# 基于篇章内容分析的文本信息处理系统差异性探析

**Exploring the Difference among Information Processing Systems Based on Text Content Analysis**

化 柏 林

(中国科学技术信息研究所　北京 100038)

**摘　要**　以篇章内容分析在知识抽取、自动文摘、自动问答、文献自动综述的作用为切入点,剖析这四类信息处理系统的分类、主要流程、关键技术。然后分析出知识抽取、自动问答、自动文摘、文献自动综述在处理对象、处理结果、处理过程、分析层面、分析粒度等方面的区别,总结基于篇章内容分析的四类信息处理系统之间的共性与发展现状,并探讨它们的发展趋势。

**关键词**　篇章内容分析　信息处理系统　知识抽取　自动问答　自动文摘　文献自动综述

**中图分类号**　TP391　G31

篇章内容分析[1-4]的研究得到了较多的关注,它的应用前景与现实意义也越来越被看好。一篇文献的内容分析,包括标题、关键词、摘要与正文。从自然语言处理的层面上看,由浅入深,从形式分析到句法分析再到语义分析,最终实现对文献的内容理解。按照语言构成的单位,由小到大,从词语分析到句段分析,最终进行篇章分析。

篇章内容分析与情报研究方法里的内容分析法有所不同,因为内容分析法是一种基于定性研究的量化分析方法[5],定性只是一种针对性的定性,最终的结果还是定量。而篇章结构内容分析一般不区分定性或定量,只有统计与规则之分,最终的结果是为了理解而不是定量,如机器翻译需要篇章结构内容分析,但没有人称之为内容分析法。内容分析法既然是一种方法,就是用来解决其它问题的,篇章结构内容分析本身就是一种处理,一般不作为一种方法来认定。篇章结构内容分析更注重自然语言处理,而内容分析法对自然语言的处理一般只涉及到词的层面。

知识抽取、自动问答、自动文摘与自动综述都是图书情报的前沿领域,是图书情报领域需要解决的前沿问题。无论是知识抽取、自动问答还是自动文摘、自动综述,按照领域覆盖度都分为通用领域与受限领域。前者适用于某一领域,概念比较集中,因此概念体系显得不重要,重点在于句型的分析;后者适用于所有领域,主题概念庞大繁杂,概念体系显得较为重要,而句型分析相对简单。从语言的处理粒度上(主要是对源文献的分析处理)分为词语级自动＊＊、句子级自动＊＊、段落级自动＊＊、篇章级自动＊＊;从语言的处理层面分为语形级自动＊＊、语法级自动＊＊、语义级自动＊＊、语用级自动＊＊。

## 1　知识抽取

知识抽取是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[6])逐条抽取出来,对知识的属性进行标记,以一定形式存入知识库中。文献的主体是用自然语言来描述的,因此知识抽取过程无法完全回避自然语言处理。面向知识抽取的自然语言处理包括句子切分、自动分词、词性标注、词义标注、句法分析、句义分析、语段分析及语用分析八大模块,其中前四个模块是基础,句法分析与句义分析是核心,语段分析与语用分析是扩展。在这八个模块的运行过程中,需要关键词词库、概率词典、语义词典、句法规则、领域叙词表与领域本体等六类资源的支撑。知识抽取过程主要是在自然语言处理的基础上,从文献表述的角度进行识别与抽取,包括论文类型分析、篇章结构分析、知识抽取、知识表示四大模块,其中前两个模块是基础,知识抽取是核心,知识表示是扩展。在这四个模块的运行过程中,需要论文内容元数据、指示词与句子功能关系、句型与句子功能关系、文献学本体以及语言学本体五类资源的支撑[7]。

利用标点符号与段落标记把文章的正文切分成句子,然后到句子库进行匹配分析,滤掉学术抄袭与科学引用的句子,得到可能含有新知识的句子,利用向量分词等方法对这些新句子进行分词。经过分词以后,利用规则及隐马尔科夫模型等方法进行词性标注。词性标注以后,进行词义标注,根据语境、语义词典等信息标记出每个词在句子中的语义。在以句子为操作重点的知识抽取中,句法分析必不可少,从成分结构与功能结构两个层面进行分析。

知识抽取旨在对学术论文的写作结构、写作风格、句型

作者简介:化柏林,男,1977 年生,助理研究员,硕士,研究方向为文本信息分析。

结构等特征进行规律性的总结,建立定义、分类、发展历史、关键技术、应用前景、发展趋势等内容元数据。然后根据关联词和主题概念确定段内句子间的关系,确定句型与句子的功能关系、指示词与句子的功能关系,在此基础上实现对论文类型分析和篇章结构分析。根据不同类型的论文内容元数据,选择不同的知识抽取模式,如对定义的抽取只是从句子的线性表达中抽取,对分类的抽取要借助于主题词表、概念体系结构等。从语段中抽取了知识以后,把用自然语言描述的句子通过知识表示转换成计算机可理解的形式,基本概念采取面向对象知识库,概念之间的关系采用语义网表示,论点、结论等采用命题逻辑表示方式,处理流程、实验过程等采用产生式表示方式[7]。

## 2　自动问答

自动问答按照答案来源分为基于常见问题集的问答系统[8]、基于百科知识的问答系统[9]、基于文献的问答系统。基于常见问题集的问答系统根据事先积攒的问题及答案[10],如《十万个为什么》、百度的"知道"等,查找相似问题与答案,把答案返回用户。此类系统问题数量有限,答案较为集中、分析过程较为简单,只是对问句有相似性判定的分析,对目标文献的分析很少,不涉及答案的组织与生成问题,适宜网上咨询系统。基于百科知识的问答系统把现行的百科知识的问答导入知识库,如《中国大百科全书》、《维基百科》,根据问题查找相关知识,然后把相关知识以答案形式返回给用户。基于文献的问答系统是经过对问句进行分析以后,提取关键词,根据词表(叙词表或轻量级本体)把关键词扩展成一系列词组,根据与问句中的关键词的相关关系为每个词分配权重,构成几个检索表达式,利用检索引擎搜索目标文献,再从目标文献中抽取答案。基于常见问题集的问答系统只涉及提问式与现有问句的匹配问题,不涉及答案的分析操作。基于百科知识的问答系统涉及提问式与现有问句的匹配问题,也涉及提问式与答案的匹配问题。基于文献的问答系统不涉及提问式与现有问句的匹配问题,主要涉及提问式与目标文献的相关性计算,以及提问式与目标文献中相关内容的匹配问题。

问答系统一般包括三大模块:问题分析、文档检索和答案抽取。一般的问答系统对提问式进行深度分析,对目标文献进行浅层次分析,并对答案进行验证。首先到问题库里去查找,如果有相同问题,就直接返回答案。如果没有相同问题,就对问题进行详细分析,包括分词、词性标注、句法分析、问题分类、答案类型分析。然后到问题库里查找是否有相似问题,如果有就替换相关内容并返回答案。如果没有相似问题,就从提问式中抽取关键词,对关键词进行扩展构成检索式执行文档检索[11],根据返回结果抽取相关段落或句子,对句子进行排序与重组,以及关联处理,以达到摘要的要求。如果答案要求是词或短语型的,就进行信息抽取,识别出答案并进行验证[12]。对答案进行验证主要是答案的句子本身是否合法、合义、合用等情况进行验证。合法指句子本身没有语法错误,是合法的句子;合义指句子的语义表达没有问题;合用指句子符合语境,避免出现答非所问的情况。

自动问答的关键包括识别问题的焦点、对问题进行合理分类并确定答案的类型,从问题集、知识库或文献中搜索相关知识点并形成答案。

## 3　自动文摘

自动文摘按照处理的技术路线分为基于统计的摘要(又称为机械式自动文摘)与基于规则的自动文摘(又称为理解式自动文摘)。按照摘要句的来源形式分为抽取式摘要与生成式摘要。抽取式摘要利用各种分析方法确定与文章的主题强相关的句子,并把它们抽取出来,以一定形式组合起来形成摘要。生成式摘要是在理解文章内容的基础上,按照摘要的写作规范,生成一些代表文章内容的句子,并把这些句子连续组合在一起形成摘要。按照对原文献的处理入口又可分为三类,即基于形式的自动文摘、基于结构的自动文摘[13~14]和基于理解的自动文摘。基于形式的自动文摘根据词频、句长、句子位置、词与句子关系等信息直接从文章中抽取一些句子组合起来形成文摘,依靠词频与句子的统计关系等形式化的信息抽取文摘句是基于形式的自动文摘的显著特征。基于结构的自动文摘充分分析文章句子之间的关系,然后根据句子的功能从文章中抽取相关句子组成文摘,以篇章结构和语用分析为突破口是基于结构的自动文摘的显著特点。基于理解的自动文摘又称基于语义的自动文摘,是充分分析句子的内部结构和句子的语义概念,以期理解句子的意思,然后根据句子与主题的相关度确定文摘句。以主题概念和句法结构来分析句子语义是基于理解的自动文摘的关键特征。

一个通用的自动文摘系统往往不是使用单纯的一种方法和技术,而是综合运用几种方法和技术,针对不同的处理对象,在不同的处理阶段使用不同的技术方法,以期达到最佳效果。一般来讲,自动文摘主要包括原文形式分析、篇章结构分析、文摘语言的生成以及文摘的可读性处理。首先进行原文形式分析与原文结构分析,通过原文的形式分析与结构分析获取文摘候选句并确定文摘的类型,对其进行句法分析甚至语义分析,确定最终的文摘句。最后把这些句子组合起来,并进行适当的处理,以提高摘要的连续性与可读性。原文形式分析包括分词、加权计算等过程。原文结构分析主要对句子之间的关系进行分析以及文章的写作结构进行分析。对原文进行形式分析与整体结构分析,对文摘句进行句法分析、语义分析甚至语用分析。对原文进行粗浅分析能够保证处理的速度,对文摘句进行精深分析保证文摘句的质量。

自动文摘的关键技术是句子相似度的计算,句子相似

度计算的方法有基于公共子串的方法[15~16]、基于编辑距离的方法[17~18]、基于语义依存的方法[19]、基于辞典的方法、基于语法结构的方法等。

## 4 文献自动综述

目前,综述型文章大都是人工完成的。如果运用句子匹配分析技术,把相关主题的文章综合到一起,进行句子级的滤重与重组,就可以实现综述型文章的自动完成,即文献自动综述。文献自动综述可满足学习型搜索与观点型搜索。学习型搜索的检索结果只有一条,不再显示所有符合检索需求的成千上万篇文章,而是一篇综合了所有满足检索需求的文章,文章会有发展历史、主要分类、使用技巧与方法、关键技术实现、发展趋势等多个主题。这样就由阅读多篇文章变成了阅读一篇文章的不同部分,实现了内容的滤重与重组。学习型搜索是文献自动综述的典型应用[20]。

观点型搜索是查询与某一观点一致或相反的文章,如查询持有观点"数据挖掘不同于知识发现,而是知识发现的一个阶段"的文章。当然也可以按观点进行聚类检索,如"数据挖掘与知识发现的关系",或者"知识管理的各种流派",检索结果按观点或流派进行聚类。观点型搜索是指根据某观点进行搜索,以自然语言形式输入,搜索含有某个观点的文章,或者关于某个知识点的所有观点,如查所有数据挖掘和知识发现的关系,其结果将不再是一篇一篇的文章,而是一个列表。列表显示几种不同的观点,数据挖掘是知识发现的一个步骤,有 x 篇文章;数据挖掘也就是知识发现,是同一个概念,有 y 篇文章;数据挖掘与知识发现是完全不同的两个概念,有 z 篇文章。这样查到的不是成百上千篇关于数据挖掘与知识发现关系的所有文章,而是三种观点,也就是三条记录,每种观点分别有多少人论述。如果想详细了解某一种观点时,就点击相应记录,系统会显示关于这种观点有哪些论述方式,分别是如何来论述的,也就是真正的知识链[20]。

文献自动综述不同于多文档自动摘要[21]。多文档自动摘要不需要覆盖文章的全部内容,只需要把文章的关键体现出来即可。多文档自动摘要的结果只是一个摘要,摘要的篇幅小于文章。而自动综述的结果相当于一本书,其篇幅大于文章。多文档自动摘要关注出现频率比较高的句子,而文献自动综述关注每篇文章中的新句子。

## 5 四类文献信息处理系统的差异性分析

自动文摘实现从大到小的变化,自动综述实现从多到少的变化。自动文摘需要覆盖每篇文章的全部内容,自动综述要实现多篇文章里所有非重复内容的重组。自动文摘结果的篇幅量远小于文章的篇幅量,而自动综述的篇幅量要大于其中任意一篇文章的篇幅量,而远小于所有(被综述的文章)文章篇幅量的总和。从入口来看,自动文摘关注文章内的重点句子,自动综述关注文章内的新句子。从出口来看,自动文摘必须反映整篇文章的主要内容,自动综述需要反映整个学科或整个研究领域的主要内容。

在现有的自动问答系统中,分析一般侧重于对问句的分析,而对目标文献及答案的分析相对来讲少一些。基于词法层面的问答系统只进行词法分析,主要包括分词过程,然后提取疑问词、主题词等词语,根据这些词语检索文献、抽取答案。自动文摘从语言处理层面也分为基于词法层面的自动问答、基于句法分析的自动问答、基于句义分析的自动问答。在现有的自动文摘系统中,分析一般侧重于对重点句的分析,分析遵循分析层面越来越深,分析内容越来越少。这四类基于篇章内容分析的文本信息处理系统差异性比较如表 1 所示。

表 1 四类基于篇章内容分析的文本信息处理系统比较

| | 处理对象 | 处理结果 | 处理过程 | 分析粒度 | 分析层面 | 用户 |
|---|---|---|---|---|---|---|
| 自动文摘 | 文章正文 | 文摘 | 从大到小 1∶1 | 词、句并重 | 语形为主、语法为辅 | 人用为主 |
| 自动综述 | 文章正文 | 大文章 | 从多到少 M∶1 | 句子段落并重 | 语形、语法并重,涉及语用 | 人用为主 |
| 自动问答 | 提问式、文章正文 | 短语、句子、段落 | 从大到小 1∶M | 句子为主、词语与段落为辅 | 语法、语义并重 | 为人服务 |
| 知识抽取 | 文章正文 | 知识元 | 从大到小 1∶M | 句子为主,段落为辅 | 语义为主、涉及语用 | 多为系统 |

从处理对象来讲,自动文摘、自动综述、知识抽取都是针对文章正文进行分析,处理过程不受用户干预。自动问答除对正文分析外,还涉及对提问式的分析。从处理结果来讲,自动文摘的结果是文摘,文摘明显小于正文,而且一篇文章抽取一篇文摘,是 1∶1 的关系,文摘不受用户干预。自动综述的结果是一篇更大的文章,把很多篇文章综合成一篇文章,实现从多到少的变化,处理对象与结果是 1∶M 的关系。自动问答的结果是问题答案,根据问题类型,答案可以是短语、句子,也可能是段落。但无论是哪种形式,结果都要明显小于处理对象,因此也是实现从大到小的变化。一篇文章可以抽取很多问题的答案,因此是 1∶M 的关系。知识抽取的处理结果是知识元,一篇文章可以抽取多条知识元,因此是 1∶M 的关系。知识抽取不依赖于使用者,自动问答因提问式的变化而变化。

自动文摘以词法分析为重点,词的权重计算为核心,适当涉及句法,从形态与语法两个层面进行分析。自动综述以句子与段落分析为重点,主要进行形态及语法方面的分析,适当涉及语用。自动问答以句子分析为主,以词语与段落分析为辅,重语法与语义是其特征。知识元的最小表示单位是句子,因此知识抽取不对词进行计算,而以句子分析为主,段落分析为辅,以语义分析为主,适当涉及语用分析。所有的系统分词是必需的。但有的系统对分词结果进行计算,有的系统分词是为了后续的句法分析。

自动文摘与自动综述的用户主要是人,但用户并不参与到处理过程中,处理过程与人无关;自动问答是直接面向用户的一种应用,其用户是人,应用实践过程需要用户参与,处理结果因人而异,也就是说针对同一问题,不同的用户有不同的提问方式,不同的提问方式影响到问题分析与返回结果。知识抽取不直接面向用户,抽取的结果一般由计算机来用,是面向系统的。

## 6 结　语

无论哪一种应用,其系统的实现都取决于系统需求以及技术的成熟程度,技术的成熟度包括支撑资源的拥有情况以及算法的复杂性与处理速度。每一种处理都或多或少地用到自然语言处理,都属于自然语言处理的应用。对文献进行篇章结构内容分析需要从主题、结构、表达方式等三个维度入手。对主题内容进行分析主要考虑词、句子、参考文献等特征;对逻辑结构进行分析主要考虑文章的逻辑框架、段落之间的关系、句子之间的关系等;对表达方式进行分析主要考虑长短句的运用,书面语的规范程度,数学公式、图书、程序、算法的使用偏好等。围绕句子分析是处理的重点,包括句子内部结构及语义表达[22]、句子之间的关系以及句子在篇章中的作用是篇章内容分析的核心。文献的主体都是用自然语言描述的(还有部分内容是用数学语言、程序语言或图形语言来描述的),注重自然语言的分析处理才能更好地解决这四类系统的问题。分析过程中,除了需要语言知识的支撑以外,还需要领域知识以及文献知识的支撑。

自动文摘已经研究了很多年,有着明确的应用需求和成熟的处理方法。自动问答是近几年的新热点,是搜索引擎之后一种新的扩展,是针对搜索引擎返回结果较多的一种有效补充。知识抽取是一件基础性的工作,只是把文献中现有的知识抽取出来,实现序化,体现不出多少创新,可是如果实现从文献中抽取知识的自动化,很多问题便会迎刃而解。如果说搜索引擎很好地解决了人们获取信息的方便性与快捷性,那么知识抽取将会实现人们获取知识的有效性与非重复性。自动综述的研究还没有起步,它与自动文摘一样是一件非常具体的事情,解决一个特定的问题,随着文献的指数增长与信息的爆炸,自动综述的需求也会越来越强。

## 参 考 文 献


1 王德亮. 语篇脉络理论述评 —宏观语篇处理[J]. 现代外语,2006,29(3):309 - 316

2 沈玮杰. 基于文献结构的自动文摘的初探[J]. 现代图书情报技术,2002(3):23 - 28

3 薛翠芳,郭炳炎. 汉语文本结构的自动分析[J]. 情报学报,2000,19(4):319 - 325

4 金　博,史彦军,滕弘飞. 基于语义理解的文本相似度算法[J]. 大连理工大学学报,2005,45(2):291 - 297

5 邱均平,王曰芬,颜端武. 内容分析法研究与发展综述[A]. 情报学进展(第六卷). 北京:国防工业出版社,2006

6 温有奎,徐国华,赖伯年等. 知识元挖掘[M]. 西安:西安电子科技大学出版社,2005

7 化柏林. 基于 NLP 的知识抽取系统架构研究[J]. 现代图书情报技术,2007(10):38 - 41

8 秦　兵,刘　挺,王　洋等. 基于常问问题集的中文问答系统研究[J]. 哈尔滨工业大学学报,2003,35(10):1179 - 1182

9 Jon Curtis, Gavin Matthews, David Baxter On the Effective Use of Cyc in a Question Answering System[EB/OL]. [2007 - 9 - 4]. http://www.cyc.com/doc/white-papers/KRAQ2005.pdf

10 樊孝忠,李宏乔,李良富等. 银行领域汉语自动问答系统 BAQS 的研究与实现[J]. 北京理工大学学报,2004,24(6):528 - 534

11 E M Voorhees. Overview of the TREC2003 Question Answering Track. In: Proceeding of the 12th Text Retrieval Conference (TREC). NIST. Gaithersburg. MD. 2003:54 - 68. http://trec.nist.gov/pubs/trec12/papers/QA.OVERVIEW.pdf(2007 - 9 - 4)

12 郑实福,刘　挺,秦　兵等. 自动问答综述[J]. 中文信息学报,2002,16(6):46 - 52

13 傅间莲,陈群秀. 基于规则和统计的中文自动文摘系统[J]. 中文信息学报,2006,20(5):10 - 16

14 刘　挺,王开铸. 基于篇章多级依存结构的自动文摘研究[J]. 计算机研究与发展,1999,36(4):96 - 105

15 王荣波,池哲儒. 基于词类串的汉语句子结构相似度计算方法[J]. 中文信息学报,2005,19(1):21 - 29

16 王荣波,池哲儒,常宝宝等. 基于词串粒度及权值的汉语句子相似度衡量[J]. 计算机工程,2005,31(13):142 - 144

17 李　彬,刘　挺,秦　兵等. 基于语义依存的汉语句子相似度计算[J]. 计算机应用研究,2003(12):15 - 17

18 Yasuhiro Akiba, Kenji Imamura, Eiichiro Sumita. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. http://www.eamt.org/summitVIII/papers/akiba.pdf(2007 - 9 - 4)

19 车万翔,刘　挺,秦　兵,李　生. 基于改进编辑距离的中文相似句子检索[J]. 高技术通讯,2004(7):15 - 19

20 化柏林,张新民. 从检索技术的实现看三大全文数据库的发展[J]. 图书情报工作,2007,51(10):13 - 16

21 秦　兵,刘　挺,李　生. 基于局部主题判定与抽取的多文档文摘技术[J]. 自动化学报,2006,30(6):905 - 910

22 雷　强,黎　林,赵　英. 基于语义的数字资源整合研究[J]. 中国信息导报,2007(4):34 - 36


(责编:贺晓利)