摘　要

知识产权信息，尤其是专利信息蕴藏丰富的技术、法律、经济和战略情报，在知识产权的创造、保护、管理和商业化的过程中都发挥着至关重要的作用。专利信息是指某项技术在谋取专利权过程中的各种信息，它具有重要的战略价值，是国家科技信息系统中重要的组成部分，是信息资源开发的重点。如何科学地使用专利信息和做好专利分析工作，是目前专利研究领域的重要课题。

本文从专利检索用户的角度出发，对美国专利数据属性进行了重新规划。根据专利信息专利权人（申请人）属性的特点，利用信息抽取、关联分析等技术，提出基于关联规则的同指消解抽取模型。同时，利用此方法对通信专利数据进行模型训练，从中抽取出可不断扩充的同指辞典。该辞典可用于建立专利检索中的申请人公司树，从而提高专利在申请人检索方面的查全率。

另外，根据专利信息发明人属性的特点，利用信息抽取、聚类分析等技术，构建基于聚类分析的异指消解抽取模型，提出了一套全新的命名实体识别模型及其算法，并选择合适的抽取结果输出方式。然后，通过实证数据进行模型实验，从中抽取出可维护和可扩展的异指库，以便建立专利检索中的发明人异指标引，提高专利在发明人检索方面的查准率。

本文有机地结合辞典、规则和统计模型方法，提出了基于关联规则的同指消解模型和基于聚类分析的异指消解模型，并在此基础上进行了大量的人工指导和机器学习训练。实验结果表明，本文所设计的信息抽取系统基本令人满意。

关键词：信息抽取　数据挖掘　专利信息　同指 异指

ABSTRACT

Intellectual property information, especially patent information which contains technical, legal, economic and strategic intelligence, plays a crucial role in the creation, protection, management and commercialization of the process of intellectual property rights. Patent Information contains all kinds of information in the process of figure of patent right. It is of important strategic value; it is important component of the National Science and Technology Information System. How to make full use of patent information and do a good job in patent research is an important subject.

On the patent search users' view, we re-plane the United States patent data. We focus on extraction in assignee and inventor by using of natural language, information extraction, data mining and so on, in order to build Co-reference glossary and De-reference glossary. That effectively improves the efficiency of patent search and the country's patent strategy and decision-making patent services.

At first, this article makes use of sophisticated information retrieval technique to customized download American communication patent. Then, pretreatment for patent data and loading the patent information blocks into database would be illustrated. Consequently, accidence analysis is presented that extracting the patent abstract to finish participles of dictionary rules and part of speech label based on to obtain the sign sequence of the patent abstract. By the fourth step, respectively on the establishment of the association rules Based on that model and with the clustering of different rules that model the fifth step is to complete the rule-based and the candidate word from the entities to identify. Finally, the keyword which accords with integrity and logicality has been filled into the result library.

This article from organically integrated, rules and statistical model put up a model of based on the proposed rules associated and model of based on cluster analysis. On this basis, we play a lot of manual guidance and machine learning training. From the experimental results, the recall and precision of information extraction

system designed in this article are acceptable.

Keyword: Information Extraction; Data Mining; Patent Information; Co-reference; De-reference

目　录

图目录

表目录

1 绪　论

1.1 选题背景

自中国入世以来，市场的全球化要求我国企业必须遵循以知识产权为核心的国际竞争规则。而相对处于弱势的我们，在知识产权领域已连遭重创并面临日益严峻的挑战。我国企业迫切需要站在战略的高度，来认识和处理知识产权问题，制定适合自身发展的知识产权战略，以增强国际竞争力，实现可持续发展[1] [2]。

战略合理、有效的制定离不开全面、准确的信息。知识产权信息，尤其是专利信息蕴藏丰富的技术、法律、经济和战略情报，在知识产权的创造、保护、管理和商业化的过程中都发挥着至关重要的作用。专利信息的有效利用直接关系到知识产权战略的制定及实施[3]。

专利信息是指某项技术在谋取专利权过程中的各种信息，它具有重要的战略价值，是国家科技信息系统中重要的组成部分，是信息资源开发的重点。专利信息的分析研究正在国内外广泛开展。总的说来，对于专利的分析主要从定性和定量两个角度展开。定性分析主要从专利信息的内容着手，通过分析专利中的某些特定信息项以获得相关专利分析情报。定量分析则主要对一些专利中的固有标引项目指标进行统计分析，再从技术和经济的角度对有关统计数据的变化进行解释，以获得动态发展趋势的分析结果[4]。

Field Code	Field Name	Field Code	Field Name
PN	Patent Number	IN	Inventor Name
ISD	Issue Date	IC	Inventor City
TTL	Title	IS	Inventor State
ABST	Abstract	ICN	Inventor Country
ACLM	Claim(s)	LREP	Attorney or Agent
SPEC	Description/Specification	AN	Assignee Name
CCL	Current US Classification	AC	Assignee City
ICL	International Classification	AS	Assignee State
APN	Application Serial Number	ACN	Assignee Country
APD	Application Date	EXP	Primary Examiner
PARN	Parent Case Information	EXA	Assistant Examiner
RLAP	Related US App. Data	REF	Referenced By
REIS	Reissue Data	FREF	Foreign References
PRIR	Foreign Priority	OREF	Other References
PCT	PCT Information	GOVT	Government Interest
APT	Application Type		

图 1.1 专利基本信息图

如图 1.1 所示的专利标引项示例可知，一条完整的专利包含了 31 项标引内容[5]。目

前的专利研究主要围绕着专利的申请日期、发明人、专利权人（所属机构）、国家、IPC分类号和引文关系的统计分析展开，但是却存在着诸多不足：比如记录着关键技术信息的专利摘要一直得不到有效地利用；发明人存在的同名同姓现象无法区分；相同机构的不同名称无法合并等。究其原因是由于目前专利分析仅仅是一些简单的统计分析，如针对领域专利数量、申请者、所在机构、申请国家的分析，不具备自然语言的功能，因此无法对其进行有效分析。不能有效处理包括专利摘要、发明人、专利权人（所属机构）等属性在内的专利文本信息，直接影响了专利信息的利用率，也制约着专利分析向更深层次的内容挖掘方面发展。为了解决当前专利信息分析所面临的问题，本文创新性的将数据挖掘和信息抽取技术引入到专利信息的分析应用中，以便有效地分析和处理专利信息，从中获得专利技术信息，填补目前专利信息分析研究中的空白，将定性与定量分析方法结合起来，为我国专利信息分析的发展提供有益的参考[6]。

1.2 研究的主要内容和意义

在本课题中，我们将研究重点放在专利信息的有效利用上，把信息抽取技术、数据挖掘技术应用在专利信息分析中，充分发挥信息抽取和数据挖掘技术在处理海量文本信息方面的优势，以期实现自动地抽取申请人、发明人等的重要信息，并尝试融合先进的专利信息分析方法，建立一套全新的专利信息分析系统以替代传统的人工分析，从而提高专利信息分析工作的质量和效率，为国家的专利战略服务。首先，文本理解不是本文研究的重点，所以本文所提到的方法很少涉及深层次的自然语言理解问题，只是应用数据挖掘和自然语言处理过程中相关的统计方法。另一方面，本文主要研究将发明人、专利权人（所属机构）进行同指和异指关系关联和区别，再通过人工指导训练和机器学习相结合的方式设计同指和异指信息抽取的实验平台。

本文的研究目的是设计基于关联规则的同指信息抽取模型和基于聚类方法的异指信息抽取模型，主要工作归纳如下：

（1）对信息抽取和数据挖掘中的关联规则和聚类分析进行了综述，并描述信息抽取的评价方法，设计了基于关联规则的同指信息抽取模型和基于聚类分析方法的异指信息抽取模型。

（2）根据专利数据源的特征，在数据准备阶段利用知识发现与数据分析实验室的专利自动下载工具从网上专利数据库下载的原始专利数据，再对获取到的专利数据进行清

洗、非相关主题信息过滤、专利信息分块、数据库导入等操作，从而积累了大量真实有效的专利结构化信息。

通过以上研究内容显示，将信息抽取技术应用于专利信息分析中，对于专利信息分析有以下几点意义：

（1）体现了专利分析工作的时效性。对于公开的专利资源，目前分析者常常是通过纸质或互联网粗略收集专利信息，专利中大量关键的技术信息还得通过人工过程加以识别和分析。信息检索技术虽然为找到目标信息提供了很好的支持，但还得根据它提供的地址去访问每一个页面，工作量大且浪费时间。信息抽取技术通过智能化处理过程大大缩短了专利信息的分析处理时间，体现了专利分析工作的时效性。

（2）实现了专利信息的动态监测。信息抽取技术的使用为专利信息的快速分析和传递提供了可能，更有效地实现了专利信息的动态监测。

（3）实现智能化的信息处理。原有的信息获取技术实现的是单纯的信息获取，在信息的识别、判断和分析处理方面明显不足。信息抽取技术本质上是一种信息获取技术，但它在某种程度上实现了信息的自动识别、判断和分析处理。

（4）专利定性和定量分析方法的结合。通过信息抽取将专利摘要中的技术关键词定性的提取出来，就可以进行技术关键词分类、关联分析和统计研究，从而将定量分析方法有机结合起来。

（5）实现规范化的管理。传统的管理方式散乱、不易查找，信息抽取最后结构化的表达方式易于理解且方便管理。充分利用这种现代信息技术，使需要的专利技术信息得到及时、准确的处理，并实现数据库管理的自动化、规范化。

因此，进行专利的信息抽取和数据挖掘研究应用，可以丰富专利信息分析研究方法，提高专利信息利用率，不仅具有理论研究价值，其实践应用价值也非常高。

1.3 论文结构与安排

本文根据结构安排，共分为五个章节：

第一章：绪论包括本文的选题背景、主要研究内容与意义、论文的结构安排以及文章创新点设计

第二章：信息抽取技术综述主要介绍信息抽取技术的概念、研究对象、研究历史及发展现状、信息抽取的类型、方法设计与流程、抽取模型选择和信息抽取系统的性能评

价；数据挖掘技术综述主要介绍数据挖掘的概述、发展、关联规则、聚类技术的介绍。

第三章：基于关联规则的同指消解技术的提出。根据专利信息的特点设计了一个抽取模型，主要包括专利数据源分析、专利数据获取、专利数据存储、专利信息抽取、专利信息服务探讨等以便建立新的理论和方法模型。同时，利用此方法通过通信专利数据进行模型的实验，把准备好的专利数据信息结合人工指导和机器学习训练从中抽取出同指库，并将抽取结果生成基于同指的专利辞典。该辞典可用于建立专利检索中的申请人公司树，从而提高专利在申请人检索方面的查全率。

第四章：基于聚类分析的异指消解技术的建立。专利异指抽取模型的总体框架与流程设计，解决数据准备问题，对专利数据进行预处理，设计辞典、规则与统计方法相结合的分析，提出了一套全新的命名实体识别模型及其算法，并选择合适的抽取结果输出方式。然后，通过实证数据进行模型的实验，结合人工指导和机器学习训练，从专利中抽取出异指库，并将抽取结果生成基于异指关系的专利辞典，以便建立专利检索中的发明人异指标引，提高专利在发明人检索方面的查准率。

第五章：总结和展望总结全文，概述研究工作成果及意义，提出本文的创新之处，明确当前研究的不足和下一步的工作方向。

2 信息抽取和数据挖掘技术综述

2.1 信息抽取

信息抽取是面向结构化、半结构化和非结构化文本所进行的浅层的或者说简化的文本理解技术，其定义为从一段文本或一处信息中抽取指定的一类信息并将其形成结构化的数据填入一个数据库中供用户查询使用的过程[7]。即它从文本中抽取用户感兴趣的事件、实体和关系，然后进入数据库，分析趋势，或进行在线服务。信息抽取还可以看作是信息检索的进一步深化，研究指定信息的查找、理解和抽取，并将指定信息以适当的方式输出。信息抽取已经发展成为自然语言处理领域的一个重要分支，涉及到了深层次的语言理解、篇章分析与推理、多语言文本处理、WEB 信息抽取、名实体识别等自然语言研究领域[8]。

2.1.1 信息抽取的概述

信息抽取(Information Extraction，IE)技术正是这样一种新型的能满足上述要求的自然语言处理技术，它通过对原文档信息内容的分析抽取出有意义的事实生成满足用户要求的简洁的信息[9]。信息抽取系统不仅能帮助人们方便地找到所需信息，而且信息的内容经过合理的分析和组织人们可以高效地获取所感兴趣的信息内容[10]。一方面信息抽取系统从文档(例如 Web 文档)中抽取指定领域的信息并使用信息模板来刻画原文档信息；另一方面信息抽取系统将非结构化的文本结构化,并将结构化的信息组织存储到信息库中使用户能够方便地进行进一步的数据分析和查询工作[11]。信息抽取的任务就是将源文档所包含的信息内容抽取出来并按模板的结构组织存储形成结构化的信息库。在信息抽取得到的结构化信息库的基础上，可以进一步完成信息搜索(Information Search)、数据挖掘(Data Mining)、机器翻译(Machine Translation)、文本生成(Text Generation)等后续信息处理[12][13]。

2.1.2 信息抽取的发展

通过调查我们发现目前信息抽取在专利信息分析方面的应用研究在国内外都还处于起步阶段，而我们将信息抽取技术应用于专利信息的分析更是一项全新的尝试。从另一个方面讲，这也是科学研究中多学科交叉、多技术融合大前提下的发展必然[14]。

虽然尚没有直接以信息抽取应用于专利信息分析的先例，但是信息抽取的概念已经

出现在了很多相关专利信息分析的工作中并发挥着重要的作用：

从自然语言文本中获取结构化信息的研究最早开始于 20 世纪 60 年代中期，这被看作是信息抽取技术的初始研究，它以两个长期的、研究性的自然语言处理项目为代表。一个是美国纽约大学开展的 Linguistic String 项目，开始于 60 年代中期并一直延续到 80 年代。另一个相关的长期项目是由耶鲁大学 Roger Schank 及其同事在 20 世纪 70 年代开展的有关故事理解的研究[15]。从 20 世纪 80 年代末开始，消息理解系列会议（MUC）的召开标志着信息抽取研究蓬勃开展起来。近几年，信息抽取技术的研究与应用更为活跃。以美国国家标准技术研究所（NIST）组织的自动内容抽取正在推动信息抽取研究进一步发展[16][17]。

在研究方面，主要侧重于：利用机器学习技术增强系统的可移植能力、探索深层理解技术、篇章分析技术、多语言文本处理能力、Web 信息抽取（Wrapper）以及对时间信息的处理等等[22]。在应用方面，信息抽取应用的领域非常广泛，除自成系统以外，还与其他文档处理技术结合建立功能强大的信息服务系统[18]。

目前国外现有的比较典型的信息抽取系统主要包括：

ATRANS 系统是早在 1981 年由 Cowie 研究出来关于动植物正规结构描述数据库的系统及其商用化产品。该系统采用了概念句子分析技术，通过一些简单的语言处理技术能够完成限制在小规模，特定专业领域的信息抽取任务[19]。

美国 GE 研究与开发中心的 Lisa F.Rau 等研制的 SCISOR（System for Conceptual Information，Organization and Retrieval）。SCISOR 首先采用关键词过滤和模式匹配的方法对待处理文献进行主题分析，以便判定该报道的内容是否与“公司合并”有关；然后采用自底向上的分析器识别句子结构，生成概念表示；最后应用自顶向下的预期驱动分析器提取预期内容[20]。

美国加里福尼亚斯坦福研究所人工智能中心从 1991 年开始开发的一个基于多层、非确定有限状态自动机模型的自然语言文本信息抽取系统 FASTUS（Finite State Automaton Text Understanding System）[21]。

LaSIE、TIPSTER 系统，分别采用统计学的方法进行词汇标注和语法分析与使用一组通用的文本处理模块满足不同的文本处理应用的需要[22][23]。

由德国人工智能研究中心语言技术实验室（DFKI-LT）在 Paradime 项目中所开发的

一个联机的德语文档信息抽取智能系统 SMES（Saarbrücken Information Extraction System）[24]。

在中文信息抽取领域，国立台湾大学和新加坡肯特岗数字实验室参加了 MUC-7 中文命名实体识别任务的评测。Intel 中国研究中心的 ZHANG Yi-Min 和 ZHOU Joe F 等人在 ACL-2000 上演示了他们开发的一个抽取中文命名实体以及这些实体间相互关系的信息抽取系统。近年来包括中国科学院、北京大学、哈尔滨工业工学和上海交通大学等一批高校和研究机构也在中文抽取方面开展了大量的工作，并且取得了一定的研究成果。但是中文信息抽取方面的研究相对起步较晚，纯粹的基于中文的信息抽取系统在国内仍处于空白，主要的研究工作集中在对中文命名[25]。

信息抽取技术是当前的热门研究方向学术会议很频繁其中最重要的一个会议是 Message Understanding Conference(MUC)它是一个由美国政府资助的为推动 IE 技术发展的重要的系列工程，迄今为止已经举办了七届 MUC 采用竞赛的方式每一届都提供标准的语料并定义了各种不同的子任务来对参赛的信息抽取系统进行评估，其难度也是越来越大，MUC 吸引了全世界越来越多的研究机构。参加 1998 年的 MUC-7 是最近的一次 MUC，它的信息抽取任务涉及抽取文档中的专名(人名组织名和地点名)、同指项、确定模板元素之间的关系如地点关系、雇佣关系和生产关系等，抽取文档中的事件文档包含多语种的新闻稿。训练用的文档专业领域是关于飞机坠毁报道，而测试用的文档专业领域是关于发射事件报道。信息抽取的发展趋势有：在抽取内容方面由单语种向多语种发展；由简单的领域实体抽取向实体的属性和实体间关系事件的抽取发展；在抽取方法方面，由单一的基于规则的系统向结合机器学习和统计方法的多策略系统发展；由表层的句子级的语言处理向深层的篇章级的语言处理发展；在实际应用方面，由早期的理论研究和技术探讨逐渐向实际应用系统的开发发展[24]。

2.1.3 信息抽取处理的研究对象

狭义的信息抽取，其处理对象主要是各种文本信息，包括结构化文本信息、半结构化文本信息和自由文本信息。而广义上的信息抽取处理对象则还包括了语音、图像和视频等多媒体数据信息。在这里，主要研究的是狭义的信息抽取技术[26]。

信息抽取的最初目的是开发实用系统，从自由文本中抽取有限的主要信息。处理自由文本的信息抽取系统通常使用自然语言处理技巧，其抽取规则主要建立在词和词类间

句法关系的基础上。需要经过的处理步骤包括：句法分析、语义标注、命名实体识别和抽取规则。

结构化文本信息是一种存储于数据库里的文本信息，或者根据事先规定的严格格式生成的文本信息。从这样的文本信息中抽取信息是非常容易的，准确度也很高，通过描述其格式即可达到目的。

半结构化文本信息是一种介于自由文本信息和结构化文本信息之间的数据信息，通常缺少语法，也没有严格的格式。自然语言处理技术对于这样的文本信息处理不一定有效，因为其可能不是由完整语句构成；同时由于其非格式化的特点导致用来处理结构化文本信息的规则方法也不能奏效。因此，半结构化文本信息的抽取模式通常依赖字符和类似 HTML 标记的分隔符号，以从中抽取出所包含的一些结构化信息。本文研究的专利信息就是该种形式。

2.1.4 信息抽取的类型

信息抽取系统要在更多的自然语言处理技术支持下，把需要的信息从文本中提取出来，再用某种结构化的形式组织起来，提供给用户（人或计算机系统）使用。信息提取技术一般被分解为五个层次[27]：第一是识别专有名词（Named Entity），主要是人名、地名、机构名、货币等名词性条目，以及日期、时间、数字、邮件地址等信息的识别和分类；第二是模板要素（Template Element），即应用模板的方法搜索和识别名词性条目的相关信息，这时要处理的通常是一元关系。第三是模板关系（Template Relation），即应用模板的方法搜索和识别专有名词与专有名词之间的关系，此时处理的通常是二元关系。第四是同指关系（Co-reference），要解决文本中的代词指称问题。第五是脚本模板（Scenario Template），是根据应用目标定义任务框架，用于特定领域的信息识别和组织。

2.1.5 信息抽取的方法设计与流程

信息抽取系统设计主要有两大方法[28]：一是知识工程方法（Knowledge Engineering Approach）；二是自动训练方法(Automatic Training Approach)。知识工程方法主要靠手工编制规则使系统能处理特定知识领域的信息抽取问题。这种方法要求编制规则的知识工程师对该知识领域有深入的了解，且开发的过程可能非常耗时耗力。自动训练方法系统主要通过学习已经标记好的语料库获取规则，并且经训练后的系统能自动学习处理

新的文本。这种方法要比知识工程方法快，但需要足够数量的训练数据，才能保证其处理质量[29]。

信息抽取的工作流程可以表述为：用一组信息模式（Information Patterns）描述感兴趣的信息；对待抽取文本信息进行“适度的”（浅层、非完整的）词法、句法及语义分析，并作各种文本标引；使用模式匹配方法识别指定的信息；进行上下文关联、指代、引用等分析和推理，确定信息的最终形式；输出结果。根据信息抽取的一般工作流程并结合其基本体系结构可以得到一个通用的信息抽取系统模型[30]，如图 2.2 所示。

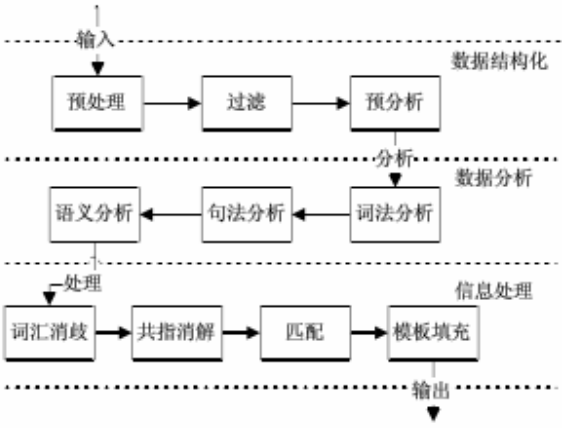


图 2.2 信息抽取模型图

2.1.6 信息抽取系统的性能评价

信息抽取系统的性能可从两个方面来进行评测：信息抽取的全面性和准确性以及信息抽取系统的可移植性。但是，为了更为独立客观的评测信息抽取的全面性和准确性，对信息抽取系统所应对的领域/任务的复杂度也要进行评测[31]。

一般采用三个指标来评测信息抽取系统的全面性和准确性：召回率 R、准确率和综合指标。召回率和准确率是从信息检索的两个性能评测指标沿用而来的。

R=系统返回的正确抽取结果个数/可能存在的正确结果个数；

P=系统返回的正确抽取结果个数/系统返回的所有结果个数。

P 和 R 的值域为[0,1]，它们的最优值为 1。一般的，对于一个信息抽取系统，单独追求一个指标的提高而忽视另一个指标的提高是无意义的，应该同时追求较大的召回率和准确率。事件抽取中，召回率和准确率一般是针对事件的各个角色来讲的，而不是针对整个事件来讲的。如果待抽取的事件较为复杂，事件所包含的角色数目较多，则常常

出现某个事件角色的召回率和准确率都很高，但整个事件的召回率和准确率却较低的情形。实际应用时，为了评价的方便，常常将 R 和结合在一起形成一个综合指标 F，用来衡量信息抽取系统的整体性能[32]。

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \qquad \text{(公式 2.1)}$$

其中，β 决定了 R 之于 P 的重要程度。若 β =1，则将 R 和 P 视为同等重要；若 β =2，则将 R 的重要程度视为 P 的两倍；若 β =0.5，则将 R 的重要程度视为的一半。

对实现了同样的领域/任务的信息抽取系统，它们的性能可以通过各自的 R、P、F 较为准确的刻画出来。但是，对于不同领域/任务的两个或多个信息抽取系统，要评测它们的性能，不仅要参考它们各自的 R、P、F，还要将它们各自所应对的领域/任务的复杂度也考虑进去。领域/任务越复杂，要提高系统的性能就越困难。领域/任务的复杂度主要体现在其中所描述事件的复杂度。

2.1.7 半结构化的信息抽取和非结构化的信息抽取

（1）Web 页面信息抽取

目前，网络半结构化数据日趋丰富。针对于 Web 的信息抽取技术越来越成熟，而针对于 www 网页的信息抽取技术，成为 Web 信息抽取最主要的研究方向[33]。

www 网页主要分为两种[34]：①HTML(Hypertext Markup Language)——由于其在目前网络资源描述格式中所占的比例最高，所以有关研究持别多。②XML(extensible Markup Language)——作为一种新的网上数据交换的标准，正在引起人们极大的关注。

Section expression①	Description②	Examples③
〈h1-6〉# 〈/h1-6〉	Title font section④	〈h2〉Invited Papers〈/h2〉
〈b〉# 〈/b〉	Physical bold-faced section⑤	〈b〉Innovation in Database Management: Computer Science ... 〈/b〉
〈strong〉# 〈strong〉	Logical bold-faced section⑥	〈strong〉Area〈strong〉
〈fontsize= 1-7〉# 〈/font〉	Section with font size⑦	〈font size= 7〉good weather〈/font〉
〈ul〉〈li〉# 〈/ul〉	List section without order⑧	〈ul〉〈li〉〈aname= " Jacobs97" href= "../../indices/a-tree/y..." 〉〈/a〉:..〈/ul〉
〈ol〉〈li〉# 〈/ol〉	List section with order⑨	〈ol〉〈li〉Today〈li〉Tomorrow 〈/ol〉
〈table# 〉〈tr〉# 〈th〉# 〈td〉# 〈/table〉	Table section⑩	〈table〉〈tr〉〈th〉Food〈/th〉〈tr〉〈td〉A〈/td〉〈td〉B〈/td〉〈/table〉

①段表达式, ②描述, ③例子, ④标题字体段, ⑤物理粗体字的段, ⑥逻辑粗体字的段,
⑦有字体大小的段, ⑧无序列表段, ⑨有序列表段, 10 表格段.

图 2.2 Web HTML 页面的结构模式

上图就是 Web HTML 页面的结构模式，针对于 Web 页面的信息抽取主要是根据这些 Web 标记来设计模板的。

对于 Web 页面的信息抽取，虽然由于它们实现时采用的技术细节不同，但是它们的工作过程是基本一样的，主要包括以下三个方面：

(1) 将信息进行分类整理。Robot 提取的网页将被放入到数据库中以便建立索引，不同的搜索引擎会采取不同方式来建立索引。

(2) 文本分析。文本分析包括去除文本中无用的一些标记，对汉语进行分词等。

(3) 语义分析。语义分析包括名词识别、归类等，例如识别人名，公司名，数字（如日期，年龄），职位名称等。对于处理方式相同的词放入一类进行处理。

(4) 结构化生成。信息抽取以后，需要以结构化的数据存放在特定的地方，如数据库等。

现在最流行的 Web 信息抽取方法被称为包装器（Wrapper），它能解析源文档并将源数据转化为结构化的或者半结构化的形式。如果转化为结构化的形式，可以直接使用查询语句如 SQL 来查询抽取的信息。如果转化为半结构化形式，那么需要使用特殊的查询语句。Wrapper 可以手工编写或者半自动化编写而成。

（2）非结构化文本的信息抽取

这是完全针对自然语言的信息抽取技术。自然语言处理过程一般可归为：语音、词、词形、语法、语义、篇章、语用 7 个不同的抽象级别[35]。自然语言理解所需的知识量是惊人的，20 世纪 80 年代初，耶鲁大学研制的 BORIS 系统，在知识工程方面经过三年的努力，只能对两段描述性的文本进行深入的分析，对其他段落则无能为力。这也初步表明自然语言理解方法不完全适合于进行广泛的信息抽取[36]。

2.2 数据挖掘技术

2.2.1 数据挖掘的概述

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据集中识别有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。它是一门涉及面很广的交叉学科，包括机器学习、数理统计、神经网络、数据库、模式识别、粗糙集、模糊数学等相关技术[37]。

数据挖掘可粗略地理解为三部曲：数据准备（data preparation）、数据挖掘，以及结果的解释评估（interpretation and evaluation）。根据数据挖掘的任务分，有如下几种：分类或预测模型数据挖掘、数据总结、数据聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等等[38]。

2.2.2 数据挖掘的发展

KDDM 是近年来一个十分活跃的研究领域。从数据库中发现知识(Knowledge Discovery in database，简称 KDD)一词首先出现在 1989 年举行的第十一届国际联合人工智能学术会议上。到目前为止，由美国人工智能协会主办的 KDD 国际研讨会已召开了 8 次，规模由原来的专题讨论会发展到国际学术大会，仅以 1999 年为例，就有近 20 个国际会议列有 KDDM 专题[39]。

这两年国内也有相当多的数据挖掘和知识发现方面的研究成果，许多学术会议上都设有专题进行学术交流。目前， KDDM 的研究重点逐渐从发现方法的研究转向实际的系统应用，国际上有影响的典型数据挖掘系统有 SAS 公司的 Enterprise Miner，IBM 公司的 Intelligent Miner，SGI 公司的 Set Miner 等。从国内外目前的研究进展来看，各学科的研究自成一派，没有突破各个领域的技术界限；没有融合各领域的不同方法；尤其是未将并行优化的诸方法集成用于数据库中的数据挖掘，从而提高实时性，并解决随机的、动态的、不完全的及混沌数据的数据挖掘，即所谓智能数据挖掘。而且以往多数技术都是在驻留于内存的数据之上进行挖掘，没有把这些技术与数据库技术相集成[40]。

同时，计算机技术的另一领域－人工智能（AI: Artificial Intelligence）自 1956 年诞生之后取得了重大进展。经历了博弈时期、自然语言理解、知识工程等阶段，目前的热点是机器学习。用数据库管理系统来存储数据，用机器学习的方法来分析数据，挖掘大量数据背后的知识，这两者的结合促成了数据库中的知识发现（KDD：Knowledge Discovery in Databases）的产生。数据库中的知识发现是一门交叉性学科，涉及到机器学习、模式识别、统计学、智能数据库、知识获取、数据可视化、高性能计算、专家系统等多个领域。从数据库中发现出来的知识可以用在信息管理、过程控制、科学研究、决策支持等许多方面。1989 年 8 月在美国底特律召开的第 11 届国际人工智能联合会议的专题讨论会上首次出现知识发现（KDD）这个术语。此后，由美国人工智能协会主办的 KDD 国际研讨会已经召开了 8 次，规模由原来的专题讨论会发展到国际学术大会，研究

重点也逐渐从发现方法转向系统应用，注重多种发现策略和技术的集成，以及多种学科之间的相互渗透[41]。

2.2.3 关联规则

关联规则(Associate Rule)在数据挖掘中占有及其重要的地位,是数据挖掘的主要任务之一。它是通过对数据库中数据的分析处理,发现不同属性的数据之间的关系。他的概念最早于 1993 年由 Agrawal R 等人提出。Agrawal R 等人提出了经典的 Apriori 算法。它是利用“在给定的事物数据库中,任意强项集的子集都是强项集,任意弱项集的超集都是弱项集”这一原理去发现频繁集。但是，该算法也存在不少的缺陷，后人给了不少改进的方法，如：基于 Hash 的项集计数、减少交易记录、分割、采样、动态项集计数等方式，就涌现出用 Partition 技术对 Apriori 算法的优化有效的提高了算法的效率和可伸缩性,DHP 以及 DIC 算法也都从一定程度上提高了执行效率[42]。

基本定义如下：X = > Y 解释为“满足 X 中条件的数据库元组多半也满足 Y 中条件”。每个发现模式都应该有一个表示其有效性或值得信赖性的确定性度量。

对于形如“A=>B”的关联规则，对其确定性度量叫做置信度，其中 A 和 B 是项目的集合。给定一个任务相关的数据元组集合(或事务数据库事务的集合)，“A=>B”的置信度定义为置信度(A=>B)=包含 A 和 B 的元组，包含 A 的元组=P(A∪B)。关联规则的支持度是模式为真的任务相关的元组(或事务) 所占的百分比。对于形如上式的关联规则，支持度定义为支持度(A=>B)=包含 A 和 B 的元组，元组总数=P(A∪B)。

Apriori 算法是一种以概率为基础的具有影响的挖掘布尔型关联规则频繁项集的算法。其利用循序渐进的方式，找出数据库中项目的关系，以形成规则。其过程分为两步:一为连接(类矩阵运算) ,二为剪枝(去掉那些没必要的中间结果)。在此算法中常出现项集的概念。项集(item set)简单地说就是项的集合。包含 K 个项的集合为 k2 项集。项集的出现频率是包含项集的事务数，称为项集的频率。如果项集满足最小支持度，则称它为频繁项集，频繁 k2 项集的集合计作 Lk[43]。

2.2.4 聚类技术

将物理或抽象对象的集合分组成为有类似的对象组成的多个簇的过程被称为聚类。聚类通过把目标数据放入少数相对同源的组或“类”（cluster）里。分析表达数据，①通过一系列的检测将待测的一组基因的变异标准化，然后成对比较线性协方差。②通过

把用最紧密关联的谱来放基因进行样本聚类，例如用简单的层级聚类（hierarchical clustering）方法。这种聚类亦可扩展到每个实验样本，利用一组基因总的线性相关进行聚类。③多维等级分析（multidimensional scaling analysis,MDS）是一种在二维 Euclidean “距离”中显示实验样本相关的大约程度。④K-means 方法聚类，通过重复再分配类成员来使“类”内分散度最小化的方法[44]。

聚类方法有两个显著的局限：首先，要聚类结果要明确就需分离度很好（well-separated）的数据。几乎所有现存的算法都是从互相区别的不重叠的类数据中产生同样的聚类。但是，如果类是扩散且互相渗透，那么每种算法的的结果将有点不同。结果，每种算法界定的边界不清，每种聚类算法得到各自的最适结果，每个数据部分将产生单一的信息。为解释因不同算法使同样数据产生不同结果，必须注意判断不同的方式。对遗传学家来说，正确解释来自任一算法的聚类内容的实际结果是困难的（特别是边界）。最终，将需要经验可信度通过序列比较来指导聚类解释。第二个局限由线性相关产生。上述的所有聚类方法分析的仅是简单的一对一的关系。因为只是成对的线性比较，大大减少发现表达类型关系的计算量，但忽视了生物系统多因素和非线性的特点。

从统计学的观点看，聚类分析是通过数据建模简化数据的一种方法。传统的统计聚类分析方法包括系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、有重叠聚类和模糊聚类等[44]。采用 k-均值、k-中心点等算法的聚类分析工具已被加入到许多著名的统计分析软件包中，如 SPSS、SAS 等。

从机器学习的角度讲，簇相当于隐藏模式。聚类是搜索簇的无监督学习过程。与分类不同，无监督学习不依赖预先定义的类或带类标记的训练实例，需要由聚类学习算法自动确定标记，而分类学习的实例或数据对象有类别标记。聚类是观察式学习，而不是示例式的学习。

从实际应用的角度看，聚类分析是数据挖掘的主要任务之一。就数据挖掘功能而言，聚类能够作为一个独立的工具获得数据的分布状况，观察每一簇数据的特征，集中对特定的聚簇集合作进一步地分析。

聚类分析还可以作为其他数据挖掘任务（如分类、关联规则）的预处理步骤。

数据挖掘领域主要研究面向大型数据库、数据仓库的高效实用的聚类分析算法。聚类分析是数据挖掘中的一个很活跃的研究领域，并提出了许多聚类算法。这些算法可以

被分为划分方法、层次方法、基于密度方法、基于网格方法和基于模型方法。

3 基于关联规则的同指消解技术

同指(Co-reference)是比指代更广的概念[45]。它指篇章中的语言单位包括名词(词组)和代词，指向同一个实体。这些名词(词组) 代词被统一称为对象。信息抽取的第四个层次是同指关系（Co-reference)，本文主要建立基于关联规则的同指消解模型。

3.1 同指消解定义

同指(Co-reference)是指篇章中的语言单位包括名词组和代词指向同一个实体，这些名词(词组)代词被统一称为篇章对象，设有同指对象(A1、A2)A1 位于 A2 前面 A1 称为先行语(antecedent)A2 称为照应语(anaphor)[46]。同指消解(Co-reference resolution)就是确定照应语与先行语的同指关系，指代消解是同指消解的重要组成部分，同指关系的语料库标注 MUC，同指关系标注模式(MUC Annotation Scheme for Co reference Relation)是 MUC-6 初步定义，MUC-7 进一步修改后形成的一个标注模式同指关系是：

用通用标记语言标准(Standard for General Markup Language, SGML)标记的设有先行语 A1 和照应语 A2 它们具有同指关系则基本的同指关系标注为

<COREF ID=”100”>A1</COREF>

<COREF ID=”101”TYPE=”IDENT”REF=”100”>A2</COREF>

属性 ID 是用来区别每个篇章对象的唯一标示符 REF 属性指出了哪一个对象与当前标记的对象，同指 TYPE 属性表示先行语和照应语的关系。

目前可用的属性值只有 IDENT 它表示标注的关系，只有身份(identity)关系可以标注的对象。

设有 3 个对象 A1 A2 A3， A1 与 A2 同指，A2 与 A3 同指，则根据同指的定义 A1 A2 A3 相互同指如图 3.1 所示(双向箭头代表同指关系)

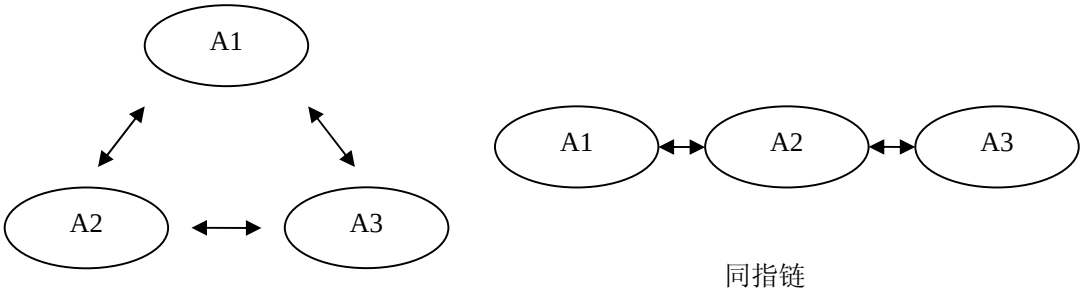


图 3.1 同指关系图

由于同指具有传递性、相互性，为了叙述的方便，我们将用图来表示 A1、A2、A3 的同指关系，我们称它们形成一个同指链。

MUC 对同指消解结果的技术评估有两个重要标准查准率 P(Precision)和查全率 R(Recall)，查准率 P 是同指消解结果中正确的对象数目占实际消解的对象数目的百分比它反映的是信息抽取系统的准确程度；查全率 R 是同指消解结果中正确的对象数目占消解系统应消解的对象总数它反映的是信息抽取系统的完备性。

评估公式

$$p = Nc / Nr$$
$$R = Nc / Nk \qquad \text{(公式 3.1)}$$

注

P 查准率；

R 查全率；

Nk 消解系统应消解的对象总数；

Nr 实际消解的对象数目；

Nc 同指消解结果中正确的对象数目；

3.2 基于关联规则的同指消解模型的提出

首先对美国专利的网页信息进行本地化下载，将 WEB 网页的非结构化数据利用信息抽取技术转化为结构化数据形式，通过消除数据冗余、去除噪音词、数据格式转化等方式进行专利数据清洗，以实现对专利数据的预处理。

然后，对需要分析的属性-申请人和从专利摘要利用文本挖掘技术挖掘出来的摘要关键词进行关联分析，以便找出不同的申请人研究的方向存在相关性，这些相关的申请人有可能存在同指关系；接着计算关联度，并利用关联度确定阈值，以便确定范围，从预定范围中构建候选先行语集合，对候选先行语的集合是设定系统首先出现的一段区域内词语的集合；接着利用辞典库与先行语进行同指规则判断，选出满足辞典库先行语，经过辞典库筛选后如果得到唯一满足同指规则的先行语则不予考虑，否则基于规则和基于属性统计的识别判断，经过分析后将还不能设为唯一的英文词填充到同指辞典中，不断扩充的英文同指辞典可以通过一定的方式进行对照和影射，实现可以中英文互换的同指辞典。

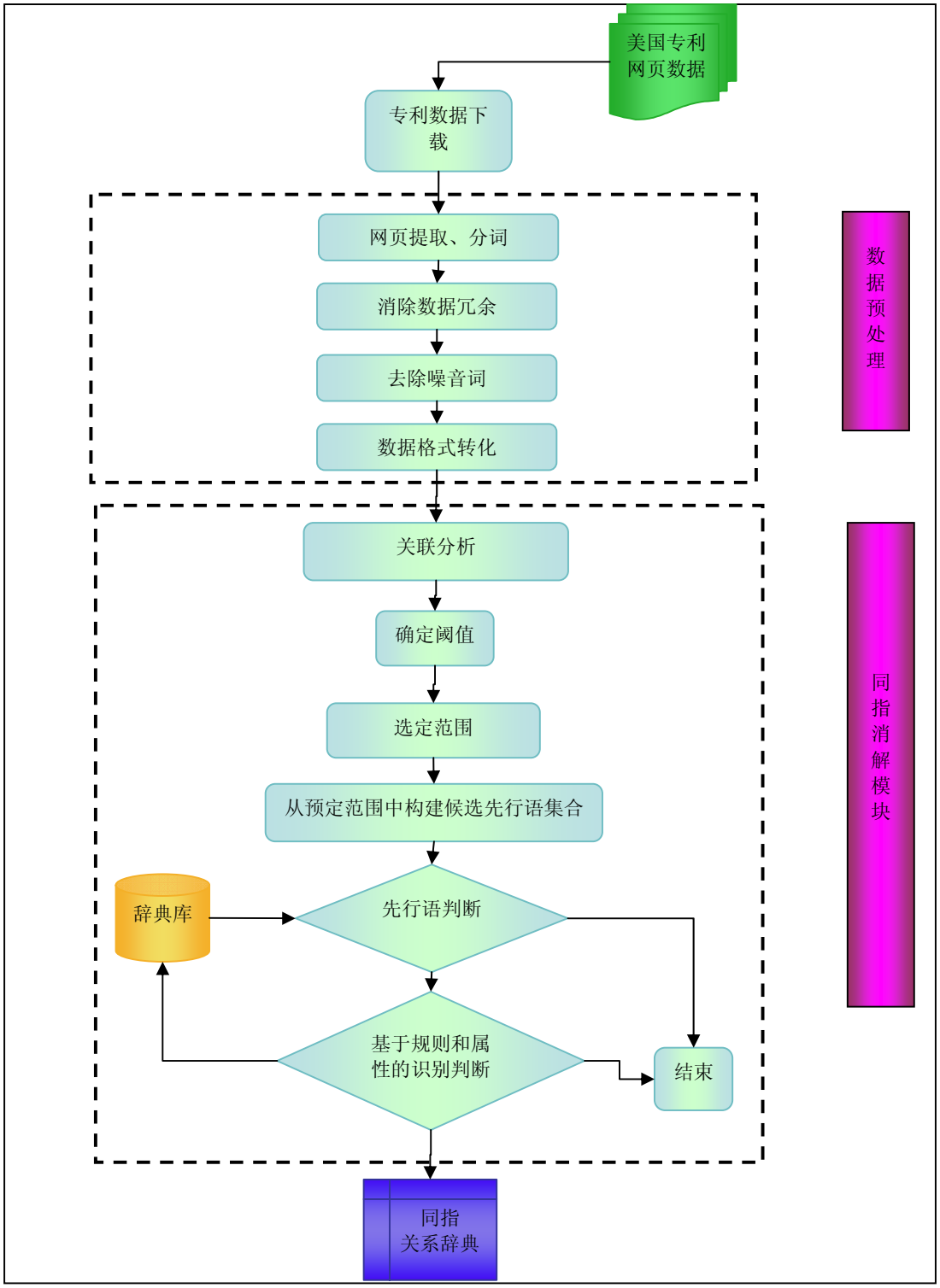


图 3.2 基于关联规则的同指消解图

3.2.1 数据的选择

本文的专利数据采取网络下载的形式，网络专利数据库是一个复杂的、不规则性极强的数据源，要访问、分析这些数据是一项具有挑战性的工作。网络专利数据库涉及多个数据库，形式和内容十分广泛，包括各种类型的数据源。因此解决数据质量的问题尤其重要，本论文研究首先要解决如下两个问题：

◆ 异构数据库环境

Web 上的每一个站点就是一个数据源，每个数据源之间都是异构的。因为每一站点间的信息和组织都不一样，这就构成了一个巨大的异构数据环境。因此，要想对这些数据进行数据挖掘和分析，就要解决好站点之间的异构数据集成问题，还要解决 Web 上的信息采集问题。

◆ 半结构化的数据源

半结构化数据是指在物理层上缺少结构的数据或者在逻辑层上缺少结构的数据。本文研究的网络文献数据库中的大量信息，分散于各个站点的 HTML 文件中，没有严格的结构和类型定义，这些都是逻辑层的半结构化数据。但是，Web 上的数据具有动态可变性，所以半结构化是 Web 上数据的最大特点。

3.2.2 网络专利数据库 Web 内容抽取

Web 内容抽取的核心是从 Web 页面所包含的非结构化或者半结构化的信息中识别用户所感兴趣的数据，并将其转化为更结构化、语义更清晰的格式，其目的是识别 HTML 文档中数据的语义，并建立映射关系。所以，一个信息抽取系统的关键是包装器的实现。包装器就是提供 Web 内容到用户需求信息的中间信息转换机制的功能模块。

一般情况下人们会采用网页结构分析的方法，判定特征关键词的位置，然后固定模板，提取相关信息。然而当需要加入新的数据源（即获取新的结构不同的网页内容）；或者原网页的结构发生了变化；网页中的特征关键词发生改变。所有这些改动，都会导致源程序无法工作，需要对源程序进行修改，重新编译，以满足新的要求，使系统缺乏通用性，导致系统的适应性很差。

因此，对于那些非标准 HTML，可以首先转换为结构化的 XML 格式达到信息抽取的目的。首先，XML 已经成为一种通用的数据交换规范，各种软件系统或者平台都可以使用 XML 数据；其次是 XML 是结构化的数据，可以方便的与关系数据库进行转换。通过分析

Web 信息抽取的过程和目的，可以设计一个更好的包装器。可以方便的处理以下信息：Web 网页集的 URLs 描述、Web 网页的结构描述、目的数据模式结构的定义、待抽取信息和目的模式之间的映射关系及抽取规则的生成。

然而某些网络专利数据库无法在 Web 页面中获取所用的信息。有可能是服务提供商从安全的角度出发，预防信息被自动化的 Robot 窃取。但是它们也要为广大用户提供所有的信息，因此它们的变通方法是让用户导出到特定的格式文件。本论文针对这种情况采取的是特征信息识别的方式对特定格式的文件进行信息的抽取。

3.2.3 数据预处理

从目标数据集中除去明显错误数据和冗余的数据，去除噪声或无关数据，进行数据清洗，并利用数据库重构方法创建一些便于分析利用的数据字段，为今后的数据挖掘和分析做好准备工作。

主要的工作有：

数据清理(data cleaning) 通过填写空缺的值，平滑噪声数据，识别、删除孤立点，并解决不一致来“清理”数据。

数据集成(date integration) 集成多个数据库、数据立方体或文件中的数据。

数据变换(data transformation) 数据的规格化或聚集操作。

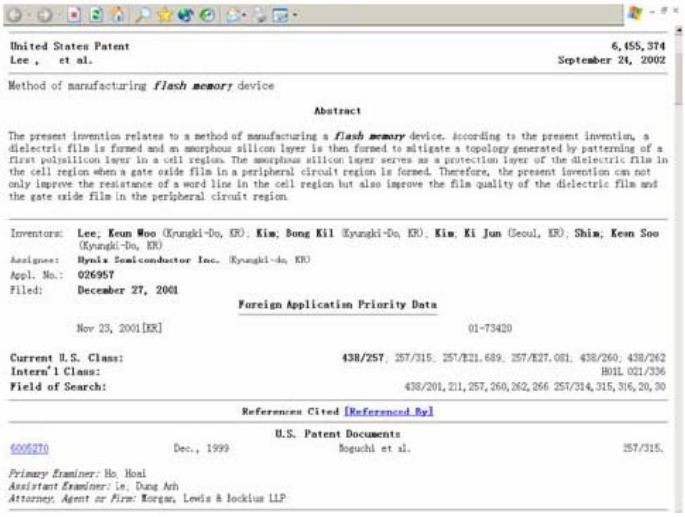
数据规约(data reduction) 在不影响分析结果的前提下得到数据集的压缩表示。它有数据聚集、维规约、数据压缩和数字规约等多种形式。

数据离散化(Data discretization) 通过将属性划分为区间，减少给定连续属性值的个数。

目前主要的专利数据源为美国专利和商标局提供的美国专利数据库，通过专利获取工具下载获得的是未经处理的 HTML 格式的专利文献信息。信息抽取系统的设计与待处理文本的特性是分不开的，针对 HTML 格式的专利信息，数据预处理的目的是过滤各种冗余标记和信息，识别出文本中的简单实体片断，并将其整理成以标点符号为分块依据的文本块作为分词和标注系统的输入，从而减轻分词和标注任务的负担，降低产生歧义的可能。

图 3.3 所示的美国专利为例进行文本特征分析。由于是 HTML 格式，文本主要由 HTML 标记和文字组成，通过对文本内容的分析可以知道专利数据的排列都是按照标引项名称

和标引项内容分布的，因此预处理的重点是清除网页标记、提取标引项内容，将专利信息分块。



United States Patent 6,455,374
Lee , et al. September 24, 2002

Method of manufacturing *flash memory* device

Abstract

The present invention relates to a method of manufacturing a *flash memory* device. According to the present invention, a dielectric film is formed and an amorphous silicon layer is then formed to mitigate a topology generated by patterning of a first polysilicon layer in a cell region. The amorphous silicon layer serves as a protection layer of the dielectric film in the cell region when a gate oxide film in a peripheral circuit region is formed. Therefore, the present invention can not only improve the resistance of a word line in the cell region but also improve the film quality of the dielectric film and the gate oxide film in the peripheral circuit region.

Inventors: **Lee; Keun Woo** (Kyungki-Do, KR); **Kim; Dong Kil** (Kyungki-Do, KR); **Kim; Ki Jun** (Seoul, KR); **Shim; Keun Soo** (Kyungki-Do, KR)
Assignee: **Hynix Semiconductor Inc.** (Kyungki-do, KR)
Appl. No.: 026957
Filed: December 27, 2001

Foreign Application Priority Data

Nov 23, 2001 [KR] 01-73420

Current U.S. Class: 438/257; 257/315; 257/E21.689; 257/E27.081; 438/260; 438/262
Intern'l Class: H01L 021/336
Field of Search: 438/201, 211, 257, 260, 262, 266 257/314, 315, 316, 20, 30

References Cited [Referenced By]

U.S. Patent Documents

6005270 Dec., 1999 Noguchi et al. 257/315.

Primary Examiner: Ho, Hoai
Assistant Examiner: Le, Dung Anh
Attorney, Agent or Firm: Morgan, Lewis & Bockius LLP

图 3.3 美国专利示意图

从专利文本中找到待处理的专利标引项，以此为目标关键字进行文本预处理标记操作，如图 3.4 所示。

```
<TABLE width="100%">
  <TBODY>
  <TR>
    <TD align=left width="50%"><B>United States Patent </B></TD>
    <TD align=right width="50%"><B>6,455,374 </B></TD></TR>
  <TR>
    <TD align=left width="50%"><B>Lee ,   et al.</B> </TD>
    <TD align=right width="50%"><B>September 24, 2002 </B></TD></TR></TBODY></TABLE>
<HR>
<FONT size=+1>Method of manufacturing <B><I>flash memory</I></B> device
</FONT><BR><BR>
<CENTER><B>Abstract</B></CENTER>
<P>The present invention relates to a method of manufacturing a <B><I>flash memory</I></B> device. According to the present invention,
formed and an amorphous silicon layer is then formed to mitigate a topology generated by patterning of a first polysilicon layer in a
amorphous silicon layer serves as a protection layer of the dielectric film in the cell region when a gate oxide film in a peripheral
Therefore, the present invention can not only improve the resistance of a word line in the cell region but also improve the film qualit
and the gate oxide film in the peripheral circuit region. </P>
<HR>

<TABLE width="100%">
  <TBODY>
  <TR>
    <TD vAlign=top align=left width="10%">Inventors: </TD>
    <TD align=left width="90%"><B>Lee; Keun Woo</B> (Kyungki-Do, KR); <B>Kim; Dong Kil</B> (Kyungki-Do, KR); <B>Kim; Ki Jun</B> (Seoul
      Keon Soo</B> (Kyungki-Do, KR) </TD></TR>
  <TR>
    <TD vAlign=top align=left width="10%">Assignee: </TD>
    <TD align=left width="90%"><B>Hynix Semiconductor Inc.</B> (Kyungki-do, KR) </TD></TR>
  <TR>
    <TD vAlign=top noWrap align=left width="10%">Appl. No.: </TD>
    <TD align=left width="90%"><B>026957</B></TD></TR>
  <TR>
    <TD vAlign=top align=left width="10%">Filed: </TD>
```

图 3.4 网页格式的专利文本

通过对美国专利页面特征分析，为了达到专利信息分块的要求，本文使用一个简化的基于多知识的网页信息抽取方法模型（Multi Knowledge Information Extraction，简称 MKIE 方法）来完成文本预处理。如图 3.5 所示。

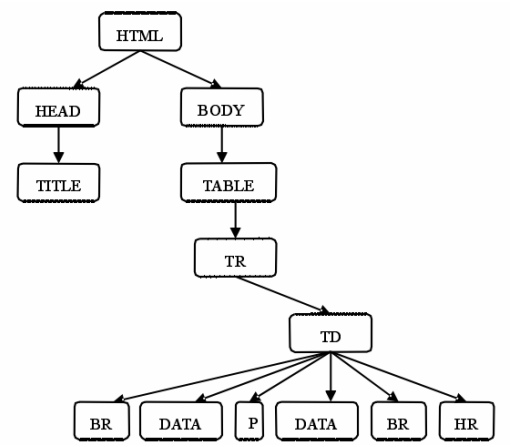


图 3.5 基于 MKIE 方法的文本预处理流程

文本预处理之后将专利分块信息导入数据库中，根据进一步处理的需要，本文设计了包含专利标引项的存储表，表的结构具体如下所示。

表 3.1 专利基本信息表

列名	数据类型	是否为空	说明
ID	Integer	NO	专利号
TITEL	Text	NO	专利名称
ABSTRACT	Text	NO	专利摘要
INVENTORS	Text	NO	专利发明人
IPC	Text	NO	国际专利分类号

3.2.4 关联规则

关联模式反映一个事物与其他事物之间的相互依赖性或相互关联性，如果两个或多个事物之间存在关联，那么，就能从其他已知事物中预测到其中一个事物。关联分析通过搜索系统中的所有事物，从中找到出现条件概率较高的模式。挖掘关联实际上就是数据对象之间相关性的确定，用关联找出所有能将一组数据项和另一组数据项相联系的规则，这种规则的建立并不是确定的关系，而是一个具有一定共生关系的可能值，即事件发生的关联度。

技术组（群）自动识别和分类技术是专利数据分析的关键性技术。本文通过专利技术关键词共生关系分析对专利组内的技术组群、作者组群、机构组群、主题变迁组群进行了深层次的、具体的微观研究。以便确定挖掘专利申请人的同指关系氛围。

技术关键词在技术组（群）自动识别和分类中起着决定性的作用。我们对一技术领域进行监测，如通信技术，并假定从专利数据库得到包含 8000 篇有关通信技术专利的一专利组，围绕这一专利组应用技术组（群）自动识别和分类方法后，期望能得到通信技术领域下的各分支技术组（群）。

首先，在本项目研究中，我们通过对技术关键词的共生关系（Terms Co-occurrences）计算来识别、确定一组专利内部所包含的技术组（群）。

假定我们有 n 篇专利，这 n 篇专利包含有 m 个技术关键词，则我们就建立了{n 篇专利 $*$ m 技术关键词}的关联矩阵 X

$$X = \begin{bmatrix} & Term_1 & Term_2 & \cdots & Term_i & \cdots & Term_j & \cdots & Term_m \\ D_1 & b_{11} & b_{12} & \cdots & b_{1i} & \cdots & b_{1j} & \cdots & b_{1m} \\ D_2 & b_{21} & b_{22} & \cdots & b_{2i} & \cdots & b_{2j} & \cdots & b_{2m} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ D_i & b_{i1} & b_{i2} & & b_{ii} & & b_{ij} & & b_{im} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ D_j & b_{j1} & b_{j2} & \cdots & b_{ji} & \cdots & b_{jj} & \cdots & b_{jm} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ D_n & b_{n1} & b_{n2} & \cdots & b_{ni} & \cdots & b_{nj} & \cdots & b_{nm} \end{bmatrix}$$

- X 矩阵中,专利 D_i 的关键词 $Term_i$ 的权值,用布尔代数值表示,当 $Term_i$ 在 D_i 专利（或专利）中出现时取 1,否则取 0。
- n 是专利组内包含专利的总数。
- m 是专利组内所有关键词总数。

基于这个{专利 × 关键词} X 矩阵，我们进一步得到{关键词 × 关键词}共生的关联矩阵 T：

$$T = X^T \bullet X \qquad \text{(公式 3.2)}$$

{关键词 × 关键词}共生的关联矩阵 T

$$T=\begin{bmatrix} & Term_1 & Term_2 & \cdots & Term_i & \cdots & Term_j & \cdots & Term_n \\ Term_1 & t_{11} & t_{12} & \cdots & t_{1i} & \cdots & t_{1j} & \cdots & t_{1m} \\ Term_2 & t_{21} & t_{22} & \cdots & t_{2i} & \cdots & t_{2j} & \cdots & t_{2m} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ Term_i & t_{i1} & t_{i2} & & t_{ii} & & t_{ij} & & t_{im} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ Term_j & t_{j1} & t_{j2} & \cdots & t_{ji} & \cdots & t_{jj} & \cdots & t_{jm} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ Term_m & t_{m1} & t_{m2} & \cdots & t_{mi} & \cdots & t_{mj} & \cdots & t_{mm} \end{bmatrix}$$

技术共生关系分析是通过对反映专利主题内容的词进行关联性或相异性定量分析，来研究专利内在联系和科学结构的一种方法，其基本出发点是：

(1) 科学研究的热点是由一系列在内容上密切相关的研究课题和概念组成的，这些热点是众多科学研究人员注意和跟踪的对象。

(2) 热衷或从事于某一科学热点研究的科学家，无论其社会和知识背景如何，在很大程度上，对于同一研究课题和概念，所使用的词汇是基本一样的。

也就是说，在技术关联性很强的一些申请人里，也存在很强的同指关系。

例如：

假定我们有 6 篇专利，这 6 篇专利共包含有 5 个技术关键词，则我们就建立了{6 篇专利 *5 技术关键词}的关联矩阵 X

$$\begin{bmatrix} & Keyword_1 & Keyword_2 & Keyword_3 & Keyword_4 & Keyword_5 \\ D_1 & 1 & 0 & 1 & 1 & 0 \\ D_2 & 1 & 0 & 1 & 0 & 1 \\ D_3 & 1 & 1 & 0 & 0 & 0 \\ D_4 & 0 & 0 & 1 & 0 & 1 \\ D_5 & 1 & 0 & 1 & 1 & 0 \\ D_6 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

其中，1 代表 $Keyword_j$ 在 D_i 中出现，0 代表 $Keyword_j$ 不在 D_i 中出现。

基于这个{专利 X 关键词}X 矩阵，我们进一步得到{关键词 X 关键词}共生的关联矩阵 CO：

$$CO = X^T \bullet X \qquad \text{(公式 3.2)}$$

我们用下式计算CO的各个元素：

$$co_{xy} = \sum_{i=1}^{m} b_{ix} \cdot b_{iy} \quad \text{(公式 3.3)}$$

式中，x，y 代表关键词，m 是专利的数目，b 代表每列关键词组成的向量。这样，我们计算出了关键词之间的关联矩阵

	$Keyword_1$	$Keyword_2$	$Keyword_3$	$Keyword_4$	$Keyword_5$
$Keyword_1$	5	2	3	2	1
$Keyword_2$	2	2	0	0	0
$Keyword_3$	3	0	4	2	2
$Keyword_4$	2	0	2	2	0
$Keyword_5$	1	0	2	0	2

从上表可以看出，每次$Keyword_4$在文档中出现时，$Keyword_3$也同时出现，这说明$Keyword_3$和$Keyword_4$的关联度较大。

3.3.4.1 关联度计算方法

（1） 专利组的各专利之间关联度计算方法

我们可以使用一般聚类分析方法中常用的一些关联度计算方法，如关联系数、距离系数、内积系数等；也可以根据需要，自己定义关联性测度方法。关联性测度方法没有严格的标准。具体问题中哪种测度方法最好，要用聚类和关联分析的结果是否符合实际情况来验证。

专利组中D_i和D_j之间的关联度计算根据以下公式：

$$Sim(D_i, D_j) = \sum_{k=1}^{m} b_{ik} \times b_{ik} \quad \text{(公式 3.4)}$$

（2） 专利组内各关键词之间关联度计算方法

专利组中$Term_i$和$Term_j$之间的关联度计算根据以下公式：

$$Sim(Term_i, Term_j) = \sum_{k=1}^{m} t_{ik} \times t_{ik} \quad \text{(公式 3.5)}$$

（3） 专利组内各主要成分元素之间关联度计算方法

一篇专利还包含作者、研究机构、作者所在国家、专利发表年份等其他信息，如果

是专利其中还包含法律状态、分类号、公开日。我们根据专利组的各专利之间关联度计算方法提出了这些元素之间在主题内容意义之间的关联关系计算方法。

对两个成分元素之间关联性大小的量度，称为关联度。严格意义上关联度指系统发展过程中因素间相对变化的情况，也就是变化大小、方向及速度等指标的相对性。如果两者在系统发展过程中相对变化基本一致，则认为两者关联度大；反之，两者关联度就小。在本文研究中，我们提出通过对各元素所涉及技术关键词的共生关系来计算元素之间的关联关系。限于篇幅，以下我们以确定研究申请人之间的关联关系为例进行讨论。

假定专利组中包含有 n 个申请人即研究机构，这 n 个研究机构包含有 m 个技术关键词，则我们就建立了{n 个研究机构 * m 技术关键词}的关联矩阵 E：

$$E=\begin{bmatrix} & Term_1 & Term_2 & \cdots & Term_i & \cdots & Term_j & \cdots & Term_m \\ A_1 & e_{11} & e_{12} & \cdots & e_{1i} & \cdots & e_{1j} & \cdots & e_{1m} \\ A_2 & e_{21} & e_{22} & \cdots & e_{2i} & \cdots & e_{2j} & \cdots & e_{2m} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ A_i & e_{i1} & e_{i2} & & e_{ii} & & e_{ij} & & e_{im} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ A_j & e_{j1} & e_{j2} & \cdots & e_{ji} & \cdots & e_{jj} & \cdots & e_{jm} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ A_n & e_{n1} & e_{n2} & \cdots & e_{ni} & \cdots & e_{nj} & \cdots & e_{nm} \end{bmatrix}$$

- E 距阵中，研究机构 A_i 的关键词 $Term_j$ 的权值,用 e_{ij} 表示，e_{ij} 的取值为 $Term_j$ 在 A_i 研究机构发表的专利中出现的频率。
- n 是研究机构的总数。
- m 是专利组内所有关键词总数。

专利组中研究机构 A_i 和 A_j 之间的关联度计算根据以下公式：

$$Sim(A_i, A_j)=\sum_{k=1}^{m} e_{ik} \times e_{ik} \qquad \text{(公式 3.6)}$$

3.3 基于关联规则的同指消解模型设计与实验

3.3.1 专利数据获取

本章的研究对象为通信技术专利数据，因此在数据准备阶段首先确定专利数据源为

美国专利商标局 USPTO 网上专利数据库。专利数据的获取流程为：

3.3.1.1 通信技术关键词的确定

关键词的确定对专利数据的获取有着直接的影响。为了提高关键词的科学性和完整性，国家科技评估中心通过专家咨询和发放调查问卷两种方式来确定通信技术领域的专利检索式。

表 3.2 通信技术领域英文检索式

子方向		英文检索词
无线移动	研究方向一	（wireless OR mobile OR Adaptive OR OFDM OR MIMO）AND（channel coding OR channel decoding OR modulation/demodulation OR multiple users detection OR link adaptation OR adaptive modulation OR smart antennal OR RF）
	研究方向二	(wireless OR mobile) AND (MAC OR RRM OR radio resource scheduling OR SDR OR E2R OR network architecture OR cooperative radio OR channel modeling OR cognitive Radio)
光通讯	研究方向一	(Optical transmission or WDM OR Optical communication or Fibre Communication) and (transmission fiber OR modulation format OR broad band optical amplification OR dispersion compensation OR polarization mode dispersion compensation OR dispersion management OR coherent detection OR balance detection OR electrical equalization OR gain equalization OR forward error correction)
	研究方向二	(Optical access or Optical network or Optical communication) AND (burst mode transmitter OR burst mode receiver OR media access control OR bandwidth allocation OR security authentication OR ranging OR protection OR restoration)
	研究方向三	(Optical switching or optical burst switching or Optical communication or multicast) AND (optical switches OR optical circuit switching OR optical packet switching OR optical switching fabric OR optical add/drop multiplexing OR reconfigurable OADM OR Multi-granularity optical switching OR scheduling OR contention resolution OR tunable)
	研究方向四	（Optical networking or Optical network or Optical communication）AND (control plane OR protection OR restoration OR multiple protocol label switching OR routing OR signaling OR GMPLS OR optical interconnect OR contention resolution OR layering OR Cross-Domain OR UNI OR NNI OR interface)
下一代网络	研究方向一	(NGN or soft switching OR ASON) and (IMS OR QoS OR security OR user access OR authentication OR control or resource OR architecture OR network convergence OR service convergence OR terminal convergence OR internetworking OR transit OR SIP OR MSTP)
	研究方向二	(NGI or IPV6 OR Router) and (security OR server OR service aware OR multimedia application OR manageable OR controllable OR architecture OR network measurement OR dual-stock OR codec OR multicast OR P2P)

参加关键词咨询会议的专家有：北京邮电大学的伍剑老师、张平老师和陆月明老师，电信科技研究院的何建伟老师，信息产业部电信传输研究所的续合元老师。

调查问卷的对象主要是 863 通信技术领域的课题负责人，调查内容是请他们提供通信技术领域的专利主检索词和辅助检索词，最后共收回问卷 24 份。

在专家咨询和调查问卷的基础上，结合关键词在国家知识产权局专利数据库（www.sipo.gov.cn）的试检结果，通过往复几轮的讨论与咨询，最终确定的通信技术领域检索式见表 3.2。

根据专家的意见，将通信技术领域划分为无线移动通信技术、光通讯和新一代网络信息技术三个方向，其中无线移动通信技术方向的主检索词为无线、移动、自适应、正交频分复用和多输入多输出；光通讯方向又细分为四个子方向，主检索词包括光传输、光接入、光交换、光联网和光网络等；新一代网络信息技术方向细分为两个方向，主检索词为下一代互联网、软交换、自动交换光网络和 IPv6、路由器等。需要说明的是，在光通讯研究方向，考虑到专利申请撰写时并不区分“光通信”与“光通讯”，所以把两个词都作为主关键词对待，以保证专利数据检索的全面性。

3.3.1.2 通信技术专利数据处理

表 3.3 专利数据信息

字段名称	中文名称	字段类型	备注
ID	序列号	int	自动针，主键
PatNumber	专利号	Char(20)	
Title	标题	Vchar(1000)	
Abstract	摘要	Text	
Inventors	发明人	Vchar(100)	
Assignee	所属机构	Vchar(1000)	
AppNo	申请号	Char(20)	
FiledTime	申请时间	date	
PctTime	PCT 时间	date	
PctNO	PCT 号	Char(20)	
USClass	US 类	Vchar(100)	
IpcClass	IPC 类	Vchar(100)	
Claims	权利要求	Text	

通过对专利数据下载，可以获得以通信技术相关专利数据。由于所有的数据都是由机器自动下载完成，得到的专利数据集中必然包含非研究对象的干扰数据，因此在需要对原始数据进行清洗、过滤和分块处理。专利数据处理的第一个工作是对数据集进行清洗。主要是将与研究主题关系不相关的专利数据清除掉。例如在实际操作中，本文设定关键词包括“Optical transmission”在专利全文中作为整体出现的频率超过 3 次的专利为有效专利，作为整体出现次数小于 1 次的专利视为非相关专利，作为整体出现 1～2 次的专利提交由实验人员判断处理。

专利数据处理的第二个工作是完成专利数据文本的标记过滤。主要根据上面提到的数据预处理方法，过滤无关网页标记和内容，将专利数据由 HTML 格式转换成为标准文本格式。根据研究需要，建立的数据库表结构如表 3.3 所示。

表 3.4　通信技术专题美国专利数据库数据状况

子方向		专利公开库		
		检索结果	数据去重	
无线移动	研究方向一	1525 件	1926 件	3072 件
	研究方向二	428 件		
光通讯	研究方向一	73 件	496 件	
	研究方向二	117 件		
	研究方向三	120 件		
	研究方向四	307 件		
下一代网络	研究方向一	25 件	660 件	
	研究方向二	639 件		

专利数据处理的第三个工作是将经过清洗和过滤的专利数据分块。本阶段工作仍然使用第二章数据预处理的 MKIE 方法，通过对网页内容的分析处理，找到专利分块标记，将专利信息按名称、摘要、专利发明人、所属机构等不同内容分块，并分别导出美国专利，商标局专利数据库（www.uspto.gov）是美国专利数据获取的数据源。美国专利数据库中包括授权专利库和公开专利库两部分。美国授权专利库收录了自 1790 年以来的美国专利授权文献，其中 1790 年至 1975 年的数据只有全文图象页，1976 年 11 月 11 日以后的数据除了全文图象页外，还提供了可检索的 Html 格式文本。

利用表中的检索式，在美国美国公开数据库中进行检索，检索结果见表 3.4。由于

数据源的限制，公开专利数据的时间跨度为为 2001 年至 2007 年 1 月。数据去重之后，通信技术领域美国公开专利数据为 3072 条。

在原始专题专利数据库的基础上，利用北京理工大学知识发现与数据分析实验室自主开发的专利分析系统，对原始数据进行数据预处理，将文本数据转化为适合进行专利情报分析的可靠的精确的数据，形成通信技术专题美国专利数据库。其过程包括：数据清洗、数据去重、数据整合、科技分词和数据格式化等。

3.3.2 同指数据库设计

数据库根据功能设计需要主要分为规则库、辞典库、专利信息库、公司库，待处理专利信息库等。由于规则主要通过算法实现，因此，这里主要展示的是作为数据库设计重点的辞典库、专利数据库、结果及其分析应用数据库。

表3.5 辞典库

字段名称	中文名称	字段类型	备注
ID	序列号	int	自动针，主键
WordID	词代码	Char(1000)	唯一
Word	词	Vchar(100)	
FatherID	父类	char(1000)	
type_layer	结构代码	char(12)	

其中，序列号是主键，而词代码唯一标示词，父类为同指的上一级词代码，如果该词为第一级则它的父类为空，结构代码限定 5 层,初始值为 000000000000000，每三位代表一层，即一层中允许 999 个子类，结构代码是类别的先序遍历，主要为减少检索数据库的次数。

按照这样的表结构，我们来看看上面例子记录在表中的数据如表 3.6 所示。现在按 type_layer 的大小来检索一下：SELECT * FROM wordtable ORDER BY type_layer。列出记录集如表 3.7 所示。

现在列出的记录顺序正好是先序遍历的结果。在控制显示类别的层次时，只要对 type_layer 字段中的数值进行判断，每 3 位一组，如大于 0 则向右移 3 个空格。

本文根据判断的需要设计了公司辞典库，即初始的同指公司表。词表中的原始公司树状词取自汤姆森科技所提供的德温特世界专利的 Delphion、Aureka、Dement 分析家软

件，在信息抽取过程中随着新词的发现和识别，公司辞典库将随着实验进程不断扩充，同时又为标识和抽取新的同指公司词库提供帮助，从而达到机器学习的实验目的。技术关键词表的逻辑结构如表 3.5 所示。

表3.6 树状表

WordID	Word	FatherID	type_layer
1	总类别	0	00000000
2	类别 1	1	01000000
3	类别 1.1	2	01010000
4	类别 1.2	2	01020000
5	类别 2	1	02000000
6	类别 2.1	5	02010000
7	类别 3	1	03000000
8	类别 3.1	7	03010000
9	类别 3.2	7	03020000
10	类别 1.1.1	3	01010100

表3.7 树状显示表

WordID	Word	FatherID	type_layer
1	总类别	0	00000000
2	类别 1	1	01000000
3	类别 1.1	2	01010000
10	类别 1.1.1	3	01010100
4	类别 1.2	2	01020000
5	类别 2	1	02000000
6	类别 2.1	5	02010000
7	类别 3	1	03000000
8	类别 3.1	7	03010000
9	类别 3.2	7	03020000

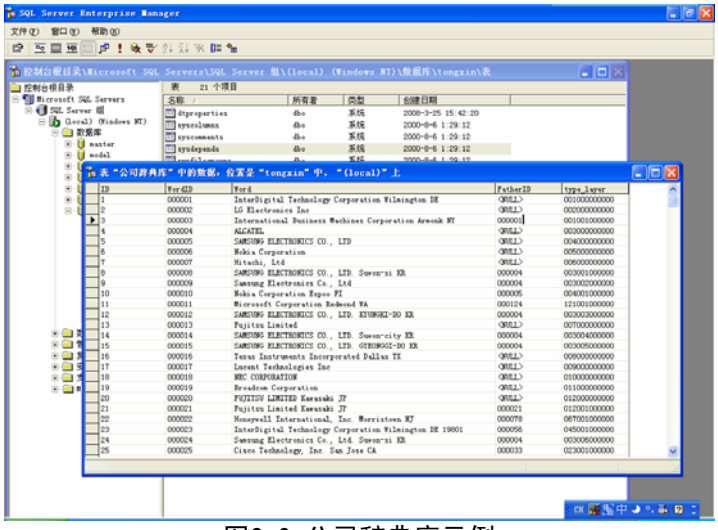


图3.6 公司辞典库示例

专利数据库主要保存经过处理后的专利信息和专利摘要和专利所述机构内容，通过数据预处理、非相关主题信息过滤之后得到有效专利信息共 3072 条。根据专利信息库的设计格式，将处理好的专利分块信息导入专利信息数据表中，所得结果如图 3.7 所示。

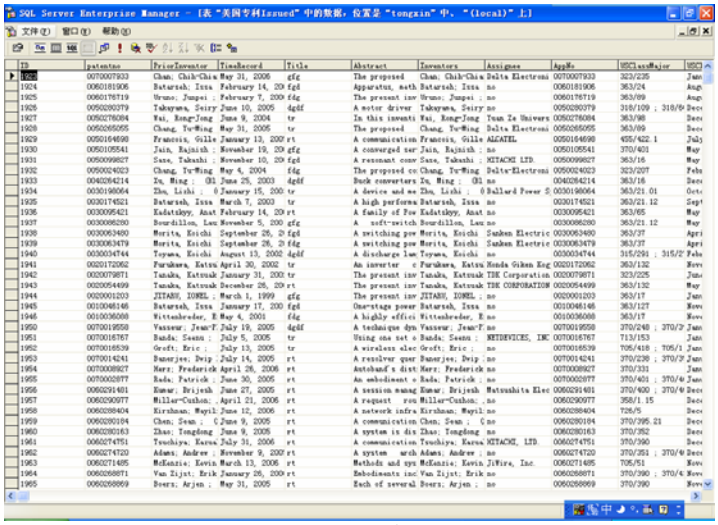


图3.7 专利信息（全）

专利信息数据表中存储的是机构化的专利分块信息，由于本文的研究对象为基于专利摘要关联规则的同指模型，因此将专利摘要和专利编号单独提取出来导入待处理专利库中，以便下一步信息的操作和处理。获得的结果如表 3.8 所示。

待处理专利信息库中的词代码与公司同指库的词代码保持一一对应，处理之后的抽取结果仍然与专利保持对应关系，以便建立公司同指库和待处理的专利文献进行关联分

析。

表 3.8 待处理专利数据（同指）

字段名称	中文名称	字段类型	备注
ID	序列号	int	自动针，主键
PatNumber	专利号	Char(20)	
Abstract	摘要	Text	
Assignee	所属机构	Vchar(1000)	
WordID	对应公司辞典库代码	Char(1000)	
Assignee1	规则 1 变化后的机构	Vchar(1000)	
Assignee2	规则 2 变化后的机构	Vchar(1000)	
Assignee3	规则 3 变化后的机构	Vchar(1000)	
Assignee4	规则 4 变化后的机构	Vchar(1000)	
Assignee5	规则 5 变化后的机构	Vchar(1000)	
Assignee6	规则 6 变化后的机构	Vchar(1000)	
Assignee7	规则 7 变化后的机构	Vchar(1000)	

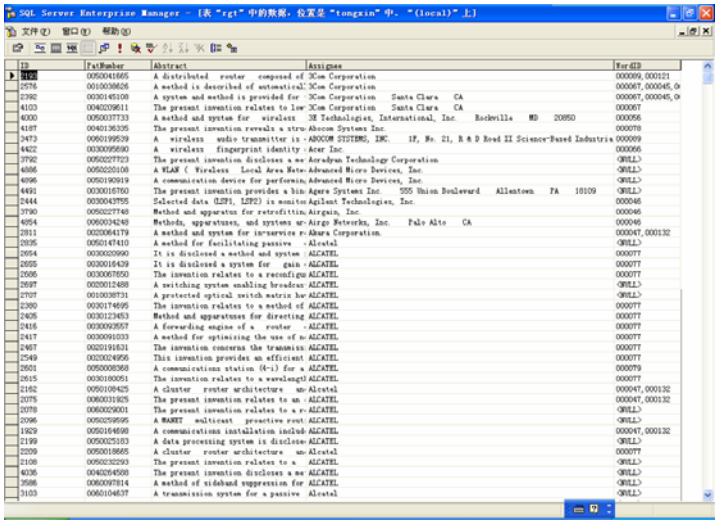


图3.8 待处理专利数据

模型中专利信息号与原专利信息产生关联，每一条专利记录对应一条或者多条公司词记录，同时也可以通过其他字段与另外的专利分析结果相关联，从而形成专利数据的关联分析网，提高了专利信息的利用率。

ID	PatNumber	Abstract	Assignee	WordID
1	00500 41665	A distributed router ….	3Com Corporation,	000047
2	00500 41665	A distributed router ….	Advanced Micro Devices, Inc.	000089

ID	WordID	Word	FatherID	type_layer
75	000047	3Com Corporation		045000000
76	000089	Advanced Micro Devices, Inc.		049000000

图3.9 表的关联图

3.3.3 基于关联规则的同指模型设计

表 3.9 高产公司表

序号	高产公司名称	数量
1	InterDigital Technology Corporation Wilmington DE	53
2	LG Electronics Inc	36
3	International Business Machines Corporation Armonk NY	30
4	ALCATEL	27
4	SAMSUNG ELECTRONICS CO., LTD	27
6	Nokia Corporation	25
7	Hitachi, Ltd	23
7	SAMSUNG ELECTRONICS CO., LTD. Suwon-si KR	23
9	Samsung Electronics Co., Ltd	18
10	Nokia Corporation Espoo FI	15
11	Microsoft Corporation Redmond WA	13
11	SAMSUNG ELECTRONICS CO., LTD. KYUNGKI-DO KR	13
13	Fujitsu Limited	12
14	SAMSUNG ELECTRONICS CO., LTD. Suwon-city KR	11
15	SAMSUNG ELECTRONICS CO., LTD. GYEONGGI-DO KR	9
15	Texas Instruments Incorporated Dallas TX	9
17	Lucent Technologies Inc	8
17	NEC CORPORATION	8
19	Broadcom Corporation	7
19	FUJITSU LIMITED Kawasaki JP	7

本文研究的是专利所述机构即公司的同指关系，首先对待处理的专利数据中所属机构属性进行统计，统计得到 633 个不同的公司或机构，其中这 633 个公司或机构中将存在同指关系，其中申请最多的公司如表 3.9 所示。

通过分析我们发现，公司同指的关系有如下几种：

标点去除：如果去掉标点之后，完全一致的

英文大小写

分公司：所属地区不同，CORPORATION、corporation、Limited、Co.，Ltd、CO.，LTD、Inc 等词的前部分相同，后部分不一致

别名：正式的或规范的名称以外的名称，如 Samsung Electronics Co.，Ltd、Samsung Co.， Ltd

简称：较复杂的名称的简化形式，如 Nippon electronic company 简称为 NEC

经过规则判断，我们得到以下几个典型地规则关系

（1） 标点去除一致：

表3.10 标点规则去除

序列号	原数据	去除标点空白符规则检验后
1	SBC Knowledge Ventures L.P. Reno NV	SBC Knowledge Ventures L P Reno NV
2	SBC Knowledge Ventures, L.P Reno NV	
3	SBC Knowledge Ventures, L.P. Reno NV	
4	Samsung Electronics, Co.,Ltd	Samsung Electronics Co Ltd
5	Samsung Electronics Co. Ltd	
6	Samsung Electronics Co., Ltd. Suwon City Kr	Samsung Electronics Co Ltd Suwon City Kr
7	Samsung Electronics Co., Ltd. Suwon-City Kr	

通过标点去除后，原来 633 个公司合并成 521 个公司。并在数据库“待处理专利数据”中的 Assignee1 属性中填入转化后的内容。

（2） 英文大小写

表3.11 英文大小写规则去除

序列号	原数据	去除大小写规则检验后
1	MOTOROLA INC	Motorola Inc
2	Motorola Inc	
3	FUJITSU LIMITED Kawasaki JP	Fujitsu Limited Kawasaki Jp
4	Fujitsu Limited Kawasaki JP	
5	Samsung Electronics Co Ltd Suwon si Kr	Samsung Electronics Co Ltd Suwon si Kr
6	SAMSUNG ELECTRONICS CO LTD SUWONSI KR	
7	Samsung Electronics Co Ltd Suwon si KR	
8	ALCATEL	Alcatel
9	Alcatel	
10	ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE	Electronics And Telecommunications Research Institute
11	Electronics and Telecommunications Research Institute	

通过对不规格的英文大小写转化，原来 521 个公司合并成 499 个公司。并在数据库“待处理专利数据”中的 Assignee2 属性中填入转化后的内容。

在专利信息的摘要中使用英文 NLP 的方法，对专利信息进行关联分析，专利关联分析是指通过计算专利情报中相关关键词之间的共生关系来确定其技术体系中各研究主题之间的关联程度。在关联图中，主题关键词之间联系用线条表示，线条长度表示两主题间的关联程度。一个主题与其它主题连线短而多，一方面说明该主题与其它主题有着较强的关系，另一方面也说明该主题的开放程度高，主题的外延在不断发展。在图 3.9 中，球的大小代表了该技术的专利个数，线的长短、虚实则反映的是相关技术间的关联关系。虚线表示两者的关联最弱，而连线越短两者之间关联关系越强。根据多次的实验经验我们选择关联度为 0.6 来确定关联范围，在这有关联的范围内有可能存在同指关系。

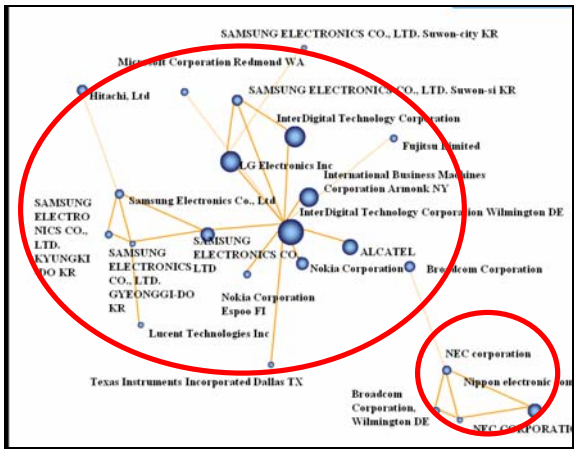


图3.9 关联规则消解

根据关联规则的关系我们把 499 个公司分成 2 大类，以便于后面的分析。接着，我们对这 2 类数据进行基于规则和基于辞典库的分析，首先判断该公司词是否在辞典库中存在，如果存在则直接在待处理专利中标注其词标号；若不存在，则进行规则判断，该公司的规则判断有如下几种情况：

（3） 别名

表3.12 别名规则去除

序列号	原数据	别名规则检验后
1	At T Corp	At T Corp
2	At T Corporation	
3	At T Co Ltd	

通过对不规格的别名转化，原来 499 个公司合并成 492 个公司。并在数据库“待处

理专利数据”中的 Assignee3 属性中填入转化后的内容。

（4） 简称

表3.13 简称规则去除

序列号	原数据	简称规则检验后
1	IBM Corporation	IBM Corporation
2	International Business Machines Corporation	
3	Nippon electronic company	NEC
4	NEC	

简称规则检验需要利用辞典库来完成，若辞典库中存在该类的简称，则直接在在数据库“待处理专利数据”中的 Assignee4 属性中填入检验后的内容；否则根据关联的紧密程度来判断它们之间存在简称关系的概率。根据几次实验经验，我们选定关联度大于或等于 0.9 的公司存在简称关系。若出现新的简称关系，需要回填到公司辞典库中。

(5) 分公司

表3.14 分公司规则去除

母公司	一级分公司	二级
Samsung Electronics Co Ltd	Samsung Electro Mechanics Co Ltd 314 Maetan 3 Dong Youngtong Ku Suwon Kyungki Do Kr	
	Samsung Electronics Co Ltd Suwon si Kr	Samsung Electronics Co Ltd Suwon Si Kr Postech Foundation Pohang Si Kr
		Samsung Electronics Co Ltd Suwon Si Kr Seoul National University Industry Foundation Seoul Kr
		Samsung Electronics Co Ltd Suwon Si Kr Korea Advanced Institute Of Science And Technology (Kaist) Yusong Gu Kr
		Samsung Electronics Co Ltd Suwon Si Kr Yonsei University Seoul Kr
	Samsung Electronics Co Ltd City University Of New York	
	Samsung Electronics Co Ltd Gyeonggi Do Kr	
	Samsung Electronics Co Ltd Kyungki Do Kr	
	Samsung Electronics Co Ltd Suwon City Kr	
Alcatel	Alcatel Paris Fr	
Broadcom Corporation	Broadcom Corporation Irvine CA	Broadcom Corporation 16215 Alton Parkway Irvine CA 92618-3616
		Broadcom Corporation 16215 Alton Parkway Irvine CA 92618
		Broadcom Corporation Irvine CA 92618
	Broadcom Corporation, a California Corporation	Broadcom Corporation, a California Corporation Irvine CA

分公司规则检验需要利用辞典库来完成，若辞典库中存在该类公司的树状结构，则

直接在在数据库“待处理专利数据”中的 Assignee5 属性中填入检验后的内容；否则首先根据公司名称前半段是否一致，来判断其父子关系，其次根据关联的紧密程度来判断它们之间存在父子关系的概率。根据几次实验经验，我们选定关联度大于或等于 0.9 的公司存在父子关系。若出现新的父子关系需要回填到公司辞典库中。

本文利用 3072 条美国通信领域的专利数据进行研究，首先通过简单的统计共有 633 个不同的公司，在通过关联规则分析和基于规则库和辞典库的检验后，结果如下：

表3.15 规则数据表

步骤	规则	数据
1	样本	633 条
2	标点规则检验	521 条
3	大小写规则检验	499 条
4	关联分析	182 条，97 条
5	别名规则检验	177 类，85 类
6	简称规则检验（基于辞典库和关联度大于或等于 0.9）	167 类，80 类
7	分公司规则检验（基于辞典库和关联度大于或等于 0.9）	67 类，33 类

3.3.4 实验结果分析

基于关联规则的同指消解模型设计是在建立在完善的关联分析和命名实体识别的基础上，根据模型的设计思想，本文将抽取过程定义为两个阶段，抽取的结果分别如下：

（1）样本训练阶段

选取词性标注第二阶段所使用的 633 篇训练样本进行基于关联规则的同指消解模型训练，获得的实验结果如下表所示。

表3.15 样本训练数据表

测试样本编号	样本包含机构（公司）	抽取到的机构（公司）	结果中正确机构
1	633	282	147

根据第二章信息抽取综述中介绍的信息抽取系统的性能评价指标，本文对实进行了统计和计算，如表所示。

召回率 R=抽取到的技术关键词/样本包含技术关键词=282/633=44.55%

准确率 P=结果中正确的技术关键词/抽取到的技术关键词=147/282=52.12%

综合评价指数

$$F=\frac{(\beta^2+1)PR}{\beta^2P+R}=\frac{(1+1)52.12\%*44.55\%}{1*52.12\%+44.55\%}=24.02\% \qquad (\beta取值为1)$$

表3.16 训练阶段信息抽取模型性能评价指标

模型召回率	模型准确率	F 指数
44.55%	52.12%	24.02%

训练结果的统计由人工来完成，利用公式计算召回率、准确率和 F 指数之后，重新修正标注样本，并根据新的统计概率数据对原模型的概率参数进行修正，同时将训练中发现并确认的新技术关键词更新到词典库中。完成模型反馈和修正工作之后，再重复对训练数据进行全过程的抽取处理，检验输出结果的改变情况。

（2）模型测试阶段

通过在样本训练阶段的机器学习和人工修正之后，选取词性标注第三阶段所使用的633 篇训练样本进行词法、句法分析和命名实体识别测试，得到的测试结果如表所示。

表5.17 模型测试数据表

测试样本编号	样本包含机构（公司）	抽取到的机构（公司）	结果中正确机构
2	633	372	246

根据第二章信息抽取综述中介绍的信息抽取系统的性能评价指标，本文对实进行了统计和计算，如表所示。

召回率 R=抽取到的技术关键词/样本包含技术关键词=372/633=58.77%

准确率 P=结果中正确的技术关键词/抽取到的技术关键词=246/372=66.13%

综合评价指数

$$F=\frac{(\beta^2+1)PR}{\beta^2P+R}=\frac{(1+1)58.77\%*66.13\%}{1*58.77\%+66.13\%}=31.11\% \qquad (\beta取值为1)$$

最后，本章将训练阶段与测试阶段的基于关联规则的同指消解模型性能数据作对比如下：

表4.1 测试阶段信息抽取模型性能评价指标

	模型召回率	模型准确率	F 指数
训练阶段	44.55%	52.12%	24.02%
测试阶段	58.77%	66.13%	31.11%
性能提高	31.92%	26.88%	29.52%

通过观察分析可以发现使用了合理的命名实体识别模型规则和经过训练改进之后的

模型在召回率和准确率方面都得到了非常明显的提高，可以认为本文所设计的人工指导和机器学习相结合的训练方法取得了成功。

3.4 在专利检索中的应用

根据专利相关技术检索使用的工具和检索系统的功能，可以将专利检索分为手工检索和机器检索。手工检索由于效率极低，随着网络的普及，正在逐渐淡出。计算机检索主要是以专利文献信息特征为依据，应用字段检索、通配符检索、布尔逻辑检索、位置检索、范围检索、跨字段逻辑组配检索等方法，对选定的专利数据库，依靠计算机查找全部专利信息的过程。

根据检索目的不同，专利检索又可以分为专利技术信息检索、专利性检索、侵权检索、专利法律状态检索、专利族检索、技术贸易检索、专利战略信息检索等。但无论哪种检索，首要目标就是找到全部相关专利信息，然后根据需求进行进一步的专利情报分析。

用户在进行专利检索时，通常以某一项或多项专利信息特征作为检索依据，这种专利信息特征主要包括：

专利文献号：具有唯一性的索取专利文献的检索依据。包括：公开号、公告号、专利号。

专利权人：是从与专利有关的人（自然人或法人）的角度检索专利的依据。包括：专利申请人、专利受让人、专利权人、专利出让人、专利发明人、设计人等。

主题词：是专利纪录中的实词，分为标引词和非标引词。标引词是指从专利文摘和全文中筛选出来的词，非标引词也称自由词，在有的专利检索系统中自由词还包括字。

专利分类号：是从技术主题角度检索专利的依据。包括：国际专利分类号、各国专利分类号、特定出版物使用的特定专利号等。

专利公布日期或专利申请日期：通常不单独使用，而与其他检索条件组合进行检索。

3.4.1 专利权人的公司树建立

在专利检索中，由于每个公司出于对自我的保护，再申请专利时有可能会使用不同的简称、别名来申请，这使得我们在搜索该公司的专利情报时，往往会出现很大的误差，专利的查全率非常低。

基于专利检索模型，我们利用基于关联规则的同指消解模型建立公司树索引，从而

提高专利专利权人检索精度。在用户使用专利检索时，模型首先处理用户输入的检索条件，将检索专利权人关键词在公司树模型中进行匹配，然后该类公司全部的公司名称与专利信息表进行匹配，返回匹配成功的专利文献。若不能得到匹配成功的返回项，则提示用户重新输入检索关键词并给出最接近的公司名称供用户选择。

根据上文专利检索的模型设计思想，本文将专利技术关键词索引的工作模式设计为：

（1）检索条件处理。主要针对用户输入的检索词语的处理，包括语句解析、错误分析、类比推测等，使得用户输入得以转化为机器能够识别的指令信息。

（2）检索条件匹配。将处理过的检索条件与公司树索引进行匹配分析，主要方法是：检索词组的整体匹配、将词组切分成短语进行匹配、以单词形式进行匹配。

（3）检索结果生成。根据检索条件与公司树索引的匹配程度对结果进行排序生成，排序的权衡指标为该检索条件与索引的关联度大小和匹配次数的多少。

（4）用户反馈。将结果输出给用户，接收用户的反馈信息，修正检索过程。

3.4.2 公司树检索的意义

我们将信息抽取技术、数据挖掘技术应用在专利信息分析中，充分发挥信息抽取和数据挖掘技术在处理海量文本信息方面的优势，以实现自动地抽取申请人公司树的重要信息，作为检索主题，弥补了当前专利检索方式的不足，降低了专利检索的应用难度，并有效提高了相关技术领域的专利检出率。同时，信息抽取与信息检索的结合也为专利信息分析工作的提供了有力的帮助。

3.5 本章小结

本章首先描述同指消解的定义和评估系数，然后根据专利信息的特点，设计了一个基于关联规则的同指消解模型，主要包括专利数据选择、专利数据获取、专利数据预处理、关联规则主要是关联度的计算等以便建立新的理论和方法模型。

同时，利用此方法通过通信专利数据进行模型的实验，把准备好的专利数据信息结合人工指导和机器学习训练从中抽取出同指库，并将抽取结果生成基于同指的专利辞典。本章将训练阶段与测试阶段的基于关联规则的同指消解模型性能数据作对比，可以发现使用了合理的命名实体识别模型规则和经过训练改进之后的模型在召回率和准确率方面都得到了非常明显的提高。该辞典可用于建立专利检索中的申请人公司树，从而提高专利在申请人检索方面的查全率。

本章的主要创新点是基于关联规则的同指消解模型的提出。基于语言模型的专利相关技术检索，是一门广泛的技术，再加上信息抽取技术在我国还是一门前沿学科，本章还有很多值得研究的地方。

4 基于聚类分析的异指消解技术

4.1 异指消解定义

异指反映了篇章的语言单位，包括所有命名实体，相同的实体指向或代表了不同的实体意义。典型的如同一个姓名可能代表不同的人。由于专利信息量非常大，具有同名同姓的专利权人（申请人）和发明人在所难免，因此可以设计合理的基于聚类分析的异指关系模型，结合模式抽取和匹配、结构信息分析等方法用在发明人的关系学习和信息抽取上。

使用基于聚类分析技术和基于统计学方法对存在异指关系的属性值进行标注。在规则消解过程中使用基于 K-MEANS 的聚类表示形式规则可以独立程序进行添加修改和删除，使该异指消解系统具有很好的可维护性和可扩展性。

4.2 基于聚类分析的异指消解模型的提出

针对专利发明人属性的特点，本章提出基于聚类分析的异指消解模型，具体的流程如图 4.1 所示：

首先对美国专利的网页信息进行本地化下载，将 WEB 网页的非结构化数据转化为结构化数据形式，通过消除数据冗余、去除噪音词、数据格式转化等方式进行专利数据清洗，以实现对专利数据的预处理。

然后对需要分析的属性-如发明人进行基于统计分析的分类排序，这些相同的发明人有可能存在异指关系，将可能存在异指关系的发明人，对其他属性包括专利摘要利用文本挖掘技术挖掘出来的摘要关键词进行聚类分析，以找出相同的申请人是否存在研究的方向相异性，接着对该关系确定阈值，以便确定是否存在异指关系，如果在此阈值内聚为一类，则我们将认为它不存在异指关系，结束；否则，可能存在异指关系，再对该类信息的其他属性进行基于规则库和统计的属性分析，存在差异者认为存在异指关系，对该类属性进行标注并填充到英文异指辞典。

本章提出了一套全新的命名实体识别模型及其算法，然后选择合适的抽取结果输出方式。接着，我们通过实证数据进行模型的实验，结合人工指导和机器学习训练，从专利中抽取出异指库，并将抽取结果生成基于异指关系的发明人属性标注，以便在专利检索提高发明人的查准率。

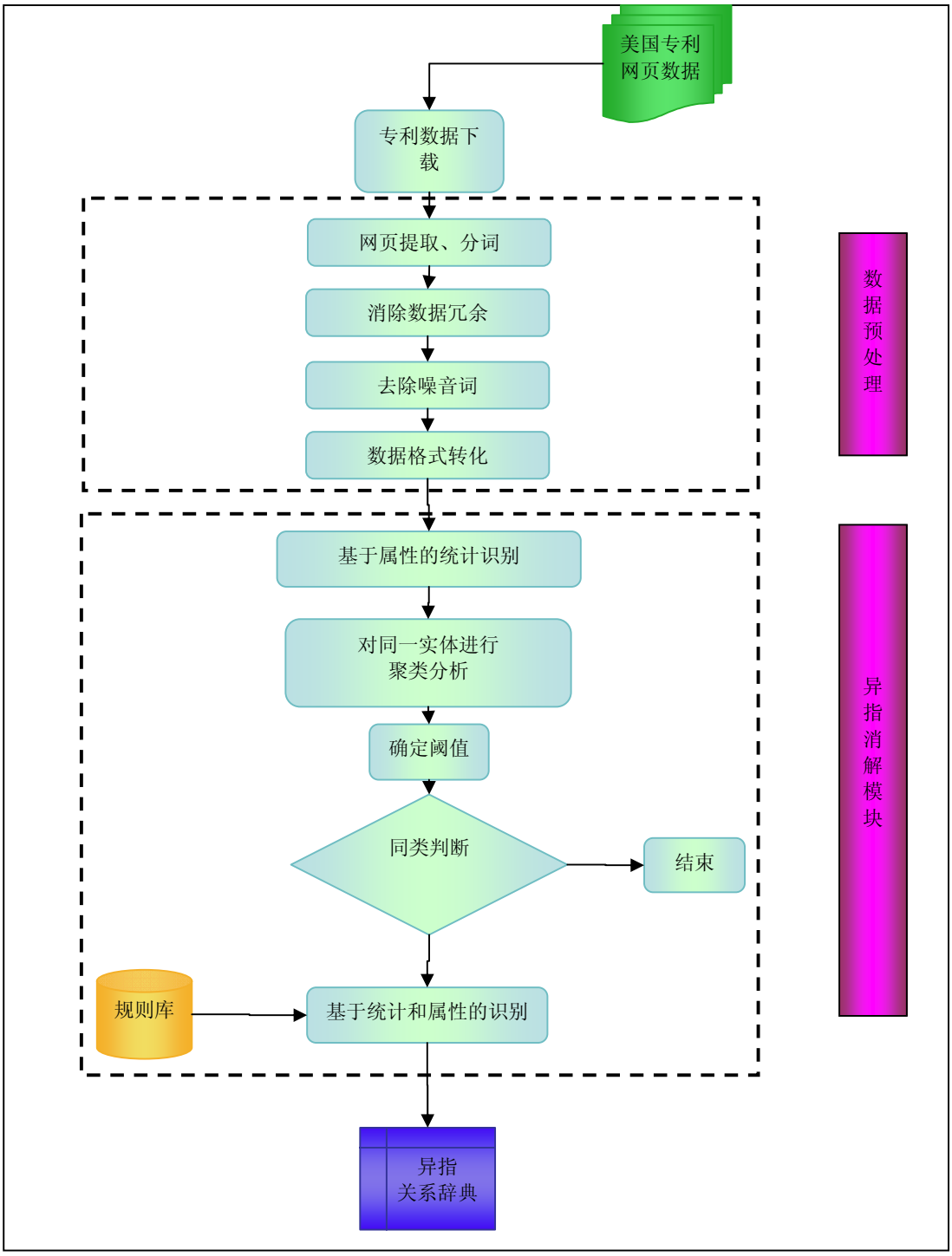


图 4.1 基于聚类分析的异指消解模型

4.2.1 k-means 聚类方法

k-means 算法是一种基于样本间相似性度量的间接聚类方法，属于非监督学习方法。

此算法以 k 为参数，把 n 个对象分为 k 个簇，以使簇内具有较高的相似度，而且簇间的相似度较低。相似度的计算根据一个簇中对象的平均值（被看作簇的重心）来进行。此算法首先随机选择 k 个对象，每个对象代表一个聚类的质心。对于其余的每一个对象，根据该对象与各聚类质心之间的距离，把它分配到与之最相似的聚类中。然后，计算每个聚类的新质心。重复上述过程，直到准则函数会聚。k-means 算法是一种较典型的逐点修改迭代的动态聚类算法，其要点是以误差平方和为准则函数。逐点修改类中心：一个象元样本按某一原则，归属于某一组类后，就要重新计算这个组类的均值，并且以新的均值作为凝聚中心点进行下一次象元素聚类；逐批修改类中心：在全部象元样本按某一组的类中心分类之后，再计算修改各类的均值，作为下一次分类的凝聚中心点[39]。

k-means 算法的工作过程说明如下：首先从 n 个数据对象任意选择 k 个对象作为初始聚类中心；而对于所剩下其它对象，则根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的（聚类中心所代表的）聚类；然后再计算每个所获新聚类的聚类中心（该聚类中所有对象的均值）；不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数. k 个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。其定义如下：

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2 \qquad \text{（公式 4.1）}$$

这里的 E 是数据库中所有对象的平方误差的总和，p 是空间中的点，表示给定的数据对象，m 是簇 C 的平均值（p 和 m 都是多维的）。这个准则试图生成的结果簇仅可能地紧凑和独立。

- **输入** 期望得到的簇的数目 *k*，*n* 个对象的数据库。
- **输出** 使得平方误差准则函数最小化的 k 个簇。
- **方法**
  - 选择 k 个对象作为初始的簇的质心；
  - repeat
  - 计算对象与各个簇的质心的距离，将对象划分到距离其最近的簇；
  - 重新计算每个新簇的均值；
  - until 簇的质心不再变化。

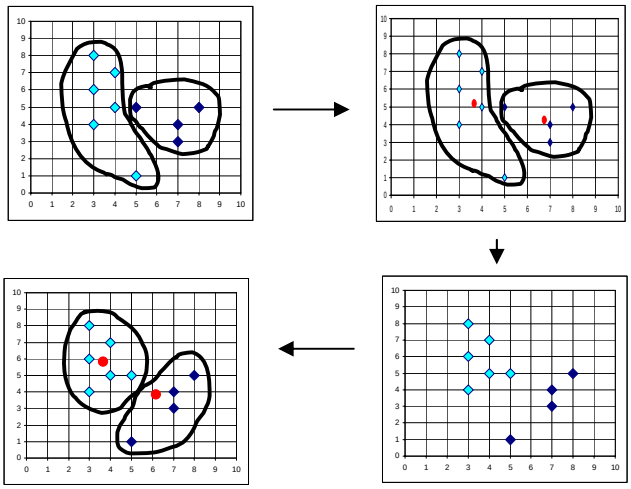


图 4.2 聚类算法图

这个算法尝试找出使平方误差函数值最小的 K 个划分。当结果簇是密集的，而簇与簇之间区别明显时，它的效果较好。对处理大数据集，该算法是相对可伸缩和高效率的，因为它的复杂度是 O（nkt），其中 n 是所有对象的数目，k 是簇的数目，t 是簇的数目。通常地，k<<n，且 t<<n。这个算法经常以局部最优结束。

一个常常有助于获得好的结果的策略就是首先应用自下而上层次算法来获得聚类数目，并发现初始分类；然后再应用循环再定位（聚类方法）来帮助改进分类结果。

K－Modes 算法：该算法通过用模来替换聚类均值、采用新差异性计算方法来处理符号量，以及利用基于频率对各聚类模进行更新方法，从而将 K－Means 算法的应用范围从数值量扩展到符号量。

K－Means 算法和 K－Modes 算法结合到一起，就可以对采用数值量和符号量描述对象进行聚类分析，从而构成了 K－Prototype 算法。

而 EM（期望最大化）算法又从多个方面对 K－Means 算法进行了扩展。其中包括：它根据描述聚类所属程度的概率权值，将每个对象归类为一个聚类，不是将一个对象仅归类为一个聚类（所拥有）；也就是说在各聚类之间的边界并不是非常严格。因此可以根据概率权值计算相应的聚类均值。该算法明确地使用某种概率性方法来确定某个数据点存在于某个分类中的概率。本文使用 EM 算法进行实证研究。

4.3 基于聚类规则的异指消解模型设计与实验

4.3.1 专利数据获取

同第三章所使用的专利数据一样，本章的研究对象同样为通信技术专利数据，因此

在数据准备阶段包括数据的下载和预处理同样利用了北京理工大学知识发现与数据分析实验室自主开发的专利分析系统，对原始数据进行数据预处理，将文本数据转化为适合进行专利情报分析的可靠的精确的数据，形成通信技术专题美国专利数据库。其过程包括：数据清洗、数据去重、数据整合、科技分词和数据格式化等。在此不作详细介绍。

4.3.1 异指数据库设计

数据库根据功能设计需要主要分为规则库、辞典库、专利信息库，专利技术关键词索引库等。由于规则主要通过算法实现，因此，这里主要展示的是作为数据库设计重点的辞典库、专利数据库、结果及其分析应用数据库。

表 4.1 待处理专利数据（异指）

字段名称	中文名称	字段类型	备注
ID	序列号	int	自动针，主键
PatNumber	专利号	Char(20)	
Inventors	发明人	Vchar（1000）	
Title	标题	Vchar（1000）	
Assignee	所属机构	Vchar（1000）	
USClass	US 类	Vchar（100）	
IPC	IPC 类	Vchar（100）	
country	国家	Vchar（100）	
InvenType	发明人标示	Char（100）	当出现同名但不指向同一个人时在这里标示

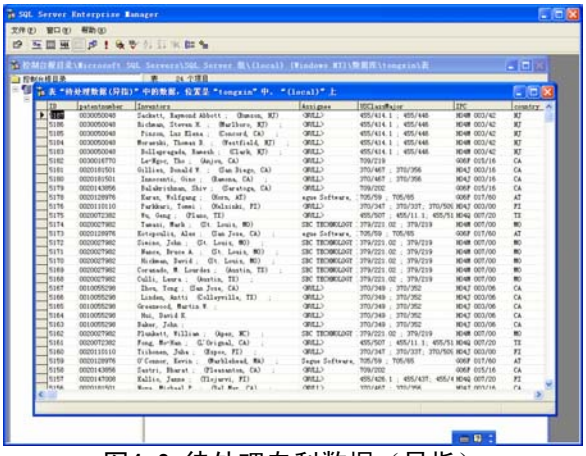


图4.3 待处理专利数据（异指）

模型中专利信息号与原专利信息产生关联，每一条专利记录对应一条或者多条发明人记录，同时也可以通过其他字段与另外的专利分析结果相关联，从而形成专利数据的

关联分析网，提高了专利信息的利用率。

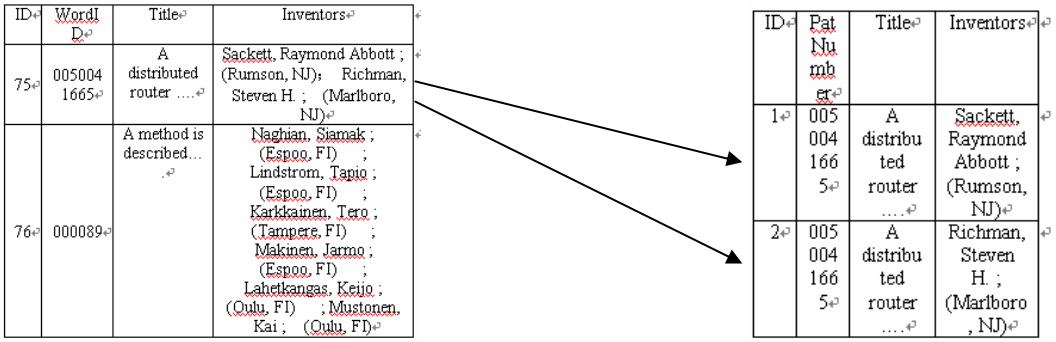


图4.4 信息对应

4.3.2 基于聚类分析的异指模型建立

本文研究的是专利发明人的异指关系，首先对待处理的专利数据中发明人属性进行统计，在 3072 条通信专利中统计得到 7262 个发明人，其中这 7262 个发明人中将存在异指关系，其中申请最多的发明人如表 4.2 显示。

表 4.2 高产发明人

序号	高产发明人	数量
1	Kim	57
2	Lee	54
3	Park	29
4	Tsuchiya, Kazuaki	18
5	Choi	17
6	David	16
6	Michael	16
6	Terry, Stephen E	16
9	Jin	15
9	Marc	15
11	Chen	13
11	Dick, Stephen G	13
11	Wang	13
14	Alain	12
14	Liu	12
14	Li	12
17	Arseneau	11
17	Chang	11
17	Charette	11
17	Jean	11

例如：申请量最多的 Kim 是否是存在异指关系即是否存在同名同姓不同人的情况，首先我们利用专利的标题、申请机构等属性对其进行 K-Means 聚类分析。本文利用 SQL Analysis Services 软件进行聚类分析。

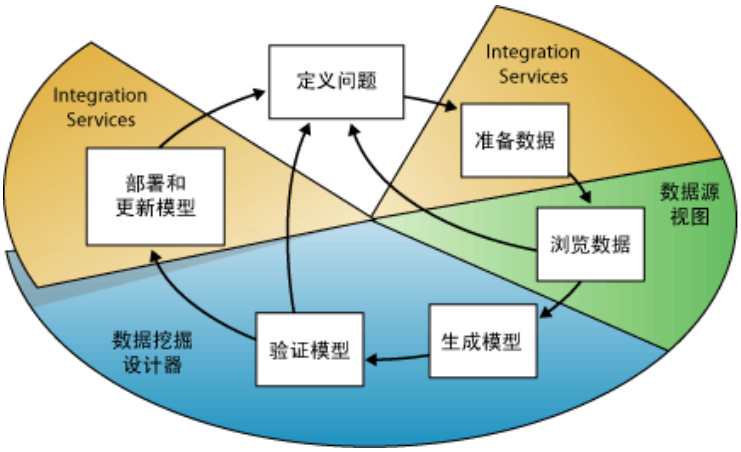


图 4.5 聚类分析过程图

此过程包括从定义模型要解决的基本问题到将模型部署到工作环境的所有事情[47]。此过程可以使用下列六个基本步骤进行定义：

定义问题：该步骤包括分析业务需求，定义问题的范围，定义计算模型所使用的度量，以及定义数据挖掘项目的最终目标。本次的目标就是分析 Kim 申请的专利已获得该人名是否存在着异指关系。

图 4.6 数据挖掘数据方案建立

准备数据：内容包括转换到自动执行数据清除和合并。根据上面数据库的建立，我们已经建立起有关 Kim 有关的数据库。我们对管理关系数据以进行多维使用的最常用的方式是星型架构。星型架构由一个事实数据表和链接到该事实数据表的多个维度表组成。

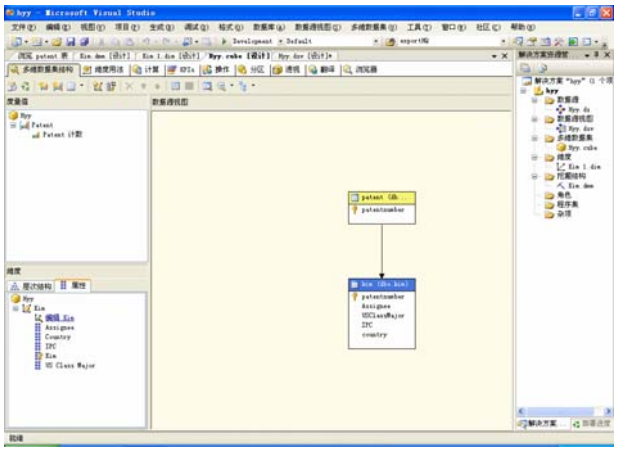


图 4.7 多维数据集建立

浏览数据：浏览技术包括计算最大值和最小值，计算平均偏差和标准偏差，以及查看数据的分布。浏览完数据之后，便可确定数据集是否包含缺陷数据，然后制订纠正这些问题的策略。

生成模型：在生成模型之前，必须随机将已准备的数据分离到单独的定型数据集和测试数据集。使用 KIM 数据集生成模型，并通过创建预测查询来使用测试数据集测试模型的准确性。同时使用 Integration Services 中的百分比抽样转换来拆分数据集。

我们使用从浏览数据步骤中获得的知识来帮助定义和创建挖掘模型。模型通常包含多个输入列、一个标识列以及一个可预测列。定义完聚类挖掘模型的结构之后，需要对其进行处理，使用说明模型的模式来填充空结构。这称为“定型”模型。模式通过利用数学算法计算原始数据而得。

首先，建立小型的数据仓库 DB_Kim 和多维数据集，多维数据集是数据的一种多维结构。多维数据集由维度和度量值的集合进行定义。根据本文的要求，我们建立了关于 Kim 的多维数据库模型，主要分为专利基本信息维、专利异指分析维、专利类型分析型维、专利时间维、专利地区维。其中，专利异指分析维是我们聚类分析的重点，在专利异指分析维里有如下几个度量值：所属机构、IPC 类。我们可以选择任意整数个对象作为初始的簇的质心进行聚类分析。

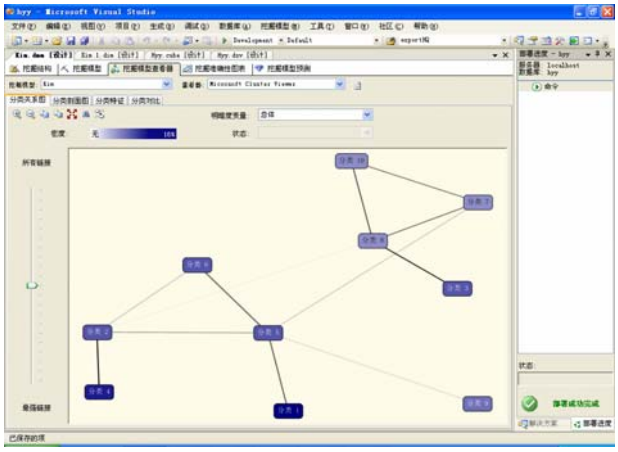


图 4.8 聚类分析结果

浏览和验证模型：当不希望在事先没有测试模型性能的情况下将模型部署到生产环境。同样，也许已经创建了数个模型，并且必须确定性能最佳的模型。如果在创建模型步骤中创建的所有模型都无法正常工作，则必须返回到此过程的上一个步骤，重新定义问题或重新调查原始数据集中的数据。

我们利用 BI Development Studio 中数据挖掘设计器内的查看器来浏览算法发现的趋势和模式。还可以使用该设计器中的工具（如提升图和分类矩阵）来测试模型创建预测的性能。这些工具要求使用在模型生成步骤中从原始数据集内分离的测试数据。

图 4.9 分类矩阵分析结果

部署和更新模型：当生产环境中部署了挖掘模型之后，便可根据需求执行多次任务。包括模型的预测和深入挖掘管理。如图 4.11 所示，我们可以深入挖掘到每个类别的具体内容，以及每个类别的概率属性。

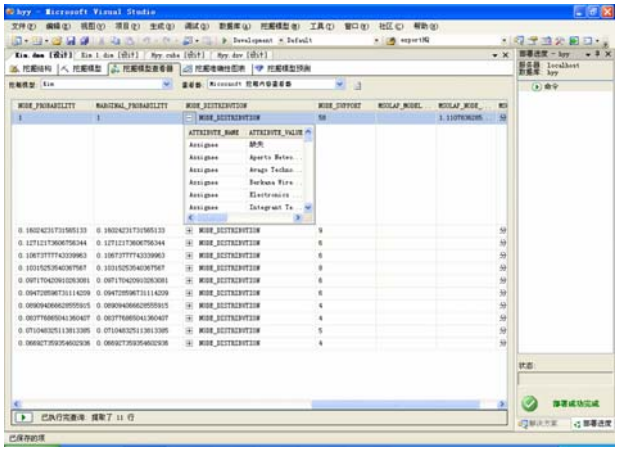


图 4.10 深入挖掘分析结果

尽管关系图中所示的过程是一个循环过程，但是每个步骤并不需要直接执行到下一个步骤。创建数据挖掘模型是一个动态、交互的过程。浏览完数据之后，我们发现数据不足，无法创建适当的挖掘模型，因此必须查找更多的数据。我们再次生成数个模型，但发现这些模型无法回答定义问题时所设定的问题，因此必须重新定义问题，在部署模型之后对其进行更新，又出现了更多的可用数据。

在“内容详情”窗格中，在分段树区域中，颜色代表事例的密度。颜色越深则节点中包含的事例就越多。单击“全部”节点。

我们通过改变输入列和预测列的属性进行多角度地比较分析，发现 Kim 申请的专利大部分都居于一类，只有少数几条数据被孤立处来。

表 4.3 孤立点分析

ID	Patent Number	Inventors	TimeRecord	USClass	IPC	filedyear	country
2341	0060154282	Park; Tae-sik ; Kim; Young-il ; Kang; Jung-ho ;	December 21, 2005	35/6 ; 340/870.18; 435/287.2; 702/19	C12Q 1/68 20060101 C12Q001/68;	2006	IL
4834	0060095615	Kim; Jong Won ; Choi; Sangsung ; Park; Kwang Rob ;	December 20, 2004	710/62	G06F 13/38 20060101 G06F013/38	2006	FI
2710	0060214131	Lee; Sang-Goo ; Kim; Min-Chan ; Choi; Byung-Ju ; Shin; Min-Chul ; Yu; Su-Han ;	July 30, 2003	252/62	E04B 1/74 20060101 E04B001/74	2006	TW

通过研究专利的摘要和全文，我们发现这几条专利无论从专利申请的内容、IPC 以及国别上都大不相同，所以我们有认为这几条专利的 Kim 与其他 53 条专利的 Kim 非同一

个人，存在同名同姓的情况，所以在待处理专利数据（异指）数据库中的 InvenType 属性进行标示。

4.3.3 实验结果分析

基于聚类规则的异指消解模型设计是在建立在完善的 K-Means 分析和命名实体识别的基础上，根据模型的设计思想，本文将抽取过程定义为两个阶段，抽取的结果分别如下：

（1）样本训练阶段

选取词性标注第二阶段所使用的 56 篇训练样本进行基于聚类规则的异指消解模型训练，获得的实验结果如下表所示。

表4.4 样本训练数据表

测试样本编号	样本包含发明人	抽取到的发明人	结果中正确机构
1	56	16	3

根据第二章信息抽取综述中介绍的信息抽取模型的性能评价指标，本文对实进行了统计和计算，如表所示。

召回率 R=抽取到的发明人/样本包含发明人=16/56=28.58%

准确率 P=结果中正确的发明人/抽取到的发明人=3/16=18.75%

综合评价指数

$$F=\frac{(\beta^2+1)PR}{\beta^2P+R}=\frac{(1+1)18.75\%*28.58\%}{1*18.75\%+28.58\%}=11.32\% \qquad (\beta\text{取值为}1)$$

表4.5 训练阶段信息抽取模型性能评价指标

模型召回率	模型准确率	F 指数
28.58%	18.75%	11.32%

训练结果的统计由人工来完成，利用公式计算召回率、准确率和 F 指数之后，重新修正标注样本，并根据新的统计概率数据对原模型的概率参数进行修正，同时将训练中发现并确认的新技术关键词更新到词典库中。完成模型反馈和修正工作之后，再重复对训练数据进行全过程的抽取处理，检验输出结果的改变情况。

（2）模型测试阶段

通过在样本训练阶段的机器学习和人工修正之后，选取词性标注第三阶段所使用的 56 篇训练样本进行词法、句法分析和命名实体识别测试，得到的测试结果如表所示。

表4.6 模型测试数据表

测试样本编号	样本包含发明人	抽取到的发明人	结果中正确机构
2	56	20	7

根据第二章信息抽取综述中介绍的信息抽取模型的性能评价指标，本文对实进行了统计和计算，如表所示。

召回率 R=抽取到的发明人/样本包含发明人=20/56=35.71%

准确率 P=结果中正确的发明人/抽取到的发明人=7/20=35%

综合评价指数

$$F=\frac{(\beta^2+1)PR}{\beta^2P+R}=\frac{(1+1)35\%*35.71\%}{1*35\%+35.71\%}=17.67\% \qquad (\beta\text{取值为}1)$$

本文将训练阶段与测试阶段的信息抽取模型性能作对比如下：

表4.7 测试阶段信息抽取模型性能评价指标

	模型召回率	模型准确率	F 指数
训练阶段	28.58%	18.75%	11.32%
测试阶段	35.71%	35%	17.67%
性能提高	24.95%	86.67%	56.10%

通过观察分析可以发现使用了合理的命名实体识别模型规则和经过训练改进之后的模型在召回率和准确率方面都得到了非常明显的提高，可以认为本文所设计的人工指导和机器学习相结合的训练方法取得了成功。

4.4 在专利检索中的应用

目前比较先进的计算机检索主要是以专利文献信息特征为依据，应用字段检索、通配符检索、布尔逻辑检索等检索等方法依靠计算机查找全部专利信息的过程。但是，目前的专利检索主要针对申请日期、申请人等简单的计算机匹配检索。由于在发明人的名称中存在同名同姓的现象非常普遍，这使得专利的查准率也不高。

4.4.1 发明人标引的建立

基于专利检索模型我们利用基于聚类分析的异指消解模型建立发明人标引，从而提高专利发明人检索准确度。在用户使用专利检索时，模型首先处理用户输入的检索条件，将检索专利权人关键词在异指词典库中进行匹配，然后该人名全部的名称与专利信息表进行匹配，返回匹配成功的专利文献。若不能得到匹配成功的返回项，则提示用户重新

输入检索关键词并给出最接近的人名供用户选择。

根据上文专利检索的模型设计思想，本文将专利技术关键词索引的工作模式设计为：

（1）检索条件处理。主要针对用户输入的检索词语的处理，包括语句解析、错误分析、类比推测等，使得用户输入得以转化为机器能够识别的指令信息。

（2）检索条件匹配。将处理过的检索条件与发明人异指库进行匹配分析，主要方法是：检索词组的整体匹配、将词组切分成短语进行匹配、以单词形式或者区分大小写形式进行匹配。

（3）检索结果生成。根据检索条件与发明人的匹配程度对结果进行排序生成，排序的权衡指标为该检索条件与索引的关联度大小和匹配次数的多少。

（4）用户反馈。将设有标引的结果输出给用户，供用户选择，接收用户的反馈信息，修正检索过程。

4.4.2 发明人标引的意义

我们信息抽取技术、数据挖掘技术应用在专利信息的发明人分析中，充分发挥聚类分析技术在文本信息方面的优势，以实现自动地抽取发明人异指标引的重要信息，作为检索主题，弥补了当前专利检索方式的不足，降低了专利检索的应用难度，提高了专利检索的查准率。作为检索发明人，弥补了当前专利检索方式对发明人检索的不足，提高了专利检索的查准率。

4.5 本章小结

本章首先描述异指消解的定义，然后根据专利信息的特点，设计了一个基于聚累分析的异指消解模型，主要包括专利数据选择、专利数据获取、专利数据预处理、聚累分析主要K-MEANS算法模型的方法等以便建立新的理论和方法模型。

然后，利用此方法通过通信专利数据进行模型的实验，把准备好的专利数据信息结合人工指导和机器学习训练从中抽取出异指库，并将抽取结果生成基于异指的专利辞典。本章将训练阶段与测试阶段的基于聚类分析的异指消解模型性能数据作对比，可以发现使用了合理的命名实体识别模型规则和经过训练改进之后的模型在召回率和准确率方面都得到了非常明显的提高。该辞典库可用于用于专利检索中的发明人的标引，从而提高专利在发明人检索方面的查准率。

本章的主要创新点是基于聚类分析的异指消解模型的提出。基于语言模型的专利相

关技术检索，是一个庞大的项目，其中有许多工作值得深入研究，本章的研究只是起到了抛砖引玉的作用。

5 总结

5.1 研究工作总结

信息抽取和数据挖掘技术都是一项应用前景十分广阔的技术，它能够使人们免于陷入信息的汪洋之中，从大量冗余的信息中迅速发现对自己有用的信息，还在一定程度上揭示信息与信息之间的关联，产生出用户以前未曾意识到的有用信息，与传统的信息检索技术相比，超越了字面的限制，定位信息更加准确、具有针对性。将信息抽取技术和数据挖掘技术引入到专利信息的分析研究中是一项全新的尝试，实现了专利数据内容一级分析的自动化，丰富了专利信息分析研究方法，同时也为专利分类检索、关联分析提供了有力的支持。本文的研究工作为利用自然语言处理方法进行专利信息分析提供了有益的参考。

5.2 本论文的创新之处

本论文主要是对专利信息进行有效的管理，通过数据挖掘和信息抽取技术应用在专利信息分析中，充分发挥其在处理海量文本信息方面的优势，以期实现对专利权人、发明人等属性重新整合，对专利信息进行深层次的挖掘，以提高专利信息的利用率，帮助企业和国家准确制定专利战略。主要从以下几个方面进行研究：

(1)将海量的WEB网页转化为结构化的数据。

(2)利用数据挖掘技术、信息抽取从申请人属性中抽取同指关系辞典，主要通过关联分析、共指消解、模板填充等信息抽取的关键技术的应用来实现。

(3)通过信息抽取、自然语言对发明人等属性进行异指关系标注，主要使用到命名实体识别、聚类分析、异指关系等将作为研究的核心。

概括起来，本文在以下两方面取得了一定创新：

(1) 提出了基于关联规则的同指消解模型。首先对经过预处理的数据进行关联规则分析，初步确定相关即可能同指的群组范围，然后使用基于规则分析和基于统计学方法的显著因子法提取同指关系的辞典。关联分析法可以将具有基本确定性的同指关系迅速提取出来，而显著规则分析可以将关联分析法不能很好处理的同指现象，利用规则、统计学的方法提取出来，二者互相补充相辅相成。

(2) 提出了基于聚类规则的异指消解模型。使用基于聚类分析技术和基于规则统计

分析方法的显著因子法对存在异指关系的属性值进行标注。在规则消解过程中使用基于聚类分析规则，使该异指消解模型具有良好的可维护性和可扩展性。

5.3 研究限制

虽然本文在对信息抽取技术应用于专利信息分析的研究过程中做了大量工作，完成了数据准备、模型选择、方法设计、模型实现和实验，并得到了一定收获，但研究过程中出现也出现了一些问题，由于本文研究涉及的知识面较宽，而信息抽取技术在我国还是一门前沿学科，当前研究过程中的一些瓶颈问题影响了此项技术的进一步发展。

5.4 下一步的工作

鉴于上文中提到的研究限制，后续的工作将围绕以下几个方面展开：

（1）继续针对专利信息的特点进行研究，改进专利获取工具和方法，以期得到更完整和准确的目标专利数据集。

（2）考虑将信息抽取模型与信息检索系统集成，将系统向专利全文分析推广。

致　谢

本论文是在我的导师朱东华教授的悉心指导下完成的，我衷心的感谢朱老师这两年来的辛勤培养。朱老师渊博的学识，高瞻远瞩的学术视野、敏锐的洞察力、严谨求实的治学态度、勤奋忘我的精神给我树立了榜样。无论从生活、学习还是研究上，朱老师对我们的培养都兢兢业业，始终高标准、严要求，使我受益终身。论文从选题、修改到完善的过程无不渗透着朱老师的心血，在此，我谨以最诚挚的心情向朱老师表示衷心的感谢和崇高的敬意。

衷心感谢管理与经济学院各位老师的培养。从大一到现在近六年的时间里，是你们渊博扎实的知识使我对专业领域从懵懂走向成熟，也为我今后的学习、工作打下了坚实的基础，谢谢你们。

衷心感谢知识发现与数据分析实验室的吕琳博士后、任智军博士、刘玉琴博士和已去南开大学任教的胡望斌博士，谢谢你们平日里对我学习、生活上的照顾以及为本论文的完成提供的帮助和建议，感谢硕士生谢菲，谢谢你们平日对我的诸多帮助。

本论文的研究过程参考了大量中外文献，在此一并向所有给我以知识和启迪的作者们表示感谢。

谨以此文献给所有关心和帮助过我的人！

攻读硕士期间发表的学术论文

1. 黄圆圆，朱东华，任智军，张诚.专利情报分析方法及实证研究.科技管理研究.2006.12

2. 黄圆圆，朱东华，任智军，张诚.对比分析在专利分析中的应用研究.现代图书情报技术. 2006.10

3. 张诚，朱东华，黄圆圆.技术监测方法在美国国防科技情报分析中的应用研究. 图书情报工作.2006.09

参考文献

[1] 张代民．专利信息在企业竞争中的作用．[EB/OL]．[2007-03-16]．http://www.fjinfo.gov.cn/publicat/qbts/021/21.htm.2004

[2] Line Eikvil 原著，陈鸿标 译．网上信息抽取技术纵览[M]．上海：复旦大学出版社，2003

[3] 廖希明，苑成生．专利信息的利用及其 Internet 获取[J] ．科技情报开发与经济，2003，(10)：43-44

[4] 黄红华，俞勇．从半结构化中抽取信息的归纳规则方法[J]．上海交通大学学报，2003，37(3)：426-433

[5] 佚名．专利信息在技术创新中的应用．[EB/OL]．[2007-03-16]．http://www.66vision.com/news/807.htm.2003

[6] 任小燕，陈永锋，官东．浅谈专利信息管理的若干问题[J]．荆门职业技术学院学报．2001，（11）：46-48

[7] 李保利，陈玉忠，俞士汶．信息抽取研究综述[J]．计算机工程与应用，2003，（10）：63-66

[8] Douglas Appelt, Jerry Hobbs, John Bear, David Israel, Mabry Tyson. FASTUS: A finite-state processor for information extraction from real-world text[C]. In Proc. 13th Int'l Joint Conf. Articial Intelligence, 1993: 1172-1178

[9] Steven Soderland. Learning information extraction rules for semi-structured and free text[J]. Machine Learning. 1999, (1): 1-44

[10] 郑家恒，王兴义，李飞．信息抽取模式自动生成方法的研究[J]．中文信息学报，2004，18（1）：48-54

[11] Wilks Yorick. Information Extraction as a Core Language Technology[J]. International Summer School. 1997, (10): 123-139

[12] Soderland G. Learning text analysis rules for domain-specific natural language processing [D]. Amherst: University of Massachusetts, 1997

[13] 刘开瑛，郭炳炎．自然语言处理[M]．北京：科学出版社，1991

[14] Paik, Woojin. Chronological information Extraction System (CHESS)[D]. Syracuse University, 2000

[15] Gerald F. DeJong. An overview of the FRUMP system[J]. Strategies for Natural Language Processing, 1982: 149-176

[16] Lim S S, Jung S W, Kwon H C. Improving Patent Retrieval System using Ontology[C].The 30th Annual Conference of the IEEE Industrial Electronics Society. Bussan, Korea, 2004, 3：2646-2649

[17] Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, Ralph Weischedel, and the Annotation Group. Algorithms that learn to extract information BBN: Description of the SIFT system as used for MUC-7[C]. Proceedings of 7th Message Understanding Conference, 1998: 1147-1156

[18] Zhang N R. Hidden Markov Models for Information Extraction[R]. Technical Report. Stanford Natural Language Processing Group, 2001: 239-244

[19] ZhaoHui Tang，Jamie MacLennan 著．数据挖掘原理与应用/(美) [M]．北京：清华大学出版社，2007

[20] Hobbs J, Appelt, D.Bear J et al. FASTUS：A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text[C]. In Roche, Schabes eds.Finite State Devices for Natural Language Processing, MIT Press, Cambridge MA, 1996: 246-256

[21] Douglas E. Appelt, David J. Israel. Introduction to Information Extraction Technology[C]. A Tutorial Prepared for IJCAI-99, 1999: 456-462

[22] Wilks Yorick. Information Extraction as a Core Language Technology[D]. International Summer School, 1997

[23] Paik, Woojin. CHronological information Extraction SyStem(CHESS)[D].Syracuse University, 2000: 24-28

[24] Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, Ralph Weischedel, and the Annotation Group. Algorithms that learn to extract information BBN: Description of the SIFT system as used for MUC-7[C]. Proceedings of 7th Message Understanding Conference, 1998：123-135

[25] 蔡自兴，徐光祐．人工智能及其应用（第二版）[M]．北京：清华大学出版社，1996

[26] Han Jiawei, Kamber M. Data Mining-concepts and tech-niques[M]. San Francisco, USA:

Morgan Kaufmann Publishers, 2001

[27] 姚天顺．自然语言理解[M]．北京：清华大学出版社，1995

[28] 李向阳，苗壮，肖江．无结构文本信息抽取综述[J]．军事通信技术，2003，25（2）：32-39

[29] 廖乐健，曹元大，李新颖．基于 Ontology 的信息抽取[J]．计算机工程与应用，2002，（23）：110-113

[30] Cowie, J. W. Lehnert. Information Extraction[J], Special NLP Issue of the Comm.ACM, 1996：39(1)：80-91

[31] Palmer, David Donald. Modeling uncertainty for information extraction from speech data[D]. University of Washington, 2001: 2-3

[32] Jun-Tae Kim, Dan I. Moldovan. Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction[J]. IEEE Transactions on Knowledge and Data Engineering, 1995,(2)：713-724

[33] 郭志红．基于 WEB 资源的信息抽取技术[J]．情报科学，2002，20(12)：1282-1284

[34] 赵黎明，李海霞，韩宇．基于数据挖掘的专利引文研究与知识发现[J]．预测，2002，21(6)：6-9

[35] GuoDong Zhou, Jian Su. Named Entity Recognition using an HMM-based Chunk Tagger. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics[C], 2002: 473-480

[36] Aone, Chinatsu & Scot W. Bennet. Applying machine learning to anaphora resolution. Connectionist, statistical and symbolic approaches to learning for Natural Language Processing[J]. Berlin: Springer, 1996: 302-314

[37] 袁玉波．数据挖掘与最优化技术及其应用[M]．北京：科学出版社，2007

[38] (美) Joy Mundy，Warren Thornthwaite，Ralph Kimball 著．数据仓库工具箱:面向 SQL Server 2005 和 Microsoft 商业智能工具集[M]．北京：清华大学出版社，2007

[39] 章兢，张小刚等编著．数据挖掘算法及其工程应用[M]．北京：机械工业出版社 2006

[40] Steven Soderland. Learning information extraction rules for semi-structured and free text[J]. Machine Learning, 1999: 1-44

[41] Jun-Tae Kim and Dan I. Moldovan. Acquisition of Linguistic Patterns for Knowledge-Based

Information Extraction[J]. IEEE Transactions on Knowledge and Data Engineering, 1995,7(5): 713

[42] 王德兴，胡学钢，刘晓平，王浩．基于概念格和 Apriori 的关联规则挖掘算法分析[J]．合肥工业大学学报(自然科学版)，2006，(10)：67-74

[43] Thabta, F. A. Cowling, P. I. A greedy classification algorithm based on association rule[J]. Applied Soft Computing, 2007, 7（3）：21-30

[44] Kuo, J. -J. Hsin-Hsi Chen. Cross-document event clustering using knowledge mining from co-reference chains[J]. Information Processing & Management, 2007, 43（2）：45-56

[45] 张威，周昌乐．汉语语篇理解中元指代消解初步[J]．软件学报，2003，13（4）：23-30

[46] 孔祥勇．多语种同指消解系统的研究和实现[D]．上海交通大学，2003

[47] (美) Edward Melomed [等] 著．SQL Server 2005 Analysis Services 标准指南：中文版[M]．北京：电子工业出版社，2008