基于概念获取的多文档主题划分研究*)

孔庆苹 刘宗田 廖 涛
(上海大学计算机科学与工程学院 上海 200072)

摘 要 对多个相关文档进行主题划分对于信息检索、自动摘要等研究领域都有重要的应用价值。当前流行的文本主题划分技术中,多采用词频向量进行文本表示,而研究表明将特征向量映射到概念级,将改善多文档主题划分的效果。本文提出了一种应用知网(HowNet)来获取多文本的概念作为特征向量,再应用聚类的方法对文档集中的相似段落进行归类,得到主题划分的结果,解决了多文档的结构分析问题。实验结果表明该方法对多个相关文档的主题划分取得了良好的效果。

关键词 信息处理,知网,概念,主题聚类,主题划分

Study on Topic Partition Based on Concept Retrieval in Multiple Documents

KONG Qing-ping LIU Zong-tian LIAO Tao
(School of Computer Science and Engineering, Shanghai University, Shanghai 200072, China)

Abstract Topic partition is a significant problem during text structuring in many domains such as information retrieval and automatic summarization. The popular technique is using the frequency of words to express the documents, but using the concept will improve the efficiency of topic partition in multiple documents. The paper presents a method that uses the HowNet to get the concepts, and then uses the technique of clustering to segment the paragraphs of the documents. And this method solves the problem of text structuring in multiple documents. The experimental results show that this method is more efficient for topic partition in multiple documents.

Keywords Information processing, HowNet, Concept, Document clustering, Topic partition

1 引言

相关文档的子主题切分主要是将以自然段为基础的文本的物理结构转换成以意义段为基础的文本的逻辑结构,这样可以提高中心主题的覆盖率并在一定程度上去除了多文档间的信息冗余。子主题划分对信息检索,自动摘要,问答系统等研究领域都有重要的应用。

目前在国内外对文本主题划分的相关研究中,多采用词频向量[1]对文本进行表示。由于自然语言的多样性,传统的基于统计词频的获取特征量的方法,只是仅仅依靠了特征词的重复频率信息,这是不够的。有的研究通过使用建立段落向量空间模型,根据连续段落相似度进行文本主题划分的算法,来解决文章的篇章结构分析[3,4],但这种算法只能对连续段落进行主题划分,灵活性不高。文献[7]提出了使用 TF * PDF 算法从日文新闻中提取主题,文献[2]采用基于匹配和统计的方法从中文中抽取主题,这些方法都没有考虑表达同一主题的不同词之间的语义关联,对于相关多文档的主题划分效果也不明显。

由于概念空间比词空间小而且各分量间相对独立,因此,概念特征比词特征更适合用来表示文本内容[5]。将特征量映射到概念级,通过概念映射和文档的概念词扩展处理,将有助于使同类文档间的相似度加大,而使不同类文档间的相似度减小。因此,采用概念作为文本特征将会对文本主题划分有较好的效果。美国 NEC 国家实验室则提出一种概念分解的聚类方法。他们主要采用 WordNet 作为知识源来获取语义,证明效果不错。

本文提出了一种基于概念获取的多文档主题划分方法。后面的内容组织如下:第 2 节介绍概念和基于知网的概念获取;第 3 节介绍用概念作为特征向量,用聚类方法进行文本主题划分;第 4 部分为实验和结果分析;最后对全文进行了总结。

2 概念获取

2.1 知网简介

知网(HowNet)是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库[5]。知网采用义原作为概念单位,义原是最基本的,不易于再分割的意义的最小单位,并利用知识词典为每个词定义一个 DEF 词条,DEF 词条由一个或者多个义原构成,这是我们进行概念获取的依据,可以通过提取词的 DEF 作为文档特征来获得文章内容在属性上的规律,以此来获取更加适合的概念作为文本的特征选项。

HowNet 中对于词语的描述格式如下:

W-C:本概念对应的汉语词汇;

G-C:表示汉语词汇的词性;

E-C:用汉语表述的关于本概念的例子;

DEF:对本概念进行语义描述的定义项,是描述概念的核心字段。

*)基金项目:本文受国家发改委项目基金(SNMCFIP-2006S001)资助。孔庆苹 硕士研究生,从事数据挖掘、自然语言处理等研究;刘宗田 博士生导师,教授,主要研究领域为人工智能和软件工程等;廖 涛 硕士研究生,从事 Web 信息抽取、文本分类等研究。

2.2 概念获取

2.2.1 文本预处理

首先文本进行分词处理后，过滤掉一些对文章影响不大的介词、虚词、数词等词语，只对一些关键的名词，动词等重要词语进行处理，得到一个候选词集。

2.2.2 概念获取

本文所指的概念指的是在文章中词义相关的基本词义单元，由于自然语言的多样性，一个概念可以对应文档中的一个词语，也可以对应多个词义相似的词。对于文章中的单义词并且在 HowNet 词典中存在的词语来说，只需要给该词语一个相应的概念就可以了。

对于多义词问题，本文提出了采用关联词汇链匹配的方法来消除多义词的歧义。这里的词汇链(Lexical Chain)[6]是指一个主题之下的一系列相互搭配的词之间共同组成的词系列。Morris 和 Hirst[9]最初引进词汇链用于文本分割，即分析文本的结构。在 HowNet 中，对每个词语给出了定义，并收录了义项的常用搭配，在文章当中，一个词语无论是什么义项，在同一个语法环境下，都必须与其它词语搭配使用才能表达出一个完整的意思，找到这些词语也就找到了文章的结构。因此，对于多义词，采用取其概念以及每个概念常用的搭配结构，加入链中，组成一个关联词汇链，然后，在文档中查找这个词语的义项及其邻近搭配，与关联词汇链进行相似匹配，判断与文中哪个义项的搭配结构最接近，记录相似匹配的个数，最后通过排序决定何种义项作为多义词语的实际义项。

例如：对于“爆炸”一词：

爆炸	定义	词性搭配
DEFI	FormChange|形变：StateFin＝{OutOfOrder}坏掉	
WordList1	飞机 核电厂 炸药	N+V
DEF2	Lighting|点燃：purpose＝{CauseToDo|使动：ResultEvent＝{FormChange|形变：StateFin＝{OutOfOrder|坏掉}}}	
WordList2	原子弹 核弹 堡垒 大楼	V+N

对于上面出现的两个词语，本文在相关文档中查找匹配的结构，若文档中与 wordlist1 这种 N+V 的结构匹配的次数多于 wordlist2 的 V+N 的匹配结构，则说明文档中描述的是发生了一件事情，表示一种状态，发生了一种形变，则选择 DEF1 作为在本文档中的概念。反之，则选择 DEF2 作为在本文档中的概念。

对于未登录词语，采用计算每个未登录词在多篇文档中出现的次数是否大于预先设定的阈值，如果是，则标注为另外一种不存在于概念获取中已经发现的概念；否则，删除那些在文章中出现次数小于阈值的未登录词语。

以概念及其重要度作为文本的特征向量，把文本表示成概念特征向量的形式，就可以对文本进行主题划分了。

3 划分文本主题

子主题的划分主要是通过合适的方法，把表达相似意思的句子或者段落聚成一个类，这样每一类就形成一个子主题，表达相近的信息。众所周知，多文档集合最大的特点就是信息的冗余，而聚类是消除冗余的一种有效策略[11]。

3.1 子主题的定义

从物理结构来看，多文档集合可以理解为多篇文档的集合，每一篇文档可以表示为文本单元的集合，即 $D=\{d_i|i=1,2,\cdots,n\}$，$d_i=\{p_{i,j}|j=1,2,\cdots,m\}$。从逻辑结构来看，一个主题是由不同侧面的信息组合而成的，每一个侧面信息称之为一个子主题(T_i)。因此，多文档集合又可以理解为由多个子主题构成的，$D=\{T_i|i=1,2,\cdots,l\}$，每个 T_i 是一个段落集合。实际上就是将多篇相同主题的文档界限打破，按照段落表达意思的相近程度进行重新组合。将多文档集合以子主题的形式表示出来。

3.2 用聚类方法划分文本主题

作为一种无监督的机器学习方法，聚类技术已经成为对文本信息进行有效的组织、摘要和导航的重要手段[10]。目前，有多种文本聚类算法，大致可以分为两种类型：一是以 GHAC 等算法为代表的层次凝聚法[8,10]；另一种以 k-means 等算法为代表的平面划分法[8,10]。层次聚类算法的计算复杂度一般为 $O(n^2)$，其中 n 表示输入文档的个数，随着信息量的增长，无疑这种算法在效率上欠佳，基于划分的聚类算法是目前最常用的聚类算法，其时间复杂度低，出于运行效率的考虑，本文中我们采用聚类效率较高的 k-means 聚类算法来划分文本主题。

3.2.1 相似度定义

本文使用欧氏距离度量文本之间的相似度。设文本 d_i，$d_j\in D$，则：

$$\mathrm{dist}(d_i,d_j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2}$$

x 表示文本向量中概念的重要度。

3.2.2 k-means 主题划分

假设 N 维样本空间的每个样本点分别对应于文档中的一个段落向量。这样，段落的聚类问题便可以直观化为 N 维样本空间中的 M 个样本点的聚类问题(N：文本中的特征概念数目，M：文档集中的段落数目)。

在本文中，首先进行概念约简，根据重要度来选取一定数目的概念，作为最终用于聚类的段落特征概念，这在一定程度上，降低了向量空间的维度，而实验证明，减少一些对文档主题中心影响不是很大的概念，并不会影响文档的聚类效果。

对文档进行子主题划分，最根本的目的是使得一个主题中的段落是相似的，而不同主题中的段落是不相似的，如果能够寻找到 m 个初始中心，它们分别代表了相似程度较大的段落集合，那么就找到了与段落在空间分布上相一致的初始聚类中心。

为了找到与文档段落在空间分布上相一致的、相似程度较大的段落集合，本文采取下列方法计算段落样本两两之间的距离，找出距离最近的两个点，形成一个样本集，并将它们从总的样本集中删除，最后得初始聚类中心，再应用聚类算法，做相关文档的主题划分。

该算法描述如下(m 初始化为 1)：

1) 计算任意两个段落样本间的距离 $\mathrm{dist}(p_{i,j},p_{t,n})$，找到样本集合 D 中最近的两个点，形成聚类集合 T_m，从样本集合 D 中删除这两个点。

2) 在样本集合 D 中找到距离聚类集合 T_m 最近的段落样本点 $p_{i,j}$，把 $p_{i,j}$ 加入 T_m，把段落样本点从 D 中删除，更新样本中心。

3) 重复第二步，直到距离聚类集合 T_m 中的样本中心的距离大于 $\alpha(0<\alpha<1)$。

4) 假设 D 中还有样本，m++，再从 D 中寻找距离最近的两个样本点，形成聚类集合 T_m，并从集合 D 中删除这两个点，return(2)。

5) 将最终形成的 m 个样本点分别进行算术平均，形成 m 个初始聚类中心。

6) 从这 m 个初始聚类中心出发，使用 k-means 聚类算法

进行子主题聚类，生成最后的子主题。

α 的取值因实验数据的不同而不同，α 值过小，则可能使几个初始聚类中心在同一区域内得到，α 值过大，则可能使初始聚类中心偏离聚类密集区域。从实验来看，α 取 0.75 时效果是比较好的。

4 实验及结果分析

由于文本主题没有绝对的答案，目前文本主题划分还没有理想的评估办法，而且最好是能有大量经专家进行了主题划分的文档用于测试，但是目前这方面的语料还比较缺乏。多篇文章如果没有各子标题的指引，不同人也可能会有不同的主题划分方法，得到的结果中主题数目与主题的分界都有可能不完全相同。

4.1 实验语料

大多数科技文献或篇幅较长的文章，为了使读者对整篇文章的组织结构有一个清晰的了解，便于读者对文章的总体把握，通常都含有各级标题。因此，本文在实验中采用了这种方法，我们从语料库中随机地抽取 5 组文献，选取其中 5 篇含有子标题的文献，组成 5 组，然后将子标题去掉，用本文的方法进行主题划分，这样效果评价就有了一定的可操作性。

4.2 实验结果及评价

分词工具使用的是中国科学院计算技术研究所研制的基于多层 HMM 模型的汉语词法分析系统 ICTCLAS，该系统具有中文分词、词性标注等功能。概念获取使用的是知网(HowNet)提供的概念集，在评测指标上，我们使用准确率来评价。定义如下：

表 1

语料集	文档 ID	段落数	子标题数	划分正确	划分后的主题数	准确率
Corpus1	d10000601	10	7	8	5 个主题	0.868
	d10000602	8	6	8		
	d10000603	9	7	8		
	d10000604	11	7	9		
Corpus2	d10100301	6	4	6	4 个主题	0.848
	d10100302	7	5	5		
	d10100303	5	7	5		
	d10100304	7	5	6		
	d10100305	8	6	6		
Corpus3	d10110701	8	6	7	5 个主题	0.853
	d10110702	7	5	6		
	d10110703	9	7	7		
	d10110704	8	6	8		
	d10110705	9	7	7		
Corpus4	d10111605	5	3	4	3 个主题	0.867
	d10111606	6	4	5		
	d10111607	5	4	5		
	d10111608	7	4	6		
	d10111609	7	5	6		
Corpus5	d10101405	4	2	4	2 个主题	0.857
	d10101406	5	3	4		
	d10101407	3	2	3		
	d10101408	5	3	4		
	d10101409	4	3	3		

$$P=n/N*100\%$$

其中 n 表示划分正确的段落数，N 表示总的段落数。

表 1 给出了 25 篇标准的文档测试得到的主题划分结果。

4.3 结果分析

从实验及其结果可以看出：

1) 实验中提取的概念的义项数目比词频方法大大减少，说明更多的词语包含于更少的概念之中，这样就可以获取那些词频统计中出现次数较少而表达的是文章一个重要概念的词语，从而能提高文章的查全率。

2) 这种算法对于多文档的主题划分准确率基本可以达到 85%以上。可以看到对于多数文章，基本上大部分相似的段落都可以正确地被划分到同一主题下。

3) 与基于连续段落相似度主题划分算法[4]相比，本算法不受相邻段落的限制，对一些主题分散的相关文档的主题划分效果更好。与基于词频主题划分[1]算法相比，用概念作为特征词，主题划分的准确率更高。

结束语 本文提出了一种通过获取多篇相关文本的概念作为文本的特征向量，然后对多文档进行主题聚类来进行文本主题划分的方法。实验结果表明，该方法对多文档的主题划分取得了良好的效果。这种主题划分方法，可以应用到信息检索、自动摘要、问答系统等多个研究领域。

本文的研究工作还有一些问题需要解决，如多文档划分的颗粒度问题，初始聚类中心的确定问题，知网中收录的词语有限等等，这些都是我们进一步的研究要解决的问题。

参 考 文 献

[1] 康恺，林坤辉，周昌乐. 基于主题词频数特征的文本主题划分[J]. 计算机应用，2006，26(8)：1993-1995

[2] 冯晋，李春平. 基于统计学和语义信息的中文文本主题识别技术[J]. 清华大学学报(自然科学版)，2005，45(S1)：1791-1794

[3] 傅间莲，陈群秀. 自动文摘系统中的主题划分问题研究[J]. 中文信息学报，2005，19(6)：28-35

[4] 傅间莲，陈群秀. 基于连续段落相似度的主题划分算法[J]. 计算机应用，2005，25(9)：2022-2024

[5] 廖莎莎，江铭虎. 中文文本分类中基于概念屏蔽层的特征提取方法[J]. 中文信息学报，2006，20(3)：22-27

[6] 索红光，刘玉树，曹淑英. 一种基于词汇链的关键词抽取方法[J]. 中文信息学报，2006，20(6)：25-30

[7] BUN KK, ISH IZUKA M. Topic extraction from news archives using TF * PDF Algorithm//The Third International Conference on Web Information Systems Engineering[C]. Singapore, 2002：73-82

[8] Hotho A. WordNet improves Text Document Clustering[A]// Proc. of the SIGIR 2003 Semantic web Workshop[C]. Toronto, Canada, 2003

[9] Morris, J Hirst G. Lexical Cohesion Computed by Thesaural relations as an Indicator of the Structure of Text[J]. Computational L linguistics, 1991, 17(1)：21-48

[10] XU R, Donald Winch II. Survey of Clustering Algorithms[J]. IEEE Transactions on Neural Networks, 2005, 16(3)：645-655