情 报 学 报 ISSN 1000 - 0135
第 33 卷 第 9 期 979 - 985, 2014 年 9 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 33 No. 9, 979 - 985 September 2014

doi:10.3772/j.issn.1000-0135.2014.09.010

学术文本的结构功能识别

——功能框架及基于章节标题的识别[1)]

陆 伟 黄 永 程齐凯

(武汉大学信息资源研究中心,信息检索与知识挖掘研究所,武汉 430072)

摘要 当前学术文本挖掘研究大多数是采用基于词汇、窗口、全文的方法,往往忽略了学术文本的内在结构,导致了很多歧义性问题。本文针对当前研究不足,提出一种研究性论文的结构功能框架,对学术文本的章节功能和逻辑结构进行了定义。在此基础上本文从三个不同层次(基于章节标题、基于章节内容和标题、基于段落)论述了结构功能的自动分类问题,并从第一个层次(基于章节标题)采用词表与序列标注相结合的方法进行了结构功能的自动分类实验,取得了令人满意的效果。

关键词 文本挖掘 结构功能 自动分类

The Structure Function of Academic Text and Its Classification

Lu Wei, Huang Yong and Cheng Qikai

(*Center for Studies of Information Resources, Wuhan University, Wuhan* 430072)

Abstract The current academic text mining research is mostly based on the word, the window and the full text. It ignors the internal structure, leading to a lot of ambiguity problems. In view of the current lack of research, this paper puts forward a kind of framework that gives definition about the structure function of the research papers' chapter. On this basis, from three different levels (based on the section headers, based on the section content and header, based on the paragraph) the automatic classification problem of structure function is discussed, and from the first level (based on the section header) by adopting the combination of vocabulary and sequence tagging method the automatic classification experiment of structure function is conducted, the satisfactory results have been achieved.

Keywords text mining, structure function, automatic classification

1 引 言

学术文本挖掘是当前研究的一个热点,随着研究的深入,研究对象呈现更细粒度化,研究内容更具有语义性。当前基于学术文本的研究与一般文本挖掘一样大都采用基于词汇、窗口、或者全文的方法,往往忽略了学术文本中重要的章节结构信息。与一般文本不同,学术文本具有一定的规范性,有着严密的内部逻辑结构,这些特性主要通过章节结构进行

收稿日期:2014 年 6 月 3 日

作者简介:**陆伟**,男,1974 年生,武汉大学信息资源管理学院情报学系,博士,副院长,教授,主要研究方向:信息检索、知识管理、数据挖掘等,E-mail :reedwhu@ gmail.com。**黄永**,男,1991 年生,武汉大学信息管理学院情报学系,博士研究生,主要研究方向:信息检索、数据挖掘。**程齐凯**,男,1989 年生,武汉大学信息管理学院情报学系,博士研究生,主要研究方向:信息检索、数据挖掘。

1) 本文系国家自然科学基金面上项目"基于语言模型的通用实体检索建模及框架实现研究"(项目编号:71173164);教育部人文社会科学基地重大项目"面向细粒度的网络信息检索模型及框架构建研究"(项目编号:10JJD630014)的研究成果之一。

体现。将文档各个章节的内容同等对待,会造成一些歧义问题,如在学术知识演化中一些常用词汇如“算法”在不同的学科领域中大量出现,使得所有的学科领域都因此相关[1],文献[2]也在研究中提出这样的问题;在引文分析研究中,传统的引文网络一直备受争议的原因就是无视参考文献被引的次数、章节位置等因素将所有参考文献同等对待,但实际上参考文献中对原作者撰写论文时有影响的文献只占少数[3],不同的引用位置、引用次数都是引文分析研究中需要考虑的重要因素;此外,在学术搜索中,由于忽略了某些关键概念在原文中出现的位置,导致用户无法获取到想要的结果。例如,用户想通过搜索SVM模型以获取与SVM提出或基于SVM进行算法改进相关的论文,但是搜索引擎往往会返回许多采用SVM模型解决具体问题的应用性文章。

本文认为学术文本中的每个章节都具有特定的功能,如某些章节用于介绍学术研究的背景,某些章节阐述了实验方法等。词汇、引用等在不同章节的出现也可能表现为不同的功能和作用。学术文本的章节结构在一定程度上承载了作者的研究思想和行文思路,章节功能的识别有助于实现对学术文本的语义理解。

目前,学术论文特别是研究性论文的章节功能和逻辑结构研究在语言学领域已经得到了一定的关注[4],虽然这些研究多停留在语言学分析上,但也在一定程度上揭示了研究性论文的基本结构包括引言、相关研究、方法、实验、结论等。本文以研究性论文(是指针对某一个(些)问题、现象进行深入分析、讨论并得到有意义结论的文章,不包括综述性论文、评论等)为对象,提出了一个针对其的结构功能框架及结构功能识别的三个不同层次,并在此基础上,利用CRF序列标注思想,结合自定义词表,实现了第一个层次基于章节标题的学术文献结构功能识别。

本文后续的结构如下,第二章对相关研究进行评述,然后在第三章对结构功能框架进行描述,并且论述学术文本的章节结构功能分类的三个层次:章节标题、章节内容、章节段落。第四章则从章节标题的层次,以序列标注的思想使用条件随机场模型(CRF)与词表相结合的方法初步探讨了结构功能的自动分类,最后对章节标题层次的分类中的不足进行分析,提出下一步的工作。

2 相关研究

本文提出的章节结构功能是每个章节在学术文献的内容层次上的功能作用。本文相关研究领域是文档的逻辑结构分析与识别。与本文研究不同,文档的逻辑结构分析与识别的目的是对文档的成分分析与识别,即分析和识别出文档内哪些文字是作者、标题、正文,哪些文字是页头、页尾等关于文档结构的组成成分[5]。文档的逻辑结构分析与识别研究相对成熟,主要包括布局分析和逻辑结构分析两个部分内容[6]。根据语料的结构化程度该任务划分为不同的等级,其中扫描件语料等级最高,难度最大,XML等结构化半结构化资源等级最低,难度最小[7]。布局分析通过对原始的PDF文档转化而来的图片的分析,将图片分割成为具有相同成分的片段。逻辑结构分析是使用位置特征、字体特征、布局特征以及OCR之后的文字特征判断出上述片段所属的类别(标题、正文、作者、页头、页尾等)。文献[6,8]基于PDF转换的图片使用人工神经网络方法完成文档的逻辑结构提取,文献[9]基于文本内容使用CRF方法完成逻辑结构识别工作。

相比较于文档的逻辑结构分析,本文研究的是章节在文献内容中所表达的功能,目前在这方面的研究较少。文献[10]在其引文分布的研究当中使用具有统一的四部分结构(引言(Introduction)、方法(Method)、结果(Results)、结论(Conclusion))的学术文献,对文献中的引文在上述四部分中的分布进行了可视化分析。文献[11]将文献结构分为6种,摘要(Abstract)、引言(Introduction)、相关研究(Literature Review)、方法(Method)、结果(Results)、结论(Conclusion),分别统计了参考(Reference)和提及(Mention)两种引文权重计算方式在这6种结构中的分布,并且使用自建词表和人工验证相结合的方式将学术文本的章节映射到上述6种结构中,最终的准确率也不是很高。上述研究中都使用不同结构文献同构化的方法,但是没有提出这样做的原因,并且所使用的分类方法过于简单不具有通用性。

Parscit[12]是一个学术文本的参考文献解析工具,其主要的功能有:论文的逻辑结构提取,即通过文本内容,以行为单位识别该行是标题、正文、页眉、页脚、注释、图表等文本的逻辑结构;参考文献的结构提取[13],即根据每一条参考文献文本,提取出其对应的作者、题目、期刊、年份等参考文献基本内容,并将该参考文献在文章中的引用上下文进行提取;通用结构分类(Generic Section Classification)[9],即给学术文本中章节赋予通用标签,如引言、方法、结论等。Parscit虽然提出了通用结构这个概念和对应

的分类方法,但是没有对这个概念做出具体的解释和定义。

可见,已有研究虽然意识到了学术文本统一逻辑结构功能的重要性,并在其分类方法、应用上进行了尝试,但仍然存在诸多局限性,如缺少对结构功能的定义、分类方法过于简单等。本文针对以上问题,首先对学术文本正文中章节的结构功能进行定义,接着提出了学术文本的章节结构功能框架,并从三个层次对结构功能的自动分类问题进行定义,最后在第一个层次进行了分类实验。

3 结构功能框架

研究性论文往往具有严谨的逻辑结构,从研究问题引入、研究背景介绍、解决方法的提出、验证到最终得出结论,各个章节都具有很强的目的性和功能性,本文称之为章节的结构功能。根据研究性论文的流程,综合前人研究成果[9,11],本文侧重于对学术文本的正文内容的结构分析,本文将章节的结构功能分为5种,即引言、相关研究、方法、实验、结论。具体如下:

(1)引言,指研究的引入、研究动机、研究目的等具有研究问题引入作用的功能。

(2)相关研究,具有该功能的章节主要是论述研究综述、相关工作、背景等介绍本文研究相关的文献和背景。

(3)方法,具有该功能的章节是论述论文中所提出方法、框架、概念、假设、设计,采用的工具等,主要是对论文研究思路的表达。

(4)实验,该功能对应章节主要论述实验数据、过程、评测、结果、发现、讨论,系统的设计、实现等实验过程。

(5)总结,具有该功能的章节是指论文的结论及描述未来的工作、应用等。

这五种功能既反映了章节的逻辑功能和目的,又构成了研究性论文的逻辑结构。本文将其定义为结构功能的主要目的就是要强调说明每一个章节对于论文的内容表达都具有特定的功能性作用,并且在具有不同功能的章节中的内容或者知识单元具有不同的作用和功能,因此结构功能的识别有很大的潜在应用价值。例如,在学术搜索中对具有不同结构功能的内容就进行加权,从而改善搜索结果;在学术文本知识挖掘中,不同结构功能当中相同的词汇可能具有不同的作用,代表不同的含义,表达不同的功能,从而为语义消歧工作提供一些线索;在共引、共现、引文分布等学术文本的科学计量学分析中,可以利用结构功能信息构建基于文章结构功能的计量标准等。

学术文本章节结构功能标注是一个分类问题,能够自动的进行章节的结构功能分类,是本框架应用于其他研究的关键。本文从整体到局部将结构功能分类问题分为三个层次。

第一,基于章节标题的结构功能分类。在只有学术文本章节标题的情况下,根据章节标题给出其对应的结构功能。学术文献中很多章节的标题可以直接体现出章节的结构功能,如章节标题是“引言”,就表明它所想要表达的是引言功能,其它类似的如“方法”、“模型”、“实验”、“结论”等,这时使用模式匹配的方法能够快速的给出章节的结构功能。然而一旦出现未登录词时该方法的作用将大打折扣。本文将以序列标注和词表相结合的方法进行基于章节标题的结构功能分类实验。

第二,基于章节全部内容的结构功能分类。由于章节标题的高度概括性,有时不能给出足够的信息,从而影响了分类效果。这时需要使用章节的内容对标题进行补充,从章节全文的角度来判断其结构功能,从而提升结构功能分类的准确率。

第三,基于章节中段落内容的结构功能分类。与基于章节全部内容的结构功能分类不同,基于章节中段落内容的结构功能分类考虑的是在只给出学术文本中一段文字的情况下,能否确定该段落应该给予什么样的结构功能标签。

这三个层次中,第一层次的结构功能分类简单、易行,能够快速应用于其他的研究当中;第二层次的结构功能分类要求更多的数据,能够为分类实验提供更多的线索;第三层次是在减少数据的情况下完成结构功能分类,使得结构功能的分类框架更具有一般性和适用性。

4 基于章节标题的结构功能分类

根据本文在上一部分提出的结构功能框架及其分类的三个层次,该部分将探讨第一个层次——基于章节标题的结构功能分类的方法和思想。

4.1 方法描述

基于章节标题的结构功能分类的任务是在只提供研究性论文章节标题的情况下,根据标题内容给

章节赋予结构功能标签。一般的分类算法忽略了上下文之间的关系,由于结构功能具有很强的顺序性,所以本文将基于标题的结构功能分类问题转化为一个序列标注问题,使用条件随机场模型[14](CRF)对问题进行训练和预测。线性CRF的表示如下:

$$p(y \mid x;w) = \frac{1}{Z(x,w)}\exp\sum_{i=1}^{n}\sum_{j} w_j f_j(y_{i-1},y_i,x,i)$$

$$Z(x,y) = \sum_{Z_{i,N}}\exp\sum_{i=1}^{n}\sum_{j} w_j f_j(y_{i-1},y_i,x,i)$$

其中,w为带估计的参数,f_j为特征函数,i表示对输入序列f_j在特征函数f_j上求和,这样可以保证对于变长的输入f_j有估计j数目的特征函数值。虽然在理论上来说,特征函数f_j可以与所有的x产生关系,但是在实际使用时,考虑到复杂性以及实际问题中输入之间的关系特征,可能选择的仅仅是当前输入以及前后一两个输入作为该特征函数的自变量。CRF的一个优点在于,不用假设输入x之间的独立性关系就能计算$P(y|x)$。而输入与输出之间的关系是通过CRF的使用者在特定的任务中指定的特征函数f_j以及CRF自动学习的参数w_j来体现。线性链CRF则对CRF有一定的条件限制:当前输出y_i除了与x有函数关系以外,只能与前一个输出y_{i-1}有关。在本文的结构功能分类中,功能之间有明显的顺序关系,前一个标题的功能的确定对下一个标题的结构功能有影响,这与线性CRF的特性正好吻合。

CRF的优点是考虑对全局的特征进行优化,但在本实验中一些标题可以根据其包含关键词直接确定,这时使用CRF进行优化时,其他的特征反而会对该标题的判断造成影响,所以本文在CRF的训练过程中加入自定义词表。

本文提出的方法步骤如下:

(1)根据自定义词表,首先将含有词表中的关键词的类别确定,如一个标题为“引言”,则其结构功能标为“引言”,一个标题中含有“实验”关键词,则该标题结构功能标记为“实验”;

(2)根据确定的章节标题的位置,计算各个标题与确定的类别之间的距离,增加到CRF使用的特征中去,然后进行CRF训练;

(3)CRF模型训练完成后,在CRF预测结果中,同样根据自定义词表对结果进行判断,将判断结果替换CRF的结果。

上述方法加入了自定义词表的目的就是为分类训练过程增加先验知识,从而达到提高分类效果的目的。本文将在下一小节叙述本文使用的词表和CRF训练所使用的特征。

4.2 特征选择

本节将叙述本文提出方法中所用到的自定义词表以及CRF模型训练所使用的特征。

(1)自定义词表

学术文献中很多章节的标题可以直接体现出章节的结构功能,含有特定关键词的章节有些就是为了表达特定的结构功能。如一个章节中的标题中含有“引言”这个词,则其对应的结构功能就是引言,含有关键词“实验”的章节标题的对应的结构功能就是实验。如表1所示,如果一个标题包含表中的关键词,则该标题就标记为对应的结构功能。使用关键词词表是为了给实验增加先验知识,使得能够根据关键词分类正确的章节标题不会被CRF分类错误。在本文中只是使用了表1中简单的几个关键词,是为了防止使用过多关键词导致分类器过拟合不具有通用性。

(2)序列特征

CRF的学习和预测是在样本的多个特征上进行的,因此,本文使用学术文本中的章节一级标题(去除摘要、致谢等不是学术文本正文的章节标题)作为实验对象,并从这些标题中提取以下特征。

1)标题所在的绝对位置和相对位置。绝对位置使用的是每一个标题对应的序号,相对位置是指将所有标题的位置分为10份,将标题所在的位置作为其相对位置。之所以使用章节标题的位置特征,是因为论文中章节结构功能的顺序性,如表达“引言”功能的标题一般处于文章第一个,描述“方法”功能的一般处于文章的中间。

2)标题中的前两个词。本文中使用前两个词作为特征,如果标题的长度小于2,则使用0进行补足。

表1 不同结构功能对应的关键词词表

结构功能	引言	相关研究	方法	实验	结论
关键词	introduction	literature reviewrelated work	method	experimentresultcase study	conclusion

3)整个章节标题。使用整个标题是记住数据集中已经标注过的实例。

4)与确定功能的标题之间的距离。是指在使用上述词表中的关键词确定文章中一些标题的结构功能之后,计算该标题与确定结构功能的标题之间的距离,如果一篇文章中没有根据上述特征可以确定的标题,则该特征设置为 -9999。

4.3 实验及结果

4.3.1 实验数据及其标注

本文在 Journal of the Association for Information Science and Technology(JASIST) 2000 ~ 2012 年 12 年的数据中随机抽取了 300 篇研究性论文,作为实验数据。如图 1,网页上 JASIST 论文的全文样式,红色框内标明的是章节标题,蓝色框内标明的是章节内容。抓取对应的网页后根据论文全文中元素各自的样式将网页转换为 XML 格式。

在转换后的 XML 格式的全文中提取出每一篇论文章节标题,并将标题写入文件中,每一个标题一行,两篇论文之间标题空一行表示分割。

在得到论文的章节标题之后,请武汉大学信息管理学院的研究生进行标注,标注中在上述的标题之后加上标签的代号,如图 2 所示,每一个标题后面标注 1 ~ 5 中的一个数字,分别代表本文提出的 5 种结构功能。

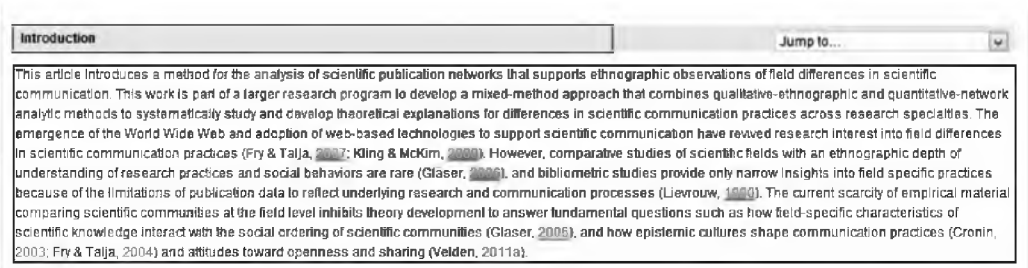
4.3.2 特征抽取

得到标注数据之后,将章节标题中的词转化为小写,并进行词干提取,之后对第三章中提出的用于 CRF 序列标注的序列特征进行抽取。如图 3 中,五种结构功能在标题的相对位置中的分布比例,可以看出标题的相对位置与五种结构功能之间具有很强的相关性。本文使用了标题的前两个词作为词汇特征,并将不存在的位置使用 0 代替,如表 2 所展示四种特征中词频大于 20 的词,可以看出章节标题特征中的高频词分布还是比较稳定的,这也是文献[11]的实验中能够使用关键词进行简单预分类的原因,本文同样只使用了本文上节给出的词表在上述的训练语料中进行分类实验,最终只能得到 61% 的准确度,并且其余部分都不能识别标题的类别。

4.3.3 实验设置及结果分析

本文使用 Parscit[12] 工具中通用结构分类方法[9]在本文标注数据集上进行实验,使用了位置信息、词汇信息作为特征,并且使用了 CRF 进行训练,并将该实验作为本文提出方法的对比实验。

笔者使用 CRF + + 工具包对本文提出的思路进行实验,并使用准确率、召回率以及 F 值作为评测标准与 Parscit 的结果进行对比。经过 5 折交叉检验,得到表 3 的结果。从表中可以看出上述两种方法相比较于基于词表的方法都有很大的提升。在本文方法结果中引言、结论两种结构功能的 F 值都在 96% 以上,这两种结构功能已经达到了很高的正确率,因此本文方法相对于 Parscit 只有很小的提升。结构功能中实验的 F 值在 91% 以上,相对于 Parscit 有小幅度提升。相关研究、方法两种结构功能是本实验的难点,相较于 Parscit,本文方法有较大幅度提升,提升幅度分别为 3.22% 和 3.65%,但从绝对值来看,这两者的效果仍有较大的提升空间。


Introduction

Jump to...

This article introduces a method for the analysis of scientific publication networks that supports ethnographic observations of field differences in scientific communication. This work is part of a larger research program to develop a mixed-method approach that combines qualitative-ethnographic and quantitative-network analytic methods to systematically study and develop theoretical explanations for differences in scientific communication practices across research specialties. The emergence of the World Wide Web and adoption of web-based technologies to support scientific communication have revived research interest into field differences in scientific communication practices (Fry & Talja, 2007; Kling & McKim, 2000). However, comparative studies of scientific fields with an ethnographic depth of understanding of research practices and social behaviors are rare (Glaser, 2006), and bibliometric studies provide only narrow insights into field specific practices because of the limitations of publication data to reflect underlying research and communication processes (Lievrouw, 1990). The current scarcity of empirical material comparing scientific communities at the field level inhibits theory development to answer fundamental questions such as how field-specific characteristics of scientific knowledge interact with the social ordering of scientific communities (Glaser, 2006), and how epistemic cultures shape communication practices (Cronin, 2003; Fry & Talja, 2004) and attitudes toward openness and sharing (Velden, 2011a).

图 1 JASIST 全文的网页格式

Two Kinds of Questions 1
The Core of Discovery System 2
How Homer Works 3
Does Homer Work?4
Discussion 4
Speculations 4
Conclusion 5

图 2 已标注的章节标题片段

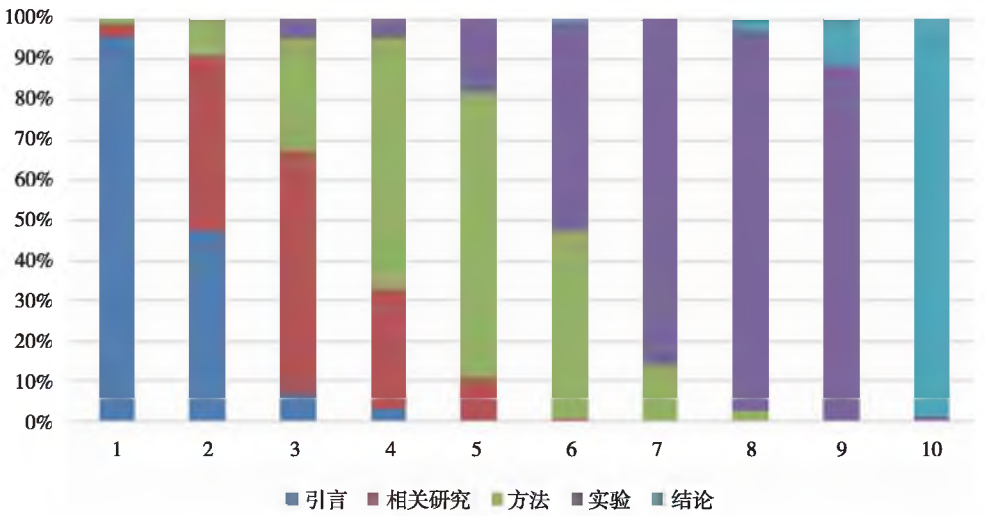


图3 五种结构功能在标题相对位置中的分布比例

表2 章节标题前两个词的高频词

词位置	引言	相关研究	方法	实验	结论
第一个词	introduction	background literature relate	method research methodolog	result experiment experi	discus conclus
第二个词	0	review 0 work	design 0 and	0 of and	and 0

表3 实验结果

	Parscit			本文方法			F值的相对提升
	准确率	召回率	F值	准确率	召回率	F值	
引言	0.979 718	0.955 681	0.967 55	0.986 324	0.952 347	0.969 038	0.15%
相关研究	0.795 36	0.790 043	0.792 693	0.831 701	0.805 237	0.818 255	3.22%
方法	0.781 943	0.804 793	0.793 204	0.811 051	0.833 52	0.822 132	3.65%
实验	0.897 538	0.907 378	0.902 431	0.904 775	0.927 53	0.916 011	1.50%
结论	0.989 998	0.968 185	0.978 97	0.986 938	0.974 435	0.980 647	0.17%

4.3.4 实验分析

实验后,笔者对错误的实验分类结果进行分析,发现存在如下问题:

(1)一篇论文中连续多个章节标题都是未登录词。这里未登录词是指在我们训练集中没有出现的词,在一篇论文中连续几个章节标题中大部分词汇都是未登录词,导致多个标题分类错误。如果只有单个章节标题出现未登录词,根据句子的位置特征、上下文特征,分类器可以对其正确的标注。但在这种情况下,章节标题不能够给分类器带来足够的分类线索,导致分类错误,因此可能需要使用章节内容对章节标题进行补充来增加特征中的词汇线索,从而提升分类器的分类准确率,这种情况就需要在第二个层次下进行。

(2)章节标题中包含了多种功能,如"Introduction and Literature Review"的章节标题中既包含了引言功能也包含了相关研究的功能。这类问题可能的解决方法有:①对上述类似标题给出多类标记,如标题为"Introduction and Literature Review"的章节标注为"引言"和"相关研究"两种类别都应该是正确的;②出现上述情况的另外一个原因是因

为标题层面的粒度不足以解决章节中包含多种结构功能的问题,所以可能需要结合章节的具体内容进行,从分类问题的第三个层次去解决,此时结构功能分类的对象不单单是章节标题,而是针对章节中的自然段,对章节中的自然段进行结构功能分类。

(3)一篇论文中的章节没有顺序性。本文假设研究性论文中结构功能顺序为引言、相关研究、方法、实验、结论,大多数的研究性论文也是采用这样的逻辑顺序。但是很多会议论文中,相关研究在论文的最后进行论述,或者很多其他研究性论文在开篇就对自己的理论进行阐述,所以基于章节标题的CRF分类器,在不具有明显顺序性的论文的章节结构功能中分类效果稍差。

5 结论及展望

本文提出了一种学术文本的结构功能框架,从基于章节标题、基于章节内容和标题、基于段落三个层次论述了结构功能的自动分类问题,并且从第一个层次对结构功能分类做了具体实验探索,得到了令人满意的效果。下一步的工作将结合当前存在的问题,从结构功能分类的第二、第三个层次入手,提升学术文本结构功能分类的准确率和通用性,并进一步探索结构功能框架在其他问题中的应用。

参 考 文 献

[1] Qikai Cheng, Xiaoguang Wang, Wei Lu, et al. NEViewer: A New Software for Analyzing the Evolution of Research Topics [J]. Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics, 2013: 1307-1320.

[2] 王晓光,程齐凯. 基于NEViewer的学科主题演化可视化分析[J]. 情报学报, 2013, 32(9): 900-911.

[3] Xiaodan Zhu, Peter Turney, Daniel Lemire, et al. Measuring academic influence: Not all citations are equal [J]. Journal of the Association for Information Science and Technology, 2014, doi:10.1002/asi.23179.

[4] Carole Slade. Form and Style: Research Papers, Reports, Theses[M]. Houghton Mifflin Company, 1997.

[5] Song Mao, Azriel Rosenfeld, Tapas Kanungo. Document structure analysis algorithms: a literature survey [C]. International Society for Optics and Photonics, 2003: 197-207.

[6] Simone Marinai, Marco Gori, Giovanni Soda. Artificial neural networks for document analysis and recognition [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005, 27(1): 23-35.

[7] Koji Nakagawa, Akihiro Nomura, Masakazu Suzuki. Extraction of logical structure from articles in mathematics [C]. Springer, 2004: 276-289.

[8] Belaïd A, Rangoni Y. Structure extraction in printed documents using neural approaches [M]//Machine Learning in Document Analysis and Recognition Springer Berlin Heidelberg, 2008: 21-43.

[9] Luong M T, Nguyen T D, Kan M Y. Logical structure recovery in scholarly articles with rich document features [J]. International Journal of Digital Library Systems (IJDLS), 2010, 1(4): 1-23.

[10] Hu Zhigang, Chen Chaomei, Liu Zeyuan. Where are citations located in the body of scientific articles? A study of the distributions of citation locations [J]. Journal of Informetrics, 2013, 7(4): 887-896.

[11] Ying Ding, Xiaozhong Liu, Chun Guo, et al. The distribution of references across texts: Some implications for citation analysis[J]. Journal of Informetrics, 2013, 7(3): 583-592.

[12] Isaac G Councill, C Lee Giles, Min-Yen Kan. ParsCit: an Open-source CRF Reference String Parsing Package [C]. 2008:

[13] Kiat N Y. Citation parsing using maximum entropy and repairs[R]. Tech. rep., National University of Singapore, 2005.

[14] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]//Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc, 2001: 282-289.

(责任编辑 马 兰)

情 报 学 报 ISSN 1000 - 0135
第 33 卷 第 9 期 986 - 993, 2014 年 9 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 33 No. 9, 986 - 993 September 2014

doi:10.3772/j.issn.1000-0135.2014.09.011

专利领域本体概念语义层次获取[1)]

李军锋 吕学强 李 卓

(北京信息科技大学网络文化与数字传播北京市重点实验室,北京 100101)

摘要 根据专利领域本体构建的需求,提出一种专利领域本体概念语义层次获取方法。通过分析专利领域技术主题概念在形式化时的构词规律以及上下位关系的表现方式,利用相对修饰度和关联规则识别上下位关系。然后分析上下位关系的特性,总结关系冗余和关系冲突的消除规则,构建出专利领域概念语义层次图。实验结果表明,上下位关系识别方法具有较高的准确率和召回率,构建概念语义层次图的方法取得了较好的关系冗余和关系冲突的消除效果,证实了本文方法的有效性。

关键词 专利领域 本体 上下位关系 概念语义层次

Deriving Concept Semantic Hierarchy of Ontology in Patents

Li Junfeng, Lv Xueqiang and Li Zhuo

(*Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing* 100101)

Abstract For the demand of ontology construction in patent domain, we propose a concept semantic hierarchy induction approach. For this purpose, the work isdiscomposed to two dimensions. First, we analyze the word-formation rules and hierarchical relation presentation forms of technology theme concepts in patent domain. Based on this, a relative decoration based approach and an association rule based approach areproposed to hierarchical relation extraction. Second, characteristics of hyponymy relations areanalyzed to achieve those rules to eliminate redundancies and conflictions in the extracted relation. The experimental result shows that the approach of hierarchical relation extraction can achieve high accuracy and recall rate, the approach of concept semantic hierarchy induction can achieve satisfied elimination result. The resultprovesthe validity of the approach in this paper.

Keywords patent, domain ontology, hierarchy relationship, concept semantic hierarchy

1 引 言

专利文献作为技术信息最有效的载体,囊括了全世界 90% 以上的最新技术情报,相比一般技术刊物所提供的信息早五至六年。而且 70% ~80% 的发明创造只通过专利文献公开,并不发布于其他科技文献[1]。因此,相对于其他的文献形式,专利文献更加新颖和实用。

专利文献摘要是对专利主要内容的概括性描述。由中华人民共和国国家知识产权局对于专利文献摘要的撰写规定[2]可知,规范的专利文献摘要中

收稿日期:2014 年 7 月 2 日

作者简介:**李军锋**,男,1990 年生,硕士研究生,主要研究方向:中文与多媒体信息处理,E-mail:lijunfeng1990@live.cn。**吕学强**,男,1970 年生,博士,教授,主要研究方向:中文与多媒体信息处理。**李卓**,男,1983 年生,博士,讲师,主要研究方向:移动互联网。

1) 基金项目:国家自然科学基金项目"基于本体的专利自动标引研究"(61271304);北京市教委科技发展计划重点项目暨北京市自然科学基金 B 类重点项目"面向领域的互联网多模态信息精准搜索方法研究"(KZ201311232037);北京市属高等学校创新团队建设与教师职业发展计划项目(IDHT20130519)