基于句子匹配分析的知识抽取

化柏林◎著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

基于句子匹配分析的知识抽取

国家科学技术学术著作出版基金资助出版

基于句子匹配分析的知识抽取

化柏林 著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

· 北 京 ·

图书在版编目（CIP）数据

基于句子匹配分析的知识抽取/化柏林著．—北京：科学技术文献出版社，2014.4

ISBN 978-7-5023-8773-0

Ⅰ.①基… Ⅱ.①化… Ⅲ.①情报检索-研究 Ⅳ.①G252.7

中国版本图书馆 CIP 数据核字（2014）第 064189 号

基于句子匹配分析的知识抽取

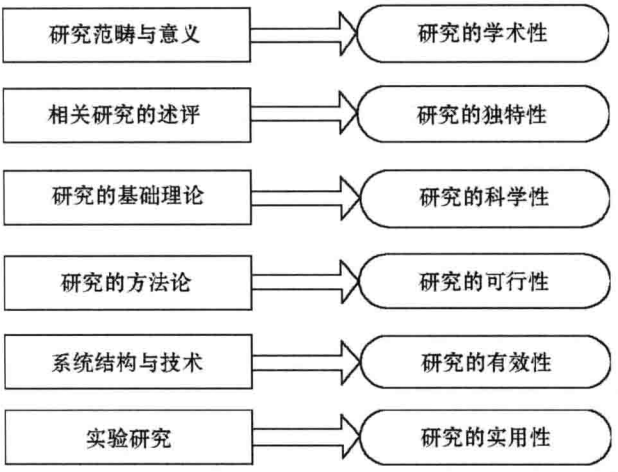
策划编辑：丁坤善 责任编辑：刘 亭 责任校对：张燕育 责任出版：张志平

出 版 者 科学技术文献出版社
地 址 北京市复兴路 15 号 邮编 100038
编 务 部 （010）58882938，58882087（传真）
发 行 部 （010）58882868，58882874（传真）
邮 购 部 （010）58882873
官方网址 www.stdp.com.cn
发 行 者 科学技术文献出版社发行 全国各地新华书店经销
印 刷 者 北京金其乐彩色印刷有限公司
版 次 2014 年 4 月第 1 版 2014 年 4 月第 1 次印刷
开 本 710×1000 1/16
字 数 161 千
印 张 12
书 号 ISBN 978-7-5023-8773-0
定 价 48.00 元

前言

本书主要针对学术文献，通过相似性判定识别文章的新句子，对文章的新句子进行句子内部结构及主题语义分析，从而确定句子的主题及语义；然后通过句子之间的关联关系分析和文章的篇章结构分析确定句子的知识元隶属，并对知识的属性进行标记，如定义、概念属性、研究方法、实验数据、研究结果、研究结论等。基于句子匹配分析的知识抽取研究不但可以解决参考文献自动标注问题，用新句子形成文献自动综述，而且可以把文献处理的颗粒度从篇章层次细分到句段层次，实现在知识单元层面上的组织、管理和利用，从而改变传统的知识组织和管理方式。

本书主要从基础理论、研究方法、技术实现和实验分析四个层面对句子级知识抽取研究进行探讨，展示了一个句子级知识抽取系统框架，在此基础上实现各种具体的抽取任务，包括定义抽取、概念属性抽取、情报方法抽取等。以理论、方法、技术、实践为四条基线，具体包括：通过交代研究的边界与意义论证本研究的学术性；通过文献综述交代本研究的价值与创新之处，论证本研究的独特性；通过理论基础研究论证本研究的科学性；通过方法研究论证本研究的可行性；通过技术研究论证本研究的有效性；通过实验研究论证本研究的实用性。其研究逻辑结构如下图所示。



本书的逻辑结构图

对于定义抽取与概念属性抽取，本书沿用了信息科学领域常用的测试方法，以《情报理论与实践》2009 年文章为训练文本构建规则，以《情报学报》2007 年和 2008 年的文章为实验文本进行了抽取实验，通过分析实验数据验证了规则构建方法以及系统的有效性和可行性。对于情报方法抽取，采用深入分析抽取结果并构建方法体系的实证方式进行验证，以《情报学报》2012 年全文为实验对象，对方法类知识进行抽取实验，对抽取的方法类句子进行知识属性的标记，包括方法的定义、方法的特点、方法的类属、方法的功能、方法的创新层次等。

基于以上逻辑，本书共 10 章内容论述，具体章节安排如下：

第 1 章，绪论。介绍了研究背景，相关概念及本书的主要研究对象；概述了本书的理论基础与研究方法。

第 2 章，相关研究述评。阐述了国内外现有的研究现状，述评了相关研究对该领域的贡献，并指出先前研究存在的缺点及不足，为本书的研究切入做好了铺垫。

第 3 章，知识抽取的理论基础。针对学术文献的句子级知识抽取涉及

知识基因理论、意义建构理论、知识谱系理论以及情报转化理论，论述了这四个相关理论的核心内容，以及对本研究的理论支撑。

第4章，知识抽取的系统架构与技术。设计了知识抽取的系统结构，包括总体设计思路、系统架构与软件结构设计；分析了知识抽取的流程，包括文献预处理、文献内容分析、知识元的抽取、知识属性标记等模块；介绍了知识抽取过程中的文献预处理技术，包括向量分词方法等，并给出了详细的处理流程、关键算法等。

第5章，知识抽取的方法研究。针对句子级知识抽取的特点，对知识抽取中的规则构建方法、句子匹配分析方法以及文献结构分析方法进行了研究。

第6章，学术定义的抽取。在剖析学术定义的表述方式及构成要素的基础上，设计并实现了学术定义抽取系统，以《情报理论与实践》2009年文章为训练文本构建规则，以《情报学报》2007年和2008年的文章为实验文本进行了抽取实验。

第7章，学术概念属性的抽取。在对属性抽取规则进行全面分析的基础上，设计并实现了学术概念属性抽取系统，以《情报理论与实践》2009年文章为训练文本构建规则，以《情报学报》2007年和2008年的文章为实验文本进行了抽取实验。

第8章，学术论文中“方法”的抽取。选取《情报学报》2012年的151篇全文数据为实验对象，利用多阶规则对论文中提及的方法进行了抽取实验，对抽取的方法进行统计汇总。

第9章，构建情报方法知识库。针对第8章抽取的实验结果，运用系统分析法进行分析，构建了情报方法知识体系。

第10章，结论与展望。总结了本书的主要研究结论和创新点，并提出了知识抽取的发展趋势与下一步研究方向的构想。

本研究的创新或价值体现在以下几个方面：

(1)对句子级知识抽取的系统架构、处理流程、研究方法进行了系统

研究，并选取小规模数据进行实验，实验结果表明，本文所采用的多阶规则的方法是行之有效的。

(2)本文采用顺排档规则与倒排档规则两套规则来提高知识属性判定的准确性，对方法的定义、方法的过程、方法的类属、方法的特点、方法的功能等知识属性进行标记。

本书内容主要来自作者近年来在该领域所做的研究工作，多数章节的内容直接来自于本人与他人合作发表的学术研究论文，部分研究成果得到国家自然科学基金的资助。另外，书稿的一些内容也体现在作者的博士学位论文中，感谢我的博士生导师中国科学技术信息研究所武夷山研究员的指导与帮助。感谢我的博士后合作导师北京大学信息管理系李广建教授。感谢中国科学技术信息研究所郑彦宁研究员、潘云涛研究员、赵筱媛副研究员的指导与帮助。感谢武汉大学信息管理学院马费成教授、南京大学信息管理学院孙建军教授、中国国防科技信息中心霍忠文研究员、北京大学信息管理系王继民副教授的指导与帮助。感谢研究团队主要成员丁君军、刘一宁、吴超等；感谢所有与我进行过合作研究的老师和同学们。在写作过程中，参考或借鉴了大量的中外文参考资料，由于篇幅所限或工作疏忽，未能一一列出，在此特向所有的参考文献作者表示衷心的感谢。此外，还要感谢科学技术文献出版社的丁坤善等老师给予的支持和辛勤付出。

本书的撰写工作虽几经努力，但限于能力和水平，缺点不足与错误疏漏之处在所难免，恳请各位专家和读者批评指正。同时，由于知识抽取在理论方法与应用实践方面都存在着较多的困难，也希望更多的学者关注知识抽取这一领域。

化柏林

2013 年 12 月

E-mail：huabolin@pku. edu. cn

目 录

1 绪 论

1.1 研究背景

1.1.1 研究环境背景

目前，信息爆炸与知识相对匮乏的矛盾日益突出，如何解决这一矛盾，是决定情报学在新环境下能否快速发展的关键。要解决这一矛盾，需要信息组织从物理层次的文献单元向认知层次的知识单元转换[1]，而知识抽取是能够完成这种转换的一种路径。

图书情报领域存在着从信息服务走向知识服务，从信息管理走向知识管理，甚至从信息科学走向知识科学的趋势。从非结构化信息中抽取、获取结构化知识或半结构化知识，是当前我国图书情报界的一个关注焦点、研究热点和实践难点。探索、发展句子级的知识抽取理论及其技术，有助于解决知识获取这一关键问题。

如何从大量的信息中抽取人们所需要的知识，是知识抽取所要研究的内容。一篇科技文献往往包括定义、特点、分类、发展历史、关键技术、应用前景、发展趋势、实验数据选取、实验结果分析等内容，我们把这些内容称为内容元数据。如何把文献中的知识按照这些内容元数据进行标记就是知识抽取研究的内容。知识抽取（knowledge extraction）通过对文献进行内容分析处理，把文献中所蕴含的知识点（也称知识元）逐条抽取出来，对知识的属性进行标记，再以一定形式存入知识库中[2]。

本研究主要针对学术论文，通过相似性判定识别文章的新句子，对文章的新句子进行句子内部结构及主题语义分析，从而确定句子的主题及语义；然后通过句子之间的关联关系分析和文章的篇章结构分析确定句子的知识元隶属，并对知识的属性进行标记。知识抽取通过自动分析文本来抽取知识点，分别以面向对象的形式和逻辑命题的形式存储到数据库里，而这种数据库就相当于“结构化的百科全书”。通过知识抽取可将文献处理的颗粒度从篇章层次细分到句段层次（从以篇章为单位转换成以知识元为单位），真正实现文献在知识单元上的组织、管理和利用，实现信息组织从物理层次的文献单元向认知层次的知识单元转换，从而彻底改变传统的知识组织和管理方式[2]。

当前我国图书情报界，基于词的研究成果众多，基于句子的研究成果寡少。本研究从句子层面入手，利用多重规则深入分析文献内容及结构，在认知和理论方面具有新颖性，同时展示了将认知和理论转化为技术的方法和途径，实现了理论与技术的统一，颇具特色，具有普适意义。

1.1.2 研究项目背景

本书是作者主持的国家自然科学基金青年项目“基于句子匹配分析的知识抽取研究与实现”这项研究及其后续研究的成果。国家自然科学基金项目“基于句子匹配分析的知识抽取研究与实现”主要针对学术论文，通过相似性判定识别文章的新句子，对文章的新句子进行句子内部结构及主题语义分析，从而确定句子的主题及语义；然后通过句子之间的关联关系分析和文章的篇章结构分析确定句子的知识元隶属，并对知识的属性进行标记，如定义、发展历史、特点、关键技术、意义、应用、未来趋势等。

该项目于2008年获批，自2009年开始执行，于2012年年初通过结题验收。在结题以后又对以下内容进行了补充与更新：

(1)对文献综述重新撰写，于2013年补充最新相关文献，并按照

更严格的要求撰写文献综述，使之更加全面。

（2）把研究重点放在方法的抽取与结果分析上，这是自项目结题以后的最新研究进展，也是体现论文创新与价值的关键部分。选取《情报学报》2012 年全文数据为实验对象进行方法的知识抽取。

1.2 研究意义

1.2.1 理论意义

本书将对论文论述的逻辑结构、知识表述的常用句式进行归纳总结，形成论文知识抽取规则。在此基础上，构建一套适用于学术论文的知识抽取理论与方法体系，为大规模知识抽取奠定理论基础与方法依据。

1.2.2 实践意义

图书情报领域存在着从信息服务走向知识服务，从信息管理走向知识管理，甚至从信息科学走向知识科学的趋势。而知识服务、知识管理、知识科学的基础是有可以操作的知识。这些知识从何而来，是知识抽取要解决的问题。知识抽取研究可用于学术文献中的知识提取，如文献自动综述；还可用于学术规范，如参考文献自动标注等。因此，开展本项研究，不仅具有一定的理论价值，而且还有较好的实际应用价值。以句子匹配分析为处理重点的知识抽取研究的应用价值表现在以下两大类典型应用系统：以相似句段或重复句段为重点的学术抄袭检测系统与参考文献自动标注系统；以新句段为重点的文献自动综述系统与知识库构建系统（见图 1－1）。

1. 学术抄袭检测与参考文献自动标注系统

抄袭检测系统是一种实用的工程系统，通过比较文章句段之间的相似性判断抄袭程度。相似度越高，引用或抄袭别人的可能性就越大；相似度越低，说明该句子作者自写的成分就越高。抄袭检测是一种后控检测，如果作者在论文投稿之前用系统检测一遍，对检测出来的重复内容