称式命名，所以用作者单位作为同名消歧的主要特征，也是唯一特征。并且假设在众多相同指称项中，同一个单位中的只有一个实体[23]。由于事先无法获得作者单位名称列表，所以这里涉及的问题属于聚类型实体消歧范畴。本人在算法设计过程中曾经尝试对作者单位字符串进行分词，建立 VSM 模型，并选用 Kmeans 算法进行聚类消歧的方法，结果效果非常不好。经过分析可能由以下三个原因导致：①由于无法事先对类别数目进行大致估计，所以 Kmeans 算法并不适用于此应用场合；②机构名称本身属于未登陆词，对机构名称进行分词相当于破坏了其原有粒度；③ TF/IDF 建立 VSM 模型的策略适用于长文本场合，而机构名称本身属于短文本，所以如果采用 VSM 模型的策略，应当事先对机构名称字符串用 boostrap[119]或者 jackknife[119]进行重采样。

基于以上分析，并参考 4.5.2.1 小节关于聚类过程的介绍，本文转换了思维角度，从而使同名消歧问题得到基本满意的解决。首先，为了避免分词不当造成机构名称字符串本身粒度被破坏以及过度概率化建模导致问题复杂话，弃用 VSM 模型表示文本的策略，而将单位字符串本身看做一个整体；其次，在聚类算法上采用不需要事先估计类别数目的、基于图论的聚类算法，之所以采用基于图论的聚类算法，另一层面的考虑是利用图具有关联推断的特性；再次，从字符串的粒度上，采用最长公共子序列（Longest Common Subsequence,LCS）以及最长非对称前缀(Longest AsymmetricPrefix,LAP)作为特征进行相似度计算。

问题建模如下：$G=(V,E,\varphi_G)$，其中V是图中结点，具有 *AttributeName*、*AttributeAddress* 两个属性。由于名字相同，所以实际上只有 *AttributeAddress* 作为区分属性使用。E是图上的边，φ_G为关联函数，$\varphi_G(e)=uv$。G的初始状态是散点图，当且仅当任意两个结点u、v之间的相似度 *Similarity(AttributeAddress(u),AttributeAddress(v))*满足一定条件时，才会在散点图

[23]这可能不完全符合实际情况，比如中国科学院自动化所综合信息中心有两个刘禹。

G上的u、v结点之间加边。最后，图G中的每一个连通分量就是一个聚类，或者说图G中的每个连通分量上的实体指称项指向现实世界中的同一个作者实体。

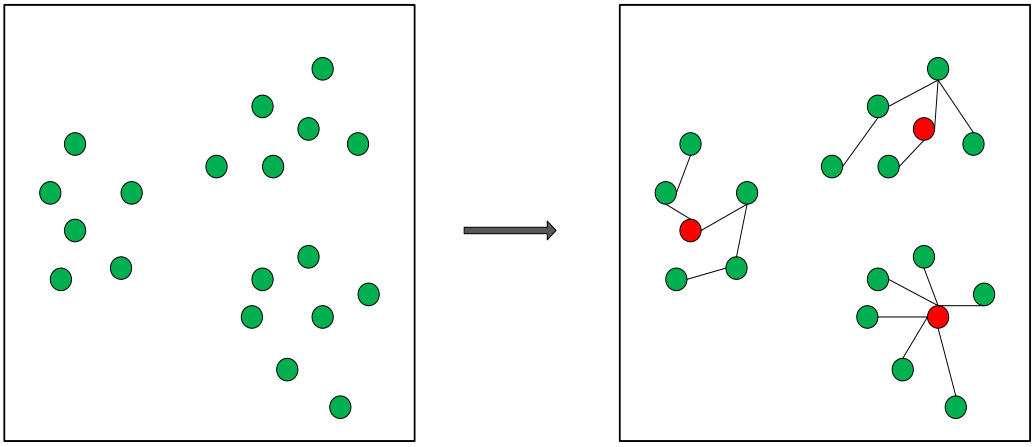


图 4-9 基于图的连通分量的聚类算法示意图

4.5.3.2 最长公共子序列

一个给定序列的子序列即为给定序列在保持原有位序的基础上去掉若干元素（也可能一个都不去掉）。如 Z=<B,C,B,D>是 X=<A,B,C,B,B,D>的子序列。两个给定序列 X,Y 的最长公共子序列（LCS）既是 X 的子序列，又是 Y 的子序列，且在所有 X,Y 的公共子序列中长度最长。LCS 问题是一个经典的动态规划问题[120]，常用于刻画两个字符串之间或者两个生物 DNA 序列之间的相似程度。

4.5.3.3 最长非对称式前缀

字符串 w 的任何连贯的符号构成的符号串称作 w 的子串，如果$w = vu$，那么子串v则称为w的前缀。我们知道如果 z 是两个字符串 s,t 的公共最长前缀，则 z 分别是 s,t 的前缀，且在 s,t 的所有公共前缀中长度最长。本文为了刻画机构字符串间的相似度，引入非对称式前缀的概念，并作如下定义：

如果 z 是两个字符串 s,t 的非对称式前缀，则可能有以下两种情况出现：

①z 是 s,t 的前缀；

②z 是 s 的前缀，是 t 的子串；

同理可定义字符串 t,s 的非对称式前缀。

如果 z 在所有 s,t 的非对称式前缀集合中长度最长，那么则称 z 是字符串 s,t 的最长非对称式前缀。同理可定义字符串 t,s 的最长非对称式前缀。

如图 4-10 所示字符串 *“abcd”* 是字符串 *s* 和 *t* 的最长非对称式前缀，“*def*”是字符串 *t* 和 *s* 的最长非对称式前缀。本文在 KMP 算法[121]的基础上设计了求两个字符串最长非对称式前缀的算法。

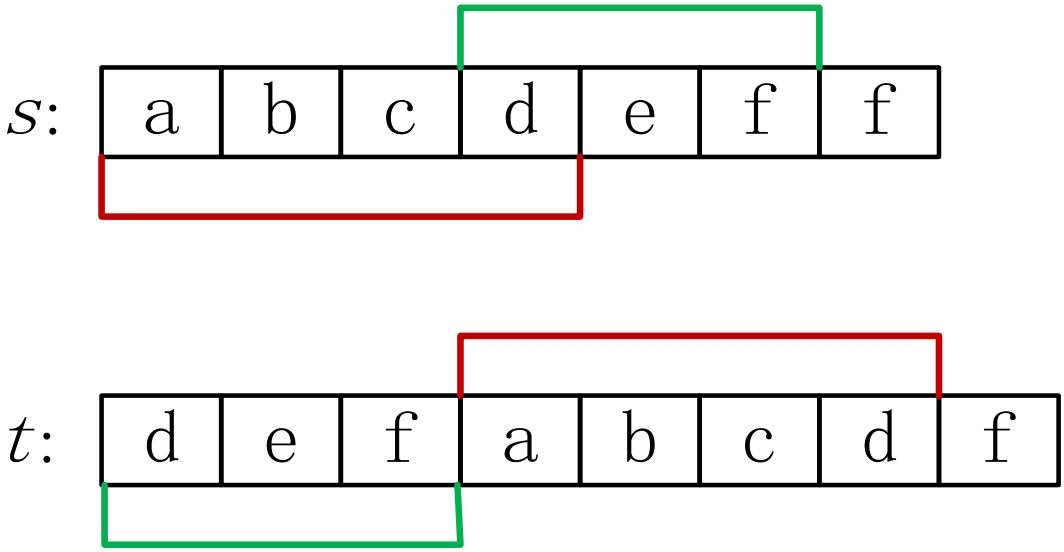


图 4-10 非对称式最大前缀

算法 4-1 求两个字符串的最长非对称式前缀

```
kmp::kmp(wstring s1,wstring s2)
{   m_target=s1;m_pattern=s2;
    int itssize=m_pattern.size()+1;
    prefixInfo=new int[itssize];memset(prefixInfo,-1,itssize*sizeof(int));
    itssize=m_target.size()+1;
    matchInfo=new int[itssize];memset(matchInfo,-1,itssize*sizeof(int));
}
void kmp:: ComputePrefixFunction()
{   int m=m_pattern.size()+1;
    const wchar_t*P=m_pattern.c_str();
    prefixInfo[0]=-1;int k=-1;
    for (int q=1;q<m;q++)
    {   while (k>-1 && P[k+1]!=P[q])
        {   k=prefixInfo[k];
        }
        if (P[k+1]==P[q])
        {   k++;
        }
        prefixInfo[q]=k;
    }
}
void kmp::KMPMatcher()
{   const wchar_t*P=m_pattern.c_str();const wchar_t*T=m_target.c_str();
    int m=wcslen(P)+1;int n=wcslen(T)+1;
    ComputePrefixFunction();
    int q=-1;//起始赋值为非法下标
    for (int i=0;i<n;i++)
    {   while(q>-1&&P[q+1]!=T[i])
        {   q=prefixInfo[q];
        }
        if(P[q+1]==T[i])
        {   q++;
            if (matchInfo[i]==-1)
            {   matchInfo[i]=q;
            }
        }
        if (q==m-1)
        {   q=prefixInfo[q];
        }
    }
}
void kmp::GetMatchInfo(vector< pair<int,int>>&matchinfo )
{for (int i=0;i<=m_target.size();i++)
    {   matchinfo.push_back(make_pair(i,matchInfo[i]));
    }
}
void kmp::GetLargestAsymmetricCommonPrefix(pair<int,wstring>&FPPInfo)
{   vector<pair<int,int>>info;wstring result;GetMatchInfo(info);
    vector<pair<int,int>>indexesInfo;vector<pair<int,int>>::iterator findit;
    vector<pair<int,int>>::iterator beginer=info.begin();
    do
    {   findit=find_if(beginer,info.end(),EqualInteger(0));
        if (findit!=info.end())
        {   pair<int,int>temp;temp.first=findit->first;
            beginer=findit;findit=find_if(beginer,info.end(),EqualInteger(-1));
            vector<pair<int,int>>::iterator assist;
            if (findit!=info.end())
            {   assist=findit-1;
            }
            else
            {   assist=info.end()-1;
            }
            temp.second=assist->first-temp.first+1;indexesInfo.push_back(temp);
        }
        beginer=findit;
    } while (beginer!=info.end());
    if(indexesInfo.size()==0)
    {   FPPInfo.first=-1;
        FPPInfo.second=L"";
    }
    else
    {   stable_sort(indexesInfo.begin(),indexesInfo.end(),isIntegerLarger);
        FPPInfo.first=indexesInfo[0].first;
        FPPInfo.second=m_target.substr(indexesInfo[0].first,indexesInfo[0].second);
    }
}
```

4.5.3.4 加边算法[24]

如算法 4-2 所示，主要有三条加边规则。第一条见算法 7-13 行，是说如果两个结点的单位字符串属性的最长公共子序列的长度与其中较小字符串长度之比大于一定阈值[25]（称之为主规则），并且两个结点的单位字符串属性之间的最长非对称式前缀大于 2（称之为辅助规则），则在两个结点之间加上一条边，本条规则依靠主规则计算单位字符串之间的相似性，借助辅助规则来避免主规则将类似于“华中理工大学”与“华南理工大学”的单位字符串归为相似的情况；第二条见算法 16-19 行，意思是如果第一条规则失效，接着检测两个结点的单位字符串的最长非对称式前缀中是否包含 RuleMain 中的模板，如果包含则在两个结点之间加上一条边；第三条规则见算法 22-25 行，意思是如果前两条规则都失效，则接着检测两个结点的单位字符串的最长非对称式前缀只能是否包含 RuleSupplement 模板，如果包含则在两个结点之间加上一条边。

算法 4-2 加边算法

```
1  lcs=LCS(AttriAddress(vi),AttriAddress(vj));
2  minlen=Min(len(AttriAddress(vi)),len(AttriAddress(vj)));
3  APrefix_ij=GetLargestAssymCommonPrefix(AttriAddress(vi),AttriAddress(vj));
4  APrefix_ji=GetLargestAssymCommonPrefix(AttriAddress(vj),AttriAddress(vi));
5  RuleMain=regex("(大学|研究院|研究所|研究中心)");
6  RuleSupplement=regex("(?!<中国科)学院");
7  if (lcs/minlen >thresholdl)
8  {
9      if(len(APrefix_ij)>2||len(APrefix_ji)>2)
10     {
11         Add an edge between vi and vj!
12     }
13 }
14 else
15 {
16     if(Match(RuleMain,APrefix_ij)||Match(RuleMain,APrefix_ji))
17     {
18         Add an edge between vi and vj!
19     }
20     else
21     {
22         if(Match(RuleSupplement,APrefix_ij)||Match(RuleSupplement,APrefix_ji))
23         {
24             Add an edge between vi and vj!
25         }
26     }
27 }
```

[24]运行加边算法之前，需要对单位字符串属性进行预处理，去掉逗号，分号，邮编等特殊字符，只留汉字为主的主要特征

[25]实验中此阈值设置为 0.9

4.5.3.5 基于图的连通分量的人名消歧算法

在依照算法 4-2 构造的图 G 上，每个连通分量分别对应于现实世界的一个作者实体，求出图上的所有连通分量，也即获得了实体指称项与实体之间的对应关系。本文以图的深度优先遍历算法[122]为基础设计的求无向图上所有连通分量算法。

4.5.4 算法有效性与局限性分析

4.5.4.1 人工观察分析

如附录 14 所示，进行同名消歧处理之前，数据库中共有 299,823 条作者记录，去重处理之后，数据库中共有 135,969 个作者实体。从数据量的规模变化上可以判定本论文提出的人名消歧算法是有一定的效果的。图 4-11 为针对图 4-1 中的名字“王伟”进行聚类消歧后的结果。由于共有 67 个聚类（见附录 15），这里仅仅节选了部分进行显示。从聚类结果来看，计算机较好地理解了“字面语义”，即通过单位字符串来识别作者是否是同一个人。值得指出的是：①如果某一人物的学术履历变迁情况通过其论文中登记的单位信息能够追溯得到，那么本论文中的算法不会因为机构变换而将同一个人物实体识别成多个人物实体，如图中聚类 6，但是如果这种履历变迁情况在论文登记的单位信息中无法追溯的话，本论文中的算法会识别成两个或者多个人物实体。②如果同一个机构在历史发展过程中存在大幅度的名称整改，那么本论文的算法可能会失效。如将“柴天佑 东北工学院”与“柴天佑 东北大学”识别成为两个不同的人物实体。

图 4-11 名字“王伟”的部分消歧结果

4.5.4.2 聚类效果评价分析

聚类评价中常用熵(entropy)、纯度(purity)[123][130]、归一化互信息(Normalized Information Gain,NMI)、准确率(Precision，P)、召回率(Recall，R)、F 值(F-measure，F)、兰德指数(Rand Index,RI)[129]来评估聚类算法的性能。

（一）熵的定义如下：

规模为n_r的簇C_r的熵为：

$$E(C_r) = -\sum_{i=1}^{q} \frac{n_r^i}{n_r} log \frac{n_r^i}{n_r} \qquad \text{公式 4-1}$$

其中q为标准答案中簇的数目，n_r^i表示簇r中属于标准答案簇i的文档数目。聚类结果的熵为聚类算法生成的每个簇C_r的熵的加权和：

$$Entropy = \sum_{r=1}^{k} \frac{n_r}{n} E(C_r) \qquad \text{公式 4-2}$$

其中 k 为聚类算法生成的簇数目，n 为总文档数。

（二）纯度的定义如下：

簇C_r的纯度用来衡量簇C_r中包含标准答案中单个簇的最大程度，定义如下：

$$P(C_r) = \frac{1}{n_r} max_i(n_r^i) \qquad \text{公式 4-3}$$

其中n_r表示簇C_r中的文档数目，$max_i(n_r^i)$表示簇C_r中含有标准答案中单个簇的最大文档数目。

聚类结果的纯度为聚类算法生成的每个簇C_r的纯度的加权和：

$$Purity = \sum_{r=1}^{k} \frac{n_r}{n} P(C_r) \qquad \text{公式 4-4}$$

其中 k 为聚类算法生成的簇数目，n 为总文档数。

（三）归一化互信息的定义如下：

$$NMI(W,C) = \frac{I(W,C)}{[H(W)+H(C)]/2} \qquad \text{公式 4-5}$$

其中，$W=\{\omega_1, \omega_2, \ldots, \omega_{k-1}, \omega_k\}$是聚类结果簇的集合，$C = \{c_1, c_2, \ldots, c_{q-1}, c_q\}$是标准答案簇的集合。

$$I(W,C) = \sum_k \sum_j \frac{|w_k \cap c_j|}{N} \log \frac{N|w_k \cap c_j|}{|w_k||c_j|} \qquad \text{公式 4-6}$$

其中 N 为文档数目，$|w_k|$、$|c_j|$、$|w_k \cap c_j|$分别为聚类结果簇w_k中的文章数目、标准答案中簇c_j中的文章数目以及在w_k和c_j中共同出现的文章数目。

$$H(W) = -\sum_k \frac{|w_k|}{N} log \frac{|w_k|}{N} \qquad \text{公式 4-7}$$

熵、纯度和归一化互信息是属于信息论范畴的聚类评价指标。如果把聚类问题看成一系列文档两两比较的决策过程，则可以定义准确率、召回率、F 值、兰德指数等指标。定义将两个相似的文档归为一簇是 true-positive(TP)决策；将两个相似的文档归为不同簇是 false-negative(FN)决策；将两个不相似的文档归为同簇是 false-positive(FP)决策；将两个不相似的文档归为不同簇是 true-negative(TN)决策。对 N 个文档进行聚类，共有$N_{TP}+N_{TN}+N_{FP+}N_{FN} = N(N-1)/2$个决策。

（四）准确率的定义如下：

$$Precision = \frac{TP}{TP + FP}$$ 公式 4-8

（五）召回率的定义如下：

$$Recall = \frac{TP}{TP + FN}$$ 公式 4-9

（六）F 值的定义如下：

$$F_b = \frac{(b^2 + 1)PR}{b^2 P + R}$$ 公式 4-10

其中 $b=1$。

（七）兰德指数的定义如下：

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$ 公式 4-11

（八）评价指标的样本加权平均公式定义如下：

$$avaEval = \sum \frac{N_m}{N_{total}} eval$$ 公式 4-12

其中 *eval* 代表评价指标，如熵、纯度、准确率、F 值等。N_m为每种样本的数目，如 45 条名字“白硕”相关的记录等，N_{total}为总样本个数。

熵、纯度、归一化互信息、准确率、召回率、F 值、兰德指数分别从不同的方面评价聚类算法的效果。熵衡量的是聚类算法生成的簇的混杂程度，因此熵值越接近 0 表示聚类算法的效果越好；而纯度衡量的是聚类算法生成的簇中包

含标准答案中单个簇的程度，所以纯度值越接近 1 表示聚类算法的效果越好；归一化互信息反映的是当已知聚类算法生成的簇后，对了解文档的真实类别的促进度，因此归一化互信息值越接近 1 表示聚类算法的效果越好；准确率、召回率、F 值、兰德系数衡量一系列聚类决策的准确程度和全面程度，因此也是值越接近 1 表示聚类算法的效果越好。

为了客观地评价本论文所设计的同名消歧算法的性能，本文针对“白硕”、“王斌”、“赵军”三个名字分别进行了手工标注，标注的基准是①常识知识（例如本人通过和中国科学院自动化所赵军老师交流得知，数据库中的“赵军 清华大学”与“赵军 中国科学院自动化所”是同一个人），②根据作者发表的论文标题、发表期刊、发表时间、论文关键词、论文摘要以及其所在单位进行综合分析判别。

“白硕”、“王斌”、“赵军”这三个名字所对应的共计 243 个条目的聚类结果和标注结果见附录 17、18、19，计算得到的熵、纯度、归一化互信息如表 4-3 所示；得到的准确率、召回率、F 值、兰德系数如表 4-4 所示；样本总体的加权平均熵、纯度、归一化互信息、准确率、召回率等如表 4-5 所示。

本论文算法所处理的原始题录信息数据库共有 45 个名字为“白硕”的实体指称项，并指向同一个人物实体。本论文算法在设计上以题录信息中作者填写的机构字符串为主要和唯一特征，而白硕老师在其学术生涯内先后就职于北京大学、国家智能计算机研究开发中心、中国科学院计算技术研究所、上海证券交易所等四家单位，所以针对名字“白硕”的评价结果也是本论文所设计的同名消歧算法的极限和下限。如表 4-3，表 4-4 所示，本论文所设计的同名消歧算法将名字为“白硕”的实体指称项归结为两个人物实体。聚类算法生成的簇的纯度是 1，熵是 0，表明每个簇内部具有高相似性；而归一化互信息为 0，表明聚类算法生成的簇对数据真实的类别情况不具有参考作用（事实上任何一个将原本的大簇划分成若干小簇的聚类算法都会得到类似的结果）；准确率为 1，而召回率、F 值、兰德指数等指标相对偏低，则表明一系列聚类决策的准确度不高。

从表 4-5 中提供的样本加权平均评测指标来看，本文所设计的同名消歧算法具有一定的普适性和实用价值，可以被应用于论文中提到的需求场景中。

表格 4-3 熵、纯度、归一化互相信息评价值

	熵	纯度	归一化互信息	文章数	生成簇数目	实际簇数目
白硕	0	1	0	45	2	1
王斌	0.1842	0.9390	0.9573	82	29	31
赵军	0.0390	0.9913	0.8985	116	25	20

表格 4-4 准确率、召回率、F 值、兰德指数

	准确率	召回率	F 值	兰德指数	决策次数
白硕	1.0000	0.6222	0.7671	0.6222	990
王斌	0.9519	0.9082	0.9295	0.9828	3321
赵军	0.9951	0.7987	0.8862	0.9376	6670

表格 4-5 样本加权平均评测指标

加权平均熵	加权平均纯度	加权平均归一化互信息	加权平均准确率	加权平均召回率	F 值	加权平均兰德指数
0.0808	0.9753	0.7963	0.9814	0,8030	0.8788	0.8945

4.6 一级机构名称识别与抽取

4.6.1 需求分析

文献计量需要统计各科研机构的科研成果，机构内部的工作人员的科研成果和合作关系等，以及不同科研机构之间的合作关系等。为此必须要获得诸如“清华大学”、“北京大学”等一级机构名称。本论文机构数据情况如图 4-12 所示，由于作者单位的写法存在许多不规范，比如一级单位名称和单位内部的子机构名称之间或存在分割符或不存在分隔符，一级单位名称可能出现在单位的内部子机构名称前面，也可以出现在子机构名称后面等，所以需要设计识别和抽取一级机构名称的算法。

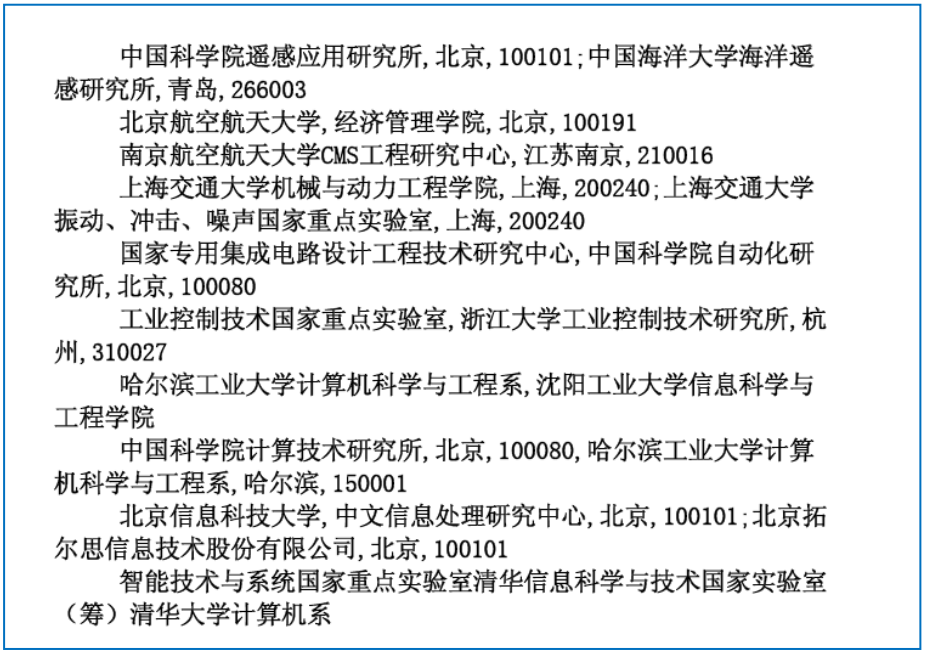

中国科学院遥感应用研究所, 北京, 100101;中国海洋大学海洋遥感研究所, 青岛, 266003
北京航空航天大学, 经济管理学院, 北京, 100191
南京航空航天大学CMS工程研究中心, 江苏南京, 210016
上海交通大学机械与动力工程学院, 上海, 200240;上海交通大学振动、冲击、噪声国家重点实验室, 上海, 200240
国家专用集成电路设计工程技术研究中心, 中国科学院自动化研究所, 北京, 100080
工业控制技术国家重点实验室, 浙江大学工业控制技术研究所, 杭州, 310027
哈尔滨工业大学计算机科学与工程系, 沈阳工业大学信息科学与工程学院
中国科学院计算技术研究所, 北京, 100080, 哈尔滨工业大学计算机科学与工程系, 哈尔滨, 150001
北京信息科技大学, 中文信息处理研究中心, 北京, 100101;北京拓尔思信息技术股份有限公司, 北京, 100101
智能技术与系统国家重点实验室清华信息科学与技术国家实验室（筹）清华大学计算机系

图 4-12 机构数据情况

4.6.2 相关工作

机构名称抽取，首先是边界识别问题，然后才是抽取问题。文献[125]通过分析和总结大学、学院类中文机构名称的构成规则，采用 prolog 实现规则模板对香港理工大学 600 万字三地语料中的 850 个大学、学院类机构名称进行实验，得到准确率、召回率分别为 84.8%和 84.6%。文献[124]事先从教育部全国普通高校名单、万方数据-企业信息网、中央企业名录中以手工统计方式获取机构规范表，然后此表对万方数据库中的作者单位字符串进行规范化处理。

4.6.3 算法设计

4.6.3.1 算法设计思路

如图 4-12 所示，原始机构名称数据由于缺乏统一的书写格式而混乱不堪。通过观察可以发现同一个人物实体的机构数据中总有一定的规律，如图 4-13、图 4-14 所示，因此可以借助于 4.5.3 节中人名消歧算法的结果。借助于 4.5.3.3 节中提出的最长非对称式前缀抽取算法，可以从同一人物的所有机构串中抽取出现频率最高的最长非对称前缀作为候选的一级机构名称字符串。至此完成了边界识别工作。之后采用模板优先队列算法从候选一级机构名称字符串中匹配

到一级机构名称，至此完成了机构名称的抽取工作。考虑到同一个人物实体可能因为工作变迁或者兼职而涉及多个工作单位，如果这些信息在题录数据库中有所体现，为了不损失掉这些信息，在采用最长非对称式前缀算法抽取之前，依然采用类似于 4.5.3 小节的人名消歧的思路，对同一个人物实体的机构字符串进行聚类，然后针对每个聚类抽取最高频的非对称式前缀作为候选的一级机构名称字符串。

	AuthorName	Alias	AuthorWorkPlace
2	马少平	Ma Shaoping	清华大学计算机科学与技术系智能技术与系统国家重点实验室,北京,100084
3	马少平		清华大学,计算机科学与技术系,北京,100084,清华大学,智能技术与系统国家重点实验室,北京,100084
4	马少平		智能技术与系统国家重点实验室,北京,100084
5	马少平		清华大学
6	马少平	MA Shao-ping	智能技术与系统国家重点实验室,清华信息科学与技术国家实验室,清华大学计算机系,北京,100084
7	马少平	MA Shao-ping	智能技术与系统国家重点实验室,清华信息科学与技术国家实验室(筹
8	马少平	MA Shao-ping	智能技术与系统国家重点实验室,清华信息科学与技术国家实验室(筹),清华大学计算机系,北京,1000...
9	马少平	MA Shao-ping	清华大学,计算机科学与技术系,北京,100084
0	马少平	MA Shao-ping	清华大学,智能技术与系统国家重点实验室,北京,100084
1	马少平	MA Shao-ping	清华大学,智能技术与系统国家重点实验室,北京,100084
2	马少平		清华大学智能技术与系统国家重点实验室,北京,100084
3	马少平		清华大学计算机科学与技术系智能技术与系统国家重点实验室,北京,100084
4	马少平		清华大学计算机系智能技术与系统国家重点实验室,北京,100084
5	马少平	MA Shao-ping	清华大学,计算机系,智能技术与系统国家重点实验室,北京,100084
6	马少平		清华大学,计算机系智能技术与系统国家重点实验室,北京,100084
7	马少平		清华大学,计算机科学与技术系,北京,100084,清华大学,智能技术与系统国家重点实验室,北京,100084
8	马少平		清华大学计算机科学与技术系智能技术与系统国家重点实验室,北京,100084
9	马少平		清华大学计算机系,智能技术与系统国家重点实验室,北京,100084
0	马少平		清华大学计算机科学与技术系智能技术与系统国家重点实验室,北京,100084
1	马少平		清华大学计算机科学与技术系,
2	马少平		清华大学
3	马少平	Ma Shaoping	清华大学计算机科学技术系;
4	马少平		清华大学智能技术与系统国家重点实验室

图 4-13 人物实体“马少平”对应的机构数据

	AuthorName	AuthorWorkPlace
7	刘群	中国科学院计算技术研究所数字化技术研究室,北京,100080
8	刘群	中国科学院计算技术研究所,北京,100080
9	刘群	中国科学院,计算技术研究所,智能信息处理重点实验室,北京,100190
10	刘群	中国科学院,计算技术研究所智能信息处理重点实验室,北京,100190
11	刘群	中国科学院,计算技术研究所,智能信息处理重点实验室,中国,北京,100190
12	刘群	中国科学院,智能信息处理重点实验室,北京,100080
13	刘群	中国科学院,计算技术研究所,北京,100080
14	刘群	中国科学院,计算技术研究所,北京,100080;中国科学院,智能信息处理重点实验室,北京,100080
15	刘群	中国科学院,计算技术研究所,北京,100080
16	刘群	中国科学院,计算技术研究所,北京,100080
17	刘群	中国科学院,计算技术研究所,智能信息处理重点实验室,北京,100080
18	刘群	中国科学院,计算技术研究所,多语言交互技术评测实验室,北京,100080
19	刘群	中国科学院,计算技术研究所,北京,100080
20	刘群	中国科学院,计算技术研究所,数字化技术研究室,北京,100080
21	刘群	中国科学院,计算技术研究所,北京,100080
22	刘群	北京大学,计算语言学研究所,北京,100871;中国科学院计算技术研究所,北京,100080
23	刘群	中国科学院计算技术研究所
24	刘群	中国科学院计算所二室;
25	刘群	中国科学院计算技术研究所,北京,100080
26	刘群	中国科学院计算技术研究所,北京,100080
27	刘群	中国科学院计算技术研究所,北京,100080
28	刘群	中国科学院计算技术研究所软件实验室,北京,100080
29	刘群	中国科学院计算技术研究所智能信息处理重点实验室;内蒙古大学计算机学院;中国科学院研究生院;
30	刘群	中国科学院计算技术研究所;中国科学院计算技术研究所 北京;

图 4-14 人物实体“刘群”对应的机构数据

4.6.3.2 算法流程

图 4-15 为一级机构名称识别与抽取的总流程。其中的“粗分割”处理过程是将原始的机构名称字符串按照半角分号进行分割；“聚类”过程可参考 4.5.3；模板优先队列的正则表达式描述形式如下：

⑴中国科学院.*?所；

⑵中国科学院.*?院；

⑶中国科学院.*?中心；

⑷中国科学院.*?台；

⑸^.*?大学；

⑹^.*?学院；

⑺^.*?学校；

⑻^.*?研究院；

⑼^.*?院；

⑽^.*?公司；

⑾^.*?厂；

⑿^.*?部队；

⒀^.*?所；

⒁^.*?中心；

⒂^.*?局。

模板优先队列的存储数据结构采用最大堆来实现[126]。在处理方式上，每个一级机构名称候选串顺次匹配模板⑴到⒂，直到第一个可以匹配上的模板为止。如“北京邮电大学信息与通信工程学院”首先匹配上“⑸^.*?大学”，则不再后续匹配“⑹^.*?学院”。这样即可保证抽取的机构名称都是一级机构名称。

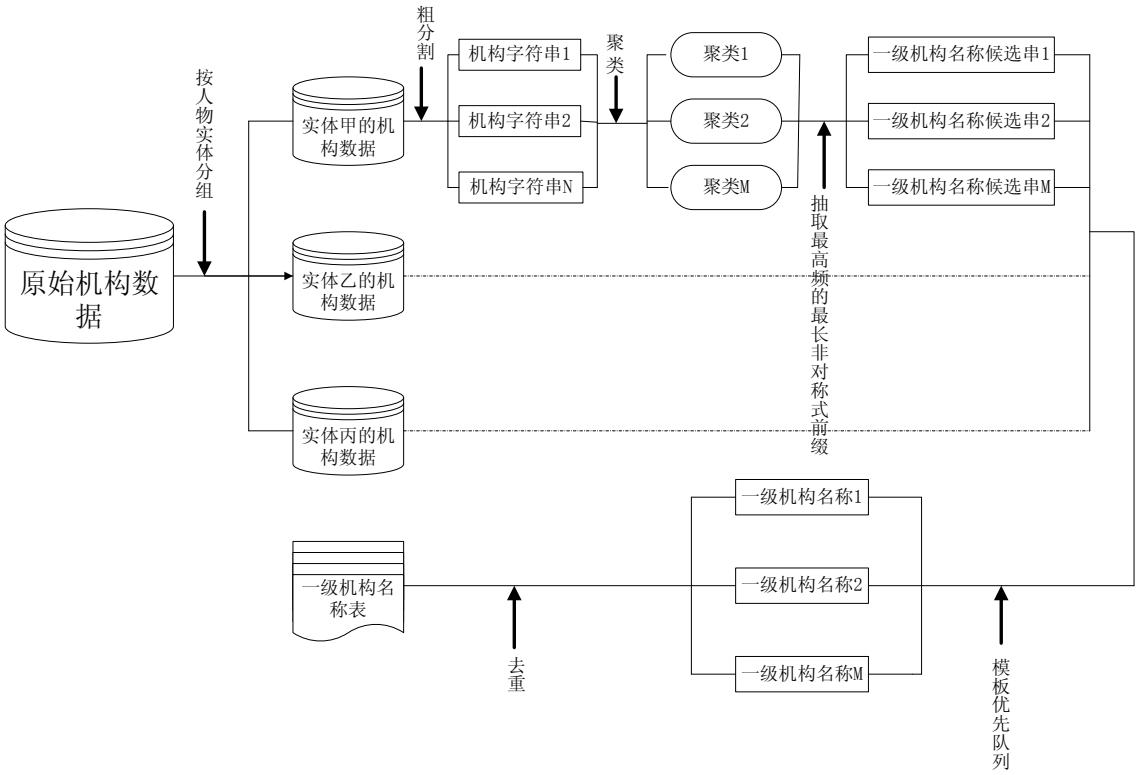


图 4-15 一级机构名称识别与抽取流程图

4.6.4 算法有效性与局限性分析

图 4-16、4-17 分别为“抽取最高频率的最长非对称式前缀作为一级机构名称候选串”算法的程序运行结果截图。分对比图 4-13、图 4-14 来看，本文设计的算法在很大程度上规范了原始机构数据，同时本文所设计的算法与文献[124]

和文献[125]相比，既不需要具有深厚语文素养的专家来总结复杂的规则模板，又不需要人工事先归纳整理出机构名称规范表，仅依靠数据源的自学习、自消歧的能力和简单的规则模板，因此可以极大减少人工工作量，从而降低人工成本。

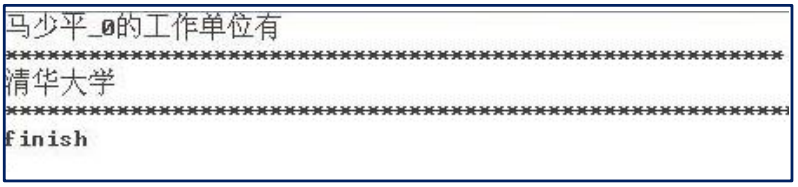


图 4-16 人物实体“马少平”的一级机构名称候选串

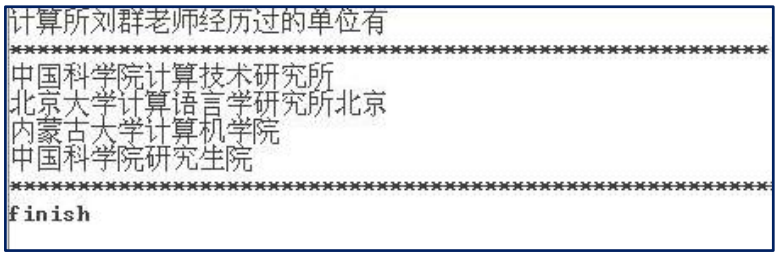


图 4-17 人物实体“刘群”的一级机构名称候选串[26]

如附录 14 所示，原始机构数据库中 60,874 个机构名称，经过算法处理后，共抽取出一级机构名称 10，865 个。由于一些机构在发展过程中存在名字变换、机构合并的情况，采用通过人工跟踪百度百科的知识的方式，对我国重点科研院所进行了手工排歧。今后可以考虑利用搜索引擎计算两个机构名称之间的检索相似度自动合并和排歧的方法[127]。

4.7 本章小结

传统的知识库构建方法主要依赖于人工操作，但是由于认知偏差、缺乏专家知识、或者疲劳操作等原因，即便是人工构建的知识库也很难保证不出差错。[128]中指出即使经过历时数月的充分训练，人工操作依然有 30%左右的错误率。尤其在大数据时代，互联网本身就是一个庞大的信息源，企图依靠人工方式从海量、异构、冗余、不规范、且含有大量噪声的网页信息中有效抽取知识，更

[26]经和刘群老师沟通得知，刘群老师并未在“内蒙古大学计算机学院”工作过，本人进一步观察题录信息原始数据发现是由于人物机构无法对齐导致的错误。在人物机构对齐算法设计一节本文提到过，由于数据源情况比较复杂，一些无法对齐的数据条目，本文采取不进行处理，直接保留的策略。

是难上加难。本文采用信息抽取技术，从半规则数据中抽取细粒度知识，进而自动构建知识库，为知识服务系统提供基础支持，极大降低了人力成本开支。无论从思维角度还是方法角度，本论文章节所设计的算法和方案都具有新颖性、有效性、实用简单的特点。此外，本论文章节所设计的算法和方案在数字图书馆领域具有一定的可行性和工业应用价值。

本论文章节的部分成果已经以开源知识库或者语料的形式发布，详情请见附录 8-12，可以通过访问自动化学科知识服务网络平台来了解本论文章节介绍的算法的效果，也可以通过下载和使用本文发布的语料和知识库资源来增加对本文工作的了解。

第五章 总结与展望

文本挖掘技术的处理对象是文本数据，旨在从海里文本数据资源中，获得有价值的信息或知识。本论文所述文本挖掘系统，面向特定领域和特定任务（自动化学科中文期刊论文题录信息），以应用为导向，涉及图 5-1 中所示的理论和技术。

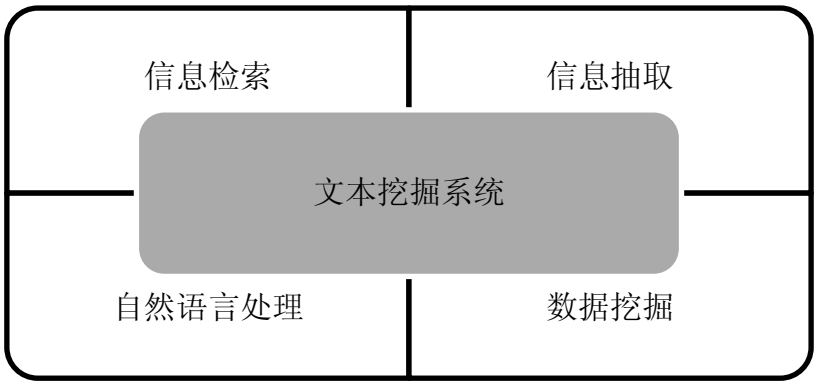


图 5-1 文本挖掘系统技术基础

5.1 工作总结

本文在本挖掘系统设计中，一方面兼顾文本挖掘算法的可移植性和泛化能力，一方面又考虑到具体问题的特殊性，在统计方法框架内揉进最具代表性，最行之有效且最具有自适应能力的启发式规则。本论文所设计的文本挖掘系统，可以直接用于计算机、电子、通信等领域的中文期刊题录信息的分析中；更换部分启发式规则可以用于对英文文献题录信息的分析中。

本文在算法论述中力求思路严谨，不仅给出算法的设计背景、设计思路、核心算法的伪代码描述形式，而且给出算法的有效性和局限性分析。其中部分算法给出数学证明；部分算法对启发式规则的设计做重点阐述；部分算法通过人工标注数据和对比实验的方式给出客观的评价。与理论研究侧重新颖性的理念有所不同，工程实践力求切实解决实际问题和选择最优解决方案，本论文中所涉及的实验可分为验证性实验和算法实验两部分。验证性实验的目标是验证已有算法是否适合在具体任务场景中应用；算法实验是对本文提出的算法做出评价。由于本论文中的实验针对项目数据进行，无公开评测语料集支持，所以

附录中对实验数据和人工标注的数据答案等进行了详细说明，此外，附录还包括对本人在毕设工作中所做的开源工作的介绍，以及没有纳入正文的技术方案等。

本文提出的基于卡方拟合优度的特征词选择算法(chifit)和传统特征词选择算法相比，能在低特征维度获得较高的分类正确率；设计的人名消歧算法平均准确率为 98.14%；设计的关键词聚类算法平均准确率为 92.14%；设计的人物机构对齐算法平均准确率为 87.6837%；设计的机构名称抽取与归一化算法减少了82.15%的冗余数据。

本论文所述算法系统面向实际应用，由于数据分布不均匀，算法中加入启发式规则，算法系统内部模块耦合级联等原因，不宜采用随机抽取小量数据进行人工标注，在标注数据集上测试算法系统的整体性能的方案。这里采用用户反馈的方式对算法系统的整体性能进行评定。

①如表 5-1 所示，文本挖掘系统大大降低了冗余信息以及歧义信息，提高了信息精度，减少了人工工作量。

表格 5-1 原始数据与处理后数据的对比

原始数据	处理后的知识数据
题录信息条目（116,642/58,235）	题录信息条目 109,788
关键词中英对照组条目(148,825)	关键词对照组条目(83,602)
作者条目（299,823）	作者条目（135，969）
机构条目（60,874）	机构条目（10，865）

②自动化学科知识服务网络平台(http://autoinnovation.ia.ac.cn)于 2011 年 11 月正式对外访问，收到了自然语言处理、数据挖掘、信息检索等相关领域研究者的反馈，大家对知识精度给予了肯定。

③为满足文献情报学研究者的数据需求，本文将部分实验数据以及算法的结果数据进行开源，地址在 http://www.datatang.com/member/5878，目前已有 500 余次的下载量，收到了良好的用户反馈。

5.2 工作展望

虽然本文的工作在文本分类的特征选择以及数字图书馆中的关键词语义聚类、作者人名消歧、机构名称规范化等方面取得了一些成果，但和很多已有方

法一样，它们或多或少的存在局限，还有很多问题和工作值得进一步讨论，主要包括以下几点：

①同名作者消歧算法，以就职机构为主要区分特征，因此无法识别同机构人物的同名问题，比如本文作者与刘禹老师同时就职于中国科学院自动化所；当同一个人物更换工作单位且从题录数据中无法挖掘到履历变迁时，或者人物的工作单位名称发生较大的变化时，会将同一个作者识别成多个作者。以上两个问题，考虑在现有算法框架内融入更多的特征，如学者共发文关系，同机构关系，论文发表刊物，论文发表时间等进一步改进。

③关键词语义聚类算法的设计初衷是为了实现术语汉英对照词典的自动化编撰功能，所以词语聚类的标准为语义相同，聚类粒度过小。知识族谱算法直接将该算法的聚类结果作为输入，因此出现了“机器学习算法”、“机器学习”成为族谱上不同结点的情况。可以考虑在构建知识族谱之前，利用话题模型对关键词再进行一次粗粒度的聚类。

④知识族谱的结构形式还比较固定化和单一，可以考虑利用有向图的有向隔离理论和马尔科夫毯理论在知识族谱的形式上进行改进。

⑤本文只能获得文献的题录信息，无法获得论文的全文内容。因此无法做更多的深入工作。希望能有机会与学术情报领域内的其他研究机构和产业机构进行多边合作，提供技术支持和技术输出。希望本文的研究和实践结果能对学术情报领域、学术知识服务领域的研究具有一定价值。

参考文献

[1] 张晓林.走向知识服务：寻找新世纪图书情报工作的增长点.中国图书馆学报,2006,26(129):32-37

[2] 吴学梯,周元.方法、创新、转变.北京:高等教育出版社,2011.

[3] 朱庆华.日本情报学研究的新动向---国立情报学研究所成立综述.图书情报工作,2001,5

[4] 刘二稳,史建华.Google Scholar 搜索工具的使用及功能简介.山东建筑大学学报,2009,24(3)

[5] Jie Tang, Jing Zhang, Limin Yao, et al. Arnetminer: Extraction and Mining of Academic Social Networks. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining,2008:990-998

[6] Ronen Feldman, James Sanger. The TextMining Handbook-Advanced Approaches In Analyzing Unstructured data.UK:Cambridge Uinversity, 2007:189-313

[7] http://www.w3school.com.cn/htmldom/dom_nodes.asp

[8] Christopher D.Manning, Prabhakar Raghavan.信息检索导论, 王斌.北京：人民邮电出版社,

[9] http://deerchao.net/tutorials/regex/regex.htm

[10] Jeffrey E.F.Friedl.精通正则表达式,余晟.北京：电子工业出版社, 2009

[11] Thomas H.Cormen, Charles E. Leiserson, Ronald L.Rivest, et al.Introduction to Algorithms. MITPress, 2002:350-356

[12] M.E.MARON Automatic indexing: an experimental inquiry. Journal of the ACM, 1967, 8(3): 404-417

[13] Hayes, P.J., Knecht, et al. A News Story Categorization System.In Proceedings of ANLP-88, 2nd Conference on Applied Natural Language Processing. Austin, JX, Association for Computational Linguistics, Morristown, 1988, NJ:9-17

[14] Hayes, P.J.,Weinstein, et al. A System for Content-Based Indexing of a Database of News Stories. In Proceedings of IAAI-90,2nd Conference on Innovative Applications of Artificial Intelligence.Boston,AAAI Press, Menlo Park,1990,CA:49-66

[15] Hayes, P.J., Andersen. Tcs: A Shell for Content-Based Text Categorization. In Proceedings of CAIA-90, 6th IEEE Conference on Artificial Intelligence Applications. Santa Barbara, CA, IEEE Computer Society Press, Los Alamitos, 1990, CA:320-326

[16] Hayes, P. Intelligent High-Volume Processing Using Shallow, Domain-Specific Techniques. In Text-Based Intelligent Systems: Current Research and Practice in

Information Extraction and Retrieval. P. S. Jacobs, ed .Hillsdale, NJ, Lawrence Earlbaum: 1992, 227-242

[17] Fabrizio Sebastiani.Machine Learning in Automated Text Categorization.ACM Computing Surveys,2002,34(1):1-47

[18] Vapnik.Statistical Learning Theory.US:John Wiley&Sons,1998

[19] Richard O.Duda,Peter E.Hart,David G.Stork.模式分类,李宏东,姚天翔.北京：机械工业出版社，中信出版社，2007,374-375

[20] Andrew Y. Ng, Michael I. Jordan. Discriminative vs generative classifier. A comparison of logistic regression and naiva bayes. NIPs, 2002

[21] Ronen Feldman, James Sanger. The Text Mining Handbook -Advanced Approaches In Analyzing Unstructured data. UK: Cambridge Uinversity, 2007:189-31

[22] Lan H. Witten, Eibe Frank. Data Mining Pratical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005, 404-405

[23] Yang Y, Pedersen J O.A comparative study on feature selection in Text Categorization. Proc of ICML'97. SanFrancisco:Morgan Kaufmann,1997:412-420

[24] Gerard Salton, Christopher Buckley. Term-weighting approaches in automatic text retrieval. Inform.Process, 1988, 24(5): 513-523

[25] Gerard Salton, The SMART Retrieval System –Experiments in Automatic Document Processing. Prentice-Hall. [122, 159, 177, 453, 461, 462]

[26] Singhal, Salton, Buckley. Length normalization in degraded text collections. Technical report, Cornell University, Ithaca, NY:[123]

[27] Christopher D.Manning,Prabhakar Raghavan.信息检索导论,王斌.北京：人民邮电出版社,2010:89-89

[28] Gerard Salton, WONG, YANG. A vector space model for automatic indexing. Commun. ACM18(11):613-620

[29] Liu Huan and Hiroshi Motoda. Feature selection for Knowledge Discovery and Data Mining. US, Kluwer Academic Publishers,1998.

[30] Luigi Galavotti, Fabrizio Sebastiani, Maria Simi.Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization.ECDL'00 Proceedings of the 4th European Conferecn on Research and Advanced Technology for Digital Libraries, 2000, 60-68

[31] Hwee Tou Ng, Wei Boon Goh, Kok Leong Low, Feature selection, perceptron learning, and a usability case study for text categorization. SIGIR'97 NewYork USA,1997, 67-73

[32] Shoushan LI, Chengqing ZONG. A New Approach to Feature Selection for Text

Categorization. Proceeding of NLP-KE’ 05, 626-630

[33] Yan Xu, Gareth Jones, JinTao Li, et al. A Study on Mutual Information-based Feature Selection for Text Categorization. Journal of Computational Information System, 2005, 203-213

[34] 李国臣.文本分类中基于对数似然比测试的特征词选择算法.中文信息学报,1999,13(4):16-21

[35] 一种高效的用于文本聚类的无监督特征选择算法

[36] 周荫清.信息理论基础.北京,北京航空航天大学出版社 2005:11-12

[37] 盛骤.概率论与数理统计.北京,高等教育出版社 2006:213-213

[38] George Casella, Roger L. Berger. Statistical Inference. China Machine Press, 2002

[39] V.I.Levenshtein .binary codes capbable of correctingdeletions,insertions, and revesals

[40] Christopher D.Manning, Prabhakar Raghavan. 信息检索导论, 王斌. 北京：人民邮电出版社, 2010:40-41

[41] Graham Cormode, S. Muthukrishnan. The String Edit Distance Matching Problem with moves. Proceedings of the 13 ACM-SIAM on Discrete algorithms. 2002:667-676

[42] Kemal oflazer. Error-tolerant Finite-state Recognition with Applications to Morphological Analysis and Spelling Correction .Computational Linguistics, 1996, 22(1)

[43] 车万翔, 刘挺, 秦生, 等. 基于改进编辑距离的中文相似句子检索. 高技术通讯, 2004, 7:15-20

[44] http://en.wikipedia.org/wiki/Damerau%E2%80%93Levenshtein_distance

[45] 刘群, 李素建. 基于《知网》的词汇语义相似度计算.第三届汉语词汇语义学研讨会, 台北, 2002

[46] 张智熊. 信息抽取技术及其在数字图书馆中的应用前景分析 .数字图书馆, 2004, 6:1-6

[47] 王婷婷.《新华字典》姓氏人名条目研究. 牡丹江教育学院学报, 2011

[48] 徐铁生.《现代汉语词典》(第 5 版)姓氏条目平议. 辞书研究, 2009, 2

[49] 《中国姓氏统计》http://bbs.huanqiu.com/thread-71314-1-1.html

[50] 郑淑花. 汉语人名常用字的统计分析. 皖西学院学报, 2010, 26(1):113-116

[51] Jim Cowie. Wendy Lehnert. information extraction. Communication of the ACM, 1996, 89(1):80-91

[52] Information filtering and information retrival: Two sides of the same coin?. Communications of the ACM, 1992, 35(12): 1-10

[53] 李保利, 陈玉忠, 俞士汶. 信息抽取研究综述

[54] Ronen Feldman, James Sanger. The Text Mining Handbook-Advanced Approaches In Analyzing Unstructured data.UK: Cambridge Uinversity, 2007: 105-105

[55] 赵军.命名实体识别、消歧和跨语言关联.中文信息学报, 2009, 23(2): 3-17

[56] 赵军.研究生院讲座:信息提取

[57] 赵军.开放式文本信息抽取.中文信息学报, 2011, 25(6)

[58] DeJong, G.F.Prediction and substantiation: A new approach to natural language processing. Cognitive Sci, 3 (1979): 251-273.

[59] Dejong,G.F. An overview of the FRUMP systemin Wendy G. Lehnert and Martin H. Ringle (eds.), Strategies for Natural Language Processing, 1982

[60] Ralph Grishman, Beth Sundhem. MessageUnder-standing Coference-6: Abrief History

[61] http://www.itl.nist.gov/iad/mig/tests/ace/[OL]

[62] http://www.nist.gov/tac/[OL]

[63] Soderland, S. Learning Information Extraction Rules for Semi-Strutured and Free Text.Machine Learning, 1999, 34(1-3): 233-272

[64] Freitag, D., Kushmerick.Boosted Wrapper Induction. In Proceedings of AAAI, 2000, Austin: 577-583

[65] Ciravegna, F..Adaptive Information Extraction from Text by Rule Induction and Generalization. In Proceedings of the 17th IJCAI, Seattle, 2001: 1251-1256

[66] A.Lavelli, M.E.Califf, F.Ciravegna, et al. A Critical Survey of the Methodology for IE Evaluation. Proceedings of LREC2004

[67] Gina-Anne Levow.T he Third International Chinese Language Processing Bakeoff: WordSegmentation and Name Entity Recognition, Proceedings of 5th SigHan Workshop, 2006: 108-117

[68] http://crfpp.sourceforge.net/

[69] http://homepages.inf.ed.ac.uk/lzhang10/maxent.html

[70] http://gibbslda.sourceforge.net/

[71] http://www.cs.princeton.edu/~blei/lda-c/

[72] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[73] YUFENG CHEN, CHENGQING ZONG. A Structure-Based Model for Chinese Orgnization Name Translation. ACM Transactions on Asian Language Information Processing, 2008, 7(1): 1-30

[74] Min Zhang, Haizhou Li, Jian Su, et al. A phrase-based context-dependent joint probability model for named entity translation.IJCNLP2005, 600-611

[75] Tadashi Kumano, Hideki Kashioka, Hideki Tanaka, et al. Acquiring bilingual named entity

translatiions from content-aligned corpora, IJCNLP, 2004

[76] Robert C. Moore. Learning translation of named-entity phrases from parallel corpora. EACL, 2003

[77] Conrad Chen, Hsin-Hsi Chen. A High-Accurate Chinese-English NE Backward Translation System Combing Both Lexical Information and Web statistics. Proceedings of the COLING/ACL2006, 81-88

[78] Ying Zhang, Fei Huang, Stephan. Mining translations of OOV terms from the web though cross-lingual query expansion. SIGIR'05

[79] Fan Yang, Jun Zhao, Kang liu. A Chinese-English Organization Name Translation System Using Heuristic Web Mining and Asymmeric Alignment.Proceedings of 47th Annual Meeting of ACL and the 4th IJCNLP,2009:387-395

[80] Abney, Steven . Parsing by chunks. Kluwer Academic Publishers, 1991 :257–278

[81] Kuhn ,H. The Hungarian method for the assignment problem. Naval Rese. Logist, Quart 2, 83-97

[82] James Munkres.Algorithms for the assignment and transportation problems. Jouranl of the Society for Industrial and Applied Mathematics, 1957, 5(1): 32-38

[83] http://www.ldc.upenn.edu/

[84] Neil R. Smalheiser, Vetle I. Torvik. AuthorName Disambiguation Surey. ARIST, 2009(43)

[85] Open Researcher&ContributorID(ORCID)[EB/OL].[2010-11-11].http://www.orcid.org/

[86] 韩先培.基于语义知识挖掘与融合的实体消歧技术研究.[博士学位论文].北京:中国科学院自动化研究所

[87] http://nlp.uned.es/weps/

[88] http://www.cipsc.org.cn/clp2010/cfp.htm

[89] Brendan J. Frey, Delbert Dueck. Clustering by Passing Messages Between Data Points. SCIENCE, 315(16), 2007: 972-951

[90] Supporting Online Material for Clustering by Passing Messages Between Data Pointshttp://www.sciencemag.org/cgi/content/full/1136800/DC1

[91] Delbert Dueck. AFFINITY PROPAGATION: CLUSTERING DATA BY PASSING MESSAGES. [doctor thesis]. Canada:Graduate Department of Electrical&Computer Engineering University of Toronto

[92] Martin Rosvall, Carl T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. PNAS, 2007

[93] Ulrike Luxburg. A Tutorial on Spetral Clustering. Journal Statistics and Computing, 17(4),

2007

[94] Michelle Girvan, M. E. J. Newman. Community structure n social and biological networks. 2001

[95] Ellis L.Johnson, Anuj Mehrotra, George L. Nemhauser. Min-cut clustering. Mathimatical programming 1993, 62(1-3), 133-151

[96] Gary William Flake, Robert E. Tarjan , Kostas Tsioutsiouliklis. Graph Clustering and Minimum Cut Trees

[97] Pattern ClassificationRichard O.Duda, Peter E. Hart, David G. Stork. 模式分类, 李宏东, 姚天翔. 北京：机械工业出版社，中信出版社，2007:455-455

[98] Oleksandr Grygorash, Yan Zhou, Zach Jorgensen. Minmum Spanning Tree Based Clustering Algorithms. ICTAI'06, 2006

[99] Amit Bagga, Breck Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. COLING'98, 1998, 79-85

[100] Ted Pedersen, Amruta Purandare, Anagha Kulkarni. Name Discrimination by Clustering Similar Contexts. CICLing'05

[101] Gideon S.Mann, David Yarowsky. Unsupervised Person Name Disambiguation.

[102] Einat Minkov, William W. Cohen, Andrew Y. Ng. Contextual Search and Name Disambiguation in Email Using Graphs. SIGIR2006

[103] Ron Bekkerman, Andrew McCallum. Disambiguatiing Web Appearance of People in Social NetWork. WWW, 2005, 463-470

[104] Minoru Yoshida, Masaki lkeda, Shingo Ono, et al. Person Name Disambiguation by Bootstrap. sigir2010, 2010, 10-17

[105] Masaki lkeda, Shingo Ono, lssei Sato.PersonName Disambiguation on the Web by TwoStage Clusteringe.WWW2009

[106] 丁海波,肖桐,朱靖波.基于多阶段的中文人名消歧聚类技术的研究.第六届全国信息检索学术会议论文集,2010

[107] Silviu Cucerzan.Large-Scale Named Entity Disambiguation Based on Wikipedia Data. EMNLP-CoNLL, page 708-716

[108] Xiaoming Fan, Jianyong Wang, Xu Pu, et al. On Graph-Based Name Disambiguation. JDIQ ,2011, 2(2)

[109] Theresa Velden, Asif-ul Haque, Carl Lagoze. A new Approach to Analyzing Patterns of Collaboration in Coauthorship Networks-Mesoscopic Analysis and Interpretation.

[110] Indrajit Bhattacharya,Lise Getooor.A Latent Dirichlet Model for Unsupervised Entity

Resolution.SIAM International Conference on Data Mining (2006)

[111] Andrew M.Dai,Amos J.Storky.Author Disambiguation: A Nonarametric Topic and Co-authorship Model.In Proceedings of the NIPS Workshop on Applications for Topic Models:Text and Beyond,Canada,2009

[112] Ricardo G. Cota, Marcos Andre Goncalves, Alberto H. F. Laender. A Heuristic-based Hierarchical Clustering Method for Author Name Disambiguation in Digital Libraries, SBBD 2007:20-34

[113] Vetle I.Torvik, Marc Weeber, Don R.Swanson, et al. A probabilistic Similarity Metric for Medline:Records A Model for Author Name Disambiguation. Journal of the American Society for Information Science and Technology, 56(2), 2005: 140-158

[114] Jie Tang, A.C.M. Fong, Bo Wang, et al. A unified Probabilistic Framework for Name Disambiguation in Digital Library. IEEE Transactions on Knowledge and Data Engineering, 2011

[115] Hui Han, Hongyuan Zha, C. Lee Giles, Name Disambiguation in Author Citations using a K-way Spectral Clustering Method.JCDL, 2005, 334-343

[116] Denilson Alves Pereira, Berthier Ribeiro-Neto, Nivio Ziviani. Using Web Information for AuthorName Disambiguation.JCDL'09, 2009 :49-58

[117] Pallika H. Kanani, Andrew McCallum. Efficient Strategies for Improving Partitioning -Based Author Coreference by Incorporatiing Web Pages as Graph Nodes. AAAI, 2007

[118] Saurabh Kataria, K. Kumar, R. Rastogi, et al. Entity Disambiguation with Hierachical Topic Models, KDD2011

[119] BRADLEY EFRON. The jackknife, the bootstrap, and other resampling plans.

[120] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, et al. Introduction to Algorithms .MITPress, 2002: 350-356

[121] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, et al. Introduction to Algorithms. MITPress, 2002: 923-930

[122] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, et al. Introduction to Algorithms. MITPress, 2002: 531-534

[123] Ying Zhao, George Karypis. Criterion functions for document clustering.

[124] 谢靖，江岚，王东波，等.基于万方数据的知识发现应用研究.情报分析与研究，2010（12）：64-69

[125] 张小衡,王玲玲,等.中文机构名称的识别与分析.中文信息学报, 1997,11(4):21-32

[126] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, et al. Introduction to Algorithms. MITPress,2 002: 138-141

[127] Rudi L. Cilibrasi, Paul M.B.Vitanyi. The Google Similarity Distance. IEEE ITSOC Information Theory Workshop 2005

[128] Will, Craig. Comparing Human and Machine Performance. In Proceedings of the TIPSTER Text Program, Phase One, Morgan Kaufmann Publishers, San Mateo, 1994.

[129] Christopher D.Manning,Prabhakar Raghavan.信息检索导论,王斌.北京：人民邮电出版社,2010:246-246

[130] Rada Mihalcea, Dragomir Radev. Graph-Based Natural Language Processing and Information Retrieval .Cambridge, 2011

[131] 李亮，邹凯，黄锋. 做好文献检索对科学研究工作的作用与意义

[132] 刘禹，刘禹，杨一平.知识谱系的可视化方法.中国，发明专利，2012100220479，20120131

[133] 吴学梯, 周元.方法、创新、转变.北京:高等教育出版社,2011:9-14

[134] 王斌.现代信息检索课件第四章：相关反馈，及查询扩展

[135] Deerwester,Scott,Susan T. ,et al.Indexing by latent semantic analysis

[136] R. Agrawal , R.Srikant. Fast Algorithms for Mining Association Rules.VLDB’94

[137] 谢彩霞，梁立明，王文辉.我国纳米科技论文关键词共现分析.情报杂志,2005(3):69-73

[138] 任艳青，陈培颖，胡蓉.科技期刊的知识服务系统—以《自动化学报》知识服务平台为例.研究与报道:688-692

[139] 谢彩霞，梁立明，王文辉.我国纳米科技论文关键词共现分析.情报杂志,2005(3):69-73

[140] 王曰芬，宋爽，苗露.共现分析在知识服务中的应用研究.数字图书馆,2006(4):29-34

[141] 王曰芬，宋爽，熊铭辉.基于共现分析的文本知识挖掘方法研究.图书情报工作，2007,51(4)

[142] Schultz, U. Investigating the Contradictions in Knowledge Management.1998

[143] Zack,M.,An Architecture for Managing Explicated Knowledge,Sloan Management Review,1998

[144] McQueen,R.Four Views of Knowledge and Knowledge Management.Proceedings of the Americas Conference of AIS,1998,609-611

[145] Carlsson, S. A., El Sawy, O.A., Eriksson, et al. Gaining Competitive Advantage through Shared Knowledge Creation: In Search of a New Design Theory for Strategic Information Systems. 4th European Conference on Information Systems.

[146] Fahey,L.,and Prusak,L..The Eleven Deadliest Sins of Knowledge Management.California Management Review,1998,40(3),265-276

[147] Ronen Feldman,James Sanger.The TextMining Handbook-Advanced Approaches In Analyzing Unstructured data.UK:Cambridge Uinversity,2007:189-313

[148] Maryam Alavi.Knowledge Management and Knowledge Management Systems:Conceptual Foundations and Research Issues.MIS,Quarterly 25(1)

[149] Fahey,L.,Prusak,et al.The Eleven Deadlist Sins of Knowledge Management.California Management Review,1998,40(3),265-276

[150] Information,Knowledge and Wisdom : The Epistemic Hierachy and Computer-Based Information System

[151] 张晓林.走向知识服务：寻找新世纪图书情报工作的增长点.中国图书馆学报,2006,26(129):32-37

[152] 廖璠，麦桂芳.近五年图书馆知识服务研究文献定量分析.图书情报工作，2006,50（5）

[153] 陈建龙，王建东，胡磊.一论知识服务的概念内涵——基于产业实践视角的考察.图书情报知识

[154] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning .US, Springer, 2001: 22-27

[155] A.McCallum, K.Nigam. A Comparison of Event Models for Naïve Bayes Text Classification. In Proc. Of the AAAI-98 Workshop on Learning for Text Categorization

[156] 卿来云.统计学习基础课件 ChiSquare Test

[157] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

个人简历

刘禹，女，1986 年 9 月出生，内蒙古呼伦贝尔市人。2005 年 9 月进入北京邮电大学理学院数学与应用数学专业学习，后转入信息与通信工程学院主修信息工程专业，2009 年 6 月获得工学学士学位。2009 年 9 月进入中国科学院自动化研究所攻读硕士学位，从事计算机技术领域的学习和研究，主攻方向自然语言处理，特别是面向半结构化数据的文本挖掘技术。

在学期间申请的专利

“知识谱系的可视化方法”，申请号：2012100220479，申请人：刘禹，刘禹，杨一平

在学期间参与的科研项目

参与国家科技部课题“自动化学科创新思想与方法研究 (2009IM020300)”，负责知识要素提取系统的算法设计和开发工作。

工作成果见：

①自动化学科知识服务网络平台：http://autoinnovation.ia.ac.cn/

②项目部分资源开源：http://www.datatang.com/member/5878

致 谢

中国科学院自动化所三年的研究经历以及生活的点点滴滴让我受益颇深，在硕士论文即告完成之际，我谨向所有关心和帮助过我的亲人、老师、同学和朋友致以最由衷的感谢。

首先，我要深深地感谢我的导师杨一平研究员。感谢自己能有机会成为杨老师的学生；感谢杨老师三年以来的谆谆教诲以及严格要求；感谢杨老师在我学业上的关心和支持。在计算机相关的技术行业，女性一直以来处于技术上的弱势地位，作为一名女生硕士，我从杨老师这里没有感受到一分一毫的性别歧视，十分感谢杨老师能够让我在工程项目中结合自己的专业所长，独当一面。没有杨老师的支持，就没有今天这份我自认为还比较满意的硕士工作答卷。

其次，我要深深地感谢任禾师兄和刘禹老师。任禾师兄是我工程能力上的领路人，感谢他在我入门初期的悉心指导，感谢任禾师兄开发自动化学科知识服务网络平台，使我的论文成果能够更直观地展现出来；刘禹老师的思路和视角总给人焕然一新的感觉，同时感谢刘禹老师对于本毕设论文的悉心指导。

再次，我还要感谢实验室的各位老师，同窗与好友：高一波老师，卢朋老师，陈琳师姐，刘贤达师兄，马良俊师兄，刘西师兄，温少欣师姐，于海涛师兄，宋丽娜同学，张峰同学，左晓晗同学，秦树鑫同学，李娜同学，陈迪师妹，谢冰师妹，宋祥龙师弟，代文师弟，感谢他们在我硕士论文撰写过程中给予的热情帮助；感谢熊佩越、许政兰老师和倪晚成师姐在生活上对我的关心；感谢综合信息系统研究中心的全体老师，感谢你们的教诲与培养！

特别感谢科技部“创新方法”课题组成员（陈琳师姐、刘贤达师兄、温少欣同学、任禾师兄和刘禹老师）以及马良俊师兄，与你们相处，我不仅仅收获了学业上的进步，更学会了如何与他人进行有效沟通和交流。

特别感谢中科院自动化所蒋永实老师，蒋老师给本篇论文从布局谋篇以及排版格式等各个方面提出了非常详细的修改建议；特别感谢中国科学院研究生院黄志蓓老师、中国科学院研究生院孙应飞老师、北京科技大学梁治国老师、中国科学院自动化所田原老师等给本篇论文提出的建设性意见和建议。

最后，感谢在中科院研究生院学习期间孙应飞老师、朱庭劭老师、刘群老

师、王斌老师、白硕老师给予我的学习方法等方面的点拨，他们的一些学术视角以及解决问题的思路方案让我受益匪浅；感谢赵军老师于 2009 年 12 月 18 在中国科学院研究生院的讲座《信息提取》，这篇报告的一些观点在我设计知识要素提取算法时给予了很大的帮助。在我即将结束学习生涯，步入职场之际，感谢一路上陪我走过的家人、伙伴，同学和老师，千言万语尽在不言之中。

附录

附录 1 论文中所采集的期刊列表

Journal of Computer Science and Technology	控制理论与应用
传感技术学报	控制与决策
传感器技术	模式识别与人工智能
传感器与微系统	软件学报
电光与控制	系统仿真学报
电机与控制学报	小型微型计算机系统
基础自动化	信息与控制
机器人	遥感学报
计算机工程	冶金自动化
计算机集成制造系统	仪表技术与传感器
计算机科学	制造业自动化
计算机学报	中国图象图形学报
计算机研究与发展	中文信息学报
计算机应用与软件	自动化仪表
控制工程	

附录 2 文本分类语料

资源下载地址：http://www.datatang.com/data/13484

本语料包含两部分资源：中文新闻分类语料库和英语新闻分类语料库，英文新闻语料库为 Reuters-21578 的 ModApte 版本。

（一）**中文新闻语料库**为采用自行设计的“基于通用模板的新闻类网页正文抽取算法”从凤凰、新浪、网易、腾讯等版面搜集，搜集时间在 2009 年 12 月—2010 年 3 月。感谢网易新闻中心、腾讯新闻中心、凤凰新闻中心以及新浪新闻中心提供新闻素材。（注：新闻著作权归以上网站所有，任何人未经上述公司允许不得抄袭）。中文新闻语料共分为 8 类：Reading、Entertainment、History、Education、Society & Law、Culture、It、Military，数据规模如下表所示：

类型	**表单名称**	**文章 ID 范围**	**类别数目**	**是否为平衡语料**

训练集	NewsTrainingCorpus	1- 13026	8	否
测试集	NewsTestingCorpus	1-3254	8	否

（二）英文新闻语料库为 Reuters-21578 的 ModApte 版本，共有 90 个类别，训练集 7769 个文档，测试集 3019 个文档，类别分别是非均匀的，最大类别有 2877 个文档，但 82%的类只有不到 100 个训练文档，33%的类中文档数甚至小于 10 个。我们把它作为单标签语料，当一个文档有多个标签时，只采用第一个标签。之后选择训练集中至少包含 10 个或以上文档的类别，并以此作为类别标准对测试语料集合进行过滤处理，处理后的语料规模如下：

类型	**表单名称**	**文章 ID 范围**	**类别数目**	**是否为平衡语料**
训练集	ReteursTrainingCorpus	1- 6950	40	否
测试集	ReteursTestingCorpus	1-2676	40	否

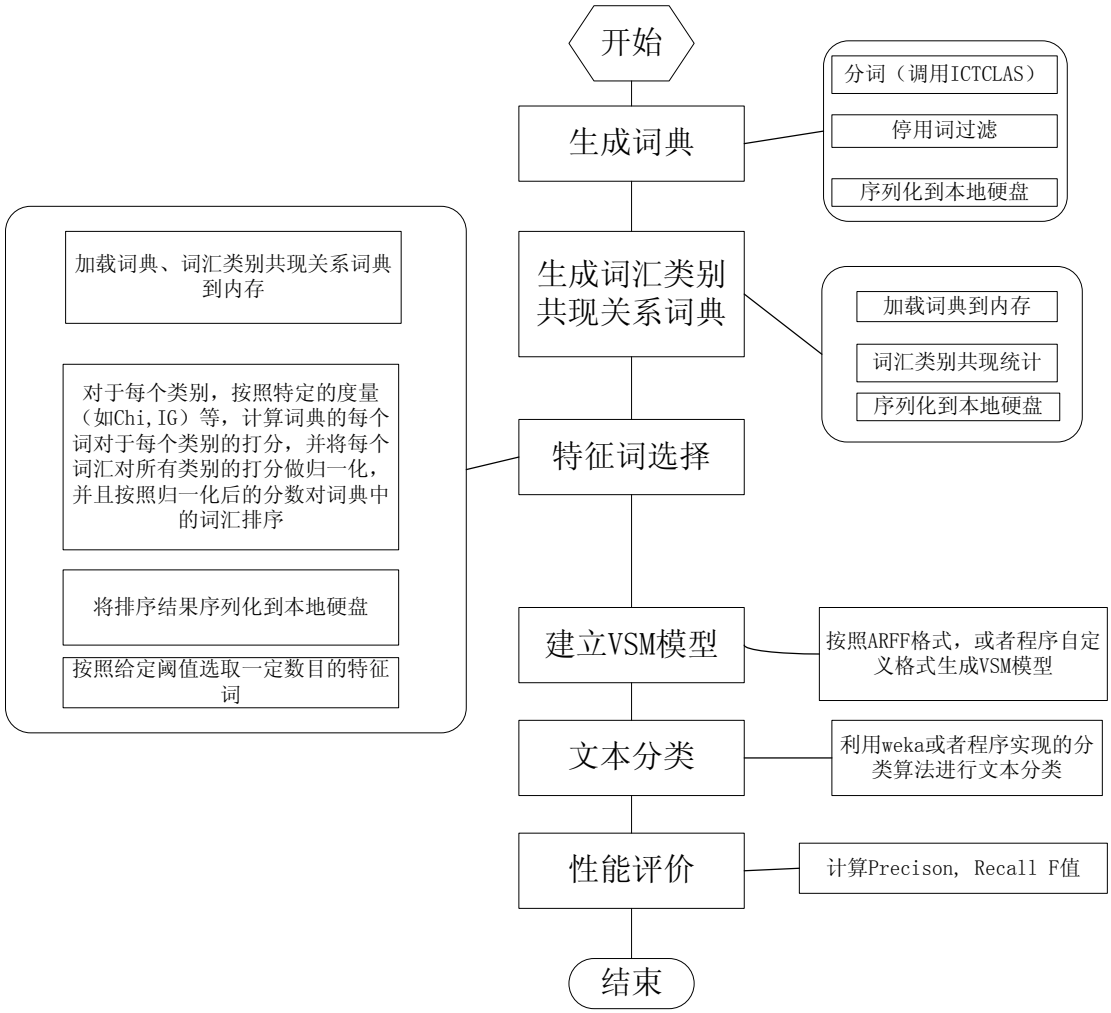
数据库名称：FinallyCorpus.mdf，FinallyCorpus.ldf

数据格式： 文本（MSSQL MDF 格式数据库）

附录 3 文本分类程序源码说明

资源下载地址：http://www.datatang.com/data/13483

此程序完整地实现了中文文本分类机制（见下图），数据格式为 C++Code，主要数据结构参见 common.h



主要数据结构：

参见 common.h

Preprocess p(beginIndex,endIndex);//文本分类预处理类

DICTIONARY mymap;//词典

CONTINGENCY contigencyTable;//词汇类别共现关系词典

FeatureWeight mymapweight;//词汇及权重

DOCMATRIX_1 trainingSet;//训练集 VSM 模型

DOCMATRIX_1 testingSet;//测试集 VSM 模型

Preprocess 类的主要工作函数

词典相关：

p.ConstructDictionary(mymap,seg,trainCorpusTable);

p.SaveDictionary(mymap,dictaddress);

p.LoadDictionary(mymap,dictaddress);

词典类别共现关系相关：

p.GetContingencyTable(mymap,labels,contigencyTable,trainCorpusTable);

p.SaveContingencyTable(contigencyTable,contigencyaddress);

p.LoadContingencyTable(contigencyTable,contigencyaddress);

特征词选择相关：

p.ChiSquareFeatureSelection(labels,mymap,mymapweight,contigencyTable,weightaddress);

p.ChiFitFeatureSelection(labels,mymap,mymapweight,contigencyTable,weightaddress);

p.InformationGainFeatureSelection(labels,mymap,mymapweight,contigencyTable,weightaddress);

p.PointWiseMIFeatureSelection(labels,mymap,mymapweight,contigencyTable,weightaddress);

建立 VSM 模型相关

（1）建立 ARFF 数据格式的测试集 VSM 模型

p.WriteHeadArff(testvsmaddress,keywordaddress,labels);

p.GetManyVSM(1,2676,testCorpusTable,mymap,testingSet,keywordaddress);

p.WriteDataBodyArff(testingSet,testCorpusTable,testvsmaddress,featuredimension[i]);

（2）建立 ARFF 数据格式的训练集 VSM 模型

p.WriteHeadArff(trainvsmaddress,keywordaddress,labels);

p.VSMConstruction(mymap,trainingSet,keywordaddress);

p.WriteDataBodyArff(trainingSet,trainCorpusTable,trainvsmaddress,featuredimension[i]);

注意事项：

i 中英停用词表：分别是程序目录下的 stopwords.txt, estopwords.txt。根据处理文本的不同，请手动定

位 MakeStopSet 函数，对其所 load 的停用词表名称进行相应的修改。

ii 程序的正确运行需要安装 boost 库，boost 的安装方法请见：

http://www.cnblogs.com/finallyliuyu/archive/2010/08/23/1806811.html

iii 关于文本分类的更详细的介绍流程请见：

http://www.cnblogs.com/finallyliuyu/archive/2010/10/04/1842261.html

程序的分词调用 ICTCLAS，如果程序调用提示过期，请到 ictclas 官方网站下载更新。

附录 4 Weka 数据格式(ARFF)的 VSM 模型

资源下载地址：

（1）以 IG 卡方等特征词选择方法生成的多维度 ARFF 格式英文 VSM 模型

http://www.datatang.com/data/13486

（2）以 IG 卡方等特征词选择方法生成的多维度 ARFF 格式中文 VSM 模型

http://www.datatang.com/data/13485

数据是应用笔者开发的文本分类程序分别对英文文新闻语料、中文新闻语料库处理后生成的 ARFF 文件。

数据文件说明：

dict.dat ：词典文件

contingency.dat：词汇类别共现频率词典

chi，ig，mi，DF 分别表示采用同名特征词选择算法生成的 VSM 模型，chifit 为笔者依据卡方思路，从另一个思考方式推导出的一种新的特征词选择算法，经验证，此种方法的效果与传统方法相比，不分伯仲。

如下图所示，在每个文件夹下面，分别为特征维度选取 50，100，150，200，250，300，350，400，450，500，600，700，800，900，1000，1500，2000，3000，5000，8000 生成的 VSM 模型。

附录 5 自动化学科知识服务网络平台涉及的实体

论文实体（PaperEntity），属性有：标题、作者、发文机构、收录期刊、发表年限、中英关键词、中英摘要；

作者实体(ScholarEntity)，属性有：作者姓名、作者单位、主要研究领域、主要学术成果；

机构实体(InsititutionEntity)，属性有：机构名称、主要研究人员、主要研究领域、主要学术成果；

知识点实体(KnowledgeEntity)，认为有四种类型的知识点：（1）理论如“贝叶斯”、“卡方检验”等；（2）方法或算法：如“贝叶斯分类算法”、“卡方特征词选择算法”等；（3）工具：如“贝叶斯分类器”、“网络采集器”等；（4）总结性知识点，它的内容涉及理论、方法或算法、工具，一般为代表研究方向的短语，如“自然语言处理”、“模糊控制”等。知识点是知识族谱上的结点，其属性包括中英关键词的表示形式

期刊实体(MagzineEntity)：期刊名称、期刊所涉及到的主要研究领域。

其中学术成果用论文实体的标题属性组成；研究领域由知识点实体的中文表示形式组成。

涉及到的关系有：

知识点间的相互关系：也即知识族谱；

知识点与作者、机构、文章、期刊的关系：针对某具体知识点，采用推荐算法给出和该知识点相关的主要机构、作者、文章、期刊。

作者合作关系：给出作者间的合作关系

附录 6 自动化学科知识服务网络平台数据表单字段说明

相关数据表单介绍：

Paper(论文表单)

字段名称	字段意义
PaperId	论文编号
title	论文中文标题
etilte	论文英文标题
Abstract	论文中文摘要
EAbstact	论文英文摘要
KeyWords	论文中文关键词
EKeyWords	论文英文关键词
AuthorString	论文发文作者
WorkPlaceString	论文作者单位
Magzine	论文发表期刊
publishTime	论文发表时间

说明：Paper 表单的数据是我们从网上多个数据源抓取数据，直接进行属性整合的结果，也是作者，机构，作者-论文-机构关系等资源的生成资源。

Institution（机构表单）

字段名称	字段意义
InstitutionId	发文机构编号
InstitutionName	发文机构名称
InstitutionDescription	发文机构描述

说明：从 paper 表 WorkPlaceString 字段中抽取的一级机构名称，如从“清华大学计算机系”抽取“清华大学”等。

Scholar（作者表单）

字段名称	字段意义
ScholarId	发文作者编号
InstitutionId	发文作者所在机构编号
ScholarName	发文作者中文名字
ScholarEnglishName	发文作者英文名字
ScholarDescription	发文作者的相关描述（预留字段）
Institutionlevel	发文作者的第几个工作单位（0--n）

说明：从 Paper 表单中 AuthorString 以及 WorkPlaceString 中抽取作者名称，以及单位名称，作者单位

对应关系等，并做**同名消歧处理**。考虑到同一作者在其学术生涯履历中工作单位历经变迁，我们用 ScholarId 唯一标识作者，并按照作者在其所在单位的发文量对其单位进行排序，并将相应信息存储在 Institutionlevel 字段，0 表示是该作者学术生涯中的主要工作单位，1，表示是该作者学术生涯中的次要工作单位，以此类推。拿白硕老师作为例子，他的工作单位涉及到中国科学院计算技术研究所（10045）与上海证券交易所(7407)，而其发表论文的主要单位是中国科学院计算技术研究所，所以储存结果如下。

	ScholarId	InstitutionId	ScholarName	institutionlevel
▶	4627	7407	白硕	1
	4627	10045	白硕	0

ScholarPaper（作者—机构—论文关联表）

字段名称	字段意义
PaperId	论文编号
ScholarId	发文作者编号
InstitutionId	发文作者所处单位编号
AppearOrder	表征该作者是论文的第几作者

说明：如果 InstitutionId=0 并且 AppearOrder=0，则表明数据抽取过程中产生了异常，但是为了照顾整个数据库系统的可用性，遂保留。

相关表单截图

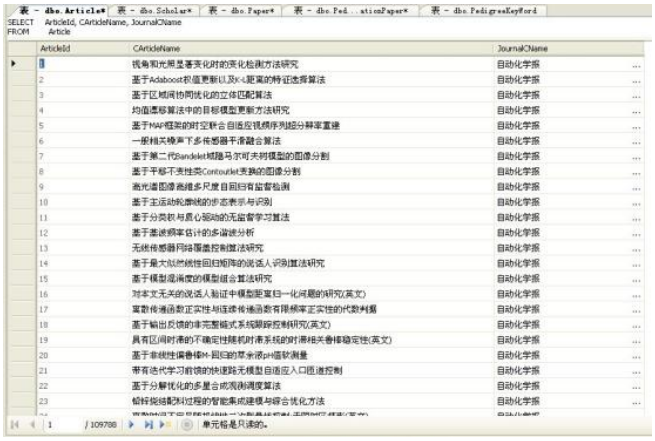


图 1 文章表

表 - dbo.Institution* | 表 - dbo.Scholar* | 表 - dbo.CETermTable | 表 - dbo.Article*

```
SELECT *
FROM   Institution
```

InstitutionId	InstitutionName	InstitutionDescription
1656	北京工贸技师学院	学院
1657	北京工商大学	大学
1658	北京工业大学	大学
1659	北京工业职业技术学院	学院
1660	北京供电设计院	研究院
1661	北京光华路北京汽车厂	工厂
1662	北京光学仪器厂	工厂
1663	北京广播电视大学	大学
1664	北京国电智深控制技术有限公司	公司
1665	北京海淀区61516部队	部队
1666	北京海淀区电子技术研究所	所级单位
1667	北京海军核化研究所	所级单位
1668	北京海思威科技有限公司	公司
1669	北京海鹰仿真中心	研究中心
1670	北京航空材料研究所	所级单位
1671	北京航空工程技术研究中心	研究中心
1672	北京航空工艺研究所	所级单位
1673	北京航空航天部五院	研究院
1674	北京航空航天大学	大学
1675	北京航空精密机械研究所	所级单位
1676	北京航空仪器厂	工厂
1677	北京航空制造工程研究所	所级单位

4262 / 10865

图 2 机构表

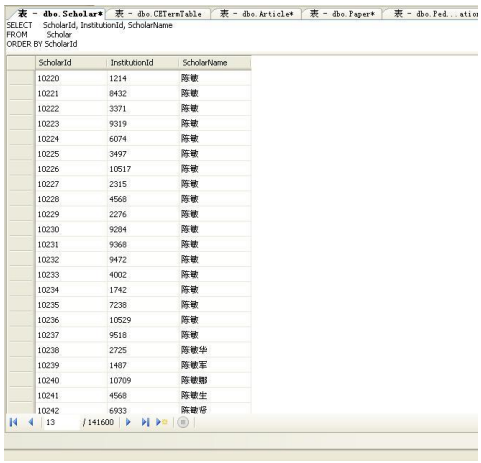


表 - dbo.Scholar* | 表 - dbo.CETermTable | 表 - dbo.Article* | 表 - dbo.Paper* | 表 - dbo.Fed... ation

```
SELECT ScholarId, InstitutionId, ScholarName
FROM   Scholar
ORDER BY ScholarId
```

ScholarId	InstitutionId	ScholarName
10220	1214	陈敏
10221	8432	陈敏
10222	3371	陈敏
10223	9319	陈敏
10224	6074	陈敏
10225	3497	陈敏
10226	10517	陈敏
10227	2315	陈敏
10228	4568	陈敏
10229	2276	陈敏
10230	9284	陈敏
10231	9368	陈敏
10232	9472	陈敏
10233	4002	陈敏
10234	1742	陈敏
10235	7238	陈敏
10236	10529	陈敏
10237	9518	陈敏
10238	2725	陈敏华
10239	1487	陈敏军
10240	10709	陈敏娜
10241	4568	陈敏生
10242	6933	陈敏容

13 / 141600

图 3 作者表

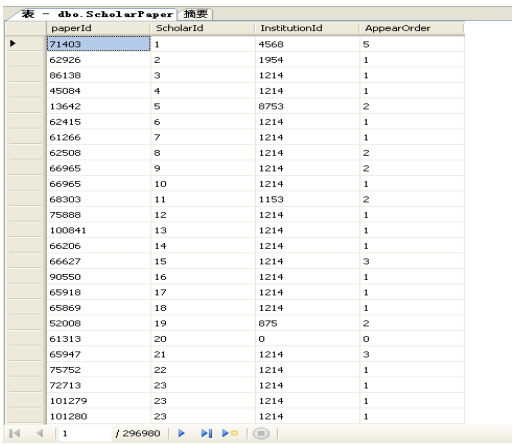


表 - dbo.ScholarPaper | 摘要

paperId	ScholarId	InstitutionId	AppearOrder
71403	1	4568	5
62926	2	1954	1
86138	3	1214	1
45084	4	1214	1
13642	5	8753	2
62415	6	1214	1
61266	7	1214	1
62508	8	1214	2
66965	9	1214	2
66965	10	1214	1
68303	11	1153	2
75888	12	1214	1
100841	13	1214	1
66206	14	1214	1
66627	15	1214	3
90550	16	1214	1
65918	17	1214	1
65869	18	1214	1
52008	19	875	2
61313	20	0	0
65947	21	1214	3
75752	22	1214	1
72713	23	1214	1
101279	23	1214	1
101280	23	1214	1

1 / 296980

图 3 论文作者关联表

附录 7 面向人物同名消歧研究的中文 DBLP 资源说明

资源下载地址：http://www.datatang.com/data/13479

数据库名称：PersonNameDisambiguation　108commonpersonnames.txt

数据规模：

表单名称	**数据条目数**
Paper	2671
Scholar	8438
Institution	1134
ScholarPaper	8981

数据产生方式：根据名字发文量对应关系，并将名字按照发文量逆序排列，从中选取 108 个常见名字，从自动化学科知识服务平台后台数据中抽取的部分子资源。

数据格式：MSSQL MDF 格式数据库

表单说明：见附录 6

附录 8 面向汉语姓名构词研究的中文人名语料库资源说明

数据下载地址：http://www.datatang.com/data/13482

数据文件名称：authornames.txt

数据规模：9,9994 个人名

数据产生方式：从自动化学科知识服务平台后台数据的 Scholar 表单中抽取的全部名字，并作去重处理。

数据格式：文本

数据预览：

附录 9 自然语言处理领域期刊的中文 DBLP 资源说明

资源下载地址：http://www.datatang.com/data/13478

数据库名称：NLP

数据规模：

表单名称	**数据条目数**
Paper	4630
Scholar	9085
Institution	976
ScholarPaper	13527

数据产生方式：借助自动化学科知识服务平台的知识族谱，以“自然语言处理”，“人工智能”、“数据挖掘”、“信息检索”等专业术语为查询结点，进行有限步拓展，从自动化学科知识服务平台后台数据中抽取的部分子资源。

数据格式：MSSQL MDF 格式数据库

表单说明：见附录 6

附录 10 面向计算机学科学术共同体的中文 DBLP 资源说明

资源下载地址：http://www.datatang.com/data/13481

数据库名称：PersonSeed

数据规模：

表单名称	数据条目数
Paper	45355
Scholar	28460
Institution	2362
ScholarPaper	138294

其中 Scholar 表单中的 ScholarDescription 字段的 0—7 分别表示为以白硕(0)、戴汝为(1)、李生(2)、吴佑寿(3)、张钹(4)、艾海舟(5)、李晓明(6)、赵军(7) 老师为种子拓展出的学术社区成员。

数据产生方式：以白硕（计算所），戴汝为、赵军（自动化所），吴佑寿、张钹、艾海舟（清华大学），李生（哈工大），李晓明（北大）等八位老师作为根节点（初始种子），固定扩展种子数为 2000，采用广度优先遍历算法从自动化学科知识服务平台后台数据中抽取的部分子资源。

数据格式：MSSQL MDF 格式数据库

表单说明：见附录 6

附录 11 万篇随机抽取论文的中文 DBLP 资源说明

资源下载地址：http://www.datatang.com/data/13480

数据库名称：RandomTenThousand

数据规模：

表单名称	数据条目数
Paper	10000
Scholar	20071
Institution	1633
ScholarPaper	30576

数据产生方式：从 1-100,000 个 paperId 中随机抽取 10,000 个，属于从自动化学科知识服务平台后台数据中抽取的部分子资源。

数据格式：MSSQL MDF 格式数据库

表单说明：见附录 6

附录 12 自动化学科知识服务网络使用说明

各位同学、老师、网友，大家好，由中科院自动化所综合信息中心承担、国家科技部支持的自动化学科数字化知识服务网络平台已经上线。网站地址是：http://autoinnovation.ia.ac.cn/,欢迎大家使用，并且给我们提出意见和建议。

下面是平台使用过程中的几点注意事项：（1）初次使用时，如果您的浏览器没有安装 silverlight 插件，请您按提示下载安装该插件；（2）如果您在使用中遇到一些小问题，可以查看网站的帮助文件；（3）该平台框架实际为实体检索系统，因此您输入检索词后，需要等待下拉菜单出现相应检索实体，选中相应检索实体，之后在点击搜索按钮，如下图所示

图 4 检索说明示意图

该平台主要包括以下组成部分：

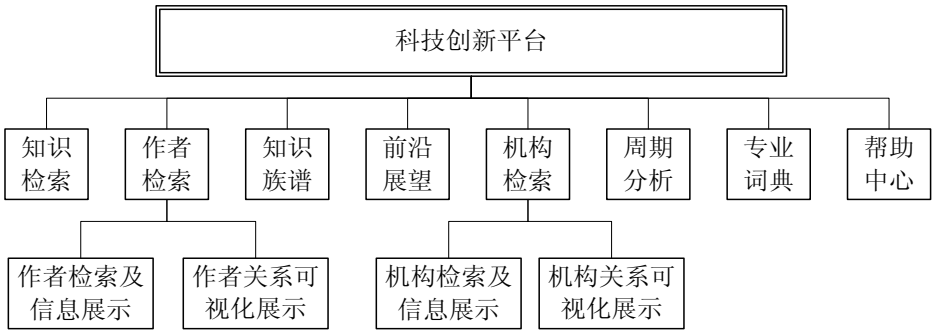


图 5 平台基本架构

它旨在挖掘、分析和展现我国自动化领域（包括部分计算机、通信的交叉领域）自 1960 年以来的学

术发展情况。我们力求展现出国内自动化领域学术活动的立体全景，对领域内的文献、学者、机构、以及研究方向、方法、理论和工具等，做了全方位的关联分析。为了更好地展现知识，我们在精心设计页面布局的基础上，使用了 Silverlight、Ajax 等技术进行网站开发；为了让展现出来的知识更加精确，我们在数据处理中使用了包括命名实体识别与排歧、文本聚类在内的多种数据挖掘技术。

该平台凝结了综合信息中心的老师、开发人员、以及多位学生的大量心血。无论是在前期设计、后台数据处理、还是前台网站开发，我们都本着精益求精的原则，团队内部经过多次尝试和试验，力求选择最佳方案。但是作为一个人员有限的开发团队，我们的思虑与广博的群体智慧相比还是有所逊色的。为此，我们热诚地欢迎各位老师、同学、工作人员向我们提出您宝贵的建议。我们欢迎大家从各个层面给我们提出意见和建议，您的意见和建议将是敦促我们进步和改进的最给力的源泉！

我们的联系方式是：

email: y.liu@ia.ac.cn

新浪微博：http://weibo.com/autoinnovation

如果您觉得方便，可以留下您的姓名和单位，我们将在我们的网站进行致谢！

附录 13 自动化学科知识服务网络平台检索示例

图 1 机构检索

图 2 作者检索

图 3 知识检索

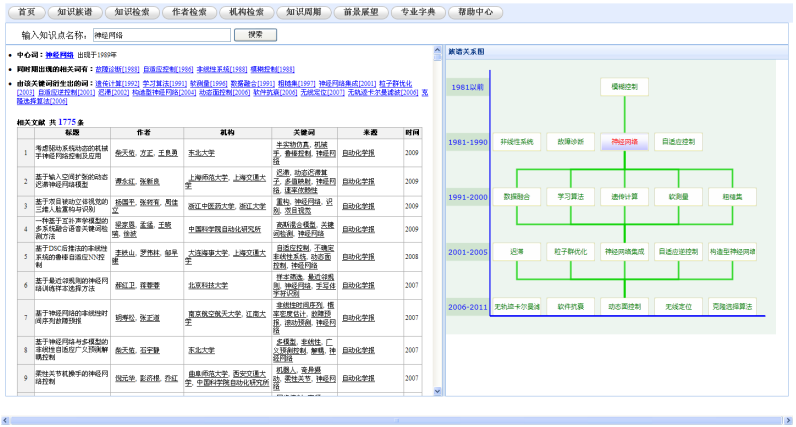


图 4 知识族谱

抽象	Abstract	交-交变频	AC-AC frequency converter	自主规划	autonomous planning
抽象代数	abstract algebra	交-交变频器	AC-AC frequency converter	自主计算	AOC
抽象体系	abstract framework	交-直-交变频	AC-DC-AC frequency conversion	自主调度执行	autonomous scheduling execution
抽象分析	abstract analysis	交-直-交型变频器	AC-DC-AC converter	自主越障	autonomous negotiation
抽象化	abstraction	交-直流变换	AC-DC Conversion	自主车辆	autonomous vehicle
抽象图归约机	abstract reduction machine	交交变频器	A.c.—a.c.cycloconverter	自主轮式机器人	autonomous wheeled robot
抽象实现结构图	Abstract implement struture diagrams	交叉耦合系数	across couple coefficient	自主运算	automatic computing
抽象层APII	abstract hierarchy APII	交变控制	alternate control	自主运行与管理	autonomous running and governing
抽象工作流	Abstract workflow	交变磁场	alternating magnetic field	自主运行分布式卫星	autonomous distributed satellites
抽象工厂模式	abstract factory pattern	交变磁场测量	Alternating Current Field Measurement	自主运行编队	autonomous satellite formation
抽象归纳	abstract generalization	交换深度	alternation depth	自主通信	autonomic communication
抽象技术	Abstraction technique	交替互补逻辑	alternating-complementary logic	自主配置	autonomous configuration
抽象指令集	abstract instruction set	交替活跃	alternate activity	自主陆地车辆	autonomous land vehicle
抽象描述模型	Abstract description model	交替游程码	alternating run-length code	自主预测	Autonomic prediction
抽象数据切片	abstract data slice	交替的ω－有穷自动机	Alternating ω-finite automata	自主飞艇	Autonomous Flight
抽象数据类型	ADT	交替类估计	Alternating cluster estimation	自主飞行器	autonomous flight vehicle
抽象数据类型（ADT）编译器	abstract data type compiler	交替迭代算法	alternative iteration algorithm	自主驾驶	autonomous driving
抽象数据类形。	abstract data type.	交替顺序重建滤波	alternating sequential filtering by reconstruction	自准直	auto-collimating
抽象文法	abstract grammar	交汇天幕	across sky screen	自制体	agent
抽象服务	Abstract service	交流传动	AC Drive	自动摘要	Automatic Summarization
抽象机	abstract machine	交流伺服	AC serve	自动	Automation

图 5 专业词典

图 6 机构标签云

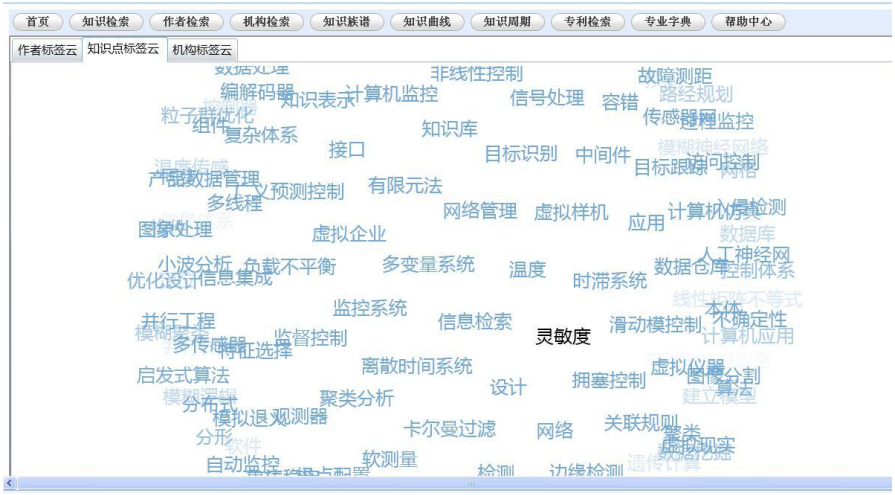


图 7 知识点标签云

图 8 作者标签云

附录 14 原始数据与处理后数据的对比

	原始数据	处理后数据
题录信息条目	知网 116,642/万方 58235	109,788
关键词中英对照组条目	148,825	83,602
作者条目	299,823	135,969
机构条目	60,874	10,865

附录 15 常用姓名出现次数及所指称的人物实体数目

排序	姓名	出现次数/关联文章数(次)	指称实体数目（个）
1	王伟	245	69
2	张伟	202	59
3	李伟	121	55
4	王勇	118	55
5	张勇	121	54
6	刘伟	85	50
7	王磊	107	49
8	张军	90	47
9	王涛	87	46
10	张鹏	81	43
11	刘军	87	41
12	张健	74	41
13	李刚	82	40

14	李军	88	38
15	张涛	82	37
16	李斌	113	37
17	王平	101	35
18	李华	84	32
19	李辉	72	32
20	王刚	111	32
21	王超	77	31
22	王强	72	30
23	刘强	69	29
24	王斌	82	29
25	陈刚	97	29
26	刘刚	61	28
27	张斌	108	27
28	杨波	78	27
29	刘芳	100	26
30	赵军	116	25
31	李平	142	23
32	王宏	94	23
33	李俊	93	22
34	张浩	78	21
35	陈明	91	21
36	吴刚	75	20
37	杨明	100	17
38	刘飞	108	14
39	吴敏	156	12
40	徐波	63	11
41	谢立	228	6
42	高文	184	6
总计		4423	1369

附录 16 自动化学科知识服务网络平台原始数据表单格式说明

表单说明：

Article 表

字段名称	**字段含义**
ArticleId	文章编号（主键）
CArticleName	文章的中文标题

EArticleName	文章的英文标题
ArticleUrl	文章题录信息 URL
ACSId	文章题录信息来源
JournalCName	文章发表刊物的中文名称
JournalEName	文章发表刊物的英文名称
Volume	文章发表在所在刊物的卷期
CKeyWord	文章的中文关键词
EKeyWord	文章的英文关键词
CAbstract	文章的中文摘要
EAbstract	文章的英文摘要
Fund	文章的基金支持
DOI	Digital Object Identifier
ClassifyNum	文章所属中图分类号
jibiaoKeyWord	机标关键词（来自万方）
jibiaoClassifyNum	机标分类号（来自万方）
AuthorString	作者字符串
WorkPlaceString	作者所在机构字符串
FirstAuthor	文章的第一作者
publishTime	文章发表年限
categorization(deprecated)	文章所属类别（自动化/计算机/其他）

说明：经过将知网和万方数据源进行融合去重共整合无重复题录信息 109,788 条，其中每个条目中文标题或者英文标题必须要有其一。

Author 表

字段名称	**字段含义**
AuthorId	作者编号（主键）
AuthorName	作者中文名字
Alias	作者英文或者拼音名字
AuthorWorkPlace	作者工作单位
trust	标识人物机构对齐算法是否可信

说明：共存储未进行人名消歧的 299,823 条作者信息。

AuthorArticle 表

字段名称	**字段含义**
ArticleId	文章编号
AuthorId	作者编号

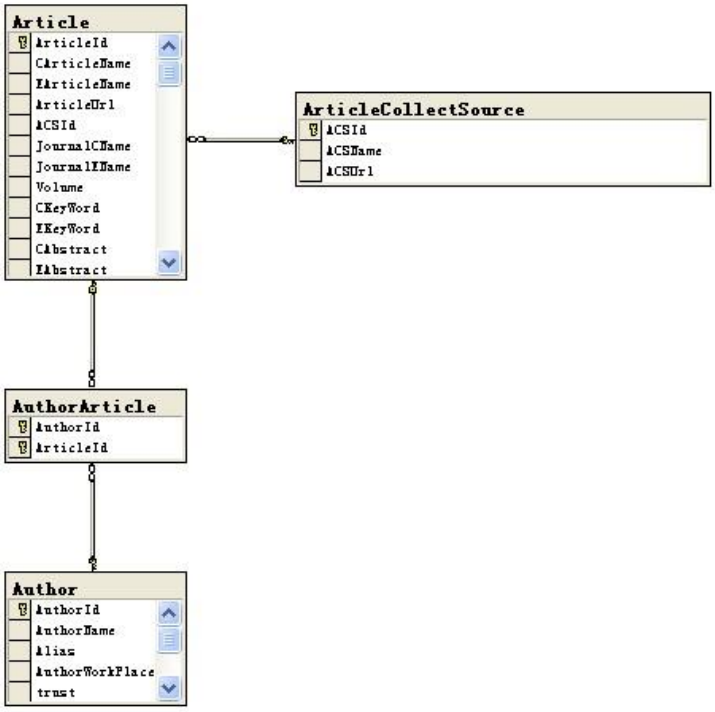
说明：存储 Article 表中 109,788 篇文章与 Author 表中 299,823 个作者之间的对应关系

ArticleCollectSource 表

字段名称	字段含义
ACSId	信息源编号（主键）
ACSName	信息源名称
ACSUrl	信息源 URL

说明：记录垂直爬虫获得的数据的数据源信息

表单直接的依赖关系：



附录 17 姓名“白硕”的聚类结果与标注结果

聚类的簇描述

clusterId	description
0	此人曾就职中科院计算所、上海证券交易所
1	此人曾就职北京大学、国家智能计算机研究开发中心

标准答案的簇描述

gold-standardId	description
0	此人曾就职北京大学、国家职能计算机研究开发中心、中科院计算所、上海证券交易所

聚类结果与标注结果对比

AuthorId[27]	**clusterId**	**goldstandardId**
8281	0	0
10622	0	0
10648	0	0
11227	0	0
12110	1	0
13491	0	0
18497	0	0
18850	0	0
19279	0	0
20087	0	0
20126	0	0
21072	1	0
21132	1	0
21147	1	0
21764	1	0
22944	0	0
23196	0	0
26960	0	0
27186	0	0
29713	0	0
30137	0	0
30934	0	0
31286	0	0
31512	0	0
37894	0	0
38317	0	0
47850	0	0
47904	0	0
48018	0	0
48691	0	0
48968	0	0
180385	1	0
183984	1	0
190019	0	0

[27]此处 AuthorId 不是屈分作者的标志，而是与与 Article 对应的，未去重的作者编号而已，见附录 17

191131	0	0
191792	1	0
191927	1	0
193203	1	0
265947	0	0
266177	0	0
270718	0	0
274192	0	0
276582	0	0
277163	0	0
279746	1	0

附录 18 姓名“王斌”的聚类结果与标注结果

聚类的簇描述

clusterId	description
0	College of Information Science and Engineering Northeastern University
1	Institute of Computing Technology;Chinese Academy of Sciences;
2	中南大学信息科学与工程学院;
3	东南大学
4	复旦大学电子工程系
5	东北大学
6	北京理工大学汽车动力性及排放测试国家专业实验室
7	西安交通大学
8	华中科技大学模具技术国家重点实验室
9	中国科学院上海光学精密机械所联合实验室
10	吉林省电力试验研究所
11	武汉大学软件工程国家重点实验室
12	西安电子科技大学电子工程学院
13	中国科学院计算技术研究所
14	中国科学院沈阳自动化研究所
15	哈尔滨电机厂有限责任公司
16	盐城工学院
17	东莞理工学院软件学院
18	上海理工大学电气工程学院
19	河北工程学院城市建设学院

20	嘉兴学院机电工程学院
21	东华大学计算机科学与技术学院
22	上海交通大学电子工程系
23	装备指挥技术学院试验指挥系;
24	中国科学院光电研究院
25	华南理工大学计算机学院
26	郑州大学电气工程学院
27	内蒙古大学计算机学院
28	一机部第九设计院

标准答案的簇描述

gold-standardId	description
0	中南大学信息科学与工程学院
1	清华大学软件学院
2	东南大学自动化学院
3	东南大学 MEMS 教育部重点实验室
4	东南大学计算机科学与工程学院
5	复旦大学电子工程系
6	东北大学信息科学与工程学院
7	北京理工大学汽车动力性及排放测试国家专业实验室
8	西安交通大学机械工程学院
9	西安交通大学计算机系
10	华中科技大学模具技术国家重点实验室
11	中国科学院上海光学精密机械所联合实验室
12	吉林省电力试验研究所
13	武汉大学软件工程国家重点实验室
14	西安电子科技大学电子工程学院
15	中国科学院计算技术研究所
16	中国科学院沈阳自动化研究所
17	哈尔滨电机厂有限责任公司
18	盐城工学院
19	东莞理工学院软件学院
20	上海理工大学电气工程学院
21	河北工程学院城市建设学院
22	嘉兴学院机电工程学
23	东华大学计算机科学与技术学院
24	上海交通大学电子工程系
25	装备指挥技术学院试验指挥系

26	中国科学院光电研究院
27	华南理工大学计算机学院
28	郑州大学电气工程学院
29	内蒙古大学计算机学院
30	一机部第九设计院

聚类结果与标注结果对比

AuthorId	**clusterId**	**goldstandardId**
260656	0	6
260814	1	15
5568	2	1
41193	2	0
42186	2	0
44052	2	0
267197	2	0
273552	2	0
274511	2	0
284445	2	0
24519	3	4
162032	3	3
162439	3	3
162914	3	3
172224	3	2
175156	3	2
60421	4	5
6246	5	6
6443	5	6
14144	5	6
15206	5	6
16189	5	6
23661	5	6
23670	5	6
23732	5	6
23757	5	6
25367	5	6
25385	5	6
25412	5	6
28727	5	6
35839	5	6

265267	5	6
268841	5	6
277020	5	6
91386	6	7
93787	6	7
118282	7	8
279228	7	9
95474	8	10
150862	9	11
218535	10	12
43588	11	13
273826	11	13
273888	11	13
13451	12	14
14720	13	15
14806	13	15
20613	13	15
22939	13	15
23332	13	15
24020	13	15
26961	13	15
47756	13	15
47849	13	15
47854	13	15
48013	13	15
48117	13	15
48203	13	15
48572	13	15
48692	13	15
48769	13	15
48844	13	15
189123	13	15
192171	13	15
272386	13	15
286105	13	15
85425	14	16
140601	15	17
68359	16	18

102556	17	19
125171	18	20
151544	18	20
144457	19	21
161225	20	22
203463	21	23
9767	22	24
224745	23	25
158717	24	26
262297	25	27
262187	26	28
269464	27	29
234999	28	30

附录 19 姓名“赵军”的聚类结果与标注结果

聚类的簇描述

clusterId	description
0	School of Electronic and Information Engineering;Dalian University of Technology;Key Laboratory of Integrated Automation of Industry;Ministry of Education;Northeastern University;
1	National Laboratory of Pattern Recognition Institute of Automation;Chinese Academy of Sciences;National Laboratory of Pattern Recognition;Institute of Automation;Beijing;China;
2	Key Laboratory of Process Industry Automation, Ministry of Education, School of Information Science and Engineering, Northeastern University, Shenyang 110004
3	重庆大学/重庆邮电学院
4	东北大学信息科学与工程学院;
5	清华大学计算机科学与技术系
6	中国科学院自动化研究所模式识别国家重点实验室,北京,100080
7	浙江大学工业控制技术国家重点实验室
8	香港科技大学计算机科学系人类语言技术中心

9	天津大学电气与自动化工程学院
10	中国计量学院计量技术工程学院
11	国防科技大学计算机系
12	燕山大学机械工程学院
13	中国兵器装备集团公司军品部
14	乌鲁木齐 21 信箱 189 分箱
15	辽宁石油化工大学理学院
16	中国科学院系统科学研究所
17	上海理工大学动力工程学院
18	武汉理工大学自动化学院
19	中国自动化学会控制论委员会
20	东南大学计算机科学与工程系
21	中国科学院软件研究所
22	北京信息与控制研究所
23	东北工学院自动控制系
24	北京航空航天大学

标准答案的簇描述

gold-standardId	description
0	重庆大学材料科学与工程学院
1	此人曾就职重庆大学计算机科学与工程学院、重庆邮电学院计算机科学与技术研究所
2	东北大学信息科学与工程学院
3	此人曾就职清华大学计算机科学与技术系、香港科技大学计算机科学系人类语言技术中心、中国科学院自动化研究所模式识别国家重点实验室
4	浙江大学工业控制技术国家重点实验室
5	天津大学电气与自动化工程学院
6	中国计量学院计量技术工程学院
7	国防科技大学计算机系
8	燕山大学机械工程学院
9	中国兵器装备集团公司军品部
10	乌鲁木齐 21 信箱 189 分箱
11	辽宁石油化工大学理学院
12	中国科学院系统科学研究所
13	上海理工大学动力工程学院
14	武汉理工大学自动化学院
15	中国自动化学会控制论委员会
16	东南大学计算机科学与工程系

17	中国科学院软件研究所
18	北京信息与控制研究所
19	北京航空航天大学

聚类结果与标注结果对比

AuthorId	clusterId	goldstandardId
172183	0	2
175115	0	2
263144	1	3
2120	2	2
2381	2	2
10104	3	1
42069	3	1
42424	3	1
226605	3	0
272458	3	1
273776	3	1
274547	3	1
274814	3	1
276074	3	1
1606	4	2
1899	4	2
1906	4	2
2600	4	2
2681	4	2
2783	4	2
3271	4	2
3347	4	2
3508	4	2
3559	4	2
75220	4	2
75961	4	2
76115	4	2
76271	4	2
76351	4	2
76460	4	2
76932	4	2
77160	4	2
77307	4	2

77456	4	2
78077	4	2
79201	4	2
79531	4	2
79681	4	2
79684	4	2
79768	4	2
111753	4	2
131260	4	2
131834	4	2
132346	4	2
132824	4	2
132982	4	2
133024	4	2
133034	4	2
133322	4	2
133407	4	2
133462	4	2
134736	4	2
135108	4	2
135306	4	2
135795	4	2
170833	4	2
170879	4	2
173401	4	2
173643	4	2
176352	4	2
176595	4	2
221650	4	2
222071	4	2
222194	4	2
222456	4	2
228571	4	2
231070	4	2
231202	4	2
12708	5	3
21731	5	3
32625	5	3

191756	5	3
286292	5	3
8178	6	3
9142	6	3
25222	6	3
48001	6	3
48056	6	3
48121	6	3
48138	6	3
48302	6	3
48329	6	3
48473	6	3
48689	6	3
48753	6	3
48924	6	3
49288	6	3
49375	6	3
286127	6	3
286140	6	3
50172	7	4
165130	7	4
11450	8	3
144880	9	5
165968	10	6
28038	11	7
44346	11	7
47537	11	7
94639	12	8
142981	12	8
143226	12	8
167000	12	8
101531	13	9
102006	14	10
134228	15	11
137044	16	12
127677	17	13
126660	18	14
133166	19	15

271161	20	16
210370	21	17
288351	22	18
288419	22	18
298306	23	2
174829	24	19
177789	24	19

附录 20　不同特征词选择算法生成的特征词

Chifit，卡方, IG, DF, MI 等特征词选择算法在中文新闻语料上生成的前 50 位特征词：

Chifit	Chi-squae	IG	DF	MI
BigNews	答疑	榜	月	%TIMES
榜	榜	摘	日	----MBA
本文	本文	精彩	年	.-.分
编辑	摘	周	人	.-分
大中小	网易	BigNews	中	.ACT
精彩	签证	头条	说	.GMAT
来源	精彩	书	后	.GRE
年	大中小	大中小	大中小	.LSAT
频道	周	频道	编辑	.SAT
日	来源	匿名	时	:newsweek
书	匿名	新闻	两	A-Level
头条	BigNews	责任	前	ACCA
新闻	头条	陶学钢	下	ACT
月	书	IP	最	AQF
摘	频道	作者	中国	ARWU
周	责任	地址	已	Aberdeen
正文	IP	隐藏	位	AcademicRankingofWorldUniversities
签证	陶学钢	本文	里	AdvancedPlacement
答疑	作者	蒋勋	新闻	AlabamaSchoolofFineArtsBirminghamAla.
责任	腾讯	美学家	次	AllianceFranaise
匿名	地址	切入	只	AmericanCollegeTest
实用	隐藏	情欲	好	AmitMenghani
网易	新闻	荐	频道	Andrews
作者	娱乐	美学	书	ArlingtonVa.
独家	onlyqshen	阐释	头条	Aston

娱乐	王勇	李磊	BigNews	AtlanticCommunity
凤凰	蒋勋	自述	成	AustinTexas
发表	切入	来源	周	Banneker
腾讯	laineyliu	伦理	条	Bath
IP	美学家	孤独	已经	Belfast
隐藏	情欲	出版社	新	BellevueBellevueWash.
地址	荐	反思	却	BellevueWash.
陶学钢	阐释	融	网	BerryvilleVa.
记者	美学	出版	评论	Bethesda-ChevyChase
出版	出版社	编辑	做	BethesdaMd.
导演	孤独	一体	精彩	BloomfieldHillsMich.
onlyqshen	自述	正文	名	BookerT.Washington
出版社	伦理	特有	家	BrianLang
王勇	李磊	追问	天	BriarcliffBriarcliffManorN.Y.
孤独	反思	面向	小	BrightonRochesterN.Y.
反思	出版	发表	凤凰	BrightonSecondaryCollege
阐释	融	批判	摘	BronxvilleBronxvilleN.Y.
美学	一体	情感	再	BusinessStudies
切入	特有	炒作	想	ButlerMatthewsN.C.
情感	追问	凤凰	万	CAMBRIDGE
蒋勋	情感	捅	第一	CARNEGIE
伦理	批判	热点	记者	CCM
荐	面向	思维	榜	Calif.
情欲	炒作	网易	认为	CaliforniaInstTech
laineyliu	正文	步步为营	种	CaliforniaInstituteofTechnology

附录 21 特征词选择算法效果验证 RI 数据

news_chi

dimension=[50,100,150,200,250,300,350,400,450,500]

testCorpusCnt=3254

naivebayes_rate_chi=[1684,1610,1517,1892,2303,2430,2515,2542,2526,2594]./testCorpusCnt;

mulbayes_rate_chi=[2495,2373,2428,2506,2688,2643,2676,2678,2671,2699]./testCorpusCnt;

dectree_rate_chi=[2726,2762,2851,2918,2960,2963,2940,2956,2971,3015]./testCorpusCnt;

knn_rate_chi=[2692,2703,2728,2722,2744,2780,2812,2788,2769,2765]./testCorpusCnt;

smo_rate_chi=[2723,2771,2830,2910,2975,2985,3022,3030,3050,3041]./testCorpusCnt;

news_chifit

dimension=[50,100,150,200,250,300,350,400,450,500]

testCorpusCnt=3254

naivebayes_rate_chifit=[1734,1803,1676,1905,2157,2451,2488,2578,2569,2567]./testCorpusCnt;

mulbayes_rate_chifit=[2632,2533,2485,2517,2556,2639,2639,2660,2664,2672]./testCorpusCnt;

dectree_rate_chifit=[2864,2908,2917,2926,2952,2951,2957,2979,2965,2955]./testCorpusCnt;

knn_rate_chifit=[2769,2791,2775,2727,2736,2794,2793,2801,2777,2772]./testCorpusCnt;

smo_rate_chifit=[2784,2865,2881,2925,2975,3018,3017,3031,3043,3047]./testCorpusCnt;

Reuters_chi

dimension=[50,100,150,200,250,300,350,400,450,500]

testCorpusCnt=2676

naivebayes_rate_chi=[1081,1120,1190,1469,1599,1670,1730,1732,1749,1799]./testCorpusCnt;

mulbayes_rate_chi=[1922,2000,2073,2087,2155,2148,2153,2150,2140,2113]./testCorpusCnt;

dectree_rate_chi=[1908,1976,2050,2065,2101,2124,2150,2140,2106,2134]./testCorpusCnt;

knn_rate_chi=[1930,1980,2049,2099,2131,2146,2150,2154,2123,2116]./testCorpusCnt;

smo_rate_chi=[1939,1996,2133,2150,2217,2218,2239,2241,2224,2230]./testCorpusCnt;

Reuters_chifit

dimension=[50,100,150,200,250,300,350,400,450,500]

testCorpusCnt=2676

naivebayes_rate_chifit=[1129,1236,1324,1545,1593,1685,1766,1765,1753,1702]./testCorpusCnt;

mulbayes_rate_chifit=[1955,1966,2024,2066,2093,2118,2113,2119,2107,2102]./testCorpusCnt;

dectree_rate_chifit=[1981,2050,2057,2115,2088,2131,2133,2138,2146,2133]./testCorpusCnt;

knn_rate_chifit=[2015,2092,2085,2104,2105,2127,2116,2116,2108,2104]./testCorpusCnt;

smo_rate_chifit=[1996,2124,2185,2219,2232,2253,2250,2249,2246,2240]./testCorpusCnt;

News_KNN

testCorpusCnt=3254;

dimension=[50,100,150,200,250,300,350,400,450,500,600,700,800,900,1000,1500,2000,3000,5000,8000];

knnrate_chifit=[2769,2791,2775,2727,2736,2794,2793,2801,2777,2772,2758,2757,2805,2766,2759,2700,2562,2

425,2102,1870]./testCorpusCnt;
knnrate_chi=[2692,2703,2728,2722,2744,2780,2812,2788,2769,2765,2773,2776,2788,2763,2752,2706,2617,2445,2111,1905]./testCorpusCnt;
knnrate_ig=[2557,2560,2669,2731,2788,2841,2820,2824,2835,2813,2772,2783,2781,2768,2741,2605,2345,2113,1821,1564]./testCorpusCnt;
knnrate_df=[2405,2613,2598,2586,2578,2565,2591,2598,2563,2560,2554,2546,2549,2496,2476,2369,1986,1445,1258,1253]./testCorpusCnt;
knnrate_mi=[2,3,4,4,4,4,4,12,12,13,9,39,69,72,76,128,128,139,618,659]./testCorpusCnt;

News_Bayes

testCorpusCnt=3254;
bayesrate_chifit=[2632,2533,2485,2517,2556,2639,2639,2660,2664,2672,2682,2695,2747,2746,2727,2772,2809,2824,2795,2711]./testCorpusCnt;
bayesrate_chi=[2495,2373,2428,2506,2688,2643,2676,2678,2671,2699,2700,2732,2758,2743,2746,2778,2811,2824,2795,2710]./testCorpusCnt;
bayesrate_ig=[2256,2247,2383,2546,2654,2648,2657,2666,2674,2676,2683,2723,2736,2732,2737,2811,2814,2805,2786,2714]./testCorpusCnt;
bayesrate_df=[2102,2392,2481,2515,2507,2468,2440,2467,2383,2290,2318,2347,2378,2376,2391,2489,2582,2632,2660,2630]./testCorpusCnt;
bayesrate_mi=[2,3,4,4,4,4,4,12,12,13,17,43,69,73,81,125,126,127,615,705]./testCorpusCnt;
dimension=[50,100,150,200,250,300,350,400,450,500,600,700,800,900,1000,1500,2000,3000,5000,8000];

Reuters_KNN

testCorpusCnt=2676;
knnrate_chifit=[2015,2092,2085,2104,2105,2127,2116,2116,2108,2104,2053,2060,2065,2066,2067,2054,2032,2012,1954,1892]./testCorpusCnt;
knnrate_chi=[1930,1980,2049,2099,2131,2146,2150,2154,2123,2116,2099,2081,2075,2067,2080,2046,2060,2116,2080,1965]./testCorpusCnt;
knnrate_ig=[1976,2045,2062,2132,2111,2139,2135,2126,2119,2136,2110,2109,2076,2055,2035,1926,1815,1737,1509,1345]./testCorpusCnt;
knnrate_df=[1834,1859,1867,1865,1873,1909,1898,1898,1869,1870,1852,1816,1805,1766,1741,1609,1515,1343,1150,1164]./testCorpusCnt;
knnrate_mi=[0,3,5,7,1,2,0,2,2,2,3,1,1,1,1,1,13,66,135,241]./testCorpusCnt;
dimension=[50,100,150,200,250,300,350,400,450,500,600,700,800,900,1000,1500,2000,3000,5000,8000];

ReutersBayes

testCorpusCnt=2676;

```
bayesrate_chifit=[1955,1966,2024,2066,2093,2118,2113,2119,2107,2102,2101,2093,2090,2085,2084,2044,2020
,1998,1954,1914]./testCorpusCnt;
bayesrate_chi=[1922,2000,2073,2087,2155,2148,2153,2150,2140,2113,2095,2080,2073,2052,2049,2027,2004,2
000,1959,1916]./testCorpusCnt;
bayesrate_ig=[1941,2058,2083,2160,2166,2176,2192,2197,2208,2203,2186,2179,2171,2160,2151,2142,2125,21
00,2040,1935]./testCorpusCnt;
bayesrate_df=[1794,1883,1928,1965,1995,2042,2033,2033,2043,2031,2047,2056,2071,2057,2063,2071,2063,20
35,2004,1962]./testCorpusCnt;
bayesrate_mi=[0,3,6,7,10,10,10,12,12,12,12,16,16,17,24,30,30,50,136,216]./testCorpusCnt;
dimension=[50,100,150,200,250,300,350,400,450,500,600,700,800,900,1000,1500,2000,3000,5000,8000];
```

附录 22 卡方与 chifit 特征词集合对称差集合大小

dimension=[50,100,150,200,250,300,350,400,450,500,600,700,800,900,1000,1500,2000,3000,5000,8000]

amount_Reuters=[26,52,54,60,50,74,66,100,80,60,144,74,116,110,144,132,164,100,142,152]

amount_News=[20,26,26,18,34,30,46,38,38,50,70,62,78,96,86,116,142,142,186,108]

附录 23 Chifit 与卡方特征词选择算法的特征集差集

Chifit 与卡方特征词选择算法 50 维度特征词集合下差异词汇情况透视说明。

Reuters 语料上只在 chifit 方法生成的特征词集合中不在 chi 方法生成的特征词集合的词汇：

词汇 ID	词汇	词汇贡献度最大的类别	该类别中包含该词汇的文档数目	该类别包含文档的总数目	词汇索引的文章总数
1	Reuter	gas	16	16	5256
2	pct	retail	16	16	2083
3	deficit	bop	24	26	188
4	seasonally	housing	12	14	74
5	wholesale	wpi	10	13	35
6	steel	iron-steel	21	31	61
7	industrial	ipi	25	31	175
8	dollar	dlr	14	15	338
9	gas	nat-gas	21	25	195
10	account	bop	21	26	163
11	gross	gnp	39	71	90
12	reserves	reserves	28	35	213
13	rules	cocoa	25	43	69

Reuters 语料上只在 chi 方法生成的特征词集合中不在 chifit 方法生成的特征词集合的词汇

词汇 ID	词汇	词汇贡献度最大的类别	该类别中包含该词汇的文档数目	该类别包含文档的总数目	词汇索引的文章总数
1	Shr	earn	1181	2698	1181
2	Manaspas	rubber	9	30	9
3	Xuto	rubber	9	30	9
4	multi-family	housing	5	14	6
5	unemployed	jobs	10	34	10
6	Paz	tin	6	18	7
7	Comibol	tin	5	18	5
8	Estenssoro	tin	5	18	5
9	Tin	tin	7	18	10
10	palm	veg-oil	23	59	34
11	grain	grain	113	334	141
12	crude	crude	121	295	182
13	Conference	rubber	10	30	13

中文新闻语料上只在 chifit 方法生成的特征词集合中不在 chi 方法生成的特征词集合的词汇

词汇 ID	词汇	词汇贡献度最大的类别	该类别中包含该词汇的文档数目	该类别包含文档的总数目	词汇索引的文章总数
1	编辑	reading	4000	4000	8962
2	年	reading	4000	4000	11156
3	日	reading	4000	4000	11848
4	月	reading	4000	4000	12063
5	实用	education	68	75	257
6	独家	education	68	75	393
7	凤凰	reading	3160	4000	5367
8	发表	reading	2986	4000	4756
9	记者	society&law	2174	3340	5173
10	导演	entertainment	462	828	863

中文新闻语料上只在 chi 方法生成的特征词集合中不在 chifit 方法生成的特征词集合的词汇

词汇 ID	词汇	词汇贡献度最大的类别	该类别中包含该词汇的文档数目	该类别包含文档的总数目	词汇索引的文章总数
1	美学家	reading	1617	4000	1622

2	自述	reading	1659	4000	1711
3	李磊	reading	1617	4000	1646
4	融	reading	1655	4000	1734
5	一体	reading	1672	4000	1793
6	特有	reading	1686	4000	1828
7	追问	reading	1684	4000	1827
8	批判	reading	1812	4000	2062
9	面向	reading	1667	4000	1802
10	炒作	reading	1666	4000	1813

附录 24 关键词形态语义聚类算法相关

语料：

2 维条码，2 维条形码，A-稳定，A-稳定性，ATM 网，ATM 网络，Ad Hoc 网络，Ad Hoc 网，Ad Hoc 网络，Ad hoc 网络，Ad hoc 网络，AdHoc 网，AdHoc 网络，Adhoc 网，Agent 组件，Agent 组织，Allan 方差，Allan 方差法，B-P 神经网络，BP 神经网，BP 神经网络，BP 神经元网络，CT 图像，CT 图象，Cache 命中，Cache 命中率，Cache 失效，Cache 失效率，D-S 证据理论，D-S 证据论，D—稳定，D-稳定性，DC 图像，DC 图象，ER 模式，ER 模型，GSM 模块，GSM 模型，Hopfield 网，Hopfield 网络，K-近邻法，K-近邻法则，L-稳定，L-稳定性，Lagrangian 松弛，lagrangian 松驰，Monte Carlo 法，Monte Carlo 方法，Monte-Carlo 法，Monte-Carlo 方法，Monte-carlo 法，NP 难，NP 完全，NP-困难，NP-难，NP-完全，NP 难度，NP 难度问题，NP 难题，NP 难问题，NP 完全，NP 完全性，Popov 超稳定理论，Popov 超稳定性理论，SAR 图像，SAR 图象，Schur 稳定，Schur 稳定性，TM 图像，TM 图象，k-错误线性复杂度，k-错线性复杂度，k-近邻，k-最近邻，α -β -γ -δ 滤波，α -β -γ 滤波，α -β 滤波，α -β 算法，β 算法，δ ++-规则，δ -规则，δ -算子，δ 算子，μ -演算，π -演算，π －演算，π 演算，χ -演算，变长度染色体编码，变长染色体编码，变尺度法，变尺度算法，变精度粗糙集，变精度粗糙集模型，变精度粗集，变精度粗集模型，不确定推理，不确定系数，不确定性大系统，不确定性动态系统，不确定性分析，不确定性时滞系统，不确定性推理，不确定性系数，不确定性系统，布局模式，布局模型，布料模拟，布料模型，步进电动机，步进电机，部分最小二乘，部分最小二乘法，采样模块，采样模拟，彩色视觉，彩色视频，彩色图像，彩色图象，参考模式，参考模型，参数不确定，参数不确定系统，参数不确定性，参数不确定性系统，词汇聚类，词汇相似度，词聚类，词相似度，词序相似度，词义相似度，词语语义相似度，单纯形法，单纯形方法，电路实现，电路实验，定理证明，定理证明器，动态规划法，动态规划算法，动态模糊，动态模拟，动态模式，短路电抗，短路电流，短期负荷预报，短期负荷预测，断层图像，断层图象，仿射不变量，仿射不变性，仿真模块，仿真模拟，仿真模型，仿真模型组件，仿真模型组态，仿真培训，仿真培养，访问控制，访问控制表，访问控制模式，访问控制模型，非负矩阵分解，非匹配不确定，非匹配不确定性，非线性不确定，非线性不确定性，非线性超声，非线性超声场，非线性特性，非线性特征，非线性系数，非线性系统，非线性最小二乘，非线性最小二乘法，客户-服务器模式，客户/服务器模式，客户/服务器模型，客户／服务器模式，客户／服务器模型，李雅普诺夫稳定，李雅普诺夫稳定理论，李雅普诺夫稳定性，李雅普诺夫稳定性理论，李亚普诺夫稳定理论，李亚普诺夫稳定性理论，力/位混合控制，力/位控制，力/位置混合控制，力/位置控制，力／位混合控制，力／位置混合控制，力传感器，力控制系统，力学传感器，力学控制系统，潜语义分析，

潜在语义分析，群决策支持系统，群体决策支持系统，文本图像，文本图象，纹理图像，纹理图象，稳定判据，稳定性判据，信息增量，信息增强，虚拟专网，虚拟专用网络，运动模板，运动模糊，运动模拟，运动模式，运动模型，运动特性，增广矩阵，增广矩阵法，耦合系数，耦合系统，安全评估，安全评价，参数曲线，参数曲面

标注答案：

(1)2 维条码，2 维条形码；(2) A-稳定，A-稳定性；(3) ATM 网，ATM 网络；(4) Ad Hoc 网络，Ad Hoc 网，Ad Hoc 网络，Ad hoc 网络，Ad hoc 网络，AdHoc 网，AdHoc 网络，Adhoc 网；(5) Agent 组件；(6)Agent 组织；(7)Allan 方差；(8)Allan 方差法；(9)B-P 神经网络，BP 神经网，BP 神经网络，BP 神经元网络；(10)CT 图像，CT 图象；(11)Cache 命中；(12)Cache 命中率；(12) Cache 失效；(13) Cache 失效率；(14) D-S 证据理论，D-S 证据论；(15) D—稳定，D-稳定性；(16) DC 图像，DC 图象；(17)ER 模式；(18)ER 模型；(19)GSM 模块；(20)GSM 模型；(21)Hopfield 网，Hopfield 网络；(22)K-近邻法，K-近邻法则；(23)L-稳定，L-稳定性；(24)Lagrangian 松弛，lagrangian 松驰；(25)Monte Carlo 法，Monte Carlo 方法，Monte-Carlo 法，Monte-Carlo 方法，Monte-carlo 法；(26)NP 难，NP-困难，NP-难，NP 难度；(27)NP 完全，NP-完全，NP 完全，NP 完全性；(28)NP 难度问题，NP 难题，NP 难问题；(29)Popov 超稳定理论，Popov 超稳定性理论；(30)SAR 图像，SAR 图象；(31)Schur 稳定，Schur 稳定性；(32)TM 图像，TM 图象；(33)k-错误线性复杂度，k-错线性复杂度；(34)k-近邻；(35)k-最近邻；(36)α -β -γ -δ 滤波；(37)α -β -γ 滤波；(38)α -β 滤波；(39)α -β 算法；(40)β 算法；(41)δ ++-规则；(42)δ -规则；(43)δ -算子，δ 算子；(44)μ -演算；(45)π -演算，π 一演算，π 演算；(46)χ -演算；(47)变长度染色体编码，变长染色体编码；(48)变尺度法，变尺度算法；(49)变精度粗糙集，变精度粗集；(50)变精度粗糙集模型，变精度粗集模型；(51)不确定推理，不确定性推理；(52)不确定系数，不确定性系数；(53)不确定性系统；(54)不确定性大系统；(55)不确定性动态系统；(56)不确定性分析；(57)不确定性时滞系统；(58)布局模式；(59)布局模型；(60)布料模拟；(61)布料模型；(62)步进电动机，步进电机；(63)部分最小二乘，部分最小二乘法；(64)采样模块；(65)采样模拟；(66)彩色视觉；(67)彩色视频；(68)彩色图像，彩色图象；(69)参考模式；(70)参考模型；(71)参数不确定，参数不确定性；(72)参数不确定系统，参数不确定性系统；(73)词汇聚类，词聚类；(74)词汇相似度，词相似度；(75)词序相似度；(76)词义相似度，词语语义相似度；(77)单纯形法，单纯形方法；(78)电路实现；(79)电路实验；(80)定理证明；(81)定理证明器；(82)动态规划法，动态规划算法；(83)动态模糊；(84)动态模拟；(85)动态模式；(86)短路电抗；(87)短路电流；(88)短期负荷预报，短期负荷预测；(88)断层图像，断层图象；(89)仿射不变量；(90)仿射不变性；(91)仿真模块；(92)仿真模拟；(93)仿真模型；(94)仿真模型组件；(95)仿真模型组态；(96)仿真培训；(97)仿真培养；(98)访问控制；(99)访问控制表；(100)访问控制模式；(101)访问控制模型；(102)非负矩阵分解；(103)非匹配不确定，非匹配不确定性；(104)非线性不确定，非线性不确定性；(105)非线性超声；(106)非线性超声场；(107)非线性特性，非线性特征；(108)非线性系数；(109)非线性系统；(110)非线性最小二乘，非线性最小二乘法；(111)客户-服务器模式，客户/服务器模式，客户／服务器模式；(112)客户/服务器模型，客户／服务器模型；(113)李雅普诺夫稳定，李雅普诺夫稳定性；(114)李雅普诺夫稳定理论，李雅普诺夫稳定性理论，李亚普诺夫稳定理论，李亚普诺夫稳定性理论；(115)力/位混合控制，力/位置混合控制，力／位混合控制，力／位置混合控制；(116)力/位控制，力/位置控制；(117)力传感器，力学传感器；(118)力控制系统，力学控制系统；(119)潜语义分析，潜在语义分析；(120)群决策支持系统，群体决策支持系统；(121)文本图像，文本图象；(122)纹理图像，纹理图象；(123)稳定判据，稳定性判据；(124)信息增量；125 信息增强；(126)虚拟专网，虚拟专用网络；(127)运动模板；(128)运动模糊；(129)运动模拟；(130)运动模式；(131)运动模型；(132)运动特性；(133)增广矩阵，增广矩阵法；(134)耦合系数；(135)耦合系统；(136)安全评价，安全

评估；(137)参数曲线；(138)参数曲面。

编辑距离二次计算方法

(1)2 维条码，2 维条形码；(2)A-稳定，A-稳定性；(3)ATM 网，ATM 网络；(4)Ad Hoc 网络，Ad Hoc 网，Ad Hoc 网络，Ad hoc 网络，Ad hoc 网络，AdHoc 网，AdHoc 网络，Adhoc 网；(5)Agent 组件；(6)Agent 组织；(7)Allan 方差；(8)Allan 方差法；(9)B-P 神经网络，BP 神经网，BP 神经网络，BP 神经元网络；(10)CT 图像，CT 图象；(11)Cache 命中，Cache 命中率；(12)Cache 失效，Cache 失效率；(13)D-S 证据理论，D-S 证据论；(14)D—稳定，D-稳定性；(15)DC 图像，DC 图象；(16)ER 模式；(17)ER 模型；(18)GSM 模块；(19)GSM 模型；(20)Hopfield 网，Hopfield 网络；(21)K-近邻法，K-近邻法则；(22)L-稳定，L-稳定性；(23)Lagrangian 松弛，lagrangian 松驰；(24)Monte Carlo 法，Monte Carlo 方法，Monte-Carlo 法，Monte-Carlo 方法，Monte-carlo 法；(25)NP 难，NP-困难，NP-难，NP 难度，NP 难题；(26)NP 完全，NP-完全，NP 完全，NP 完全性；(27)NP 难度问题，NP 难题，NP 难问题；(28)Popov 超稳定理论，Popov 超稳定性理论；(29)SAR 图像，SAR 图象；(30)Schur 稳定，Schur 稳定性；(31)TM 图像，TM 图象；(32)k-错误线性复杂度，k-错线性复杂度；(33)k-近邻；(34)k-最近邻；(35)α -β -γ -δ 滤波；(36)α -β -γ 滤波；(37)α -β 滤波；(38)α -β 算法；(39)β 算法；(40)δ ++-规则，δ -规则；(41)δ -算子，δ 算子；(42)μ -演算；(43)π -演算，π －演算，π 演算；(44)χ -演算；(45)变长度染色体编码，变长染色体编码；(46)变尺度法，变尺度算法；(47)变精度粗糙集，变精度粗集；(48)变精度粗糙集模型，变精度粗集模型；(49)不确定推理，不确定性推理；(50)不确定系数，不确定性系数；(51)不确定性大系统；(52)不确定性动态系统；(53)不确定性分析；(54)不确定性时滞系统；(55)不确定性系统；(56)布局模式；(57)布局模型；(58)布料模拟；(59)布料模型；(60)步进电动机，步进电机；(61)部分最小二乘，部分最小二乘法；(62)采样模块；(63)采样模拟；(64)彩色视觉；(65)彩色视频；(66)彩色图像，彩色图象；(67)参考模式；(68)参考模型；(69)参数不确定，参数不确定性；(70)参数不确定系统，参数不确定性系统；(71)词汇聚类，词聚类；(72)词汇相似度，词相似度；(73)词序相似度；(74)词义相似度，词语语义相似度；(75)单纯形法，单纯形方法；(76)电路实现；(77)电路实验；(78)定理证明；(79)定理证明器；(80)动态规划法，动态规划算法；(81)动态模糊；(82)动态模拟；(83)动态模式；(84)短路电抗；(85)短路电流；(86)短期负荷预报；(87)短期负荷预测；(88)断层图像，断层图象；(89)仿射不变量；(89)仿射不变性；(90)仿真模块；(91)仿真模拟；(92)仿真模型；(93)仿真模型组件；(94)仿真模型组态；(95)仿真培训；(96)仿真培养；(97)访问控制；(98)访问控制表；(99)访问控制模式；(100)访问控制模型；(101)非负矩阵分解；(102)非匹配不确定，非匹配不确定性；(103)非线性不确定，非线性不确定性；(104)非线性超声；(105)非线性超声场；(106)非线性特性；(107)非线性特征；(108)非线性系数；(109)非线性系统；(120)非线性最小二乘，非线性最小二乘法；(121)客户-服务器模式，客户/服务器模式，客户／服务器模式；(122)客户/服务器模型，客户／服务器模型；(123)李雅普诺夫稳定，李雅普诺夫稳定性；(124)李雅普诺夫稳定理论，李雅普诺夫稳定性理论，李亚普诺夫稳定理论，李亚普诺夫稳定性理论；(125)力/位混合控制，力/位置混合控制，力／位混合控制，力／位置混合控制；(126)力/位控制，力/位置控制；(127)力传感器，力学传感器；(128)力控制系统，力学控制系统；(129)潜语义分析，潜在语义分析；(130)群决策支持系统，群体决策支持系统；(131)文本图像，文本图象；(132)纹理图像，纹理图象；(133)稳定判据，稳定性判据；(134)信息增量；(135)信息增强；(136)虚拟专网，虚拟专用网络；(137)运动模板；(138)运动模糊；(139)运动模拟；(140)运动模式；(141)运动模型；(142)运动特性；(143)增广矩阵，增广矩阵法；(144)耦合系数；(145)耦合系统；(146)安全评估；(147)安全评价；(148)参数曲线；(149)参数曲面；

传统编辑距离方法

(1)维条码，2 维条形码；(2)A-稳定，A-稳定性，L-稳定；(3)ATM 网，ATM 网络；(4)Ad Hoc 网络，Ad Hoc 网络，Ad hoc 网络；(5)Ad Hoc 网，Ad Hoc 网络，AdHoc 网；(6)Ad hoc 网络；(7)AdHoc 网络；(8)Adhoc 网；(9)Agent 组件，Agent 组织；(10)Allan 方差，Allan 方差法；(11)B-P 神经网络，BP 神经网络；(12)BP 神经网，BP 神经网络；(13)BP 神经元网络；(14)CT 图像，CT 图象；(15)Cache 命中，Cache 命中率；(16)Cache 失效，Cache 失效率；(17)D-S 证据理论，D-S 证据论；(18)D—稳定；(19)D-稳定性，L-稳定性；(20)DC 图像，DC 图象；(21)ER 模式，ER 模型；(22)GSM 模块，GSM 模型；(23)Hopfield 网，Hopfield 网络；(24)K-近邻法，K-近邻法则；(25)Lagrangian 松弛；(26)lagrangian 松驰；(27)Monte Carlo 法，Monte Carlo 方法，Monte-Carlo 法；(28)Monte-Carlo 方法；(29)Monte-carlo 法；(30)NP 难，NP-难；(31)NP 完全，NP-完全，NP 完全；(32)NP-困难，NP-难；(33)NP 难度，NP 难题；(34)NP 难度问题，NP 难问题；(35)NP 完全性；(36)Popov 超稳定理论，Popov 超稳定性理论；(37)SAR 图像，SAR 图象；(38)Schur 稳定，Schur 稳定性；(39)TM 图像，TM 图象；(40)k-错误线性复杂度，k-错线性复杂度；(41)k-近邻，k-最近邻；(42)α -β -γ -δ 滤波；(43)α -β -γ 滤波；(44)α -β 滤波；(45)α -β 算法；(46)β 算法；(47)δ ++-规则；(48)δ -规则；(49)δ -算子，δ 算子；(50)μ -演算，π -演算，χ -演算；(51)π 一演算，π 演算；(52)变长度染色体编码，变长染色体编码；(53)变尺度法，变尺度算法；(54)变精度粗糙集，变精度粗集；(55)变精度粗糙集模型，变精度粗集模型；(56)不确定推理，不确定性推理；(57)不确定系数，不确定性系数；(58)不确定性大系统，不确定性系统；(59)不确定性动态系统；(60)不确定性分析；(61)不确定性时滞系统；(62)布局模式，布局模型；(63)布料模拟，布料模型；(64)步进电动机，步进电机；(65)部分最小二乘，部分最小二乘法；(66)采样模块，采样模拟；(67)彩色视觉，彩色视频；(68)彩色图像，彩色图象；(69)参考模式，参考模型；(70)参数不确定，参数不确定性；(71)参数不确定系统，参数不确定性系统；(72)词汇聚类，词聚类；(73)词汇相似度，词相似度，词序相似度，词义相似度；(74)词语语义相似度；(75)单纯形法，单纯形方法；(76)电路实现，电路实验；(77)定理证明，定理证明器；(78)动态规划法，动态规划算法；(79)动态模糊，动态模拟，动态模式；(80)短路电抗，短路电流；(81)短期负荷预报，短期负荷预测；(82)断层图像，断层图象；(83)仿射不变量，仿射不变性；(84)仿真模块，仿真模拟，仿真模型；(85)仿真模型组件，仿真模型组态；(86)仿真培训，仿真培养；(87)访问控制，访问控制表；(88)访问控制模式，访问控制模型；(89)非负矩阵分解；(90)非匹配不确定，非匹配不确定性；(91)非线性不确定，非线性不确定性；(92)非线性超声，非线性超声场；(93)非线性特性，非线性特征；(94)非线性系数，非线性系统；(95)非线性最小二乘，非线性最小二乘法；(96)客户-服务器模式，客户/服务器模式，客户／服务器模式；(97)客户/服务器模型，客户／服务器模型；(98)李雅普诺夫稳定，李雅普诺夫稳定性；(99)李雅普诺夫稳定理论，李雅普诺夫稳定性理论，李亚普诺夫稳定理论；(100)李亚普诺夫稳定性理论；(101)力/位混合控制，力/位置混合控制，力／位混合控制；(102)力/位控制，力/位置控制；(103)力／位置混合控制；(104)力传感器，力学传感器；(105)力控制系统，力学控制系统；(106)潜语义分析，潜在语义分析；(107)群决策支持系统，群体决策支持系统；(108)文本图像，文本图象；(109)纹理图像，纹理图象；(120)稳定判据，稳定性判据；(121)信息增量，信息增强；(122)虚拟专网；(123)虚拟专用网络；(124)运动模板，运动模糊，运动模拟，运动模式，运动模型；(125)运动特性；(126)增广矩阵，增广矩阵法；(127)耦合系数，耦合系统；(128)安全评估，安全评价；(129)参数曲线，参数曲面；