

硕士学位论文

领域自适应的中文实体关系抽取研究

**RESEARCH ON DOMAIN ADAPTIVE
CHINESE ENTITY RELATION
EXTRACTION**

王莉峰

哈尔滨工业大学

2011 年 6 月

国内图书分类号：TP391.2
国际图书分类号：681.37

学校代码：10213
密级：公开

工学硕士学位论文

领域自适应的中文实体关系抽取研究

硕 士 研 究 生：王莉峰

导 师：秦兵教授

申 请 学 位：工学硕士

学 科：计算机科学与技术

所 在 单 位：计算机科学与技术学院

答 辩 日 期：2011 年 6 月

授予学位单位：哈尔滨工业大学

Classified Index: TP391.2

U.D.C: 681.37

Dissertation for the Master Degree in Engineering

RESEARCH ON DOMAIN ADAPTIVE CHINESE ENTITY RELATION EXTRACTION

Candidate:	Wang Lifeng
Supervisor:	Prof.Qin Bing
Academic Degree Applied for:	Master of Engineering
Speciality:	Computer Science and Technology
Affiliation:	School of Computer Science and Technology
Date of Defence:	June, 2011
Degree-Conferring-Institution:	Harbin Institute of Technology

摘 要

随着计算机的快速普及，互联网的迅猛发展，各式各样的信息呈爆炸式增加，如何从海量数据中准确、快速地获取用户真正需要的信息成为人们关注的话题。信息抽取的主要目的是将非结构化的自然语言文本转化成半结构化或结构化的数据，方便人们准确、快速地获取关键信息。关系抽取作为信息抽取的子任务和关键技术之一，已经逐渐发展成为众多自然语言处理任务的重要支撑技术。

传统的关系抽取方法需要预先定义关系类型，依赖于大量人工标注的训练语料库，难以满足互联网海量信息处理的需求。本文提出了一种新的关系抽取研究框架，探索最大程度地避免人工参与，且具有较强领域自适应能力的关系抽取解决方案，提高关系抽取的自动化程度，增强可移植性。

首先，通过分析关系实例上下文语言现象发现，绝大多数产生语义关系的实体对均可以由其上下文中的一般动词和一般名词触发描述（统称为特征词），由此，本文提出基于特征词聚类的方法，在一定规模的未标注语料库上实现关系类型的自动发现，实验中达到了与人工预定义关系类型相当的效果；其次，针对大量待处理的关系类型，本文提出基于 Web Mining 的关系种子集自动抽取方法，充分利用搜索引擎收集和处理大规模真实数据的能力和优势，抽取具有代表性的实体关系核心网，经过在选取的 9 种关系类型上进行实验，平均准确率达到了 90.91%；再次，根据中文语言学特点，本文定义了启发式上下文模式及其泛化策略，引入 Bootstrapping 方法，以实体关系核心网作为输入，在未标注语料库上迭代地挖掘关系描述模式，并抽取关系元组，通过对采样的关系元组进行人工评价，平均准确率达到了 88.24%，满足了实用系统的需求。

最后，本文设计并实现了领域自适应的关系抽取平台 XInfo，在该平台上，研究人员可以专注于算法的改进和研究，快速进行实验，为自然语言处理相关领域研究和应用提供支持。另外，本文以人物社会关系抽取作为应用任务，开发了一套人物社会关系在线演示系统，以直观、清晰的方式展示关系抽取效果。

关键词 关系抽取；领域自适应；关系类型发现；关系种子抽取；关系描述模式挖掘

Abstract

With the rapid popularization of computers, and the Internet's rapid development, the amount of information is becoming more and more. So, how to quickly and accurately obtain necessary information from the massive data becomes a topic of concern. The main purpose of information extraction is to transform unstructured natural language text into semi-structured or structured data, easy for people to obtain key information quickly and accurately. Relation extraction as one of the subtask and key technology of information extraction, has gradually become an important supporting technique for many natural language processing tasks.

Traditional relation extraction methods required pre-defined relation types, and rely on large amount of manually annotated training corpora. So they are difficult to meet the needs of the Internet massive information processing. We propose a new relation extraction research framework to explore the maximum to avoid human intervention, and has a strong domain adaptive capacity, in order to improve the automaticity and enhance protability of relation extraction.

First, by analyzing the linguistic phenomenon of the relation instances context, we found the vast majority of the entity pairs which generating some semantic relations could be triggered or described by the general verbs and nouns (referred to as feature words), so this paper proposes the feature words clustering method, which can discover relation types from a certain amount of unlabeled corpus automatically, and can be compared with predefined result with the artificial. Second, for the large number of relation types to be processed, this paper proposes the Web Mining based relation seed extraction method, which can make full use of search engine's large-scale data collection and processing capabilities and advantages, to extract the representative relation core network. The method gets an average precision of 90.91% on selected nine relation types. Next, according to Chinese linguistic characteristics, this paper defines the general context pattern and its generalization, then introduces the bootstrapping method. The method takes the relation core network as input, then iteratively generates the relation description patterns and extracts relation tuples. Through manual evaluation on the sampling relation tuples, the average precision achieves 88.24%, meets the practical needs.

Finally, a domain adaptive relation extraction platform named XInfo is designed and implemented, on the platform, researchers can focus on algorithm improvement and research, then make rapid experiment. Also, XInfo can provide support for natural language processing research and applications. In addition, this paper takes the social relations between people as an application task, and develops an online demo system to show relation extraction results in an intuitive and clear way.

Keywords Relation Extraction; Domain Adaptive; Relation Type Discovery;
Relation Seed Extraction; Relation Description Pattern Mining

目录

摘 要	I
Abstract	II
第 1 章 绪 论	1
1.1 课题背景	1
1.2 研究目的和意义	1
1.3 关系抽取的历史	2
1.4 关系抽取的研究现状	5
1.4.1 先确定关系类型体系的方法	5
1.4.2 后确定关系类型体系的方法	10
1.5 问题的提出	12
1.5.1 关系类型体系构建困难	12
1.5.2 关系抽取标注语料库匮乏	12
1.5.3 领域自适应的关系抽取研究滞后	13
1.6 本文的主要研究内容	13
第 2 章 基于特征词聚类的关系类型发现	15
2.1 引言	15
2.2 算法流程	15
2.3 算法设计	16
2.3.1 语料库获取	17
2.3.2 种子实体抽取	18
2.3.3 特征词抽取	18
2.3.4 特征词聚类	23
2.4 实验结果与分析	26
2.4.1 实验数据	26
2.4.2 评价标准	27
2.4.3 结果与分析	27
2.5 本章小结	31
第 3 章 基于Web Mining的关系种子集抽取	33
3.1 引言	33
3.2 算法流程	33
3.3 算法设计	34
3.3.1 查询构造	34
3.3.2 查询扩展	35
3.3.3 网页检索	36
3.3.4 答案抽取	37
3.4 实验结果与分析	40
3.4.1 实验数据	40
3.4.2 评价标准	40
3.4.3 结果与分析	40

3.5 本章小结	42
第 4 章 基于Bootstrapping的关系描述模式挖掘	43
4.1 引言	43
4.2 算法流程	43
4.3 算法设计	45
4.3.1 关系实例抽取	45
4.3.2 上下文模式生成	45
4.3.3 模式泛化与过滤	47
4.3.4 元组抽取与评价	49
4.4 实验结果与分析	50
4.4.1 实验数据	50
4.4.2 评价标准	52
4.4.3 结果与分析	52
4.5 本章小结	56
第 5 章 领域自适应的关系抽取平台设计与实现	57
5.1 引言	57
5.2 关系抽取平台XInfo.....	57
5.3 关系抽取演示系统	58
5.4 本章小结	61
结 论	62
参考文献	64
攻读学位期间发表的学术论文	69
哈尔滨工业大学学位论文原创性声明及使用授权说明	70
致 谢	71

第1章 绪 论

1.1 课题背景

随着计算机的快速普及以及互联网的迅猛发展, 各式各样的信息呈爆炸式增加和病毒式传播, 例如商业信息、新闻资讯、百科知识、博客、论坛、微博等等, 这些信息以电子文本的形式在网络上迅速传播并呈现在人们面前, 但令人尴尬的是, 人们根本无法从互联网海量信息中准确、快速地获取真正需要的信息。

传统的搜索引擎(如Google¹、百度²)通过对文本进行浅层分析, 对网页结构进行链接分析计算重要性, 返回大量包含检索词的相关网页集合, 成为目前人们查找信息获取知识的重要渠道和工具, 一定程度上满足了用户需求, 但是人们仍然需要通过浏览大量的网页, 获取有用的知识。为了改变这种尴尬的局面, 迫切需要对文本进行深层理解与挖掘, 深入语义层面抽取文本中包含的关键信息, 为用户提供精准化的检索服务。

信息抽取(Information Extraction, IE)相关的研究正是在这种背景下产生的, 其主要目的就是将非结构化的自然语言文本转化成半结构化或结构化的数据, 并以数据库形式存储, 一方面可以用于对文本的快速阅读和理解, 帮助人们更方便的获取知识, 另一方面可以用于深入地挖掘分析, 对本体构建、垂直搜索、自动问答等自然语言处理相关领域起着非常重要的作用^[1]。其中, 关系抽取(Relation Extraction, RE)作为信息抽取的子任务和关键技术之一, 近年来受到了越来越多研究学者的关注, 逐渐成为了研究热点。

1.2 研究目的和意义

信息抽取系统的主要功能是从自然语言表示的文本中抽取用户关心的事实信息(Factual Information), 即实体(Entity), 这个过程我们往往称之为实体识别^[1]。例如, 从新闻报道中抽取恐怖事件的时间、地点、作案者、受害者、袭击目标、武器等。我们通常将这些事实信息分类为人名、机构名、地点、时间、日期(Date)、数词、专有名词等。

在大多数的实际应用需求中, 不仅要求识别出文本中的实体, 还需要抽取这些实体之间所表达的语义关系, 我们称之为实体关系抽取, 或者简称为关系抽取。传统的关系抽取任务中往往需要领域专家预先定义好需要处理的关系类型体系, 比如, 部分整体关系(PART-WHOLE)、人与社会关系(PER-

¹ <http://www.google.com/>

² <http://www.baidu.com/>

SOC)、地理位置关系 (PHYS) 和雇佣关系 (EMP-ORG) 等。比如, 文本中出现“... 微软公司董事长比尔·盖茨...”, 其中“微软公司”和“比尔·盖茨”分别为机构名 (Organization) 和人名 (Person) 类型的实体, 它们构成了一种雇佣关系 (EMP-ORG), 即“比尔·盖茨”受雇于“微软公司”。

通过以上介绍可知, 如果说信息抽取的主要任务是自动将自然语言表示的文本信息转化为结构化的表格数据, 实体识别确定了表格中各个元素的话, 关系抽取则是确定这些元素在表格中的相对位置的过程^[2]。可见, 关系抽取的主要目的就是在实体识别基础上, 从非结构化的自然语言文本中抽取出实体之间的语义关系, 并将之采用结构化方式存储, 供数据查询或进一步的分析利用, 是信息抽取系统重要的基础性环节。

传统的关系抽取方法主要基于预先定义好的特定关系类型, 针对不同的抽取对象, 人工标注训练语料库, 采用不同的机器学习方法训练分类模型, 以用于新关系实例的识别和关系元组抽取。在待处理的文本表达形式相对固定的情况下, 这种方法可以有效的的工作, 并且在具备足够充足的标注语料库时, 可以达到不错的抽取性能。但是, 一方面预先定义关系类型限制了可处理的关系种类, 对于新的应用领域, 需要相关的领域专家定义关系类型体系, 费时费力, 应用难免受限; 另一方面, 该过程需要大量的人工参与, 如选取领域资源, 定义标注规范, 构建标注语料库等, 可扩展性较差, 难以适应互联网异构、海量信息处理的要求。

另外, 有些研究学者通过书写知识规则或者构建少量关系种子, 然后基于 Web 或大规模语料库半自动或全自动迭代地抽取关系描述模式, 识别关系实体对, 这类方法不需要人工构建训练语料库, 大幅度降低了投入成本, 但是, 一方面对于知识规则的书写和关系种子的质量评估较困难, 另一方面当系统被用于新领域的关系抽取时, 仍然需要根据领域文本特点重新编写规则或构建高质量关系种子。

基于此, 本课题分析中文语言学现象特点, 提出了一种新的关系抽取研究框架, 探索最大程度避免人工参与, 且具有较强领域自适应能力的关系抽取解决方案, 主要解决关系类型发现、关系种子集抽取、关系描述模式挖掘和关系元组抽取等关键问题, 利用丰富的未标注语料库挖掘统计信息和启发式规则, 大大降低关系抽取系统对人工标注语料库的依赖程度, 提高关系抽取的自动化程度, 扩大适用范围, 具有重要的理论研究和实用价值。

1.3 关系抽取的历史

信息抽取研究最早开始于 20 世纪 60 年代中期, 以两个著名的自然语言处理相关研究项目为代表。第一个是由美国纽约大学开展的 Linguistic String

项目^[3]，项目从 20 世纪 60 年代中期开始，一直持续到 20 世纪 80 年代，项目旨在构建一套大规模英文计算语法。

另一个是由耶鲁大学于 20 世纪 70 年代开展的一个故事理解相关的长期性研究项目，最终他们基于期望驱动与数据驱动相结合的抽取方法设计并实现了基于故事脚本理论的信息抽取系统 FRUMP^[4]，系统可以实现自动从新闻报道中抽取如地震、工人罢工等许多场景信息，他们的方法被后来大量的研究人员借鉴和使用。

随着 20 世纪 80 年代由美国国防高级研究计划委员会（the Defense Advanced Research Projects Agency, DARPA）资助的消息理解会议（Message Understanding Conference, MUC）的召开，信息抽取的研究得到了蓬勃的发展。从 1987 年到 1998 年，MUC 会议共举办了七届，会议定义了几大类重要的任务，包括命名实体识别（Named Entity Recognition, NER），多语种实体识别（Multi-lingual Entity, MET），模板元素（Template Element, TE），模板关系（Template Relation, TR），共指消解（Co-reference, CR），情节模板（Scenario Template, ST），这些任务为后续的相关评测会议提供了重要参考价值，逐渐发展成为自然语言处理和文本挖掘领域重要的组成部分，并且其评价体系被沿用至今，成为事实上的标准。

关系抽取的研究最早在 1998 年 MUC-7^[5]会议上引入，最初定义为模板关系 TR 任务，主要目的是确定实体之间与特定领域无关的语义关系^[6]。

随着 MUC 会议的停办，美国国家标准技术研究院（National Institute of Standards and Technology, NIST）组织了自动内容抽取（Automatic Content Extraction, ACE）评测会议³，继续进行信息抽取相关的评测任务。迄今为止，ACE 评测已经成功举办了九次，最近的一次会议于 2008 年 5 月举行。ACE 会议旨在从新闻文本中自动抽取所包含的实体、关系、事件等关键信息，该会议主要定义了五大类评测任务，包括实体检测与识别（Entity Detection and Recognition, EDR）、属性检测与识别（Value Detection and Recognition, VAL）、时间检测与识别（Time Detection and Recognition, TERN）、关系检测与识别（Relation Detection and Recognition, RDR）和事件检测与识别（Event Detection and Recognition, VDR），这五项任务都制定了较为细致的类别和模板^[7]。另外，每次 ACE 评测的任务都略有不同，如 2008 年评测只有实体检测与识别 EDR 和关系检测与识别 RDR 两项任务，不仅处理单文档，还定义了多文档信息抽取，ACE 2008 RDR 任务共定义了 7 大类、18 个子类的实体关系类型体系^[8]。ACE 评测提供的语料不仅有英文，还包括中文、

³ <http://www.itl.nist.gov/iad/mig/tests/ace/>

阿拉伯文和西班牙文，其中，中文标注语料库主要来自广播新闻（40%），新闻专线（40%）和网络对话（20%）。从 2009 年开始，ACE 被归入文本分析会议（Text Analysis Conference, TAC）⁴，是 Knowledge Population 任务的主要组成部分。

另一个著名的相关评测会议是 SemEval (Semantic Evaluations)⁵，其前身是已经成功举办三次的 Senseval 评测，主要评测与语义相关的多方面任务，吸引了全球范围内很多大学和研究机构参与，具有非常广泛的影响力。在 SemEval-2007 评测中，定义了 Task 04: Classification of Semantic Relations between Nominals 任务^[9]，该任务仅提供少量英文实验语料库，定义了 7 种常见的语义关系，主要负责评测名词和名词短语之间的关系抽取。

中文信息抽取相关的研究起步较晚，近年来的一些研究工作主要集中在中文命名实体识别 NER 任务上，在设计并实现完善的中文信息抽取系统方面仍然处于探索阶段。对于中文关系抽取的研究，哈尔滨工业大学、苏州大学、微软亚洲研究院、北京大学、中科院等研究机构相继开展了一系列相关研究工作，并取得了一定的研究成果。如苏州大学自然语言处理实验室提出了基于树核的实体语义关系抽取方法，充分分析利用句法树，以更加准确的描述实体之间的关系，取得了一系列具有影响力的研究成果。哈尔滨工业大学社会计算与信息检索研究中心开展了题为《基于实体关系的文本内容挖掘与集成技术平台》的 863 课题，课题致力于建立以实体为核心的关系文本内容挖掘与集成技术平台，旨在通过相同类型实体之间的共指消解（即等价关系），和不同类型实体之间的关系抽取（即非等价关系），构建以实体为节点，实体关系为边的拓扑图，为文本信息的结构化、可视化提供有力保证，同时该课题以事件作为触发，根据事件抽取结果动态更新、维护已经建立的实体关系网，并通过事件链进行重要信息的选择，完成篇章文本的信息集成与汇总，自动生成文本文摘。他们针对 ACE 中文语料库和音乐领域自制语料库进行了有益的尝试和探索，最终，搭建了文本挖掘系统（Text Mining System, TMS）⁶。微软亚洲研究院网络搜索与挖掘组提出了对象级别（Object-Level）的概念，设计并开发了一款新型的社会化搜索引擎：人立方关系搜索⁷，它从十亿级的中文网页库中自动抽取人名、地名、机构名以及中文短语等关键信息，并计算它们之间存在关系的可能性，此外，它还可以自动找出产生关系的两个人名之间最可能的关系描述词，并索引关系来源文本等信息。

⁴ <http://www.nist.gov/tac/>

⁵ <http://www.senseval.org/>

⁶ <http://ir.hit.edu.cn/demo/tms/index.html>

⁷ <http://renlifang.msra.cn/>

1.4 关系抽取的研究现状

现有对关系抽取的研究可以按照确定关系类型的先后分为两大类：先确定关系类型后识别实体对和先识别实体对后确定关系类型。

1.4.1 先确定关系类型体系的方法

先确定关系类型体系的方法，是一种自顶向下的信息抽取策略，需要预先定义好关系类型体系，然后根据预定义的关系类型确定关系描述模式，利用关系描述模式进行实体对识别。产生关系描述模式的方法主要有以下三种：

(1) 基于知识工程的方法

该方法主要依靠语言学家手工编写规则，构建大规模知识库^[10]，用于处理特定领域的信息抽取问题。这种方法要求编写规则的知识工程师对相关领域有深入的了解，投入成本较大，且耗时耗力，可移植性差。另外，关系的描述形式具有多样性，难以单纯依靠人工定义抽取规则完成，所以逐渐产生了一系列自动获取模式的解决方法。

(2) 有指导的学习方法

基于有指导的学习方法主要分为两大类：基于特征向量的机器学习方法和基于核函数的机器学习方法。这种方法将关系抽取问题看作一个分类问题，首先需要人工标注大规模训练语料库，然后在已标注好的语料库基础上进行特征抽取和选择，通过利用不同的机器学习算法训练学习分类模型，用于抽取新的实体对。其中，任何对该知识领域比较熟悉的人都可以根据事先确定的规范标注语料库。

Kambhatla (2004)^[11]使用了词、实体类型、提及层、重叠关系、依存关系和句法树等信息作为特征，采用最大熵模型 (Maximum Entropy Model, MaxEnt)，在 ACE 2004 语料库共 24 个子类型上进行实验，取得了 F 值为 52.8% 的抽取结果。

Zhou (2005) 等^[12]在 Kambhatla 的基础上，使用支持向量机 (Support Vector Machine, SVM) 作为分类算法，提取了更多的特征，进一步提高了关系抽取的性能。

车万翔 (2005) 等^[2]通过构造不同窗口的特征向量，分别采用 Winnow 和 SVM 算法，在 ACE 2004 中文语料库的 7 大类关系上，取得了 F 值为 73.27% 的好成绩。

董静 (2007) 等^[13]分析语料库特点，提出将中文实体关系划分为：包含实体关系与非包含实体关系，针对同一种句法特征识别性能的明显差异，对这

两种关系采用了不同的句法特征集，并提出了一些适合各自特点的新的句法特征。并且，在条件随机场（Conditional Random Fields, CRFs）模型框架下，以 ACE 2007 中文语料库作为实验数据，实验结果表明其划分方法和新特征有效的提高了汉语实体关系任务的抽取性能。

Chan and Roth (2010)^[14]引入了丰富的背景知识，如本体、共指消解、Wikipedia、词聚类等，并使用整数线性规划（Integer Linear Programming, ILP）框架灵活地将特征向量与背景知识融合，大幅度提高了关系抽取性能。

Sun (2011) 等^[15]提出特征稀疏是影响有指导学习方法抽取性能的重要因素，而关系实例往往可以通过实体对上下文中的一些关键词表示，基于此，他们提出对关键词特征进行聚类，并统计选取较优的关键词聚类结果子集，用于训练关系抽取分类模型。

基于特征向量的方法需要启发式的选取特征，并构造特征向量形式的训练数据，综上，研究者们分别从不同的角度进行特征的抽取，主要包括实体对组合特征、实体对位置关系、上下文词法特征、依存路径、句法树等。特征向量的方法尽管速度很快，也很有效，但是，由于实体间语义关系表达的复杂性和多样性，要进一步提高关系抽取的性能较困难，因为很难再找出适合、关系抽取的新的有效特征。所以，一些研究学者开始使用核函数的方法。

对于基于核函数（Kernel）的方法，它最早是在 SVM 方法中被引入，后来发现多种学习方法可以使用 Kernel 的形式表示。在自然语言处理领域应用基于 Kernel 的学习算法时^{[16][17]}，不需要抽取复杂的特征向量，可以使用字符串或者语法结构树等能够直接表示关系实例的形式作为处理对象，算法只需要计算任意两个处理对象之间的相似度函数，即 Kernel。核方法在用于抽取实体关系时不再局限于有限的特征，可以扩展出大量甚至无限的特征，使得准确率和召回率均有较大程度的提高。

Haussler (1999) 等^[16]最早提出了基于字符串和树等离散结构的核函数计算方法。随后，Lodhi (2002) 等^[17]研究了如何利用字符串核函数解决文本分类问题，通过实验验证了核方法在文本分类应用中的有效性。

Zelenko (2003) 等^[18]提出将关系实例表示成连接实体对的最小公共子树，其方法主要思想是通过计算两棵子树之间的核函数确定实体对之间的是否存在某种语义关系，通过在两个简单的关系抽取任务中尝试实验，方法取得了较好的抽取效果。

Culotta and Sorensen (2004)^[19]在 Zelenko 等人^[18]工作基础上提出了依存句法树核函数的改进方法，其方法规定不同树上当前比较的两个节点必须在相同层级，且从根节点到当前节点的依存路径必须一致。这种非常严格的约束牺

牲了召回率，而提高了准确率，在 ACE RDR 2003 语料库上召回率小于 35%，准确率大于 67%。尽管作者提出并尝试了一些提高召回率的方法，但实验效果并不明显。

Bunescu and Mooney (2005)^[20]提出了基于最短依存句法路径的核函数，他们方法通过简单地统计两个关系实例依存句法路径上相同词语的个数，计算核函数，同样遇到了较高准确率、较低召回率的问题。

Che (2005) 等^[21]提出了一种改进的编辑距离核函数，分别结合 Voted Perceptron 和 SVM 进行实验，在 person-affiliation 关系上达到了 F 值为 91% 的好成绩。

近年来，研究人员还提出了基于字符串、短语结构树和依存句法树等结构的卷积核函数的关系抽取方法。所谓卷积核函数，就是通过计算两个离散结构之间的相同子结构的个数计算它们相似度^[23]。

Zhang (2006) 等^[23]提出利用卷积树核函数计算包含实体对的句法树之间的相似度，识别实体对语义关系。

Qian (2007) 等^[23]在卷积树核函数上融入了实体特征，如实体类型、子类型和提及层等，在 ACE2004 语料库上验证了方法的有效性。

基于以上研究，学者们开始考虑如何将传统的基于特征向量的机器学习方法与基于核函数的方法结合起来。于是，混合核函数 (Composite Kernel) 的方法被越来越多的使用到了研究当中。

Zhao and Grishman (2005)^[24]为了将特征空间覆盖到不同的语言学特点，为不同层次的信息建立了不同的核函数，这样就可以有针对性的综合使用各层次上有用的信息，在 ACE RDR 2004 语料上 F 值达到了 70.4%。

Zhang (2006) 等^[26]为了挖掘关系实体对的多样化结构信息，获取最短路径封闭树 (Shortest Path enclosed Tree, SPT)，而提出了基于句法树的卷积核 (Convolution Tree Kernel, CTK) 的方法，通过在 ACE RDR 2004 语料库的 7 种关系大类上实验，取得了 F 值为 67.7% 的结果。但是，这种方法存在一个严重的问题，即没有有效利用最短路径封闭树之外的上下文信息，而这些信息可能对于表征实体对之间的关系起着重要的作用。所以，如何在卷积核函数中有效引入合适的上下文信息成为了后续研究工作的重点。

Zhou (2007) 等^[27]指出基于句法树的卷积核函数，以及最短路径封闭树都是上下文无关的，没有有效利用可能有贡献的上下文信息，所以，他们在前人工作基础上，提出在最短路径封闭树基础上进行动态扩展方法，将一些有用的上下文信息和句法树路径信息等。这样以来，之前的上下文无关的卷积核函数就变成了上下文有关的卷积树核函数。同样，在 ACE RDR 2004 语料库的 7

大类关系上展开实验，最终取得了 F 值为 73.2% 的好成绩。但是，他们方法在引入有限上下文信息的同时，也不可避免地引入了一些噪声信息，如何尽可能地剔除噪声信息成为该方法的亟需解决的问题。

Qian (2008) 等^[28]借助成分依存理论产生表示实体对关系结构化信息的动态关系树，并同实体语义信息有机结合起来，进一步提高了关系抽取的性能。在 ACE RDR 2004 语料库上进行实验，方法取得了 F 值为 77.1% 的结果，达到了目前最好的中文关系抽取实验结果。

虽然在实体对上下文信息表示方面，相对于特征向量方法显式的表示信息，基于核函数的方法将更多的信息隐藏在核函数中，有较大的优势，但也存在严重的缺点，首先核函数是不利于我们灵活地控制实验过程中所需要使用的特征，这样可能导致引入一定的噪声信息，影响算法的最终性能，另外，由于核函数计算复杂度高，训练和预测速度慢，所以，不适用于处理大规模语料库。并且，由于核函数的确定没有一定的规律可循，不同的核函数在处理不同关系类型的句子时优劣各不相同，不存在单一的核函数在识别关系时具有绝对的优势，即核函数均存在一定的适用性，这样的话，如何预先定义或选取合适的核函数比较困难。

对于有指导的学习方法，无论是基于特征向量的学习算法，还是基于核函数的学习算法，主要依赖于实体对上下文中的各种词法、句法、语义等信息，或者背景知识，提高算法的性能。所以，如何挖掘和有效使用更多对关系抽取更加有用的词法、句法、语义等特征，即特征提取和特征选择两个关键过程，已经成为基于有指导的关系抽取方法的研究重点。

(3) 半指导的学习方法

又称弱指导的学习方法，主要是基于种子的 Bootstrapping 方法，该方法首先需要根据预定义好的关系类型，人工构造对应的关系实例作为种子；然后，通过模式学习方法，迭代地生成关系描述模式集。

Brin (1998) 等^[29]最早提出了基于 Bootstrapping 的半指导的关系抽取解决方案，并构建了 DIPRE (Dual Iterative Pattern Expansion) 系统，以实现从大规模 HTML 文档中抽取出结构化的数据。使用该系统人们只需要针对待处理的实体关系（如<人，出生时间>等）给定少量的初始关系种子（如<毛泽东，1893 年>等）作为输入，其方法可以自动获取与该实体关系对应的五元组描述模式和丰富的关系实例。

Agichtein (2000) 等^[30]在 Brin 的工作基础上，实现了 Snowball 系统，该系统定义了带权重的五元组关系描述模式表示方式，使用命名实体识别技术对句子进行标注，仅抽取命名实体之间的关系，给出了完善的模式和元组的评价

筛选标准，在规模为 300,000 篇的新闻语料库上验证了方法的有效性。

Etzioni (2005) 等^[31]搭建了信息抽取系统 KnowItAll，系统可以自动地从 Web 上抽取与特定领域无关的事实类信息，其输入是如“科学家”、“城市”、“电影”等之类的类别概念信息，输出是特定类别下的实例集合，该系统具有较高的抽取准确率，但是召回率较低。

Pasca (2006) 等^[32]借助语义词典对关系实例的上下文模式进行泛化，避免使用句法分析、命名实体识别等深层文本处理技术，旨在从大规模的 Web 文档中抽取出“<人，出生时间>”的关系实例，其方法在保证准确率的同时，大幅度提高了召回率。

Rosenfeld (2006) 等^[33]实现了关系抽取系统 URES (an Unsupervised Web Relation Extraction System)，该系统从关系种子集出发，在序列模式基础上，利用最佳匹配的动态规划算法对模式进一步泛化，得到软模式 (Soft Pattern)，最后使用 Soft Pattern 匹配识别新的关系实例，系统在 5 种关系类型上进行实验，最终获得了 90% 左右的准确率。但是，模式的泛化程度反映在最佳匹配中单元匹配中的分值和阈值的选择上，而系统并未给出详细的解释说明和对比实验。

Feldman 和 Rosenfeld (2006) ^[34]在 URES 系统基础上引入命名实体识别处理结果，并且基于已抽取的泛化的序列模式对系统输出进行细分类，进一步提高了 URES 系统的性能。

李维刚 (2006) 等^[35]提出使用种子集和关键词作为输入，一定程度上解决了循环依赖问题，并在此基础上提出了一种改进的模式获取和迭代策略，利用网络挖掘技术从 Web 上抽取关系元组，平均准确率达到了 98.42% 的好成绩，可以较好的满足信息抽取实际应用需求。

邓肇 (2007) 等^[36]研究了汉语实体关系抽取技术，他们在模式匹配基础上引入了基于语义词典《同义词词林》的词汇语义匹配技术，并详细比较了一般的模式匹配技术和词汇语义模式匹配技术在汉语关系抽取任务中的性能，实验结果证明，由于汉语在构词、语法、语义和时态等多方面与英文存在较大区别，以至于在英文上抽取效果较好的一般模式匹配方法，在处理中文文本时效果相对较差。

Banko (2007) 等^[37]根据是否依赖预先定义好的关系类型，将关系抽取定义为传统的关系抽取 (Traditional Relation Extraction) 和开放式的关系抽取 (Open Relation Extractio)，首次提出了 Open IE (Open Information Extraction) 的概念，旨在实现领域可移植的关系抽取，或者说开放域的关系抽取，他们对二者进行了详细深入的比较分析，在关系类型已知且具备标注好

的训练语料库时，针对特定关系类型的传统的关系抽取学习方法能够取得更好的效果，但是，在关系类型未知或者标注语料库不充分甚至缺少的情况下，可以采用开放式的关系抽取方法，他们最终还使用 Self Supervised 方法构建了第一个 Open IE 系统：TEXTRUNNER。

Banko（2008）等^[38]在之前的工作基础上，将 Traditional Information Extraction（类别已知，类别较少，语料库规模有限）和 Open IE（类别未知，Web 作为语料库）进行了折衷，构建了 O-CRF 系统，同时解决这两类问题。

Zhu（2009）等^[39]对 Snowball 系统进行了改进，构建了 StatSnowball 系统，该系统采用马尔可夫逻辑网络（Markov Logic Networks, MLNs）学习 Pattern 的权重，使用概率方法评估和选择 Pattern，而非启发式的规则，并且该系统具有很强的扩展性，可以同时解决 Traditional Information Extraction 和 Open IE 问题，目前，该方法已经用于微软人立方关系搜索（中文版）⁸和 EntityCube（英文版）⁹。

半指导的学习方法不需要人工标注语料库，所需要的只有构造初始关系种子集，然后利用 Web 或者大规模语料库信息的高度冗余性，充分挖掘对应的关系描述模式，并通过模式匹配抽取新的关系实例，准确、高效地完成关系抽取任务。但是，这种方法也存在几个关键问题，如：初始关系种子集的产生和选择方式、Pattern 的组成方式、Pattern 的质量评估、迭代过程的速度、高准确率召回率等问题。

1.4.2 后确定关系类型体系的方法

在很多情况下，构建关系类型体系很困难，甚至不可能。所以，产生另一类关系抽取方法，即先确定关系实体对再确定关系类型。该方法又称基于无指导的学习方法，或者基于聚类的方法，是一种自底向上的信息抽取策略。

Hasegawa（2004）等^[40]最早提出了基于聚类的关系抽取的研究方法，该方法基于相同的实体关系应该出现在相似的上下文中这一假设。所以，如果将实体对对应的上下文信息（实验中只取实体对中间的词）进行聚类，不同类即可表示不同的关系类别，类中所有实体对对应的上下文信息作为类的上下文信息，最后通过对上下文统计，并选取最高频词用于描述该关系。

Zhang（2005）等^[41]提出利用句法树表示实体对，并通过改进的 Tree Kernel 计算句法树之间的相似度，然后对实体对聚类，实验证明该方法不仅可以对高频实体对进行有效处理，对低频实体对也能达到不错的效果。

Rosenfeld（2006）等^[42]构建了 URIES（an Unsupervised Relation

⁸<http://renlifang.msra.cn/>

⁹<http://entitycube.research.microsoft.com/>

Identification and Extraction System) 系统, 该系统首先以高频实体对为主要研究对象, 根据实体对之间的词序列模板, 使用层次聚类算法对实体对聚类, 然后从每类中选取高准确率的实体对作为关系种子, 作为 URES 的输入, 最终将无指导和半指导方法结合起来完成关系抽取任务。

Rosenfeld (2007) 等^[43]利用高频实体对的上下文作为特征, 尝试了基于不同的聚类方法进行关系识别的对比实验, 基于单连通 (Single linkage) 层次聚类算法取得了最好的实验结果。

Davidov (2007) 等^[44]根据给定的概念和实例作为种子, 借助 Google 搜索引擎识别高频的关系模式, 采用基于模式聚类的方法抽取与特定概念相关的各种关系。

Davidov (2008) 等^[45]提出了基于模式群 (Pattern Clusters) 抽取词语之间一般性语义关系的解决方案, 该方法无需使用自然语言底层处理技术作为支撑, 而是随机从句子中选取词语, 这样以来, 他们的方法就需要超大规模语料库的支持, 在对方法的有效性进行评价时, 他们首次提出采用学术能力测试问题检验方法。

Sun (2009) ^[46]在 Snowball 基础上进行了改进, 提出了一种新颖的 Two-stage Bootstrapping 的关系抽取学习框架, 该框架只需要针对特定的实体对类型 (如 <Person, Organization>), 以少量具有代表性的关系种子集 (如 <Bill Gates, Microsoft>、<Louis Gerstner, IBM>) 和大规模生语料库作为输入, 采用 Bootstrapping 方法挖掘出关系描述模式之后, 采用一系列启发式规则选取出包含名词性关系描述词的模式, 用于抽取新的关键词, 达到关系类型自动发现的目的, 并且该方法一定程度上解决了 Bootstrapping 方法对于初始关系种子集过于依赖的问题。

Yang (2009) 等^[47]提出了基于规约和聚类的自底向上的关系抽取方法, 他们认为领域动词是对领域关系进行刻画的重要元素, 所以, 可以以领域动词为中心自动获取关系类型, 并抽取关系实例, 和传统的关系抽取技术相比, 其方法不需要预先定义关系类型, 不需要先验的领域知识, 不需要人工标注语料库, 可以应用于任何领域。然而, 实体关系往往不仅由动词表示, 还可以是名词, 不免遗漏一部分关系实例的抽取, 影响召回率。

Mesquita (2010) 等^[48]与以往工作不同, 他们首次将博客文章作为处理语料库, 采用聚类的方法构建了 SONE (SOcial Network EXtraction) 系统, 为了评价方法性能, 并且避免人工构建评测集的偏置和召回率等问题, 他们还提出了一种新颖的基于语义知识库 Freebase¹⁰ 的自动评测方法, 通过实验证明了

¹⁰<http://www.freebase.com/>

聚类方法用于博客文本实体关系抽取的可行性和有效性。

无指导的学习方法基于这样一种假设：拥有相同语义关系的实体对，它们的上下文信息较为相似，其上下文集合代表着该实体对的语义关系。抽取过程大体分为三部分：实体对及其上下文信息提取；根据上下文信息对实体对聚类；标注各个类的语义关系，即对关系类型进行描述。该方法产生的分类一般比较宽泛，并且定义合适的类别比较困难，另外，该方法对低频的实体对处理能力有限，缺乏标准的评测语料，甚至没有统一的评价标准。

1.5 问题的提出

目前，中文关系抽取研究的难点主要体现在以下三个方面：

1.5.1 关系类型体系构建困难

据我们所知，现有的两个公认的关系类型体系，一个是被广泛使用的 ACE 2008 RDR 任务定义的 7 大类、18 个子类的实体关系类型体系^[8]，不够细化，且覆盖面有限，如表 1-1 所示，另外一个 Semeval'07 定义的 Task 4，包含了 7 种常见的语义关系^[9]。

由于当前自动发现关系类型的方法还不够成熟，通常根据特定应用需要领域专家预先定义关系类型，费时费力，可移植性差。所以，如何自动或半自动地建立一套合理的关系类型体系仍然是一个亟待解决的问题。

表 1-1 ACE 2008 RDR 任务中实体关系类型
(*表示对称关系，括号中的数字是各关系在 ACE2005 语料库中的实例数目)

类型	子类型
人工制品	使用者-拥有者-发明者-制造者 (567)
一般关系	市民-居民-宗教团体-种族 (842)、机构所在地 (1067)
转喻*	无 (35)
机构关系	雇佣 (1466)、创建者 (12)、所有者 (18)、学生-校友 (56)、运动-团体 (38)、投资者-股东 (61)、会员 (340)
部分-整体	人工制品 (9)、地理 (1211)、子公司 (附属的) (863)
人-社会*	职务 (161)、家庭 (279)、持续的-私人的 (70)
物理*	处于 (1279)、临近 (218)

1.5.2 关系抽取标注语料库匮乏

一直以来，绝大多数的关系抽取相关研究人员主要集中在 ACE 英文语料库，基于 Wikipedia 构建的语料库，或者生物领域的自制语料库上进行实验，取得了具有影响力的研究成果，而中文关系抽取方面使用最多的是 ACE RDR

任务提供的中文语料库，其中包含了 8,592 个关系实例，数据规模有限，且根据上节介绍可知，其关系类型数目少且过于宽泛，不实用。所以，如何构建一套公认的较具规模的（句子级和篇章级）关系抽取语料库数据资源，用于训练学习并验证抽取方法的性能，成为待解决的重要问题。

1.5.3 领域自适应的关系抽取研究滞后

Hasegawa (2004)^[40]首次提出无指导的关系类型发现和关系抽取之后，越来越多的研究学者都转向无指导方法的研究，尽可能避免人工的参与，提高领域自适应能力，尤其是 Banko (2007)^[37]提出 Open IE 的概念，并相继发表了一系列相关的研究成果，如 TextRunner, O-CRF 等。但是，他们的工作主要集中在英文方面，中文方面的研究相对较少，而现有的研究主要集中在有指导和半指导的学习方法上，对领域自适应的关系抽取研究相对滞后。

1.6 本文的主要研究内容

本文旨在探索领域自适应的中文实体关系抽取解决方案，主要结合半指导和无指导的学习方法解决关系类型自动发现、关系种子集自动构建、关系描述模式挖掘和关系元组抽取等关键问题，大大降低对人工标注语料库的依赖程度，提高关系抽取的自动化程度，扩大适用范围。

具体来讲，整个论文的工作流程如图 1-1 所示，各章节安排如下：

第 1 章，首先介绍了本课题的研究背景，探讨了研究目的和意义，然后分析了关系抽取的发展历史，从关系抽取的相关技术和方法上综述了研究现状，并提出了目前中文实体关系抽取研究存在的问题，最后，在此基础上提出了本文的主要研究内容。

第 2 章，研究并探讨了基于特征词聚类的关系类型发现方法。我们以从真实数据中抽取关键信息为出发点，首先从 Web 上获取大规模的网页文本作为语料资源，以实体对类型为单元（如<人名—人名>、<人名—机构名>等），通过挖掘高频实体对之间的特征词（主要包括名词和动词）实现关系类型的自动发现，即借助特征词表示实体关系，而不局限于有限的几类关系。

第 3 章，提出了基于 Web Mining 的关系种子集抽取方法。对于自动发现的大量关系，采用有指导的方法人工标注语料库耗时耗力、稀疏问题严重，而直接使用半指导的学习方法，由于类型的繁多需要人工构造大量的关系种子集，投入成本较大。针对这些问题，我们提出了基于 Web Mining 的关系种子集自动抽取方法，该方法将关系实例看作三元组 $\langle e_1, e_2, R \rangle$ ， e_1 可以赋值为高频的实体， R 为关系类型，对应一个特征词集合，在二者基础上进行查询的构造和扩展，利用搜索引擎检索出相关的网页，获取页面摘要 snippet，统计学

习抽取候选答案，即另一个实体 e_2 ，获取实体关系核心网。

第 4 章，采用了基于 Bootstrapping 的关系描述模式挖掘方法。我们以每类抽取的少量高质量关系种子出发，提出了基于 Bootstrapping 的关系描述模式挖掘方法，定义了启发式上下文模式及其泛化策略，从未标注语料库中迭代地挖掘关系描述模式，抽取关系元组，整个过程仅需很少的人工干预，领域移植性较强。

第 5 章，设计并实现了一套领域自适应的关系抽取平台。在该平台上，研究人员可以集中精力进行关系抽取中算法的改进和深入研究，快速展开实验。最后，以人物社会关系为切入点，开发了可视化的在线演示系统，可以向用户提供查询和浏览服务，以直观、清晰的方式展示抽取效果。

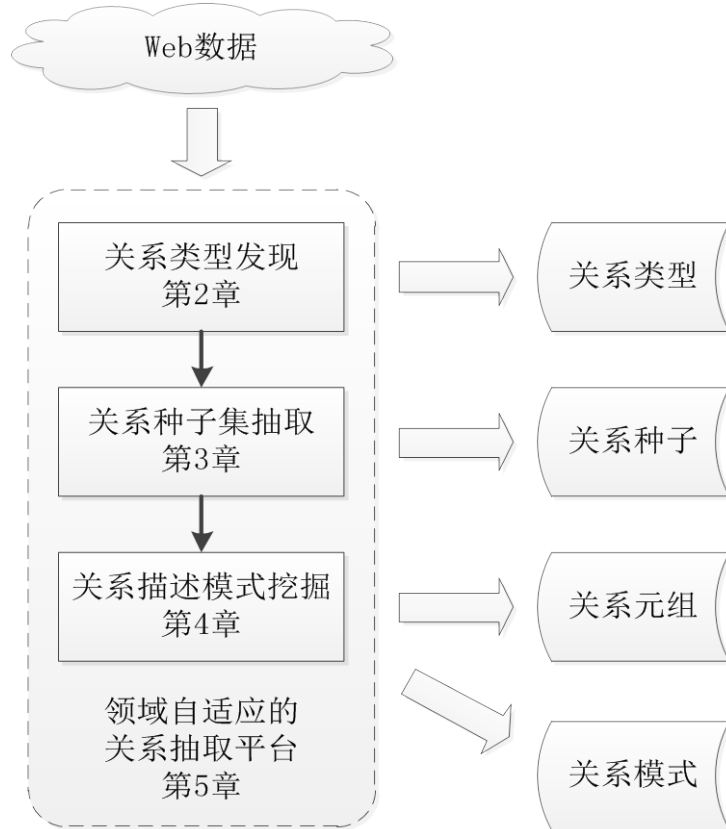


图 1-1 本文研究工作流程图

第2章 基于特征词聚类关系类型发现

2.1 引言

通过观察分析真实语料库发现，绝大多数产生关系的实体对都可以由其上下文中的一般动词和一般名词触发和描述（统称为特征词），且这些特征词均与待处理的实体对在依存句法分析树中产生有限的几类关系。

由此，首先我们以实体对类型（如“人名—人名”和“人名—机构名”代表两个不同的实体对类型）为单位，采用基于大规模语料库统计的方法抽取与特定实体对类型相关度较大的候选特征词集；然后，采用启发式通用过滤规则对候选特征词集进行过滤，如过滤低频词，保留那些在实体对一定上下文窗口中的词，经动词细分类后保留一般动词，保留一般名词，保留在依存句法分析树中与任一实体产生特定关系的动名词、保留相关度最大的 K 个词等；最后，借助语义词典计算候选特征词之间的相似度，采用层次聚类或 Affinity Propagation 等算法对候选特征词聚类，完成关系类型的自动发现，此时每类即为自动发现的一个关系类型，聚类的中心或者与特定实体对类型相关度最大的词作为对应关系类型的统一描述词。

2.2 算法流程

基于特征词聚类关系类型发现算法框架图如图 2-1 所示。

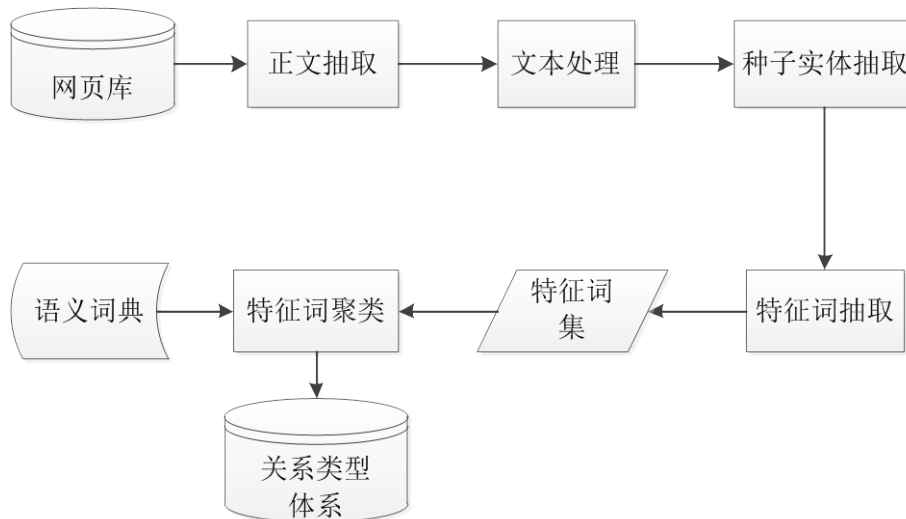


图 2-1 基于特征词聚类关系类型发现算法

首先我们对算法中涉及的一些概念进行解释说明：

特征词 (Feature Word)： 实体对上下文中可以用来描述实体之间关系的一般动词和名词，如表 2-1 中的“校长”、“合作”等。

种子实体 (Seeded Entity)： 特定实体类型在文本库中的高频实体，如

“章子怡”、“成龙”、“刘德华”经常出现在娱乐资讯中，属于人名实体类型的种子实体。种子实体可以用于后续特征词的抽取。

如图 2-1 所示，本算法以从真实网页库中发现关系类型体系为出发点，具体处理过程如下：

- (1) 正文抽取：对获取的网页准确地定位正文内容块，仅保留标题和正文内容块两部分；
- (2) 文本处理：对原始文本进行断句、中文分词、词性标注、依存句法分析、命名实体识别等底层自然语言处理操作，并将结果以 DOM 树的形式组织；
- (3) 特征词抽取：从 DOM 中读取句子的处理结果，计算实体出现频率，选取种子实体，进而，从与种子实体形成实体对的句子集中统计抽取特征词集，它们将用于描述实体关系；
- (4) 特征词聚类：由于不同的特征词可以表达相同的实体关系，所以，我们进一步利用语义词典计算特征词之间的相似度，采用不同的聚类策略（如层次聚类、AP 聚类、语义代码聚类等），聚类结果即为自动发现的实体关系类型。

2.3 算法设计

通过观察分析真实语料库发现，绝大多数产生关系的实体对都可以由其上下文中的一般动词和一般名词触发和描述（统称为特征词），如表 2-1 所示。基于此，我们提出了一种基于特征词聚类的关系类型发现，该方法主要包括语料库获取、种子实体抽取、特征词抽取和特征词聚类模块，我们以实体对关系类型“人名—人名”为处理对象，对各个模块进行阐述说明。

需要特别说明的是，我们将不同的实体对类型看作不同的“领域”，而不依赖于现实中的领域（如娱乐、体育、教育、科技等），即以实体对类型为单元，直接进行开放域的关系类型自动发现。为了更好的对方法进行验证与评价，我们选择以复杂的人物社会关系为主要实验对象，进行关系类型的自动发现探索。

表 2-1 特征词触发关系类型示例

绿色动力能源公司与“哈工大”签订两项合作协议。			
关系类型	ARG1	ARG2	说明
合作 (机构名—机构名)	绿色动力 能源公司 /Ni	哈工大/Ni	“绿色动力能源公司”和“哈工大”均为机构名，二者由“签订”、“合作”触发，存在“合作”关系
哈尔滨工业大学校长王树国荣获法国荣誉勋章。			
关系类型	ARG1	ARG2	说明
校长 (机构名—人名)	哈尔滨工 业大学/Ni	王树国/Nh	“哈尔滨工业大学”为机构名，“王树国”为人名，二者由“校长”触发，存在“从属”关系
巨星刘德华携手巩俐、胡静等人气明星打造的都市爱情大片《我知女人心》在博纳悠唐国际影城正式首映。			
关系类型	ARG1	ARG2	说明
携手 (人名—人名)	刘德华/Nh	巩俐/Nh	“刘德华”、“巩俐”和“胡静”均为人名，由“携手”触发，存在两个“携手”合作关系
携手 (人名—人名)	刘德华/Nh	胡静/Nh	

2.3.1 语料库获取

通过实现网络爬虫抓取来自 9 个门户网站 2002 年至 2009 年可供访问下载的所有娱乐新闻，包括腾讯网、新浪网、搜狐网、网易、凤凰网、中国新闻网、中国娱乐网、TOM、21CN，共获取了超过 100 万个网页（简称 RE100W），占用空间资源 26.2G。

由于我们所处理的均为正文较集中的主题类资讯网页，通过观察网页源代码中 HTML 标签和文字分布特点发现，网页中正文内容块比附加内容文本中含有更多的文本字符，也即含有更少量的 HTML 标签信息。基于此，我们提出了一种简单的基于文本行分布的正文抽取方法。

首先将网页源代码中的 HTML 格式信息删除，每行仅保留文本内容，即文本行。接下来，正文内容块的抽取可以被看成一个优化问题，即计算行 b 和 e ，用于最大化低于行 b 和高于行 e 的非文本字符数，以及在行 b 和行 e 之间的行文本字符数，相应的就是最大化对应的目标函数，如公式 (2-1) 所示。

$$\underset{b,e}{argmax}\{\sum_{i=0}^{b-1}(s_i - t_i) + \sum_{i=b}^e t_i + \sum_{i=e+1}^n (s_i - t_i) | 0 < b \leq e < n\} \quad (2-1)$$

其中， s_i 是原始网页源代码中行 i 的总字符数， t_i 是剔除 HTML 标签后行 i 的文本字符数， n 为网页源代码总行数，编号从 0 到 $n-1$ 。

算法不需要针对特定网页书写正则表达式，不需要解析 HTML 以建立

DOM 结构，不被病态的 HTML 标签所累，可以高效、准确地定位网页正文内容块。经过在随机选取的 300 篇网页上测试，准确率可达到 95%左右，满足了实际应用的需求。

正文抽取完成后，我们使用哈尔滨工业大学社会计算与信息检索研究中心研发的语言技术平台（Language Technology Platform, LTP）^[49]对正文文本进行断句、中文分词、词性标注、依存句法分析和命名实体识别等处理，在此基础上，我们还使用了开源全文检索引擎工具包Lucene 3.0.3¹¹建立句子级索引，为后续研究工作提供数据支持和操作服务。

2.3.2 种子实体抽取

我们认为高频的实体对之间更有可能产生人们关注的实体对关系，且能够包含大部分的重要的关系，所以，我们以高频实体对为限定筛选出其所处上下文信息，进行特征词的抽取。

我们经过对 RE100W 中的命名实体进行频率统计，如表 2-2 所示为 Top-20 的人名统计结果，最终选取 Top-1000 个人名集合作为种子实体（*NEList*），即接下来以与 *NEList* 中每个实体形成句子级、近距离、高共现实体对的上下文信息为特征词抽取的统计资源。

表 2-2 Top-20 的高频种子实体示例

序号	实体	频率	序号	实体	频率
1	章子怡	21746	11	张柏芝	13917
2	周杰伦	21041	12	陈冠希	13870
3	成龙	19904	13	谢霆锋	13724
4	刘德华	19285	14	陈奕迅	13602
5	范冰冰	18337	15	梁朝伟	12615
6	赵本山	17390	16	容祖儿	11942
7	蔡依林	16520	17	赵薇	11552
8	王菲	16039	18	张艺谋	11426
9	周迅	15856	19	李宇春	11384
10	刘嘉玲	14851	20	李冰冰	11027

2.3.3 特征词抽取

特征词抽取 FWE（Feature Words Extraction）算法考虑语料库中与特定实体对类型下的实例（即实体对类型相同）共现且具有特定语义关系的动词和名词，通过计算它们在整个语料库中和特定实体对类型上下文中的分布情况，抽取特征词。具体操作步骤如下所示：

¹¹<http://lucene.apache.org/>

- 1) 对每个实体 $NE_i \in NEList$ ，在 RE100W 中检索出包含 NE_i 的所有句子，保留那些同时包含 NE_i 和另一个与其形成特定实体对类型的实体的句子，得到 $\langle NE_i, SentList_i \rangle$ ，需要特别说明的是，这些句子均包含词法、依存句法分析的处理结果。
- 2) 使用布隆过滤器（Bloom Filter）对 $SentList_i$ 进行句子去重，得到 $\langle NE_i, SentSet_i \rangle$ 。
- 3) 针对每个 $\langle NE_i, SentSet_i \rangle$ 的句子集 $SentSet_i$ ，统计与 NE_i 词距离小于 $maxDistance$ 的共现实体 NE_j 的频率，得到 $\langle NE_i, NE_j, CoFreq_{i,j} \rangle$ 。
- 4) 对于每个 NE_i ，仅保留与其共现频率大于 $minCoFrequency$ 的实体，及实体对共现的句子，得到筛选后的句子集 $\langle NE_i, NE_j, SentSet_{i,j} \rangle$ 。
- 5) 以 $\langle NE_i, NE_j, SentSet_{i,j} \rangle$ 包含的所有句子为处理对象，按照如下启发式通用规则对实体对上下文中的词进行统计过滤，得到候选特征词集 $CandFWSet$ 。
 - Rule1: 特征词必须是出现在实体对之间的动词或名词
 - Rule2: 针对动词进行细分类，仅保留一般动词
 - Rule3: 针对名词进行细分类，仅保留一般名词
 - Rule4: 对句子进行依存句法分析后，动词必须满足与实体对中任一实体存在主谓关系 SBV 或动宾关系 VOB
- 6) 对每个 $w_k \in CandFWSet$ ，统计其在（4）得到的句子集中的频率分布 $Freq(w_k)$ ，仅保留频率大于常数 $minFrequency$ 的词，得到特定实体对类型对应的候选特征词的分布情况 $\langle w_k, Freq_T(w_k) \rangle$ 。
- 7) 根据候选特征词 w_k 在 RE100W 和特定实体对类型上下文中的分布信息采用公式（2-2）计算其与实体对类型的相关度 $Rel(w_k)$ ，其中， $Freq_T(w_k)$ 和 $Freq_A(w_k)$ 分别表示 w_k 在特定实体对类型上下文和语料库 RE100W 中的频率。

$$Rel(w_k) = \frac{Freq_T(w_k)}{Freq_A(w_k)} \quad (2-2)$$

- 8) 根据相关度对候选特征词进行排序 $RCandFWList$ ，根据排序位置 $RP(Rel(w_k))$ 取靠前的 Top-K 个作为特征词 $fwSet$ 。

其中，动词细分类我们采用了马金山博士论文（2007）^[50]中提出的基于最大熵模型的动词细分类研究工作（准确率为 88.3%），动词细分类标准如表 2-3 所示。由于在 2003 年国家 863 分词和词性标注一体化评测中，颁布了一

个含有 28 个词性的词性标注集，其中已经包含了对名词的细分类，所以，我们可以在LTP词法分析结果基础上直接获取。依存句法分析我们使用了哈尔滨工业大学社会计算与信息检索中心研发的基于图的依存句法分析模块¹²。

表 2-3 动词子类列表

动词	描述	举例
vx	系统词	他 是 对的
vz	助动词	你 应该 努力
vf	形式动词	他要求 予以 澄清
vq	趋向动词	他认识 到 困难
vb	补语动词	他看 完 了电影
vg	一般动词	他 喜欢 踢足球
vn	名动词	参加我们的 讨论
vd	副动词	产量 持续 增长

我们以 RE100W 语料库为输入，以“人名—人名 (Nh—Nh)”为实验的特定实体对类型，使用 *NEList* 中包含的所有实体作为种子实体，使用 FEW 算法进行特征词抽取，maxDistance, minCoFrequency 和 minFrequency 分别赋值为 10, 10, 50。我们对 *RCandFWList* 中的所有词进行人工标注（共 1,225 个标准特征词），结合词出现的句子，如果抽取的词可以用于描述两个人之间的某种社会关系，则认为抽取正确，其性能如图 2-2 所示。

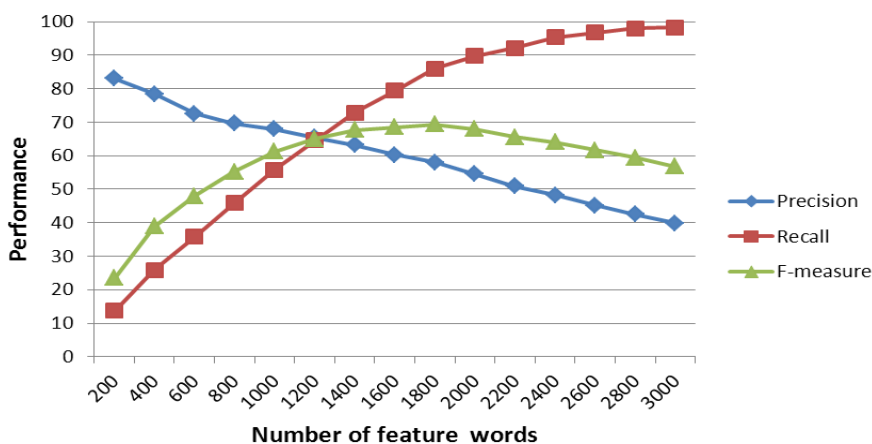


图 2-2 使用 Top-1000 实体种子时特征词抽取的性能

图 2-2 显示，随着抽取的特征词数量 Top-K 的增加，准确率不断下降，召回率不断升高，且升高幅度逐渐变慢。当抽取前 1,800 个词时达到了最好的综合性能，F 值为 69.30%，召回率达到了 85.94%。当抽取的特征词数量 Top-K 达到 3,000 时，FEW 的召回率接近 100%，即几乎包含了所有特征词。

由于抽取的特征词目的在于描述实体之间的关系，在特征词抽取算法中仅

¹² <http://ir.hit.edu.cn/ltp/>

考虑和不同实体对及具有特定词法和句法关系的动名词，因此抽取特征词的质量和数量直接受所使用的种子实体数量的影响。所以，我们又在不同数量的初始种子实体对情况下进行特征词抽取，以衡量种子实体数量对特征词抽取结果的影响，如图 2-3 所示。

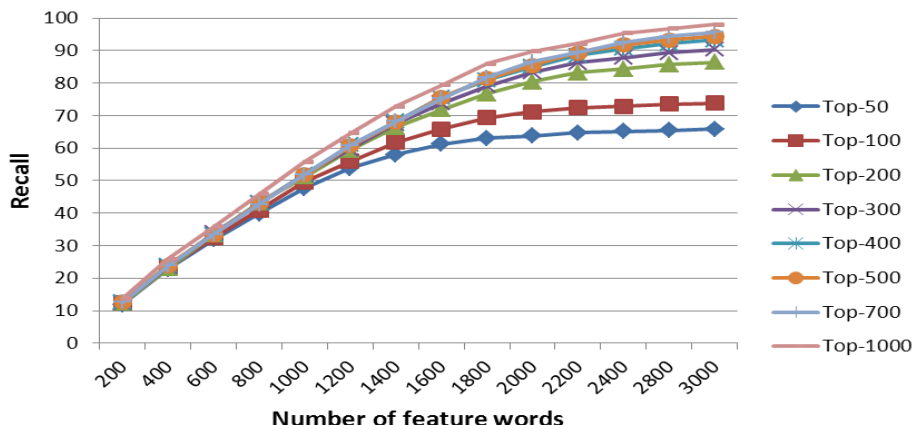


图 2-3 使用不同数量的种子实体对特征词抽取的召回率性能的影响

图 2-3 显示特征词相关度分布的质量受初始种子实体的数量影响，使用种子实体越多，抽取特征词的性能越好。当抽取前 1,000 个特征词时，即使在使用不同数量的种子实体情况下，召回率和 F 值相差也很小，主要原因首先是我们定义的特征词相关度排序规则的有效性，其次由于在特定实体对类型中存在一定数量的高频实体关系类型，规模约占关系类型总量的一半以上。使用最多的 Top-1000 个种子实体时，方法达到了最好的性能，分析主要原因在于使用更多的种子实体，将会得到更多的高频实体对，统计信息也会更加丰富，从而有效提高了特征词抽取的性能。

随着抽取特征词数量 Top-K 的增加，召回率增长速度由快变慢，达到 3,000 左右之后，性能几乎不再增加，趋于稳定，且此时在不同数量的种子实体情况下，最高召回率也存在一定的差距，即关系类型种类/数量不同，这是因为使用越多的种子实体就能得到越多的上下文信息用于统计，以发现更加丰富的关系类型。但是，当种子实体达到一定数量之后（Top-1000），关系种类几乎不再增加。

另外，即使仅使用 50 个实体作为种子，抽取特征词数量 Top-K 为 3,000 时，也能达到约为 70% 的召回率，并且随着种子实体的增加，召回率迅速上升，当使用 400 个实体作为种子时，召回率达到 93% 以上，之后随着种子实体的增加，召回率增加幅度很小，这些现象反映了相同类型的实体对之间产生关系的数量范围和稳定程度，也说明了我们开始对于“高频实体对之间更有可能产生人们关注的实体对关系，且能够包含大部分的重要关系”这一假设的正

确性和可行性。

由此说明，只需要以一定规模的生语料为输入，针对特定实体对，根据我们提出的特征词抽取算法统计筛选，最后选取 Top-K 个词作为特征词即可。当然，此时选择的特征词集合覆盖面广，但是难免存在一定的噪声，此时，可以以候选特征词为参考，人工参与过滤掉不能用于表示实体关系的词，进一步提高准确率，也避免了预先定义关系类型的盲目性。

如表 2-4 所示，我们列出了通过 FWE 算法抽取的相关度较高的特征词及其相关句子示例。

表 2-4 特征词示例

特征词	示例句子
好友	刘嘉玲邀请中田英寿、胡军等圈内 好友 前往自己的 MUSE CLUB 狂欢。
扮演者	82 版《西游记》中沙僧 扮演者 闫怀礼的遗体告别仪式，今日上午 10 时在八宝山殡仪馆东礼堂举行。
干爹	香港歌手关楚耀与女友卫诗涉嫌藏毒于日本被捕，身为关楚耀 干爹 ，谭咏麟十分痛心。
状告	赵文卓 状告 画家方力钧，要求其拆除违规房屋
老搭档	电视剧《军旗飘扬》眼下正在北京热拍，朱时茂、丛珊这对 老搭档 在剧中再次扮演夫妻。
伯父	“千亿少奶”徐子淇最近卷入 伯父 徐传文金钱纠纷
成婚	去年七月，刘嘉玲、梁朝伟在不丹 成婚 后，刘嘉玲怀孕的消息就不断传出，可惜每次都是空穴来风。
得意门生	张纪中的 得意门生 李亚鹏、胡军、黄晓明等演员已是热门的人选。
爱将	曾主演过《断箭》、《风语者》等片的吴宇森 爱将 克里斯汀·史莱特在片中饰演超自然案件侦探爱德华。
经纪人	林志玲的 经纪人 阎柔怡接受采访时说，对空穴来风的消息不会予以回应。
至交	昨晚，罗嘉良的 TVB 前同事和 至交 薛家燕、马国明、马德钟、陈敏之、艾威等已纷纷入驻酒店等待好友大婚时刻。
栽培	昨日成龙与他 栽培 的香港发明家陈克敏，出席了“同心‘研’放香港之光”记者会。
父子	影视圈里的父子档不在少数，仅青壮派里，除成龙房祖名 父子 、张国立张默 父子 、李诚儒李大海 父子 外，还有杜志国杜淳 父子 、巍子王子义 父子 、杨亚洲杨博 父子 等。
前夫	陈琳 前夫 沈永革召开记者会，痛忆当年美好情意
偕同	他们翻出柳岩去年 偕同 苏醒共同主持某男性杂志年度盛典的合影
合作	刚刚从美国回来的容祖儿赴京将与韩庚等 合作 拍摄央视剧《青春舞台》。
夫妻	《唐山大地震》票房破五亿，冯小刚徐帆 夫妻 档访台造势

2.3.4 特征词聚类

特征词（主要包括一般动词和一般名词）是体现实体关系的最为重要的词汇单元，一系列具有相同含义和用法的特征词可以体现同一种关系，因此我们认为采用基于特征词聚类的方法发现关系类型是可行的。

本课题分别基于语义词典《知网》和《同义词词林（扩展版）》提出了两类不同的特征词聚类的关系类型发现算法。

（1）基于《知网》的特征词聚类

《知网》又称 HowNet^[51]，是一个以英语和汉语的词语所代表的概念为描述对象，以词为基本单位，将揭示概念之间以及概念属性之间的关系为基本内容的常识知识库。目前其词汇已经达到十万规模，每一个词可以由多个概念表示，这里，概念是对词汇语义信息的描述，用义原来描述，义原是描述一个概念的基本的也是最小的意义单位。与《同义词词林》和 WordNet 这些一般的语义词典不同，HowNet 并不是简单地将所有的“概念”归结到一个树状的概念层次体系中，而是试图使用一系列的“义原”对每一个“概念”进行描述。不同词之间的语义相似度可以借助 HowNet 中的“义原”计算得到。

我们利用语义相似度或语义距离描述两个特征词 w_i 和 w_j 之间的相似度，其相似度值 $Sim(w_i, w_j)$ 根据 HowNet 计算得到， $Sim(w_i, w_j)$ 由 w_i 和 w_j 相同“义原”的数量除以两者的“义原”总和实现归一化得到，计算公式如（2-3）所示。

$$Sim(w_i, w_j) = \frac{2 * NC_{i,j}}{NC_i + NC_j} \quad (2-3)$$

其中， $NC_{i,j}$ 表示 w_i 和 w_j 在 HowNet 概念定义 DEF (the concept definition in HowNet) 中相同“义原”的个数， NC_i 和 NC_j 分别表示 w_i 和 w_j 概念定义中的“义原”个数。

通过 HowNet 得到特征词之间的相似度之后，我们可以使用层次聚类 (Hierarchical Agglomerative Clustering, HAC) 和 Affinity Propagation (AP) 算法对特征词进行聚类实现关系类型发现。接下来，我们对两种聚类算法进行简要介绍。

1) 层次聚类

HAC 是一种自底向上 (bottom-up) 的聚类算法，开始时把每一个特征词都当作独立的簇，聚类算法每次将两个（或可能多个）现有最相近的簇合并成一个新的簇（聚合过程），因此，每次迭代都会使簇的总数减少。当簇之间的相似度小于某个阈值或簇的个数达到一定数量时，算法停止。聚合过程主要取

决于簇间的相似度函数 $Sim(C_i, C_j)$ ，其中 C_i 和 C_j 表示两个不同的簇。目前，存在很多定义相似度函数的方式，每一个都会使最终的聚类结果具有不同的特点，较为流行的方法主要有单连通（single linkage），全连通（complete linkage）和平均连通（average linkage）。

single linkage 方法通过计算簇 C_i 中每个特征词和簇 C_j 中每个特征词之间相似度，这些相似度中最小的就是就是 C_i 和 C_j 之间的相似度，用数学公式表示如公式（2-4）所示。

$$Sim(C_i, C_j) = \min\{Sim(w_i, w_j) | w_i \in C_i, w_j \in C_j\} \quad (2-4)$$

complete linkage 和 single linkage 类似，首先计算两个簇中每一对特征词之间的相似度，但是它使用最大的相似度作为代价，而非最小相似度，如公式（2-5）所示。

$$Sim(C_i, C_j) = \max\{Sim(w_i, w_j) | w_i \in C_i, w_j \in C_j\} \quad (2-5)$$

average linkage 是 single linkage 和 complete linkage 的折中方案。和之前一样，计算出簇 C_i 和 C_j 中每两个特征词之间的相似度。正如其名字所隐含的，平均连通使用所有特征词对相似度的平均值作为两簇之间的相似度，因此，其公式如（2-6）所示。

$$Sim(C_i, C_j) = \frac{\sum_{w_i \in C_i, w_j \in C_j} Sim(w_i, w_j)}{|C_i| |C_j|} \quad (2-6)$$

2) AP 聚类

AP 聚类算法是 Brendan J. Frey 在 Science 2007 上提出的一种基于近邻信息传播的快速、有效的聚类算法^[52]。AP 算法不需要事先指定聚类数目，相反它将所有的数据点都看作潜在的聚类中心，称之为 exemplar。算法根据数据点之间的相似度进行聚类，数据点之间的相似度函数既可以是对称的，即两个数据点之间的相似度相等（如欧氏距离），也可以是不对称的，即两个数据点之间的相似度可以不等（如 KL 距离）。

假设数据点个数为 N ，数据点之间的相似度矩阵为 $S(N \times N)$ 。算法以 S 矩阵对角线 $s(k, k)$ 作为 k 点能否成为聚类中心的评判标准，该值越大， k 点成为聚类中心的可能性也就越大， $s(k, k)$ 还被称作参考度 preference。最终聚类的数目受参考度 preference 的影响，如果认为每个数据点成为聚类中心的概率相等（即没有先验知识），那么 preference 取值相同，并且，实际中我们往往取相似度矩阵的中位数或均值作为 preference 的初始值，这样可以得到中等的聚类数目，如果 preference 取值越少，得到的聚类个数越少，相反，得到的聚类

个数越多。

AP 算法通过迭代的传递和更新两种类型的信息，即 responsibility: $r(i,k)$ 和 availability: $a(i,k)$ 。 $r(i,k)$ 从数据点 i 指向候选聚类中心数据点 k ，表示 k 点作为 i 点聚类中心的吸引度， $a(i,k)$ 从候选聚类中心数据点 k 指向数据点 i ，表示 i 点选择 k 点作为其聚类中心的归属度。 $r(i,k)+a(i,k)$ 越大，数据点 k 作为聚类中心的可能性就越大，并且数据点 i 隶属于以 k 为聚类中心的聚类的可能性也就越大。AP 算法通过不断迭代更新 r 和 a 矩阵，最终，对于数据点 i ，其聚类中心为 k ，二者满足 $\underset{k}{\operatorname{argmax}}\{r(i,k)+a(i,k)\}$ 。

AP 算法的优点是不需要预先设置聚类个数，任意数据点都作为潜在的聚类中心，并且可以通过设置 preference 方便地融入先验知识。但是，AP 算法并非在所有数据集上都能收敛，这和数据点本身分布特点和迭代过程中阻尼系数等参数的设置有关，并且其初始参考度 preference 的设置难以确定和科学地解释。

(2) 基于《同义词词林（扩展版）》的特征词聚类

《同义词词林（扩展版）》语义词典又称Cilin(Extended)¹³，共收录了 8 万余词，在《同义词词林》^[53]原有的三层分类体系基础上进一步细分类和扩充，得到现在的五层分类体系，其树形结构如图 2-4 所示。在图 2-4 中，以第一层中的每大类为根均可以构成一个独立的词语树，整个语义词典由 12 棵这样的树构成，即第一层有 12 个结点。从第一层开始，按照词义不同进一步细分，依次出现第二层，第三层，直至第五层。在词语树形结构的第四层，每个结点的词语范畴相对其父结点（第三层）进一步缩小，而词语间的词义也更加接近。到了第五层（均为叶子结点），每个结点的词语同义或近似同义。其中，我们所说的词语词义之间的“同义”，既可以是语义上的，也可以是功能上的。

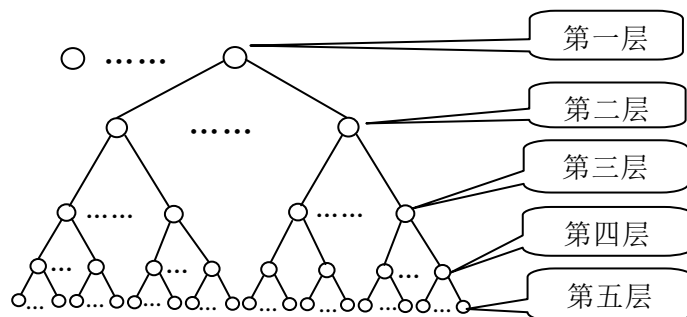


图 2-4 词语树形组织结构

¹³<http://ir.hit.edu.cn/>

一方面，我们可以基于《同义词词林（扩展版）》计算特征词之间的相似度，任意两个词 w_i 和 w_j 的相似程度由它们在树中的距离间接反应，假如两个特征词的树距离为 $Dist(w_i, w_j)$ ，则相似度 $Sim(w_i, w_j)$ ，计算公式如（2-7）所示。

$$Sim(w_i, w_j) = \frac{1}{Dist(w_i, w_j) + 1} \quad (2-7)$$

计算出特征词相似度矩阵后，我们同样可以使用 HAC、AP 等算法对特征词进行聚类，实现关系类型的自动发现。

另一方面，我们可以借助《同义词词林（扩展版）》的分层结构定义特征词之间的语义相似度，简单的将特征词在不同层别的语义代码作为直接聚类标准，而不需要使用其他聚类算法。其中，对于一词多义现象（即一个词具有多个不同层级的语义代码），我们采用将其归入最常出现的一类中，而非模糊聚类。

特征词聚类结果即为自动发现的实体关系类型体系，关系类型由特征词集合组成，为了更直观的观察和理解关系类型体系，我们又提出了一种以每个关系类型对应的特征词集中与当前实体对类型相关度 $RP(Rel(w_k))$ 最大的的一个或多个词作为该关系类型的描述词，自动对关系打标签。

2.4 实验结果与分析

2.4.1 实验数据

我们的实验在中文新闻文本中进行，使用了 2.3.1 获取的 RE100W 作为语料库，以“人名—人名（Nh—Nh）”为实验的特定实体对类型，即集中在人物之间关系类型的发现。需要特别说明的是，在实验过程中，对文本进行中文分词、词性标注、依存句法分析、命名实体识别、动词细分类等所产生的累计误差都没有被排除，使得我们的评价结果更加真实的反映了包含累计误差的数据规律。

我们在特征词抽取结果基础上，选取人工后处理标注出来的所有可以描述实体对之间特定关系的特征词集进行聚类实验，共有 1,225 个。由于聚类是一个主观性比较强的过程，可以从不同角度、不同粒度进行聚类，为了构建一个更加合理的标准集，我们组织多人同时独立的对特征词集进行归类标注，然后保留归类一致的情况，对不同意见的标注情况，经大家讨论确定。据此，我们构建出一个标准的特征词聚类评测集，共有 256 类，其中 123 类仅包含一个特征词。

2.4.2 评价标准

我们采用聚类方法中常用的两种评价方法对算法性能进行评测： F 值和纯度 $Purity$ ，它们分别从不同的方面对结果进行衡量。对于算法自动聚类的类别使用下标 r ，人工构建的标准类别使用下标 i 表示。 F 值计算方法如公式（2-8）~（2-11）所示。

$$Recall(i, r) = \frac{n(i, r)}{n_i} \quad (2-8)$$

$$Precision(i, r) = \frac{n(i, r)}{n_r} \quad (2-9)$$

$$F-measure(i, r) = \frac{2 \times Recall(i, r) \times Precision(i, r)}{Recall(i, r) + Precision(i, r)} \quad (2-10)$$

其中， $n(i, r)$ 是属于 i 类但是被自动分到 r 类中特征词的个数， n_i 是人工标注的第 i 类中特征词的个数， n_r 为自动聚类结果中第 r 类中特征词的个数， $F-measure$ 是 $Recall$ 和 $Precision$ 的调和平均。综合 $F-measure$ 如公式（2-11）所示， n 是特征词总数。

$$F-measure = \sum_i \frac{n_i}{n} \max\{F-measure(i, r)\} \quad (2-11)$$

$Purity$ 计算方法如公式（2-12）和（2-13）所示。

$$P(r) = \frac{1}{n_r} \max\{n_r^i\} \quad (2-12)$$

$$Purity = \sum_{r=1}^k \frac{n_r}{n} P(r) \quad (2-13)$$

其中， n_r^i 表示属于类别 i ，而被自动聚类到 r 中的特征词个数。

2.4.3 结果与分析

我们以每个词为一类，即不对抽取的特征词进行聚类（用 **Singleton** 表示）情况，作为 **baseline** 算法进行比较。另外，让未参与构建标准集的一个人工对特征词分类，作为 **upper bound** 进行对比评价（用 **Manual** 表示）。

（1）基于《知网》的特征词聚类实验

我们分别使用 **HAC** 和 **AP** 两种聚类算法进行实验，性能如表 2-5 所示。表 2-5 显示对特征词聚类之后比 **singleton** 的各项指标均有大幅度提升，说明聚类的必要性和有效性。**AP** 算法效果取得了最好的 F 值，达到 57.43%，且相比 **HAC** 耗时更少，但是与人工标注的结果 F 值相差 8.27%，说明自动聚类的

方法还有一定的提升空间。HAC 采用不同的类别间相似度函数，其中平均连通（average link）方法效果相对最好。

表 2-5 基于 HowNet 聚类的关系类型发现算法性能

Clustering Algorithm		F-measure (%)	Purity (%)	Running Time (s)
Singleton (baseline)		29.99	100.00	0
Manual		65.70	77.19	约 28800
HowNet-based	HAC (single link)	51.66	78.37	480.1
	HAC (complete link)	54.57	81.39	458.4
	HAC (average link)	55.98	80.98	484.3
	AP	57.43	77.28	103.5

针对 HAC 的三种类别间相似度计算方法，我们针对不同阈值对关系类型发现性能的影响进行了对比实验，如图 2-5 所示。图 2-5 显示 HAC (complete link) 和 HAC (average link) 的性能比 HAC (single link) 稳定，整体性能更好，且它们均在阈值等于 0.6 时达到了最好的性能，分别为 54.57% 和 55.98%，HAC (average link) 达到了最好的评测性能。

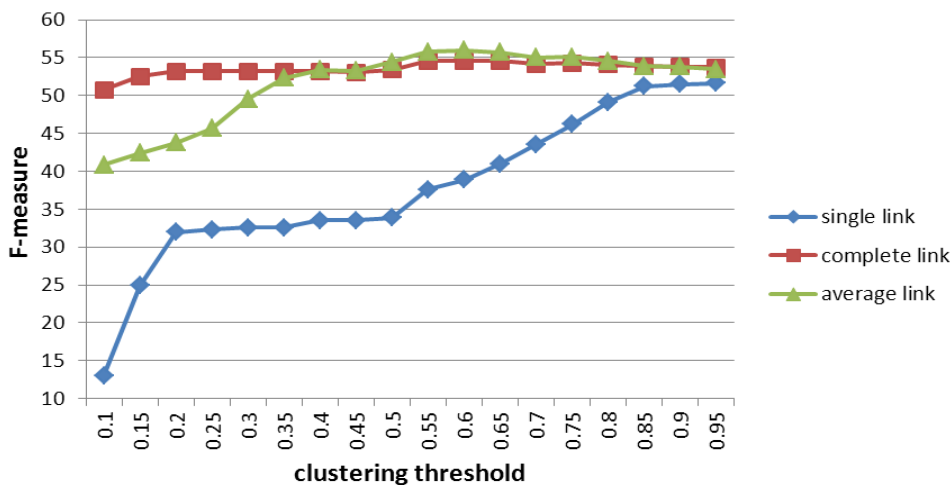


图 2-5 基于 HowNet 层次聚类（HAC）的关系类型发现算法性能

(2) 基于《同义词词林（扩展版）》的特征词聚类实验

基于词语在《同义词词林（扩展版）》上的树距离得到相似度，与上节类似，我们分别使用 HAC 和 AP 两种聚类算法进行实验，性能如表 2-6 所示。表 2-6 显示 HAC (single)、HAC (complete) 和 AP 算法取得了性能相当的结果，F 值约为 57.5%，HAC (average) 取得了最好的 F 值，达到 59.61%，但是耗时较长，纯度较低。

表 2-6 基于《同义词词林（扩展版）》聚类的关系类型发现算法性能

Clustering Algorithm		F-measure (%)	Purity (%)	Running Time (s)
Singleton (baseline)		29.99	100.00	0
Manual		65.70	77.19	约 28800
Cilin (Extended) - based	HAC (single link)	57.52	80.85	453.6
	HAC (complete link)	57.41	85.84	485.3
	HAC (average link)	59.61	85.02	1295.4
	AP	57.89	86.05	319.9

针对 HAC 的三种类别间相似度计算方法，我们对不同阈值对关系类型发现性能的影响进行了对比实验，如图 2-6 所示。图 2-6 显示 HAC (average link) 的性能比 HAC (single link) 和 HAC (complete link) 稳定，整体性能更好，且在树距离等于 4（第三层）时，达到了最好的 F 值 59.61%。

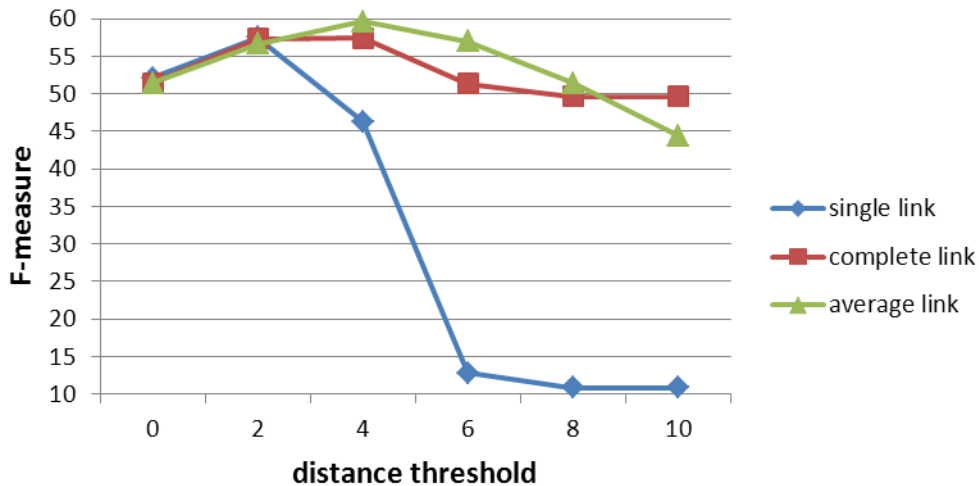


图 2-6 基于《同义词词林（扩展版）》层次聚类（HAC）的关系类型自动发现算法性能

另外，我们简单的利用《同义词词林（扩展版）》不同层级的语义代码对特征词进行直接聚类。对于在当前层级一词多义的情况，我们通过简单的取其最常见词义。对于未登录词的情况，单个词组成一类。在取不同层级的语义代码聚类时算法性能如图 2-7 所示。图 2-7 显示从第一层到第四层性能不断上升，到第五层时明显下降，在取第三和第四层时算法取得最好的聚类效果，主要原因是第一、二层分类较宽泛，而第五层的分类又太细，而第三、四层分类适中，自动分类也更倾向于将特征词分到三、四层所代表的小类和词群的级别。

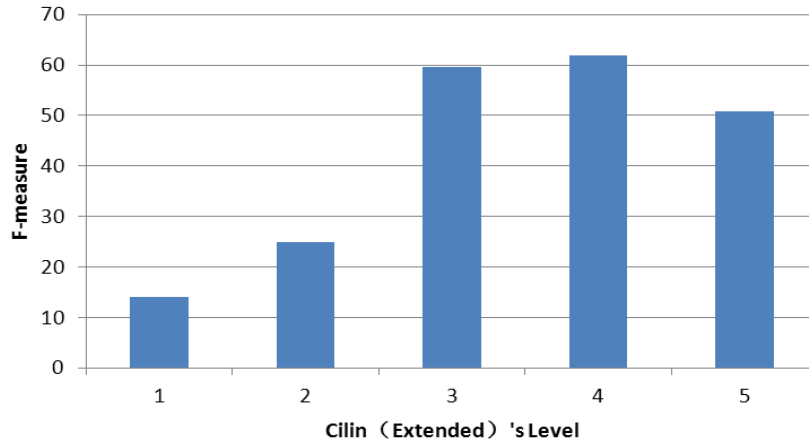


图 2-7 基于《同义词词林（扩展版）》直接聚类关系类型自动发现算法性能
如表 2-7 所示可以更加直观地比较不同关系类型发现算法的实验结果。

表 2-7 不同聚类算法的关系类型发现性能比较

Clustering Algorithm		F-measure (%)	Purity (%)	Running Time (s)
Singleton (baseline)		29.99	100.00	0
Manual		65.70	77.19	约 28800
HowNet-based	HAC (single link)	51.66	78.37	480.1
	HAC (complete link)	54.57	81.39	458.4
	HAC (average link)	55.98	80.98	484.3
	AP	57.43	77.28	103.5
Cilin (Extended) - based	HAC (single link)	57.52	80.85	453.6
	HAC (complete link)	57.41	85.84	485.3
	HAC (average link)	59.61	85.02	1295.4
	AP	57.89	86.05	319.9
	Cilin (level = 3)	59.67	76.24	0.45
	Cilin (level = 4)	61.83	89.14	0.45
	Cilin (level = 5)	50.86	94.69	0.5

表 2-7 显示基于 HowNet 的关系类型发现算法达到了最好为 57.43% 的 F 值，而基于 Cilin(Extended)的方法在依据第四层语义代码直接聚类时达到了最好为 61.83% 的 F 值，相比前者提高了 4.4%，纯度 Purity 相比前者提高了 11.86%，且聚类耗时由 103.5s 缩减到 0.45s，提高了两个数量级，与人工标注结果相比，F 值相差 3.87%，说明自动聚类方法几乎可以达到与人工直接标注相匹敌的效果。但是，这两种方法都存在一些共同的问题，如无法有效处理特征词未出现在语义词典的情况，没有考虑特征词所处上下文的语用信息，等等。

最后，我们对基于 Cilin(Extended)第四层语义代码聚类结果基础上，对聚

类结果，选取出现频率最高的特征词作为该类的描述词，如表 2-8 所示。

表 2-8 关系类型对应的特征词和关系描述词示例

关系类型	部分特征词	关系描述词
夫妻关系	{夫妇, 夫妻, 结婚, 完婚, 复婚, 抚养权, 爱人, 再婚, 改嫁, 成家, 迎娶, 老公, 丈夫, 太太, 妻, 贤妻, 妻子, 夫人, 老婆, 婚姻, 洞房, 姻缘, 同居, 婚期, 婚事, 婚礼, 新娘子, 新娘, 新郎, 新郎官}	结婚
经纪人关系	{经纪人, 经济人, 经理人}	经纪人
合作关系	{合作, 搭档, 合伙, 拍档, 合力, 联名, 联手, 搭伴, 对唱, 合唱, 对手戏, 合演, 配戏}	合作
情侣关系	{相恋, 恋情, 恋爱, 男朋友, 情侣, 男友, 女朋友, 女友, 情人, 伴侣, 女伴, 相爱, 新欢, 爱河, 拍拖}	男友
父母—子女	{母亲, 妈妈, 老母, 生母, 养母, 爹, 爸, 爸爸, 家父, 生父, 老子, 继父, 干爹, 老爸, 父母, 独生子, 公子, 儿子, 私生子, 女儿, 千金, 姑娘}	父母
好友关系	{好友, 朋友, 老友, 挚友, 老相识, 至交, 老朋友, 战友, 友谊, 友情, 闺中密友, 闺蜜, 哥们儿, 哥们}	朋友
角色扮演	{饰演, 出演, 饰, 扮演, 饰演者, 扮演者, 接演}	饰演
兄弟姐妹	{老大哥, 胞弟, 表哥, 兄弟, 堂兄弟, 胞兄, 二哥, 弟弟, 哥哥, 长兄, 妹, 妹子, 妹妹, 姊妹, 姐弟, 姐姐}	兄弟
伯乐关系	{学生, 入室弟子, 弟子, 徒弟, 门生, 学徒, 师父, 师傅, 老师, 教练, 恩师, 武师, 拜师, 师生, 师徒, 栽培, 提拔, 伯乐}	老师

2.5 本章小结

本章提出了基于特征词聚类的关系类型发现方法，方法以实体对类型为处理单位，仅需要一定规模的文本语料库以及语义词典（如《同义词词林（扩展版）》）。首先，该方法从文本中统计得到高频实体，即种子实体；然后，从与种子实体构成高频实体对所对应的上下文中抽取候选特征词（主要包括一般动词和一般名词）；接下来，通过一系列启发式过滤规则对候选特征词进行过滤；最后，通过实验基于《知网》和《同义词词林（扩展版）》的相似度计算方法，并详细对比不同的聚类算法（HAC 聚类、AP 聚类和语义代码直接聚

类), 发现较为简单的基于《同义词词林(扩展版)》语义代码直接聚类的关系类型发现方法取得了最好的实验结果, F 值为 61.83%, 与人工构建的关系类型体系效果相差不大, 验证了方法的有效性和可行性。

综上所述, 对于不同的应用领域需求, 不需要人工预先定义关系类型, 仅提供一定规模的生语料库, 基于特征词聚类的方法可以自动发现大量用户可能感兴趣的关系类型, 为研发人员提供参考, 实现方法的快速移植。尤其对于一些陌生的领域, 本章方法提供了一条快速、有效的自动发现关系类型的途径。

第3章 基于 Web Mining 的关系种子集抽取

3.1 引言

在第 2 章领域无关的实体关系类型自动发现基础上，获取的关系类型变得非常丰富，数量较大，这时完全依靠传统的人工选择具有代表性的关系种子集的方法变得愈加困难，且费时费力，可移植性差。针对该问题，本章提出了一种基于 Web Mining 的关系种子集抽取方法，旨在抽取高准确率的初始关系种子集，以构建一个高质量的实体关系核心网为研究目标。

3.2 算法流程

基于 Web Mining 的关系种子集抽取算法框架图如图 3-1 所示。

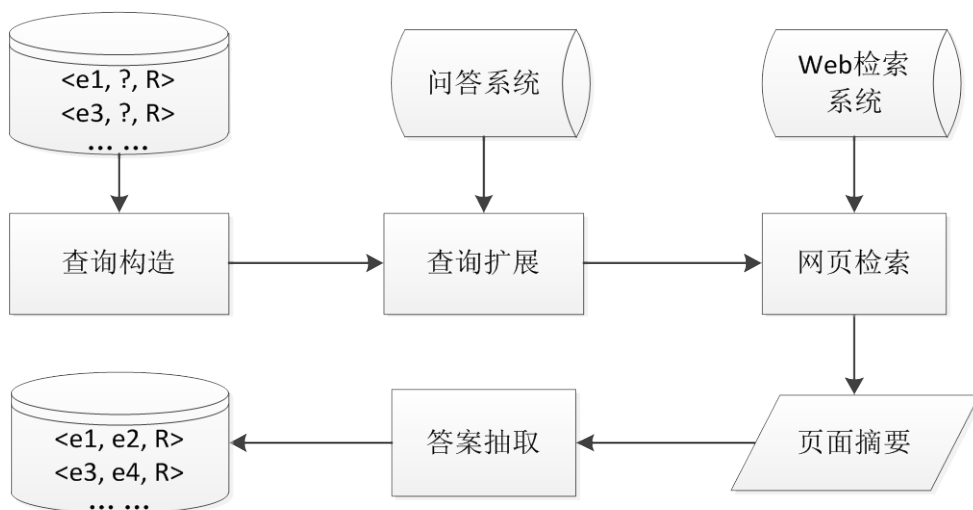


图 3-1 基于 Web Mining 的关系种子集抽取的算法

首先我们对算法中涉及的一些概念进行解释说明：

关系元组 (Relation Tuple)： 实体关系可以用三元组 $\langle e1, e2, R \rangle$ 形式化表示，其中 $e1$ 表示第一个实体， $e2$ 表示第二个实体，二者关系为 R ，关系可以是对称的和非对称的，如“朋友关系”就是对称关系，“父子关系”就是非对称关系。

关系种子 (Relation Seed)： 满足特定关系类型且具有代表性的关系元组集合，以及该关系类型对应的特征词，被用于 Bootstrapping 迭代过程的初始输入，如 $\langle \text{梁朝伟}, \text{刘嘉玲} \rangle$ 和 $\langle \text{李亚鹏}, \text{王菲} \rangle$ 都是满足“夫妻关系”的代表性实体对，“结婚”、“夫妻”、“夫妇”和“完婚”是“夫妻关系”的部分特征词，实体对和特征词组合搭配可作为关系种子。

上下文模式 (Context Pattern)： 根据关系实例抽象得到的上下文模式 (Context Pattern)，如将实体替换成槽 (SLOT)，上下文的词替换成词性，命

名实体替换成实体类型等。

如图 3-1 所示，根据第 2 章容易得到关系元组 $\langle e1, e2, R \rangle$ 中的 $e1$ 和 R ， $e1$ 是种子实体， R 为关系类型，由特征词集 $fwSet$ 组成，所以有时也写作 $\langle e1, e2, fw \rangle$ ，其中 $fw \in fwSet$ 。我们将填充 $e2$ 的问题转化为事实型（factoid）答案抽取问题，借助问答系统和 Web 检索系统进行网络挖掘，抽取得到完整的关系元组，构建实体关系核心网。具体处理流程如下：

- （1）查询构造：根据由种子实体和特征词组成的缺失的关系元组 $\langle e1, ?, R \rangle$ ，结合待填充的实体 $e2$ 类型，采用启发式规则构造基本查询；
- （2）查询扩展：借助问答系统百度知道，利用基本查询自动扩展出更加丰富、具体的查询；
- （3）网页检索：利用查询到百度网页和新闻搜索中检索，仅保留页面摘要 snippet；
- （4）答案抽取：基于频率统计和上下文模式聚类相结合的方法完成答案抽取，补全关系元组。

3.3 算法设计

基于 Web Mining 的关系描述模式挖掘算法主要包括查询构造、查询扩展和答案抽取模块，对各模块进行详细介绍。

3.3.1 查询构造

由第 2 章可以获取到种子实体，发现关系类型体系，其中每种关系类型对应一个特征词集合。每个种子实体 $e1$ 与特征词 fw 组合可以得到一个待填充的关系元组 $\langle e1, ?, fw \rangle$ ，如 $\langle \text{周杰伦}, ?, \text{父亲} \rangle$ ， $\langle \text{周杰伦}, ?, \text{老爸} \rangle$ ， $\langle \text{赵薇}, ?, \text{饰演} \rangle$ ， $\langle \text{赵薇}, ?, \text{出演} \rangle$ 等。

针对不同词性的特征词（名词或动词），和待填充的实体类型，按照不同的规则进行基本查询构造。以人物社会关系为例，查询构造规则如下：

（1）名词性特征词的查询构造规则

- $e1 + \text{“”} + fw$ ，例如：周杰伦 父亲，周杰伦 老爸
- $fw + \text{“”} + e1$ ，例如：父亲 周杰伦，老爸 周杰伦
- $e1 + \text{的} + fw$ ，例如：周杰伦的父亲，周杰伦的老爸
- $e1 + fw + \text{是谁?}$ ，例如：周杰伦的父亲是谁？，周杰伦的老爸是谁？
- $\text{谁是} + e1 + \text{的} + fw?$ ，例如：谁是周杰伦的父亲？，谁是周杰伦的老爸？

（2）动词性特征词的查询构造规则

- $e1 + \text{“”} + fw$ ，例如：赵薇 饰演，赵薇 出演

- fw + “ ” + e1, 例如: 饰演 赵薇, 出演 赵薇

3.3.2 查询扩展

我们尝试在基本查询集基础上, 借助全球最大的基于搜索的中文互动式知识问答分享平台: 百度知道¹⁴实现查询扩展。百度知道与搜索引擎结合, 旨在将用户头脑中的隐性知识转化为互联网上的显性知识, 是用户生成内容 (User Generated Content) 的代表性产品。在该平台上, 用户可以根据需求有针对性地提出问题, 通过一定的激励制度吸引其他用户回答, 分享知识和经验, 并且, 这些问题及答案数据又可以用于提供搜索服务, 根据当前查询问题, 给出相似问题推荐, 实现问答知识的分享。截止到 2011 年 06 月 03 日, 百度知道已解决问题达 136,157,55 个, 待解决问题达 2,804,637 个, 涵盖了文化、艺术、娱乐、健康、商业、生活、社会等众多领域。



图 3-2 “周杰伦 父亲”百度知道检索结果

我们使用 e1、fw 搭配得到的查询在百度知道中检索, 可以得到按照相关

¹⁴ <http://zhidao.baidu.com/>

性排序的问句集，如图 3-2 所示为“周杰伦 父亲”百度知道检索结果页面。通过观察大量数据，我们发现排序靠前的问句与我们的查询往往意图一致，这样我们选取 Top-5 的结果直接作为扩展出来的查询，如“*周杰伦的父亲是谁*”，“*周杰伦的父母是谁*”，“*周杰伦的爸爸叫什么？*”，等等。另外，进入单个问句页面，百度知道给出了最多 5 个查询意图相似的“相关内容”推荐结果，如图 3-3 所示为问句“*周杰伦的父亲*”相关内容结果，我们将这些相关问题也作为扩展查询的一部分，这样就可以最多扩展出 25 个查询（可能存在重复）。

为了保证准确率，我们过滤掉那些不含当前关系类型的任一特征词的扩展查询，并且对于重复的扩展查询增加其重要性。这样对于每种关系类型对应的一组关系元组可以得到大量的查询。

 相关内容	
周杰伦的父亲是谁	2011-4-23
周杰伦的父亲周耀中（宗）到底是物理老师还是生物老师？  7	2010-11-14
周杰伦的父亲叫什么名字？  10	2010-8-20
周杰伦不父亲叫什么名字？	2010-6-28
周杰伦的父亲叫什么	2010-6-27
更多关于周杰伦 父亲的问题>>	
查看同主题问题： 周杰伦 父亲	

图 3-3 问句“周杰伦的父亲”的相关内容

3.3.3 网页检索

我们将基于启发式通用规则构造和扩展得到的查询按照不同的类别和重要性进行检索，具体如下：

- （1）基本查询：这类查询准确率较高，我们分别将每个查询在百度网页和新闻搜索中检索，保留前 50 条页面摘要，另外，在百度知道检索中检索，保留前 10 条页面摘要；
- （2）扩展查询：一方面保留扩展查询对应的最佳答案，另一方面，将扩展查询在百度网页和新闻搜索中检索，对于重复出现的扩展查询，保留前 20 的页面摘要，对于出现一次的扩展查询，仅保留前 10 条页面摘要。

针对每个关系元组的查询集合，通过上述的检索过程，可以获取相应的文本集合，然后，对这些文本进行断句和句子去重，接着，对保留下来的句子进行中文分词、词性标注和命名实体识别等自然语言处理，并保存成 DOM 结构形式，为下阶段提供数据支持。

3.3.4 答案抽取

针对页面摘要的特点，我们主要尝试了三种答案抽取解决方案，第一种为基于频率统计的答案抽取，作为基准方法 **baseline**；第二种为基于上下文模式的答案抽取；第三种是基于频率统计和上下文模式相结合的方法。

(1) 基于频率统计的答案抽取

作为基准方法 **baseline**，算法以关系元组为单位，根据待填充实体 e_2 的类型，在 3.3.3 获取的句子集中统计该类型的实体，即候选答案的出现频率，并进行。则实体 e_2 对应的每个候选答案 e_{2_i} 可信度 $Conf_{fq}(e_{2_i})$ 如公式 (3-1) 所示。

$$Conf_{fq}(e_{2_i}) = Freq(e_{2_i}) \quad (3-1)$$

最终实体 e_2 取可信度最大且大于阈值 $minFreq$ 的候选答案（可能无解），如公式 (3-2) 所示。

$$e_2 = \underset{e_{2_i}}{argmax}\{Conf_{fq}(e_{2_i}) \mid Conf_{fq}(e_{2_i}) > minFreq\} \quad (3-2)$$

如关系元组<周杰伦，？，父母—子女>对应的候选答案排序列表如表 3-1 所示，最终抽取答案为“周耀宗”。

表 3-1 关系元组<周杰伦，？，父母—子女>候选答案排序列表

候选答案	频率
周耀宗	638
叶惠美	517
周耀中	279
周润发	145
周耀	79
...	...

(2) 基于上下文模式的答案抽取

针对特定关系类型 R ，我们将所有关系元组检索得到的句子合并在一起，仅保留同时包含实体 e_1 、关系类型 R 中任一特征词，且至少存在一个与 e_2 相同类型实体（即候选答案）的句子，用于生成上下文模式。

首先，对每个句子进行裁剪，得到实体对两侧有限的上下文信息，得到候选上下文模式 $candCP$ ，裁剪规则如公式 (3-3) ~ (3-7) 所示。

$$candCP = \begin{cases} E[\min\{L_{fw}, L_{lc}\} : L_{max}] + E[L_{max} : L_{rc} + 1], & \text{if } L_{fw} < L_{min} \\ E[L_{lc} : L_{min}] + E[L_{min} : L_{max}] + E[L_{max} : L_{rc} + 1], & \text{if } L_{min} < L_{fw} < L_{max} \\ E[L_{lc} : L_{max}] + E[L_{max} : \max\{L_{fw}, L_{rc}\} + 1], & \text{if } L_{max} < L_{fw} \end{cases} \quad (3-3)$$

$$\text{其中, } L_{\min} = \min\{L_{e1}, L_{e2}\} \quad (3-4)$$

$$L_{\max} = \max\{L_{e1}, L_{e2}\} \quad (3-5)$$

$$L_{lc} = \max\{0, L_{\min} - cL\} \quad (3-6)$$

$$L_{rc} = \min\{L_{\text{end}}, L_{\max} + cL\} \quad (3-7)$$

L_{e1} 表示实体 $e1$ 在句子中的位置； L_{e2} 表示候选实体 $e2$ 在句子中的位置； L_{fw} 表示特征词 fw 在句子中的位置； cL 表示模式上下文信息的长度； L_{end} 表示句子最后一个词的位置； L_{lc} 表示上下文模式左侧最远词的位置； L_{rc} 表示上下文模式右侧最远词的位置； $E[L_i:L_j]$ 是左闭右开区间，表示句子中 L_i 到 L_j 位置之间所有的词； $E[L_i:L_j]$ 之间的“+”表示区间的拼接，即将词串连接在一起。

在候选上下文模式 $candCP$ 基础上，我们使用[**SLOT1**]和[**SLOT2**]替换已知实体 $e1$ 和候选实体 $e2$ ，然后依据词法信息进行规则泛化，得到上下文模式 cP 。具体的，词法规则包括：

- 1) 除 $e1$ 和 $e2$ 之外的其他命名实体，仅保留命名实体类型；
- 2) 对于数词和代词，仅保留其词性；
- 3) 对于其他词性，同时保留词和词性信息。

如句子“权威媒体 *TVBS* 娱乐记者正式对外公布了周杰伦的父亲周耀中的一篇关于杰伦身世之谜的博客文章。”。经过对关系实例裁剪，得到候选上下文模式如下所示，其中 cL 取值为 3。

对外/v 公布/v 了/u 周杰伦/Nh 的/u 父亲/n 周耀中/Nh 的/u 一篇/Nm 关于/p

经过词法规则泛化后，得到最终的上下文模式为：

对外/v 公布/v 了/u [SLOT1]/Nh 的/u 父亲/n [SLOT2]/Nh 的/u /Nm 关于/p

统计上下文模式出现频率 $Freq(cP)$ ，每个关系元组 $\langle e1, e2, R \rangle$ 的候选答案 $e2_i$ 均对应一组上下文模式 $cPSet(e2_i)$ ，如公式 (3-8) 所示。

$$\begin{aligned}
 \langle e1, ?, R \rangle \left\{ \begin{array}{l} \langle e1, e2_1, R \rangle \left\{ \begin{array}{l} \langle cP_{11}, Freq(cP_{11}) \rangle \\ \langle cP_{12}, Freq(cP_{12}) \rangle \\ \dots\dots \\ \langle cP_{1p}, Freq(cP_{1p}) \rangle \end{array} \right. \\ \langle e1, e2_2, R \rangle \left\{ \begin{array}{l} \langle cP_{21}, Freq(cP_{21}) \rangle \\ \langle cP_{22}, Freq(cP_{22}) \rangle \\ \dots\dots \\ \langle cP_{2q}, Freq(cP_{2q}) \rangle \end{array} \right. \\ \dots\dots \\ \langle e1, e2_n, R \rangle \left\{ \begin{array}{l} \langle cP_{n1}, Freq(cP_{n1}) \rangle \\ \langle cP_{n2}, Freq(cP_{n2}) \rangle \\ \dots\dots \\ \langle cP_{nm}, Freq(cP_{nm}) \rangle \end{array} \right. \end{array} \right. \quad (3-8)
 \end{aligned}$$

其中，关系元组 $\langle e1, e2_i, R \rangle$ 为合法的候选元组， $\langle cP_{ik}, Freq(cP_{ik}) \rangle$ 为 $e2_i$ 对应的句子集生成的模式，及模式的统计频率。

定义候选答案 $e2_i$ 的可信度 $Conf_{cp}(e2_i)$ 如公式 (3-9) 所示。

$$Conf_{cp}(e2_i) = \sum_{cP_{ij} \in cPSet(e2_i)} Freq(cP_{ij}) \quad (3-9)$$

最终实体 $e2$ 取可信度最大且大于阈值 $minCP$ 的候选答案（可能无解），如公式 (3-10) 所示。

$$e2 = \underset{e2_i}{argmax}\{Conf_{cp}(e2_i) \mid Conf_{cp}(e2_i) > minCP\} \quad (3-10)$$

(3) 基于频率统计和上下文模式相结合的答案抽取

为了综合考虑单个关系元组对应页面摘要中候选答案的频率信息和特定关系类型下所有关系元组对应页面摘要的句子相似信息，我们在公式 (3-1) 和 (3-9) 基础上利用线性插值方法定义了新的候选答案可信度计算公式，如公式 (3-11) 所示。

$$Conf(e2_i) = w \cdot Conf_{fq}(e2_i) + (1 - w) \cdot Conf_{cp}(e2_i) \quad (3-11)$$

其中， $0 \leq w \leq 1$ 为影响因子，表示频率统计和上下文模式两种方法的重要程度。由于上下文模式引入了更多的约束条件，且考虑了其他关系元组检索句子集，准确率较高，可以赋予较高的权重。

最后，选取可信度最大的实体作为关系元组答案，如公式 (3-12) 所示。

$$\begin{aligned}
 e2 = \underset{e2_i}{argmax}\{Conf(e2_i) \mid Conf_{fq}(e2_i) > minFreq \\ \&\& Conf_{cp}(e2_i) > minCP\} \quad (3-12)
 \end{aligned}$$

3.4 实验结果与分析

3.4.1 实验数据

实验在实体对类型“人名—人名”上进行，种子实体和关系类型特征词可以在第 2 章工作基础上获得。我们选取了 Top-500 的种子实体和 9 种关系类型进行详细实验，主要包括夫妻关系、经纪人关系、合作关系、情侣关系、父母—子女、好友关系、角色扮演、兄弟姐妹和伯乐关系。

3.4.2 评价标准

由于本章的目的是自动挖掘关系种子集，抽取实体关系核心网，对关系种子的数量没有特别需求，在保证准确率的情况下，尽可能增加抽取数量即可。所以，本章仅对准确率进行评价，并且，如果某关系类型填充关系元组数目大于 100，随机选取 100 个进行人工评价，否则，将对全部的关系元组进行人工评价，而召回率通过抽取总数间接反映，准确率定义如公式（3-13）所示。

$$Precision = \frac{\text{某类抽取正确的元组数}}{\text{某类抽取的所有元组数}} \quad (3-13)$$

另外，我们定义了平均准确率作为各类综合性能的最终评测，设 N_i 为第 i 类关系类型填充关系元组个数， p_i 为第 i 类关系类型抽取元组的准确率，平均正确率 P_{avg} 定义如公式（3-14）所示。

$$P_{avg} = \frac{\sum_{i=1}^n (N_i \times p_i)}{\sum_{i=1}^n N_i} \quad (3-14)$$

3.4.3 结果与分析

我们将基于频率统计的答案抽取方法作为基准方法，用 **baseline** 表示，基于上下文模式的答案抽取方法，用 **Pattern** 表示，以及二者相结合的方法，用 **Combine** 表示。算法中定义了 4 个需要调整的参数，我们通过观察统计一定数量的真实文本，将它们设置成如表 3-2 所示的经验值。

表 3-2 实验参数设置说明

参数	取值	描述
cL	3	实体对两侧所取上下文长度（3.3.4）
$minFreq$	100	候选答案在页面摘要中出现频率的最小值（3.3.4）
$minCP$	20	候选答案对应上下文模式总和的最小值（3.3.4）
w	0.35	频率统计方法对最终结果的影响因子（3.3.4）

对于不同关系类型的关系元组抽取评价结果如表 3-3 所示。

表 3-3 关系元组抽取评价结果

关系类型	算法	关系元组	正确?		准确率 (%)
			是	否	
夫妻关系	Baseline	173	51	49	51.00
	Pattern	107	73	27	73.00
	Combine	74	64	10	86.49
经纪人关系	Baseline	128	62	38	62.00
	Pattern	73	69	4	93.32
	Combine	67	64	3	95.55
合作关系	Baseline	215	58	42	58.00
	Pattern	132	92	8	92.00
	Combine	121	94	6	94.00
情侣关系	Baseline	223	46	54	46.00
	Pattern	142	83	17	83.00
	Combine	137	82	18	82.00
父母—子女	Baseline	251	77	23	77.00
	Pattern	154	92	8	92.00
	Combine	121	96	4	96.00
好友关系	Baseline	123	75	25	75.00
	Pattern	94	93	1	98.93
	Combine	114	98	2	98.00
角色扮演	Baseline	204	48	52	48.00
	Pattern	97	89	8	91.75
	Combine	96	90	6	93.75
兄弟姐妹	Baseline	103	56	44	56.00
	Pattern	71	63	8	88.73
	Combine	68	62	6	91.30
伯乐关系	Baseline	64	38	26	59.37
	Pattern	46	39	7	78.88
	Combine	49	41	8	83.67
平均	Baseline	1484	906	578	61.05
	Pattern	916	811	105	88.53
	Combine	847	770	77	90.91

由表 3-3 易知，基准方法 Baseline 整体效果最差，抽取元组数目较多，但平均准确率较低，为 61.05%，这主要因为频率统计的方法可以有效的抽取出关系元组，尤其是那些歧义性较少的元组，抽取性能较好，但是，因为没有过多的上下文信息加以约束限制，噪声较大。Pattern 和 Combine 方法平均准确率都得到了大幅度提高，分别达到了 88.53%和 90.91%，但是获取的关系元组数目相对 baseline 有不同程度的降低。分析其原因，Pattern 进行了较强的上下文限定，抽取元组较少，但是容易得到高准确率的结果。另外，上下文模式的

答案抽取方法可以增加频率统计中候选答案实体的区分度，具有一定的互补作用，可以进一步提高抽取性能。综合以上结果，如果实体对的确存在某种关系，通过构造查询到搜索引擎中检索，可以获取许多相关文本信息（问答对和网页文本），尤其对于我们处理的高频的种子实体，网上存在充足的相关信息用于统计挖掘，保证了答案抽取方法的可行性和有效性。

为了评价查询扩展的必要性，我们给出了如表 3-4 所示增加扩展查询前后对实验结果的影响，其中，准确率代表 9 种关系类型的平均准确率。

表 3-4 显示，Baseline 所使用频率统计的方法，单纯使用基本查询情况可以得到更高的准确率，究其原因在于扩展查询在扩大查询种类，检索更多的统计文本同时，也引入了一定的噪声，而频率统计的方法过于简单，没有任何的上下文限定，反而影响了抽取结果。而 Pattern 的方法在增加扩展查询后得到了较大幅度的提高，达到了 88.53%，提高 5 个百分点，说明扩展查询对于答案的抽取存在较大的贡献，我们可以通过上下文抽取模式加以限制利用，保证准确率。融合的方法进一步提高了准确率，也同样说明了查询扩展的有效性和必要性。

表 3-4 查询扩展对实验结果的影响

算法	查询来源	准确率(%)
Baseline	基本查询	64.76
	基本查询+扩展查询	61.05
Pattern	基本查询	83.09
	基本查询+扩展查询	88.53
Combine	基本查询	87.50
	基本查询+扩展查询	90.91

3.5 本章小结

本章在第 2 章关系类型发现结果基础上，提出了基于 Web Mining 的关系种子集抽取方法，算法将关系元组抽取问题转化为事实型答案抽取问题，我们基于种子实体和特征词，通过简单的启发式通用规则进行基本查询的构造，并利用社区化知识问答平台：百度知道进行查询的扩展，在此基础上，充分利用搜索引擎收集和处理海量信息的能力，检索得到大量可能包含候选答案的页面摘要，接下来，综合使用频率统计和上下文模式挖掘的方法抽取答案，填充关系元组。最后，通过在 9 种关系类型上展开实验，得到了 90.91%的关系种子抽取准确率，且每种关系类型平均抽取了超过 90 的元组数量，为后续工作提供可靠的数据支持。

第4章 基于 Bootstrapping 的关系描述模式挖掘

4.1 引言

第 2 章通过对高频实体对上下文中的特征词聚类，实现关系类型的自动发现，即由特征词描述实体关系类型，第 3 章利用搜索引擎收集和处理大规模真实数据的能力和优势，基于 Web Mining 的方法抽取少量高准确率的初始关系种子集。本章将关系种子集（包括实体对和关系类型对应的特征词）作为输入，使用 Bootstrapping 方法从大规模语料库中迭代地挖掘关系描述模式，并抽取关系元组。根据中文语言特点，本章还在上下文模式基础上引入了泛化处理，挖掘关系类型对应的软模式集，在保证高准确率的同时，尽可能提高召回率。

4.2 算法流程

基于 Bootstrapping 的关系描述模式挖掘算法框架图如图 4-1 所示。

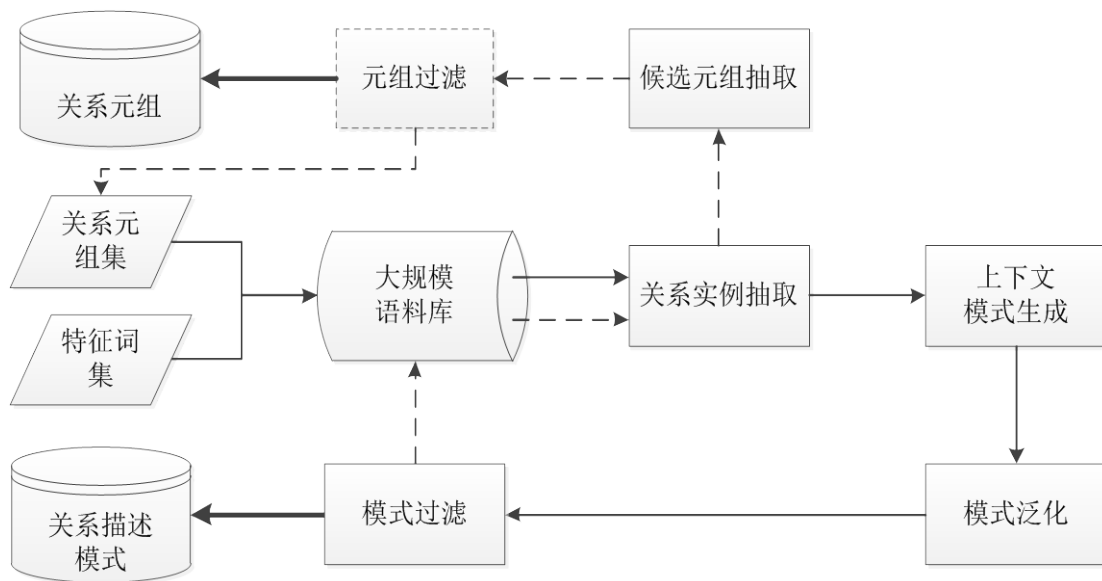


图 4-1 基于 Bootstrapping 的关系描述模式挖掘算法

首先我们对算法中涉及的一些概念进行解释说明：

关系实例 (Relation Instance)：包含特定关系类型实体对的句子，如句子“梁朝伟、刘嘉玲自本月 21 日于不丹正式结婚后，24 日早上首次公开露面。”为关系元组<梁朝伟，刘嘉玲，结婚>的一个关系实例。一个关系元组可以对应一个或多个关系实例。

软模式 (Soft Pattern)：为了提高召回率，对上下文模式进一步泛化得到的模式，详细见 4.3.3。

关系描述模式 (Relation Pattern)：根据关系实例抽象得到的用于抽取新

的关系元组的通用模式，常见的关系描述模式有上下文模式（Context Pattern）、句法树模式（Tree Pattern）和特征向量模式（Vector Pattern），本文主要使用裁剪的上下文模式，以及基于上下文模式泛化的软模式。

如图 4-1 所示，针对每种关系类型，本算法以关系特征词和少量高质量的初始关系元组集作为输入，以从大规模语料库中抽取关系描述模式和识别关系实例为目标，主要包括关系描述模式挖掘（实线）和关系元组抽取（虚线）两大主要模块。我们以“夫妻关系”为例，关系种子为<梁朝伟，刘嘉玲，结婚|完婚>，具体步骤如下：

- (1) 关系实例抽取：利用关系种子到大规模语料库中检索，根据字符串匹配技术，抽取同时包含实体对及至少一个特征词，并满足一系列启发式约束规则的句子，如句子“梁朝伟和刘嘉玲自本月 21 日于不丹正式结婚后，24 日早上首次公开露面。”，“梁朝伟与刘嘉玲 7 月 21 日将于不丹完婚。”；
- (2) 上下文模式生成：对句子进行裁剪划分，将句子片段转换成上下文模式，主要操作有将关系元组的实体替换成槽（SLOT），其他命名实体替换成实体类型，增加实词在《同义词词林（扩展版）》中的语义代码，代词、数词等替换成词性，等等，如上面的两个句子其上下文模式为：“[SLOT1]/Nh 和/p [SLOT2]/Nh 自/p /Nr 于/p /Ns 正式/a/Ed53A 结婚/v/Hj51C”和“[SLOT1]/Nh 与/p [SLOT2]/Nh /Nr 将/d 于/p /Ns 完婚/v/Hj51C 。/wp”；
- (3) 模式泛化：根据句子间相似度进一步泛化，抽取软模式，其中软模式必须包含两个槽和一个特征词，如由上面两个上下文模式泛化得到软模式：“[SLOT1]/Nh * [SLOT2]/Nh /Nr 于/p /Ns* 结婚/v/Hj51C”；
- (4) 候选元组抽取：使用软模式检索语料库，获得大量匹配模式的句子，根据软模式中的槽从句子中识别出新的关系元组，如从句子“李亚鹏和王菲在乌鲁木齐正式登记结婚。”识别出关系元组<李亚鹏，王菲，结婚|完婚>；
- (5) 候选元组评价与过滤：此过程可选，采用一定的策略对新识别的关系元组进行评价和过滤，将可信度较高的关系元组与特征词组合添加到初始关系种子集中，完成一轮迭代；
- (6) 重复执行（1）~（5），直到没有新的关系元组产生，或者不再产生新的模式，或者达到预定义的迭代次数阈值后，算法结束。

算法终止后，针对每种关系类型可以获得对应的关系描述模式集和关系元组知识库。

4.3 算法设计

基于 Bootstrapping 的关系描述模式挖掘算法主要包括四个功能模块，分别为关系实例抽取、上下文模式生成、模式泛化与过滤以及元组抽取与过滤。

4.3.1 关系实例抽取

关系实例抽取主要包括两部分，第一部分是根据关系种子匹配获取对应的关系实例，第二部分是根据关系描述模式匹配获取对应的关系实例，用于抽取关系元组。

对于第一部分，我们将关系种子表示成三元组 $\langle e1, e2, R \rangle$ ，其中 $e1$ 和 $e2$ 是满足关系类型 R 的两个实体，而 R 由特征词集 $fwSet$ 构成，则抽取过程为：

- (1) $e1, e2$ 与特征词 $fw \in fwSet$ 组合搭配得到具体的三元组，将三个元素转化为 Query 从 RE100W 正文索引库中检索出对应的网页集合；
- (2) 对网页正文进行断句，使用字符串匹配技术，保留那些同时包含三元组的句子，组成候选关系实例集；
- (3) 对候选关系实例进行中文分词、词性标注和命名实体识别，根据下列启发式约束条件对候选关系实例进一步过滤，得到最终关系实例集合；
 - 两个实体的词距离小于等于 $maxE1ToE2Distance$ ；
 - 两个实体之间不能包含多于 $maxInsertedEntity$ 个相同类型的实体；
 - 两个实体之间包含的标点符号不能多于 $maxPunc$ 个；
 - 特征词在实体对两侧时，与距离最近的实体之间的词距离小于等于 $maxEToFWDistance$ ；
 - 特征词在实体对两侧时，与距离最近的实体之间包含的标点符号不能多于 $maxInsertedPunc$ 个。

上述条件认为产生关系的两个实体之间，以及与特征词之间，往往出现在一定空间范围内，而距离较远的实体对产生关系的可能性较小，即使产生关系也过于特殊化。

假设待处理的关系类型为“夫妻关系”，其特征词为“结婚”和“完婚”，实体对为 $\langle \text{梁朝伟}, \text{刘嘉玲} \rangle$ ，抽取的关系实例如句子“梁朝伟和刘嘉玲自本月 21 日于不丹正式结婚后，24 日早上首次公开露面。”和“梁朝伟和刘嘉玲 7 月 21 日将于不丹完婚。”。

对于第二部分，将在 4.3.4 中详细介绍。

4.3.2 上下文模式生成

上下文模式以单个关系实例为单位生成，算法伪代码如图 4-2 所示。

```

For each relation type  $R$ 
  Let  $S(R)$  be the set of relation seed
  For each  $s \in S(R)$  containing a word  $fw$  from  $fwSet(R)$ 
    For each entity pair  $e1, e2 \in s$ , such that  $e1 \neq e2$ 
      and  $Type(e1) = Type(R.e1)$ , and  $Type(e2) = Type(R.e2)$ 
      Let  $candCP := Prune(s)$ 
      Let  $candCP' = candCP$  with  $e1$  changed to  $[SLOT1]$ ,
        and  $e2$  changed to  $[SLOT2]$ .
      Let  $cP := LexicalGeneralize(candCP')$ 
      Add  $cP$  to  $contextPatternSet(R)$ 

```

图 4-2 ContextPattern 生成算法伪代码

图 4-2 中 $Prune(s)$ 对关系实例（已经过中文分词、词性标注和命名实体识别处理）裁剪划分，此处，对单个句子的裁剪，我们使用了与 3.3.4 中基于上下文模式的答案抽取介绍的裁剪规则相同，不再赘述。

在候选上下文模式基础上，我们使用 $[SLOT1]$ 和 $[SLOT2]$ 替换实体 $e1$ 和 $e2$ ，然后依据词法信息使用一系列启发式规则进行简单泛化，在抽象性和特殊性之间进行一定的权衡，得到上下文模式 cP ，即图 4-2 中的 $LexicalGeneralize(candCP')$ 。具体的，词法规则包括：

- 1) 除 $e1$ 和 $e2$ 之外的其他命名实体，仅保留命名实体类型；
- 2) 对于实词（我们规定实词主要包括名词、动词、形容词和名词修饰语），增加其在《同义词词林（扩展版）》中的语义代码；
- 3) 对于数词和代词，仅保留其词性；
- 4) 对于其他词性，同时保留词和词性信息。

如句子“梁朝伟、刘嘉玲自本月 21 日于不丹正式结婚后，24 日早上首次公开露面。”和“梁朝伟与刘嘉玲 7 月 21 日将于不丹完婚。”。经过对关系实例裁剪后可以得到候选上下文模式如下所示。

```

梁朝伟/Nh 和/c 刘嘉玲/Nh 自/p 本月 21 日/Nr 于/p 不丹/Ns 正式/a 结婚/v
梁朝伟/Nh 和/c 刘嘉玲/Nh 7 月 21 日/Nr 将/d 于/p 不丹/Ns 完婚/v 。 /wp

```

经过词法规则泛化后，得到最终的上下文模式如下所示。

```

[SLOT1]/Nh 和/c [SLOT2]/Nh 自/p /Nr 于/p /Ns 正式/a/Ed53A 结婚/v/Hj51C
[SLOT1]/Nh 和/c [SLOT2]/Nh /Nr 将/d 于/p /Ns 完婚/v/Hj51C 。 /wp

```

4.3.3 模式泛化与过滤

为了进一步提高召回率，我们在上下文模式基础上，借鉴 URES 系统提出的 Best Match 模式泛化算法^[33]，结合中文语言学特点，提出了基于上下文模式对泛化的软模式 SoftPattern 生成策略，算法伪代码如下所示。

```

For each relation type  $R$ 
  For each context pattern pair  $cP_i, cP_j$  from  $contextPatternSet(R)$ 
    if  $posCompare(cP_i.e1, cP_i.e2) = posCompare(cP_j.e1, cP_j.e2)$  then
      Let  $matchLen := BestMatch(cP_i, cP_j)$ 
      Let  $J := Jaccard(cP_i, cP_j, matchLen)$ 
      if  $J \geq minJaccard$  then
        Let  $softPattern := Generalize(cP_i, cP_j)$ 
        Add  $softPattern$  to  $softPatternSet(R)$ .
  FilterSoftPattern( $softPatternSet(R)$ )
    
```

图 4-3 SoftPattern 生成算法伪代码

给定关系类型 R ，可以结合使用函数 $BestMatch(cP_i, cP_j)$ 和 $Generalize(cP_i, cP_j)$ 对任意两个实体对位置一致 ($posCompare(cP_i.e1, cP_i.e2) = posCompare(cP_j.e1, cP_j.e2)$ 保证该条件) 的上下文模式进行泛化处理，生成软模式 $softpattern$ ，软模式主要包括以下几部分：

- 槽 SLOT：包括[SLOT1]和[SLOT2]，同时保留实体类别；
- 其他命名实体：句子中涉及到的除实体对之外的其他命名实体仅用实体类型表示；
- 实词词法单元：同时保留词、词性及词在《同义词词林（扩展版）》中的语义代码（可能有多个）；
- 词法单元：主要由词、词性和语义代码三部分组成，其中词、语义代码在某些匹配情况下可以省略；
- 省略单元：用“*”表示，代表可省略的连续词法单元，泛化程度最大。

我们将基于上下文模式对求解软模式问题转化为最长公共子序列 (Longest Common Subsequence, LCS) 问题，这样以来，可以使用动态规划算法实现。传统的 LCS 是基于完全匹配求最长匹配公共子序列，而我们规定上下文模式单元之间可以进行模糊匹配，从而增加了匹配代价的概念，旨在获取最佳匹配公共子序列，得到了改进的 BestMatch 算法。

BestMatch 算法中针对不同的单元定义了不同的匹配代价，且分别对应软模式生成策略，具体如下：

- 如果待匹配单元完全相同， $cost=0$ ，生成软模式中的槽、其他实体类型或词法单元，均保持匹配单元原始状态；
- 如果待匹配单元均含有语义代码，且两个语义代码有交集，认为模糊匹配， $cost=5$ ，生成软模式中的词法单元，保留第一个匹配单元对应的词、词性及相同的语义代码；
- 如果待匹配单元词性相同， $cost=8$ ，生成软模式中的词法单元，仅保留词性信息；
- 如果待匹配单元完全不匹配， $cost=10$ ，生成软模式中的省略单元“*”。

根据匹配过程，不难发现，最终累加匹配代价 $cost$ 越小，说明上下文模式 cP_i 和 cP_j 匹配度越大，反之，匹配度越小。

为了更加直观地理解算法，我们对 4.3.2 生成的上下文模式 “[*SLOT1*]/Nh 和/c [*SLOT2*]/Nh 自/p /Nr 于/p /Ns 正式/a/Ed53A 结婚/v/Hj51C ” 和 “[*SLOT1*]/Nh 和/c [*SLOT2*]/Nh /Nr 将/d 于/p /Ns 完婚/v/Hj51C 。/wp” 执行 BestMatch 算法，具体匹配过程如表 4-1 所示，累加匹配代价为 65。

表 4-1 BestMatch 算法示例

cP_i	cP_j	代价
[<i>SLOT1</i>]/Nh	[<i>SLOT1</i>]/Nh	0
和/c	和/c	0
[<i>SLOT2</i>]/Nh	[<i>SLOT2</i>]/Nh	0
自/p		10
/Nr	/Nr	0
	将/d	10
于/p	于/p	0
/Ns	/Ns	0
正式/a/Ed53A		10
结婚/v/Hj51C	完婚/v/Hj51C	5
	。/wp	10

接下来，我们借助 Jaccard 系数判断上下文模式之间的相似度，这里将上下文模式看作集合，最佳匹配单元个数（即匹配代价小于 10 的匹配单元个数）看作集合之间的交集，则上下文模式 cP_i 和 cP_j 的 Jaccard 系数为 $J(cP_i, cP_j)$ ，定义如公式（4-1）所示。

$$J(cP_i, cP_j) = \frac{BestMatch(cP_i, cP_j)}{Length(cP_i) + Length(cP_j) - BestMatch(cP_i, cP_j)} \quad (4-1)$$

其中， $Length(cP_i)$ 表示上下文模式 cP_i 的单元个数， $BestMatch(cP_i, cP_j)$ 表示

上下文模式 cP_i 和 cP_j 最佳匹配长度。我们设定相似度阈值 $minJaccard$ ，决定上下文模式 cP_i 和 cP_j 是否可以用于生成有效的软模式，根据表 4-1 计算 $J=7/(9+9-7)=7/11=0.636$ 。如果 $J(cP_i, cP_j)$ 大于等于 $minJaccard$ ，可以仿照 LCS 构造最长公共子序列的方法构造软模式，根据表 4-1 可以得到软模式如下所示。

[SLOT1]/Nh 和/c [SLOT2]/Nh */Nr * 于/p /Ns * 结婚/v/Hj51C *

对任意两个上下文模式处理后，可以生成软模式集 $softPatternSet(R)$ ，通过观察发现，有些软模式过于泛化，包含大量噪声。所以，我们又采取了一系列过滤与合并操作，保证关系描述模式的质量。

- 1) 在模式开始和结尾各增加一个省略单元“*”，增加泛化能力；
- 2) 如果停用词两边均为“*”，删除该停用词，因为此时停用词没有限定能力，却影响召回率；
- 3) 合并连续的“*”，便于模式归一化；
- 4) 删除[SLOT1]或[SLOT2]两边均为“*”的模式，避免模式过于泛化，影响准确率。

经过模式的过滤和合并后，对于上例可以得到如下的最终软模式。

*[SLOT1]/Nh 和/c [SLOT2]/Nh */Nr * 于/p */Ns * 结婚/v/Hj51C *

4.3.4 元组抽取与评价

检索出所有包含 $fw \in fwSet(R)$ 的句子集，根据句子进行关系实例转化，得到可能包含关系元组的上下文片段，然后使用当前迭代过程新生成的模式与关系实例匹配抽取关系元组，对元组过滤后，保留下来的新关系元组即为关系种子。以模式 “*[SLOT1]/Nh 和/c [SLOT2]/Nh */Nr/ * 于/p /Ns * 结婚/v/Hj51C *” 为例，具体关系元组抽取过程如下：

- 1) 根据“结婚”检索得到句子“李亚鹏和王菲昨日于乌鲁木齐正式登记结婚，两人爱情终于修得正果！”；
- 2) 对句子进行中文分词、词性标注、命名实体识别及上下文裁剪后，识别候选关系元组<李亚鹏，王菲>，得到上下文模式“李亚鹏/Nh 和/c 王菲/Nh 昨日/Nt 于/p 乌鲁木齐/Ns 正式/a/Ed53A 登记/v/Hc15A 结婚/v/Hj51C , /wp”；
- 3) 使用 BestMatch 对上下文模式和软模式进行匹配度计算，为了保证准确率，限制省略单元“*”最多可匹配 $maxTokenMatch$ 个其他命名实体与词法单元；

- 4) 如果软模式中每个非省略单元在上下文模式中都可以找到对应的模糊匹配单元，且相对位置一致，则抽取关系元组。

根据上面的例子，可以抽取出关系元组<李亚鹏，王菲>。对所有句子使用同样的处理方法，最终可以抽取出更多关系元组。

由于网页文本的多样性，关系元组中不可避免的存在噪声，而新的关系元组要作为下一轮迭代过程的种子，噪声元组容易导致错误蔓延，产生循环依赖现象。为此，我们对关系元组进行可信度评价，根据可信度决定关系元组是否进入下一轮迭代。具体的，关系元组可信度计算公式定义如下。

$$Conf(T) = \frac{\sum_{fw_i \in fwSet(R)} OccSentence(T + fw_i)}{OccSentence(T) + 1} \quad (4-2)$$

其中， T 为待评价关系元组， $fwSet(R)$ 为关系类型 R 对应的特征词集， $OccSentence(T)$ 表示语料库中包含关系元组 T 的句子数， $OccSentence(T + fw_i)$ 表示语料库中同时包含关系元组 T 和至少一个特征词的句子数。

可信度公式 $Conf(T)$ 反映了关系元组 T 与关系类型 R 的互信息， $Conf(T)$ 越大，说明关系元组与特征词共现概率越大，元组 T 属于关系 R 的可能性越大。当 $Conf(T)$ 大于特定阈值 $minTupleConf$ 时，进入下一轮迭代过程，而可信度较低的关系元组，与特征词共现呈现一定的不确定性，这样可以有效的提高关系元组的质量，尽可能避免迭代过程中的循环依赖现象，保证迭代过程的准确性。

4.4 实验结果与分析

4.4.1 实验数据

实验主要在中文文本上进行，我们使用第 2 章获取的 RE100W 娱乐资讯网页作为大规模语料库。由于不同关系类型在互联网文本中的丰富程度不同，可能得到不同的实验结果，我们选取“人名—人名 (Nh-Nh)”实体对类型下的 9 种自动发现的关系类型作为实验对象，分别为夫妻关系、经纪人关系、合作关系、情侣关系、父母—子女、好友关系、角色扮演、兄弟姐妹和伯乐关系，如表 4-2 所示，详细列出了每种关系类型所对应的部分特征词（第 2 章自动获取）和部分初始关系种子（第 3 章自动获取）。

表 4-2 关系类型及其初始关系种子示例

关系类型	特征词	关系种子示例
夫妻关系	{夫妇, 夫妻, 结婚, 完婚, 老伴, 配偶, 结发夫妻, 新婚, 成婚, 迎娶, 婚嫁, 出嫁, 老公, 丈夫, 夫婿, 丈夫, 贤内助, 太太, 妻, 贤妻, 妻子, 老婆, 爱妻, 原配, 遗孀, 媳妇, 新娘子, 新娘, 新郎, 新郎官}	<梁朝伟, 刘嘉玲> <谢霆锋, 张柏芝> <李亚鹏, 王菲> <许晋亨, 李嘉欣> <田亮, 叶一茜>
经纪人关系	{经纪人, 经济人, 经理人}	<赵本山, 高大宽>
合作关系	{合作, 搭档, 合伙, 拍档, 联手, 共事, 同事, 老同事, 合伙人, 合作者, 伙伴, 搭伴, 同台, 舞伴, 对唱, 合唱, 对手戏}	<赵本山, 宋丹丹> <周杰伦, 方文山> <冯小刚, 葛优>
情侣关系	{相恋, 恋情, 恋爱, 男朋友, 恋人, 情侣, 男友, 女朋友, 意中人, 女友, 情人, 相爱, 伴侣, 女伴, 爱恋, 初恋, 爱恋, 相爱, 拍拖, 旧爱}	<倪震, 周慧敏> <邓超, 孙俪> <刘德华, 喻可欣>
父母—子女	{母亲, 干妈, 母女, 妈妈, 老母, 生母, 养母, 母子, 继母, 娘, 妈妈, 父亲, 义父, 养父, 父子, 爹, 爸, 老父, 老父亲, 爸爸, 家父, 生父, 继父, 干爹, 老爸, 独生子, 大儿子, 儿子, 干儿子, 次子, 养子, 义子, 幼子, 爱子, 长子, 小儿子, 私生子, 女儿, 千金, 爱女, 大女儿, 独生女, 养女, 千金, 闺女, 幼女, 长女, 女儿, 小女儿}	<谢贤, 谢霆锋> <狄波拉, 谢霆锋> <谢霆锋, 谢振轩> <沈殿霞, 郑欣宜> <黄秋生, 周杰伦> <曾志伟, 曾宝仪> <陈金飞, 刘亦菲> <郑东汉, 郑中基>
好友关系	{好友, 朋友, 老友, 挚友, 老相识, 好姐妹, 好兄弟, 知音, 忘年交, 至交, 老朋友, 世交, 牌友, 亲友, 友人, 师友, 密友, 亲朋, 战友, 友谊, 友情, 私交, 结识, 相交, 叙旧, 知己, 闺中密友, 哥们儿}	<刘嘉玲, 王菲> <刘嘉玲, 胡军> <张学友, 庾澄庆> <周杰伦, 方文山> <常宝华, 马季>
角色扮演	{饰演, 出演, 饰, 扮演, 饰演者, 扮演者}	<孙红雷, 余则成>
兄弟姐妹	{老大哥, 兄长, 表哥, 胞兄, 弟弟, 大哥, 小弟, 哥哥, 二姐, 兄妹, 哥俩, 表弟, 胞哥, 兄弟, 亲兄弟, 兄, 哥, 弟, 堂弟, 长兄, 妹妹, 胞妹, 姐弟, 姐姐, 姐妹, 大姐, 双胞胎, 龙凤胎, 兄妹}	<李亚鹏, 李亚伟> <谢霆锋, 谢婷婷> <王中磊, 王中军> <侯耀文, 侯耀华> <孙权, 孙尚香>
伯乐关系	{学生, 入室弟子, 弟子, 徒弟, 门下, 门生, 学徒, 得意门生, 高徒, 爱徒, 师父, 师傅, 导师, 老师, 班主任, 教练, 恩师, 师承, 拜师, 师生, 师徒, 伯乐, 栽培, 发掘, 力捧, 捧红}	<赵本山, 小沈阳> <侯耀文, 郭德纲> <梅艳芳, 何韵诗> <张纪中, 李亚鹏> <吴宗宪, 周杰伦>

4.4.2 评价标准

本章主要目的是从大规模语料库中抽取关系描述模式，并识别关系元组，前者可以用于新文本的实体关系抽取，后者可以直接作为关系知识库，由于二者都是从大规模语料库抽取得到，我们无法评价完全的召回率，并且关系描述模式，尤其是软模式，也无法直接评价，主要通过对识别的关系元组集合进行准确率评测，评价算法性能。由于不同关系类型信息丰富程度存在一定差异，导致最终抽取关系元组数目不同，且准确率也不同。如果最终抽取关系元组数目大于 100 的关系类型，我们随机选取 100 个进行人工评价，如果关系元组数目等于或小于 100 的关系类型，我们将对全部的关系元组进行人工评价。我们选取与 3.4.2 相同的关系元组准确率和平均准确率定义公式，不再赘述。

4.4.3 结果与分析

实验过程中，我们只是简单的对过于泛化的迭代软模式进行了过滤，其他模式全部进入下一轮迭代过程，以识别出尽可能多的关系元组。我们根据关系元组的不同过滤策略实现了以下四种抽取策略：

- **B**：即基准方法 **baseline**，在迭代过程中（包括关系描述模式挖掘和关系元组抽取），不采取任何过滤措施；
- **B+F1**：每一轮获取的关系元组全部进入下一轮迭代，迭代过程完成后，对获取的所有关系元组进行过滤；
- **B+F2**：每一轮都对获取的关系元组进行过滤，可信度超过特定阈值 *minTupleConf* 的关系元组进入下一轮迭代，而低于可信度阈值 *minTupleConf* 的关系元组直接作为最终抽取结果；
- **B+F1+F2**：每一轮都对获取的关系元组进行过滤，可信度超过特定阈值 *minTupleConf* 的关系元组进入下一轮迭代，舍弃低于可信度阈值 *minTupleConf* 的元组。

算法中定义了 9 个需要调整的参数，我们通过观察统计分析一定数量的真实文本，将它们设置成如表 4-3 所示的经验值。

对于不同关系类型的关系元组抽取评价结果如表 4-4 所示。

表 4-3 实验参数设置说明

参数	取值	描述
<i>maxE1ToE2Distance</i>	10	实体对最大词距离 (4.3.1)
<i>maxInsertedEntity</i>	2	实体对之间最大其他相同类型实体数量 (4.3.1)
<i>maxPunc</i>	2	实体对之间最大标点符号数量 (4.3.1)
<i>maxFWToEDistance</i>	8	特征词与距离最近的实体之间最大词距离 (4.3.1)
<i>maxInsertedPunc</i>	2	特征词与距离最近的实体之间最大标点符号数量 (4.3.1)
<i>cL</i>	4	实体对两侧所取上下文长度 (4.3.2)
<i>minJaccard</i>	0.42	上下文模式之间相似度阈值 (4.3.3)
<i>maxTokenMatch</i>	3	省略单元“*”最多匹配单元个数 (4.3.4)
<i>minTupleConf</i>	0.1	关系元组可信度阈值 (4.3.4)

表 4-4 显示基准方法 B 抽取关系元组总数为 7,795 个, 最低准确率为 55.00%, 最高准确率为 97.00%, 大部分均在 90% 以上, 平均准确率达到了 79.73% 的结果。分析其原因, 尽管基准方法没有对元组进行过滤, 但是, 我们所采用的关系种子与传统的方法不同, 在实体对基础上, 增加了特征词的限制, 即实体对必须与特征词在句子中共现, 使用特征词进行限制一定程度上避免了传统方法难以解决的错误蔓延问题。

由于我们从大规模互联网文本中抽取关系元组, 无法直接对召回率进行评价, 但是, 通过表 4-4 可以看到不同算法对应的关系元组数目, 可以间接的反映召回率。

后面三种方法在基准方法基础上引入了不同的过滤措施, 抽取的关系元组数目有着不同程度的下降, 即召回率降低, 以大幅度提高准确率, 最终, B+F1、B+F2 和 B+F1+F2 三种方法的平均准确率分别为 86.87%, 84.78% 和 88.24%, 这些充分说明了我们定义的关系元组过滤方法的有效性。

对三种方法对比发现, B+F1 和 B+F1+F2 准确率提升最大, 均达到了 87% 左右, 但是相比较而言, 关系元组数量下降幅度较大, 减少约 2,400 个, 付出代价较大。而 B+F2 在准确率足够高的同时, 提高了 5 个百分点, 召回率下降最少, 关系元组数据减少了 586 个, 在准确率与召回率二者之间达到了最好的平衡, 主要原因是每次迭代后基于统计的方法对关系元组进行过滤, 只保留那些可信度较高的关系元组进入下一轮迭代, 以保证抽取准确率, 而可信度较低的元组中同时包含少量错误元组和大量正确的低频稀疏元组, 将它们作为最终抽取结果, 而不进入下一轮迭代, 一方面防止错误元组带来的错误蔓延现象, 另一方面, 稀疏的正确元组不易于区分, 且它们对于关系描述模式的挖掘和新关系元组的识别贡献较小。

表 4-4 关系元组抽取评价结果

关系类型	算法	迭代次数	关系元组	正确?		准确率 (%)
				是	否	
夫妻关系	B	5	917	55	45	55.00
	B+F1	5	436	77	23	77.00
	B+F2	4	529	73	27	73.00
	B+F1+F2	4	393	85	15	85.00
经纪人关系	B	2	673	94	6	94.00
	B+F1	2	452	97	3	97.00
	B+F2	2	664	96	4	96.00
	B+F1+F2	2	445	97	3	97.00
合作关系	B	3	1192	94	6	94.00
	B+F1	3	760	97	3	97.00
	B+F2	3	1089	96	4	96.00
	B+F1+F2	3	691	98	2	98.00
情侣关系	B	3	457	90	10	90.00
	B+F1	3	380	96	4	96.00
	B+F2	3	450	91	9	91.00
	B+F1+F2	3	376	97	3	97.00
父母—子女	B	5	568	78	22	78.00
	B+F1	5	465	86	14	86.00
	B+F2	5	557	81	19	81.00
	B+F1+F2	5	458	87	13	87.00
好友关系	B	3	457	97	3	97.00
	B+F1	3	339	99	1	99.00
	B+F2	3	454	98	2	98.00
	B+F1+F2	3	338	99	1	99.00
角色扮演	B	4	3272	74	26	74.00
	B+F1	4	2442	80	20	80.00
	B+F2	4	3214	75	25	75.00
	B+F1+F2	4	2396	81	19	81.00
兄弟姐妹	B	2	175	93	7	93.00
	B+F1	2	141	95	5	95.00
	B+F2	2	173	94	6	94.00
	B+F1+F2	2	140	97	3	97.00
伯乐关系	B	3	84	76	8	90.48
	B+F1	3	67	63	4	94.03
	B+F2	3	79	73	6	92.41
	B+F1+F2	3	64	60	4	93.75
平均	B	-	7795	751	133	79.73
	B+F1	-	5482	790	77	86.87
	B+F2	-	7209	777	102	84.78

B+F1+F2	-	5301	801	63	88.24
---------	---	------	-----	----	-------

由表 4-4 不难发现，实验的 9 种关系类型经 2 到 5 次迭代后，抽取结果均不再改变，可以快速收敛。

为了深入分析迭代次数对实验结果的影响，我们以“夫妻关系”为例，对四种不同算法产生的所有关系元组进行人工评价，对比结果如图 4-4 和图 4-5 所示。

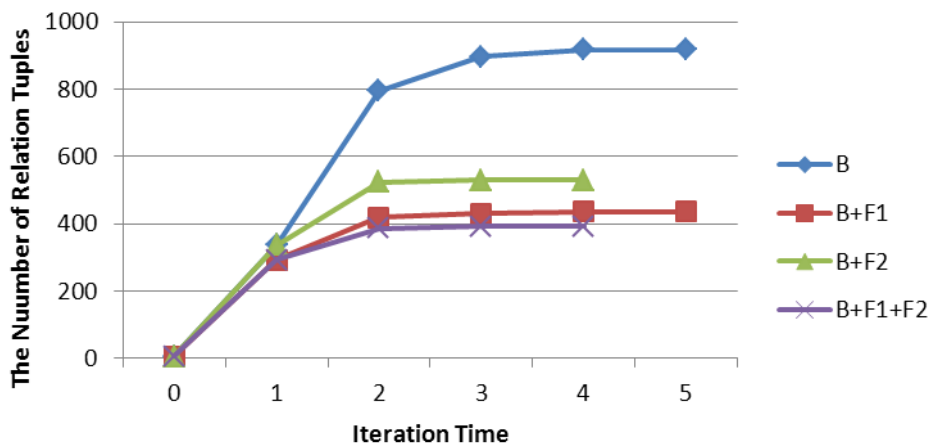


图 4-4 迭代次数对关系元组数目的影响

由图 4-4 可以直观的看出，尽管使用不同的抽取算法，抽取关系元组数量均随着迭代次数增加而不断增长，并且在前两次迭代过程中，产生的新关系元组较多，上升幅度较大，之后逐渐减缓，最终趋于稳定。分析其原因，主要是后续迭代过程识别出大量已经被识别的关系元组，新元组逐渐较少，而每轮迭代后，我们只将这一部分新的关系元组放入关系种子集中，完成后续迭代，直至算法结束。

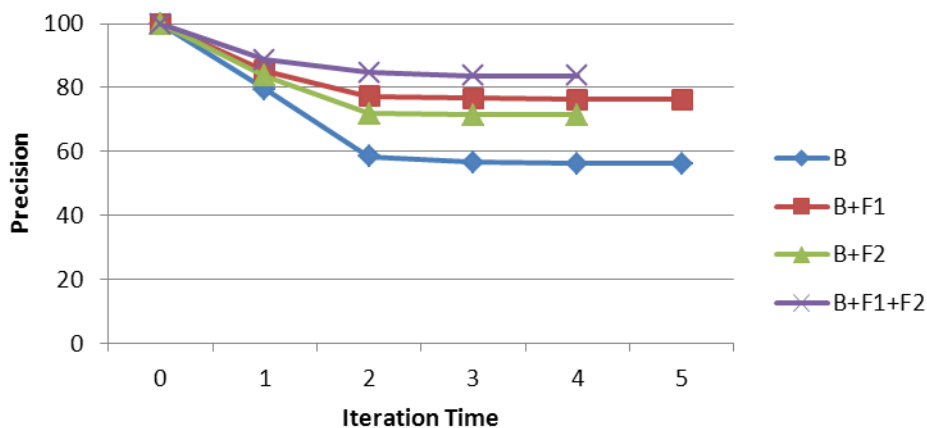


图 4-5 迭代次数对关系元组准确率的影响

图 4-5 中显示，随着迭代次数的增加关系元组准确率不断下降，并且在前

两次迭代过程中，下降幅度较大，之后逐渐减缓，最终趋于稳定。这是因为在刚开始的迭代过程中获取的一些关系描述软模式仍然过于泛化，导致识别出错误的关系元组，从而使得准确率下降较快。但是，我们发现随着迭代次数的增加，准确率下降的幅度并没有不断增加，这主要因为关系种子不仅包括关系元组，还包括关系类型对应的特征词，二者将组合在一起进入下一轮迭代，对关系实例进行了严格限制，这有效降低了错误的蔓延程度，一定程度上避免了 Bootstrapping 方法常见的循环依赖问题。

对比图 4-5 中 B 与 B+F2，以及 B+F1 与 B+F1+F2，不难发现，经过关系元组可信度过滤，使得准确率下降幅度减小，大大提高了准确率。与 B 相比 B+F2 在每轮迭代过程中只将可信度大于特定阈值的关系元组作为下一轮关系种子，对比图 4-5 中二者实验结果，不难发现，可信度较低的关系元组依然带来一定的错误蔓延问题，尽管可以抽取出更多的关系元组，但是准确率下降较多，结果较差，同样可以说明对低可信度关系元组的过滤的有效性和必要性。

4.5 本章小结

在前面章节处理结果基础上，本章采用了基于 Bootstrapping 的方法自动挖掘关系描述模式，抽取关系元组。首先，本章给出了 Bootstrapping 算法框架，介绍了具体处理流程；然后，从具体的关系实例出发，详细阐述了上下文模式和软模式挖掘方法，以及基于模式匹配的关系元组抽取和评测策略；最后，本章选取 9 类关系类型，以百万级新闻资讯为处理资源，进行实验，由于无法直接评价召回率，本章仅对关系元组准确率进行评价，并对四种关系元组抽取方法进行对比实验。实验结果表明，基准方法抽取关系元组平均准确率达到 79.73%，由于基于统计的评价和过滤措施的引入，关系元组平均准确率最高达到了 88.24%，满足了实际应用系统要求。

第5章 领域自适应的关系抽取平台设计与实现

5.1 引言

在前面的章节中，本文提出了一套新颖的领域自适应的中文实体关系抽取研究框架，针对相关的关键技术进行了深入研究，提出了基于特征词聚类的关系类型发现方法、基于 Web Mining 的关系种子集抽取方法以及基于 Bootstrapping 的关系描述模式挖掘方法。这些方法均以实体对关系类型为单元，遵循尽可能少的人工参与，不依赖语料库，不局限于限定领域，适应范围广，可移植性较强。

本章尝试将以上的关键技术整合起来，设计并实现一套领域自适应的关系抽取平台 XInfo，与学术界和产业界共享研究成果。在我们的 XInfo 平台基础上，一方面研究人员可以专注于算法的改进与研究，快速进行实验，循环往复以不断的提高关系抽取系统性能，提高关系抽取自动化程度；另一方面，开发人员可以进行成果转化，为数据分析人员提供数据支持，甚至为普通用户直接提供服务。

另外，我们以实体对类型“人名—人名”为例，开发了一套人物社会关系抽取在线演示系统，以直观、清晰的方法展示 XInfo 抽取效果。

5.2 关系抽取平台 XInfo

本章提出的领域自适应的关系抽取平台 XInfo 系统架构如图 5-1 所示。该系统只需要给定待处理的文本语料库资源，系统首先对文本进行自然语言处理，如断句、分词、词性标注、命名实体识别、依存句法分析等，将文本映射成 DOM 结构；然后，根据指定的实体对类型采用基于特征词聚类的方法进行关系类型的自动发现，得到大量用户可能感兴趣的关系类型；接着，人工可以选择性参与筛选或合并一些比较关注的关系类型，采用基于 Web Mining 的方法利用搜索引擎抽取少量关系种子集，此过程也可以根据实际情况替换为人工直接给定关系种子集，可配置，耦合性低；最后，以关系种子集（包括关系元组和特征词）为输入，采用基于 Bootstrapping 的方法从语料库中挖掘关系描述模式，抽取关系元组，前者可以用于处理新的文本，后者可以作为实体关系知识库的一部分，直接投入使用。

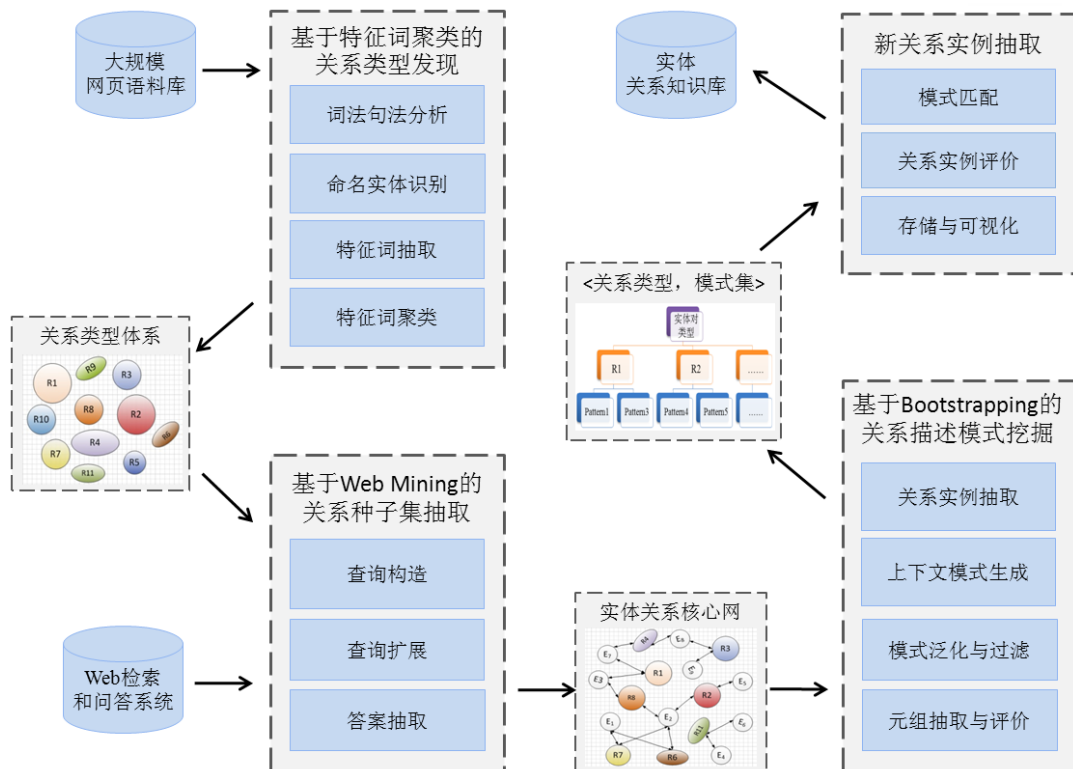


图 5-1 领域自适应的关系抽取平台 XInfo 系统架构图

5.3 关系抽取演示系统

我们在XInfo系统平台基础上，以实体对类型为“人名—人名”为处理对象，在百万级的娱乐资讯网页库RE100W中挖掘人物社会关系网络。后台使用Java语言，前台使用Flex可视化开发技术，搭建了一套在线演示系统：人脉搜索(PersonMap)¹⁵，为终端用户提供关系查询服务。

如图 5-2 所示，根据人物实体在网页库中的 $TF \times IDF$ 权重排序，在首页默认显示最近一段时间内的热点人物，如成龙、王菲、梁朝伟、赵本山等。

如图 5-3 所示，在输入框中输入人名“成龙”，检索得到与其产生关系的部分网络图，即按照关系元组可信度进行排序，选取排名靠前的关系元组进行展示。当鼠标停留在人物节点上时，右上角会显示当前关系网络中与特定人物参与的关系元组，并且每个关系元组都有对应关系类型和可信度值。关系网络图中使用连线的粗细表示可信度的高低，这些信息都可以辅助用户甄别判断，如<成龙，房祖名，合作搭档，7>，<成龙，房祖名，子女，73>等。

¹⁵ <http://ir.hit.edu.cn/personmap/>

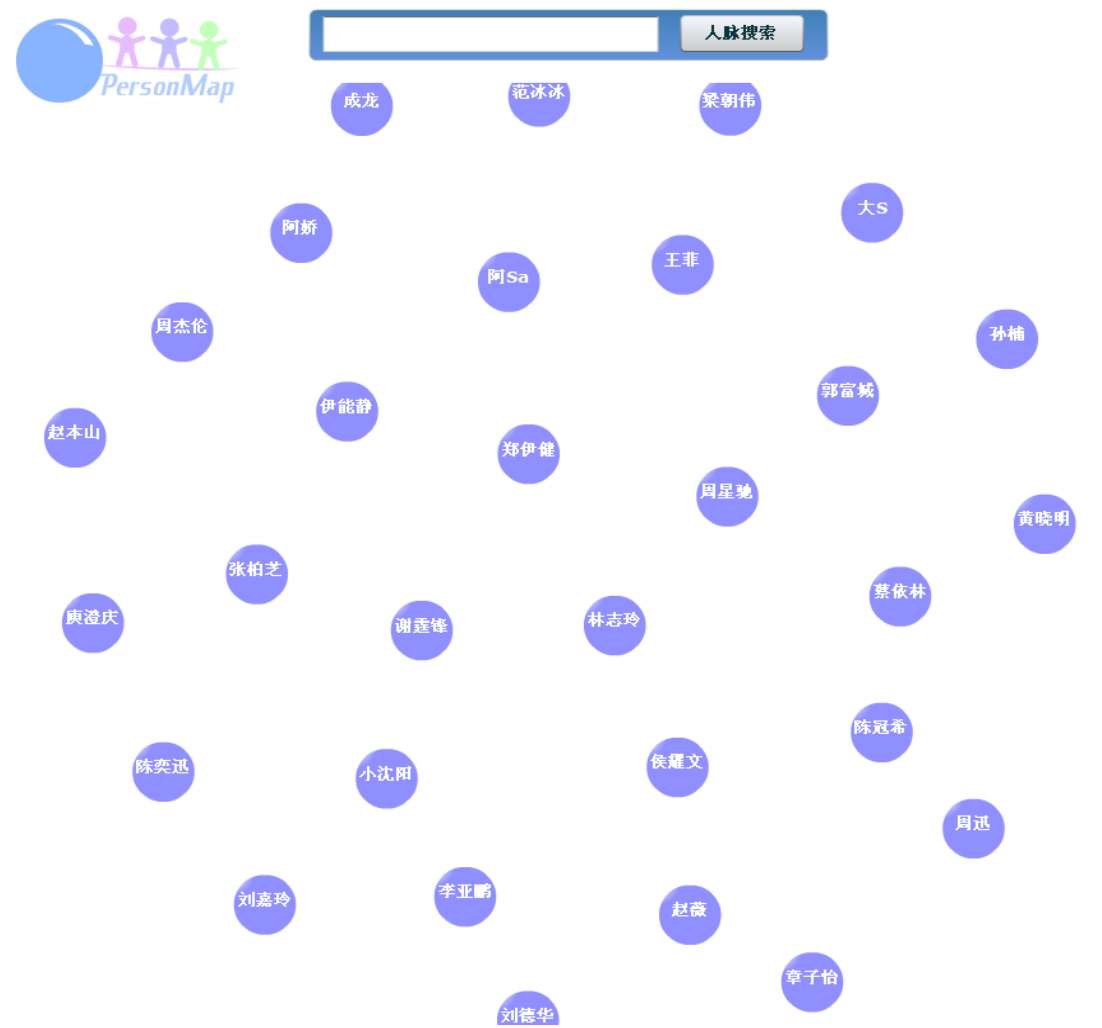


图 5-2 人脉搜索首页：热点人物

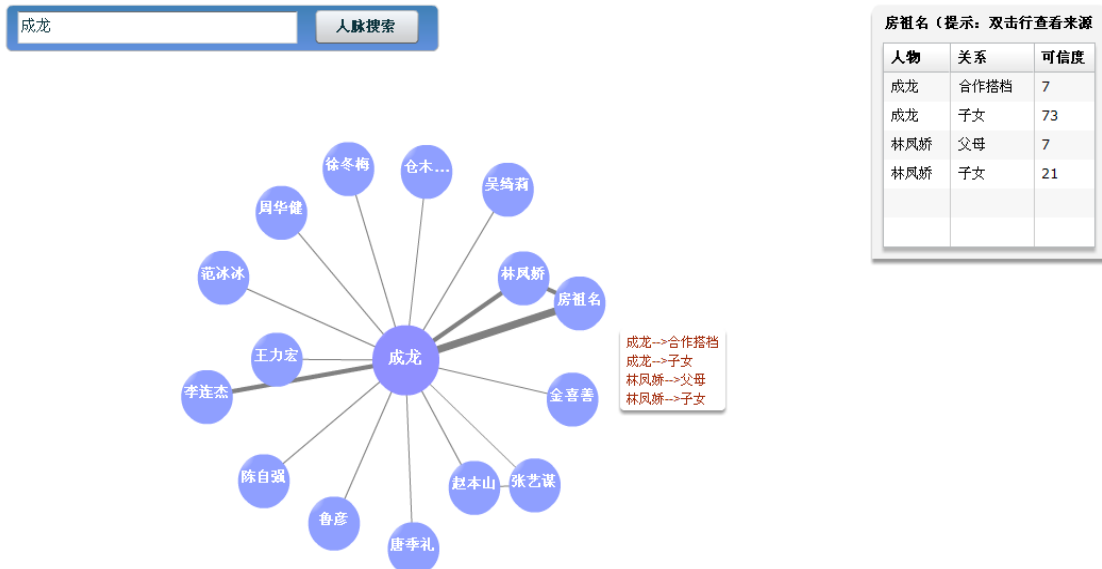


图 5-3 人物社会关系示例

如图 5-4 和 5-5 所示，当双击右上角关系元组表中的某一行时，可以查看当前关系元组出现的句子，即我们不但存储关系元组，计算其可信度，还索引了来源句子及句子所在的网页，以作为证据，给用户提供更加充分的相关信息，图 5-4 中显示了关系元组<成龙，林凤娇，夫妻>的来源文本，图 5-5 显示了关系元组<成龙，房祖名，子女>的来源文本。

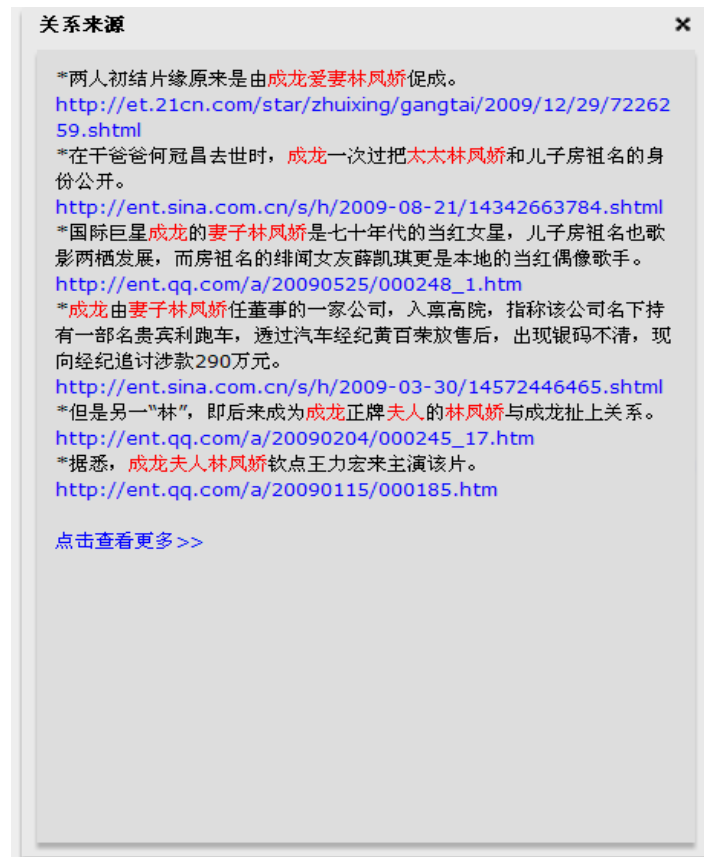


图 5-4 <成龙，林凤娇，夫妻>关系来源



图 5-5 <成龙, 房祖名, 父母—子女>关系来源

5.4 本章小结

本章在前面章节研究成果基础上，设计并实现了领域自适应的关系抽取平台 XInfo，主要包括系类型发现、关系种子集抽取、关系描述模式挖掘以及新关系实例抽取模块，各模块之间耦合性低，可配置性强。在该平台上，相关人员可以集中展开算法的改进与研究，快速实验，大大节省时间，提高工作效率，且便于研究成果的积累和转化。

在 XInfo 平台基础上，本章还实现了人物社会关系抽取在线演示系统：人脉搜索，在成果展示的同时，还可以为终端用户提供查询服务。

结 论

关系抽取作为信息抽取的子任务和关键技术之一，是自然语言处理相关领域重要的支撑技术，受到了越来越多研究学者的关注，已经发展成为了研究热点。本文针对领域自适应的中文实体关系抽取展开了一系列深入研究，提出了一套全新的领域自适应的中文实体关系抽取研究框架，探索不依赖于人工标注的训练语料库，最大程度地避免人工参与，且具有较强领域自适应能力的关系抽取解决方案，主要解决关系类型发现、关系种子集构建、关系描述模式挖掘和关系元组抽取等关键问题，以提高关系抽取的自动化程度，提高可移植性，扩大关系抽取的适用范围。

具体来讲，本文的创新点和主要贡献包括以下几个方面：

(1) 提出了基于特征词聚类的关系类型发现方法。我们以从真实的数据中抽取关键信息为出发点，首先从 Web 上获取大规模的网页文本作为语料资源，以实体对类型为单元，通过挖掘高频实体对之间的特征词实现关系类型的自动发现，即借助特征词触发、描述实体关系，而不局限于人工预定义的有限关系类型，通过在实体对类型“人名—人名”上展开实验，该方法达到了与多人协作预定义的关系类型相当的效果，在领域移植应用中，可以为研发人员提供可靠的关系类型体系参考。

(2) 提出了基于 Web Mining 的关系种子集抽取方法。对于自动发现的大量关系类型，采用有指导的方法需要人工标注语料库，耗时耗力，稀疏问题严重，可移植性差，而直接使用半指导的学习方法，由于关系类型的繁多，需要人工构造大量的关系种子集，投入成本较大。针对这些问题，我们提出了基于 Web Mining 的关系种子集自动构建方法，该方法将关系实例看作三元组 $\langle e_1, e_2, R \rangle$ ， e_1 可以赋值为高频实体， R 为关系类型，对应一个特征词集合，利用二者启发式地构造查询集合，充分利用搜索引擎收集和处理大规模真实数据的能力和优势，检索出相关的网页，获取页面摘要，统计抽取候选答案，即另一个实体 e_2 ，获取实体关系核心网，经过在选取的 9 种关系类型上进行实验，平均准确率达到了 90.91%。

(3) 采用了基于 Bootstrapping 的关系描述模式挖掘方法。以每类抽取的高质量关系种子集出发，定义了启发式上下文模式及其泛化策略，引入了 Bootstrapping 方法，可以在未标注语料库上迭代地挖掘关系描述模式，抽取关系元组，通过对采样的关系元组进行人工评价，最高平均准确率达到了 88.24%。

(4) 设计并实现了一套领域自适应的关系抽取平台。在该平台上, 研究人员可以集中精力进行关系抽取中算法的改进和深入研究, 快速进行实验。另外, 本文以人物社会关系作为应用任务, 开发了可视化在线演示系统, 可以为用户提供查询和浏览服务, 以直观、清晰的方式展示抽取效果。

尽管目前本文已经取得了一定的阶段性研究成果, 但是, 仍然存在许多需要改进的地方, 存在一些关键问题值得进一步研究, 主要包括以下几部分:

(1) 关系描述模式的改进与评价。本文定义了裁剪的上下文模式, 并使用最佳匹配算法进行泛化, 未来可考虑探索更加有效的关系描述模式, 如基于特征向量的关系描述模式, 基于依存句法树或短语结构树的关系描述模式, 等等。另外, 本文仅仅对过于泛化的关系描述模式进行了简单过滤, 未定义可信度公式进行评价, 主要侧重对关系元组的可信度计算和后处理, 下一步可以定义模式的评价标准, 并优化关系元组的结果。

(2) 关系元组的推理和冲突消解。本文将每种关系类型看作一个独立的处理对象, 而未考虑它们之间潜在的关系, 如在父子和母子关系基础上可以推理出夫妻关系, 在父子关系本身可以推理出祖孙关系, 等等。另外, 某些关系类型的实体对之间存在一定的约束条件, 如夫妻关系必须是一对一, 朋友关系可以是一对多, 等等。下一步可以考虑使用语义网 (Semantic Web) 进行数据的统一表示, 并在其基础上实现关系的推理和冲突消解。

(3) 领域自适应的关系抽取方法与传统的关系抽取方法相结合。领域自适应的关系抽取可以从未标注语料库中自动发现关系类型、挖掘关系描述模式和抽取关系元组等, 这种技术扩大了关系抽取的适用范围, 但是, 该研究方法以从大规模文本库中抽取关键信息为目标, 基于真实关系实体对频繁出现在文本库中为前提假设, 最终未以单个句子 (或关系实例) 为单位评价, 而是针对关系元组抽取结果进行评价, 相比传统的关系抽取, 这无疑会牺牲一定的召回率。所以, 针对某些应用领域已经存在标注语料库的情况下, 可以探索在不影响自动化程度的情况下, 将两种研究方法进行有效融合, 增加可扩展性, 并提高系统的整体性能。

(4) 篇章级实体关系抽取。本文主要以句子级实体之间的非等价关系为研究对象, 从而丢失了大量的代词参与的关系, 可以进一步考虑引入等价关系, 即共指消解处理结果, 通过实体之间等价关系和非等价关系的融合和简单推理实现篇章级实体关系抽取, 提高召回率, 更好地对篇章进行理解。具有重要的现实意义。

参考文献

- [1] 李保利, 陈玉忠, 俞士汶.信息抽取研究综述. 计算机工程与应用. 2003, 39(10):1-5.
- [2] 车万翔, 刘挺, 李生.实体关系自动抽取. 中文信息学报. 2005, 19(2):1-6.
- [3] N. Sager, Natural Language Information Processing: A Computer Grammar of English and Its Applications. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA. 1981:233-258
- [4] G. DeJong, An overview of the FRUMP system. Strategies for Natural Language Processing, 1982:149-176.
- [5] In Proceedings of the 6th Message Understanding Conference (MUC-7) [C]. National Institute of Standards and Technology. 1998.
- [6] Chinchor N., Marsh E. MUC-7 Information Extraction Task Definition (version 5.1). In Proceedings of the Seventh Message Understanding Conference. 1998:2-3.
- [7] The ACE 2007 (ACE07) Evaluation Plan. In Proceedings of the ACE 2007 Evaluation. 2007:1-3.
- [8] Automatic Content Extraction 2008 Evaluation Plan (ACE08). In Proceedings of the ACE 2008 Evaluation. 2008:1-3.
- [9] SemEval-2007 Task 04: Classification of Semantic Relations between Nominals. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). 2007:13-18.
- [10] Aone, C. and M. Ramos-Sanacruz. REES: A Large-Scale Relation and Event Extraction System. in Proceedings of the 6th Applied Natural Language Processing Conference. 1998:76-83.
- [11] Kambhatla N. 2004. Combining lexical, syntactic and semantic features with Maximum Entropy models for extracting relations. ACL-2004 (poster).
- [12] G.D. Zhou, SU Jian, ZHANG Jie, ZHANG Min. Exploring Various Knowledge in Relation Extraction. In Proceedings of the 43rd Annual Meeting of the ACL. 2005:427-434.
- [13] 董静, 孙乐, 冯元勇. 中文实体关系抽取中的特征选择研究. 中文信息学报. 2007,21(4):80-91.
- [14] Yee Seng Chan and Dan Roth. Exploiting background knowledge for relation extraction. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10). Association for Computational Linguistics, 2010:152-160.

- [15] Sun, A., R. Grishman and S. Sekine, Semi-supervised Relation Extraction with Large-scale Word Clustering. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics(ACL 2011), 2011:521-529.
- [16] Haussler, D., Convolution kernels on discrete structures. UC Santa Cruz Technical Report UCS-CRL-99-10, 1999.
- [17] Lodhi, H., et al., Text classification using string kernels. The Journal of Machine Learning Research, 2002. 2:419-444.
- [18] Zelenko, D. and C.A.A. Richardella, Kernel methods for relation extraction. Journal of Machine Learning Research, 2003. 3(6):1083-1106.
- [19] Culotta, A. and J. Sorensen, Dependency tree kernels for relation extraction. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics(ACL 2004), 2004:423-429.
- [20] Bunescu R. C. and Mooney R. J. A Shortest Path Dependency Kernel for Relation Extraction[J]. EMNLP 2005,2005:724-731.
- [21] Che Wanxiang, Jiang Jianmin, Su Zhong, Liu Ting. Improved-Edit-Distance Kernel for Chinese Relation Extraction [C]. The 2nd Int'l Joint Conf on Natural Language Processing (IJCNLP-05) , 2005 vol. 2651.
- [22] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, NIPS, MIT Press, 2001: 625–632.
- [23] M. Zhang, J. Zhang, J. Su and G.D. Zhou. A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association of Computational Linguistics (COLING/ACL-2006), Sydney, Australia. 2006:825-832.
- [24] Qian LongHua, Zhou GuoDong, Zhu QiaoMin, Qian Peide. Relation Extraction using Convolution Tree Kernel Expanded with Entity Features. PACLIC'21: 415-421.
- [25] Zhao Shubin and Ralph Grishman. Extracting relations with integrated information using kernel methods. In Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005). 2005:419-426.
- [26] Zhang Min, Jie Zhang, Jian Su and Guodong Zhou. A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association of Computational Linguistics (COLING/ACL-2006). 2006: 825-832.

- [27] Zhou Guodong, Zhang Ming, Donghong Ji and Qiaoming Zhu. Tree Kernel-based Relation Extraction with Context-Sensitive Structured Parse Tree Information. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL-2007). 2007: 728-736.
- [28] L.H. Qian, G.D. Zhou, F. Kong, Q.M. Zhu and Peide. Exploiting Constituent Dependencies for Tree Kernel-based Semantic Relation Extraction In Proceedings of the 22st International Conference on Computational Linguistics (COLING-2008), Manchester, August 2008:697-704.
- [29] Sergey Brin. Extracting patterns and relations from world wide web. In Proceedings of WebDB Workshop at 6th International Conference on Extending Database Technology (WebDB'98), 1998:172-183.
- [30] Agichtein, E., and Gravano, L. Snowball: extracting relations from large plain-text collections. In Proceedings of the ACM Conference on Digital libraries, 2000:85-94.
- [31] Etzioni, O., M. Cafarella, et al. Unsupervised named-entity extraction from the Web: An experimental study. Artificial Intelligence 2005,165(1): 91-134.
- [32] Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. Organizing and searching the world wide web of facts - step one: the one-million fact extraction challenge. Inproceedings of the 21st national conference on Artificial intelligence - Volume 2 (AAAI'06), 2006:1400-1405.
- [33] Benjamin Rosenfeld and Ronen Feldman. URES: an unsupervised web relation extraction system. In Proceedings of the COLING/ACL on Main conference poster sessions (COLING-ACL'06). Association for Computational Linguistics, Stroudsburg, PA, USA, 2006:667-674.
- [34] Ronen Feldman and Benjamin Rosenfeld. Boosting unsupervised relation extraction by using NER. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-2006). 2006: 473-481.
- [35] 李维刚, 刘挺, 李生. 基于网络挖掘的实体关系元组自动获取. 电子学报, 2007, 35(11):2111-2116.
- [36] 邓肇, 樊孝忠, 杨立公. 用语义模式提取实体关系的方法. 计算机工程, 2007. 33(10): 212-214.
- [37] Banko, M.; Cafarella, M. J.; Soderland, S.; Broadhead, M.; and Etzioni, O. Open information extraction from the Web. In Proceedings of the International Joint Conference on Artificial Intelligence, 2007:2670-2676.
- [38] Banko, M., and Etzioni, O. The tradeoffs between open and traditional relation

- extraction. In Proceedings of the Annual Meeting of the ACL, 2008:28–36.
- [39] Zhu, J.; Nie, Z.; Liu, X.; Zhang, B.; and Wen, J.-R. Statsnowball: a statistical approach to extracting entity relationships. In Proceedings of the International Conference on World Wide Web, 2009:101–110.
- [40] T. Hasegawa, S. Sekine, and R. Grishman. Discovering Relations among Named Entities from Large Corpora. Proc. of ACL-2004, 2004:415–422.
- [41] Zhang, M.; Su, J.; Wang, D.; Zhou, G.; and Tan, C. L. Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In Proceedings of the international Joint Conference on Natural Language Processing, 2005:378–389.
- [42] Rozenfeld, B. and Feldman, R. High-Performance Unsupervised Relation Extraction from Large Corpora. In Proceedings of the Sixth international Conference on Data Mining (December 18-22, 2006). ICDM. IEEE Computer Society, Washington, DC, 2006:1032-1037.
- [43] Rosenfeld, B., and Feldman, R. Clustering for unsupervised relation identification. In Proceedings of the ACM Conference on Information and Knowledge Management, 2007:411–418.
- [44] Davidov, D., Rappoport, A. and Koppel, M., Fully unsupervised discovery of concept-specific relationships by Web mining. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007). 2007:232-239.
- [45] Dmitry Davidov and Ari Rappoport. Unsupervised Discovery of Generic Relationships Using Pattern Clusters and its Evaluation by Automatically Generated SAT Analogy Questions. Proceedings of ACL-08: HLT, 2008:692:700.
- [46] Ang Sun. A Two-stage Bootstrapping Algorithm for Relation Extraction. Student Research Workshop, RANLP 2009, 2009:76-82.
- [47] Yuhang YANG, Qin LU, Tiejun ZHAO. A Clustering Based Approach for Domain Relevant Relation Extraction. International Journal of Information. 2009.12(2): 399-410.
- [48] F. Mesquita, Y. Merhav, and D. Barbosa. Extracting information networks from the blogosphere: State-of-the-art and challenges. In ICWSM '10: Proceedings of the 4th Int'l AAAI Conference on Weblogs and Social Media, 2010.
- [49] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010:Demonstrations. 2010.08:13-16.
- [50] 马金山. 基于统计方法的汉语依存句法分析研究[D]. 博士毕业论文. 哈尔

滨工业大学. 2007:27-34.

- [51] Z. Dong and Q. Dong. HowNet and the Computation of Meaning. World Scientific Publishing Co. Pte. Ltd. 2006.
- [52] Frey, B.J. and D. Dueck, Clustering by Passing Messages Between Data Points. Science, 2007.315: 972-976.
- [53] 梅家驹, 竺一鸣, 高蕴琦等. 同义词词林.上海:上海辞书出版社出版. 1983.

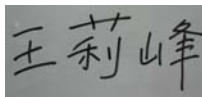
攻读学位期间发表的学术论文

哈尔滨工业大学学位论文原创性声明及使用授权说明

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《领域自适应的中文实体关系抽取研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签名：



日期：2011 年 6 月 30 日

学位论文使用授权说明

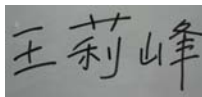
本人完全了解哈尔滨工业大学关于保存、使用学位论文的规定，即：

（1）已获学位的研究生必须按学校规定提交学位论文；（2）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（3）为教学和科研目的，学校可以将学位论文作为资料在图书馆及校园网上提供目录检索与阅览服务；（4）根据相关要求，向国家图书馆报送学位论文。

保密论文在解密后遵守此规定。

本人保证遵守上述规定。

作者签名：



日期：2011 年 6 月 30 日

导师签名：



日期：2011 年 6 月 30 日

致 谢

值此论文即将完成之际，心中感慨良多，这篇论文能够得以顺利完成，使我既体会到辛勤劳动后的喜悦，又深深感到它与大家的帮助和支持是分不开的。

感谢哈工大社会计算与信息检索研究中心的所有老师，特别感谢研究中心主任刘挺教授，感谢您为我们提供优越的工作环境和良好的学习科研氛围，感谢您给我参与讨论、发表意见的机会。您开阔的视野、敏锐的思维、严谨的学风，以及严以律己、宽以待人的高尚品质无不是我学习的楷模。

感谢我的导师秦兵教授一直以来对我的信任和鼓励。秦老师在生活上给了我无微不至的关心，在研究上给了我自由发挥的空间，让我在很多项目的研究和开发中锻炼动手和管理能力，使我学习了知识，开阔了视野，相信这些将使我终生受益。在此再次向恩师致以崇高的敬意并表示衷心的感谢！

感谢已经毕业的周蓝珺和刘龙师兄，是你们把我带到了 NLP、IR、IE 领域，与你们相处的日子里我学到了很多，成长了很多，你们广阔的视野，出色的研发能力和团队合作意识给我留下了深刻的印象，并一直影响、改变着我，祝愿周蓝珺师兄多发表高水平论文，祝愿刘龙师兄早日创业成功。

感谢“科学家”郭宇航师兄，师兄渊博的知识、严谨的态度和勤恳的作风深深感染了我，激励着我不断地把工作做得更好。

感谢 TM 组所有组员，感谢在实验室一起学习奋斗过的张牧宇、张文斌、陈鑫、丁效、付瑞吉、唐都钰、张一博、王彪、胡燊、张梅山、赵静、韩中华、康维鹏、刘安安、郭江、秦海龙以及其他成员，谢谢你们平日里热心的帮助、信任和鼓励，希望你们学习、工作顺利。

感谢我的两位好朋友、好兄弟王孝余和刘胜宇，是你们陪我渡过了两年轻松愉快的研究生生涯，我会怀念与你们吃喝玩乐，谈天说地，努力奋斗的日子，希望你们事业有成。

感谢吉林大学和哈尔滨工业大学对我的培养，希望母校蒸蒸日上，培养出更多优秀的人才，为国强民富作出更瞩目的贡献。

感谢养育我长大成人的父母家人，谢谢你们始终如一的关心和支持，这些都是我不断向前进取的重要动力和保障。尤其要感谢我的哥哥，是你一直在教育、引导和关心着我，是你教会我真诚待人，乐于交流，勇于探索，有意义的成长与生活，相信你会永远羡慕我“有这么好的一个哥哥”。

感谢所有曾经给予我关心、支持和帮助的人们，愿你们好运常伴！