文章编号: 1003-0077(2007)02-0058-05

一种基于主题的文本聚类方法

赵世奇,刘 挺,李 生

(哈尔滨工业大学 信息检索实验室,黑龙江 哈尔滨 150001)

摘 要: 现有的文本聚类方法难以正确识别和描述文本的主题,从而难以实现按照主题对文本进行聚类。本文提出了一种新的基于主题的文本聚类方法: LFIC。该方法能够准确识别文本主题并根据文本的主题对其进行聚类。本方法定义和抽取了"主题元素",并利用其进行基本类索引。同时还整合利用了语言学特征。实验表明,LFIC 的聚类准确率达到 94.66%,优于几种传统聚类方法。

关键词: 人工智能;模式识别;基于主题文本聚类;基本类索引;语言学特征

中图分类号: TP391 **文献标识码**: A

A Topical Document Clustering Method

ZHAO Shi-qi, LIU Ting, LI Sheng

(Information Retrieval Laboratory, Haerbin Institute of Technology, Haerbin, Heilongjiang 150001, China)

Abstract: Few of the existing document clustering methods can detect or describe document topics properly, which makes it difficult to conduct clustering based on topics. In this paper, we introduce a novel topical document clustering method called Linguistic Features Indexing Clustering (LFIC), which can identify topics accurately and cluster documents according to these topics. In LFIC, "topic elements" are defined and extracted for indexing base clusters. Additionally, linguistic features are exploited. Experimental results show that LFIC can gain a higher precision (94.66%) than some widely used traditional clustering methods.

Key words: artificial intelligence; pattern recognition; topical document clustering; base clusters indexing; linguistic features

1 引言

随着网络信息的快速增长,提供一种有效的机制用来组织网络文本、帮助使用者获得他们想要的信息变得愈加重要。因此,文本聚类技术被广泛研究。虽然研究者已提出多种文本聚类方法,但是他们中很少能按照主题进行聚类。在本文中,我们提出了一种新的文本聚类方法,我们将该方法命名为结合语言学特征的索引聚类法(Linguistic Features Indexing Clustering: LFIC)。

在基于主题的文本聚类方法中,一个主要的问题是如何对"主题"进行描述。我们认为一个由一系列的有紧密联系的事件组成的主题[1]应该由包括参与者、地点、时间、道具、行为等一系列主题元素来表示。例如,在"2004 年 8 月 27 号,刘翔夺得雅典奥运会金牌"这一主题中,参与者为"刘翔",地点为"雅典",时间为"2004 年 8 月 27 号",道具是"金牌",行为是"夺得"。依照上述原则,我们建立了主题元素索引。这样,具有相同主题的文本可以被索引并聚类。

我们在实验中对 LFIC 与其他几种传统聚类方法进行了比较,其中包括会聚层次聚类(Agglomerative Hierarchical Clustering: AHC),K-均值聚类(K-means Clustering: KMC)以及后缀树聚类(Suffix Tree Clustering: STC)。结果证明 LFIC 方法

收稿日期: 2006-07-12 **定稿日期**: 2007-01-18

基金项目: 国家自然科学基金资助项目(60575042,60503072,60675034);腾讯基金资助项目

作者简介: 赵世奇(1981—),男,博士生,主要研究方向为复述。

可以在维持一个可以接受的召回率的同时得到较高的准确率。考虑到网络信息极大丰富,在许多应用中准确率比召回率更重要。因此,我们认为 LFIC 是有效的。

本文后续章节组织如下: 第二节简要回顾聚类研究的相关工作。第三节对我们提出的 LFIC 方法进行详细描述。第四章给出实验及结果。最后在第五节中进行总结并对未来工作进行展望。

2　相关工作

文本聚类方法大致可以分为两种: 层次聚类(Hierarchical Clustering)和非层次聚类(Partitional Clustering)。其中,会聚层次聚类方法(AHC)是层次聚类的一种,它对待聚类文本实现自下而上的逐层聚类,并最终形成一个树形结构。具体地,AHC 方法初始以每个文本作为一个类别,在接下来的每一步中,依次合并相似度最大的两个类。AHC 方法的优点在于能够清楚地显示整个聚类过程以及中间聚类结果。K-均值聚类方法(KMC)属于一种非层次聚类方法。该方法初始定义 K 个聚类质心,然后比较每个文本和质心。每个文本将被分配到与它最近的质心所代表的类。接下来,该方法根据每个类中所含的元素重新计算质心。这一分配和重新计算质心的过程将重复进行至每个质心不再变化为止。

虽然上述方法广泛应用,但他们存在一些共同的缺点。首先,它们都需要事先确定一个停止条件。比如,层次聚类要求事先确定所要聚成的类别数,而 K-均值聚类则需要设定 K 值。上述条件往往在实际应用中很难事先确定。除此之外,它们都不能很好地描述和解释聚类的结果。再有,这些方法限定每个文本只能属于一个类,而没有考虑一个文本可能属于多个主题的情况。

为了克服上述缺点,Zamir 和 Etzioni 提出了后缀树聚类方法(STC)[2]。STC 利用一个后缀树[3]来发现文本所共同含有的短语并进而利用这些信息来构建基本类。这里,我们可以将一个类中的文本所含有的共同的短语看作是这些文本的索引。为了避免出现大量重复的或非常相似的类别,STC 合并那些高度重叠的基本类。STC 方法的优点在于不需要人为指定类别的数目,并且能够利用每个类中各个文本所含有的共同短语即索引来描述这些类。此外,STC 还允许一个文本出现在多个类别中。尽管如此,STC 不是一种基于主题的聚类方法,因为它无法保证一个类别中包含共同短语的文本都是关于同一主题的。比如,我们不能想象出那些共同含有短语"我们需要"的文本是关于什么主题的。

本文提出的 LFIC 方法借鉴了 STC 利用索引产生基本类的思想。但与之不同的是,在 LFIC 中索引是文本之间共享的主题元素。一般地说,被相同主题元素索引的文本是关于同一主题的。一些主题元素,例如,参与者、地点和时间等,会以命名实体(NE)的形式在文本中出现。其余的一些主题元素,包括道具及行为等则往往是文本中的重要名词或动词。因此,我们需要使用词性和命名实体等语言学的特征来提取主题元素。

3　LFIC

3.1　基本类索引

如前面介绍的,STC 方法有一些明显的优点。事实上,该方法的这些优点都归功于通过索引来形成基本类的这一做法。为了表示的方便,在本文的后续内容中我们称这种方法为基本类索引。基本类索引的确切含义可以表述如下:

假设 $D = \{D_1, D_2, \cdots, D_n\}$ 为一个文本集合,$I = \{I_1, I_2, \cdots, I_m\}$ 为一个被抽取出的索引集合。则文本 D_i 被置于索引为 I_j 的类当且仅当 I_j 出现在 D_i 中的次数超过一个事先确定的阈值 T。

在基本类索引中,不需要事先确定停止条件。这一点非常的重要,因为在实际应用中,由于待聚类文本数和主题数都是不确定的,因此事先准确设定停止条件是不现实的。其次,我们考虑到了一个文本可能同时关于两个甚至多个主题的情况。因此,本方法允许一个文本属于多个类。具体地,不同的主题可以通过不同的索引来体现,而一篇文本又可以被多次索引。此外,每个基本类的索引都可以看作是对这个基本类内容的描述。

3.2　结合语言学特征

我们希望能从基本索引类中受益,正如 STC 方法那样。但是,在 STC 方法中的索引只是短语甚至简单的 n 元组(N-grams),这种索引方式的缺点是明显的。一方面,许多短语或者 n 元组没有意义进而无法表征任何主题。另一方面,许多有用的信息往往在文本中不被表示为短语或者连续出现的 n 元

组,比如一个主题可能由散布在文本中的多个词共同表征。

因此,在构建索引的时候我们利用了语言学特征。在自然语言处理中,一个文本通常可以用一个词的向量来表示[4],而词的权值则通常用统计学的方法计算(例如:"tf.idf")。然而,我们认为词的语言学特征本身在度量其权重及表征文本中也是至关重要的。以词性(Part-of-Speech)为例,显然,不同词性的词在表征文本的时候其贡献是不同的[5]。通常情况下,名词和动词最为重要。形容词和副词次之。功能词或者虚词(如叹词、代词、连词等)几乎没有什么作用,所以可以像停用词一样被去掉。另外,命名实体(NE)往往比一般的词更具有判别力,应该被赋予更高的权值。在LFIC方法中,我们利用命名实体以及重要的名词和动词来形成索引。

3.3 主要步骤

LFIC包含三个主要步骤:(1)文本表示与预处理;(2)构建索引与形成基本类;(3)基本类的合并。

文本表示与预处理:待聚类文本首先经过一系列预处理,包括分词、词性标注以及命名实体识别。同时,我们去掉了停用词。这里我们使用的停用词表包含标点、高频词(如"的","我","了"等)以及一些新闻中的常见词(例如:报刊名称和新闻术语等)。在文本表示方面,我们使用向量空间模型(VSM)。其中向量的元素只包括命名实体、名词和动词。同时,我们使用tf.idf方法计算词的权值。

构建索引与形成基本类:在LFIC方法中一个索引包括两部分:命名实体部分和关键词部分。LFIC中的索引定义如下:

设 D 为一个文本,$X=\{X_1, X_2, \cdots, X_m\}$ 为一个在 D 中至少出现两次的命名实体的集合,$Y=\{Y_1, Y_2, \cdots, Y_n\}$ 为一个在 D 中的tf.idf权值超过一个事先设定的阈值 T 的关键词(名词和动词)的集合。$\forall x \in X$ 以及 $y \in Y$,,我们定义二元组 (x, y) 为文本 D 的一个索引。

显然,如果集合 X 和 Y 的元素个数分别为 m 和 n 的话,则文本 D 将被索引 $m * n$ 次(X 和 Y 中的元素的两两组合)。这使得一个文本被索引到不同的主题并进而被置于多个基本类成为可能。通过这种结合了命名实体和关键词的索引,LFIC将具有相同索引的文本合并起来以构成基本类。

基本类合并:上述形成的基本类之间有着很多的重复。例如,关于主题"美军攻打伊拉克"的文本可以分别被"美国,攻打","伊拉克,攻打","美国,伤亡","伊拉克,伤亡"等索引到。这些索引对应的基本类所含文本几乎相同。因此我们需要通过合并基本类来去除冗余,并形成更完善的类。

我们定义 C_i、C_j 为两个基本类,如果它们之间的距离小于一个阈值 $Thre$,则合并这两个类。这里,我们需要首先计算出两个基本类的质心以计算两个基本类间的距离。这里用于计算两个质心的相似度的度量方法为余弦相似度:

$$\cos(c_i, c_j) = \frac{c_i \cdot c_j}{\|c_i\| \|c_j\|} \quad (1)$$

其中,"·"表示向量点积,$\|c\|$ 表示向量 c 的长度。

进一步地,如果一个基本类 c_i 可以分别与基本类 c_{j1}, …, c_{jk} 合并,则这些基本类将被合并成同一个类。

4 实验

4.1 实验数据与评价指标

我们使用从网络上搜集的2 021篇新闻文本来作为评测数据。我们将这些文本手工地划分出266个主题,其中最大的主题包含24篇文本,最小的包含3篇文本。本文中,我们分别使用准确率和召回率作为评价指标。

给定一个含有 n_i 篇文本的主题 T_i,以及一个含有 n_j 篇文本的类别 C_j。设 n_{ij} 为 T_i 和 C_j 共同含有的文本数。则与 T_i 和 C_j 对应的准确率定义如下:

$$precision(T_i, C_j) = n_{ij} / n_j \quad (2)$$

主题 T_i 的准确率被定义为:

$$precision(T_i) = \max_{C_j \in C} precision(T_i, C_j) \quad (3)$$

整个方法的准确率被定义为各个主题的准确率的加权平均值,即:

$$precision = \sum_{i}^{N_T} \frac{n_i}{N} precision(T_i) \quad (4)$$

其中,N 为待聚类文本总数,N_T 为主题数。

同样地,本方法的召回率可以被定义为:

$$recall = \sum_{i}^{N_T} \frac{n_i}{N} recall(T_i) \quad (5)$$

其中,$recall(T_i) = \max_{C_j \in C} recall(T_i, C_j)$,$recall(T_i, C_j) = n_{ij} / n_i$。

4.2 与其他聚类方法的比较

我们做了两组实验。在第一组实验中,我们将

LFIC 与 AHC、KMC 和 STC 等传统聚类方法作了比较。其中,AHC 和 KMC 的停止条件(即最终聚成的类别数)被设置为实际的主题数,即 266。

首先,我们计算了上述四种待比较方法的准确率如图 1。正如预期的一样,LFIC 方法的准确率最高。我们认为这主要是由于 LFIC 所采用的索引聚类的方式,这种方式能够更准确地对主题进行识别。

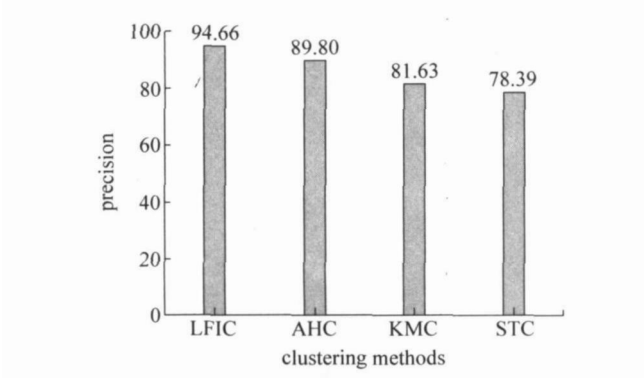

图 1　四种聚类方法的准确率

图 2 显示了四种方法召回率的比较结果。我们可以看到 LFIC 的召回率是 84.46 %,低于 AHC 的 95.60 %和 KMC 的 90.25 %。我们认为其中部分的原因在于一些文本在通过索引构建基本类时很难被检索到,因为它们很少和别的文本含有相同的命名实体和关键词。然而,考虑到网络信息的极大丰富,84.46 %的召回率对于很多应用来讲是可以接受的。

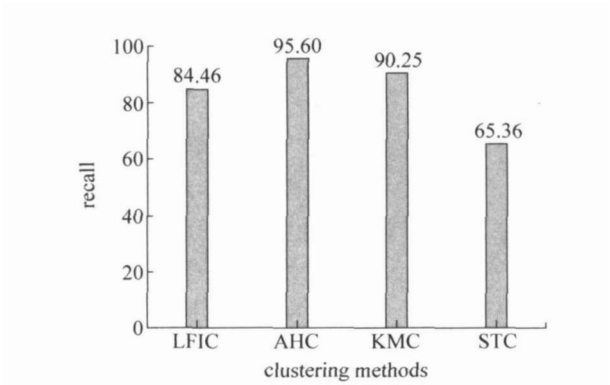

图 2　四种聚类方法的召回率

4.3　对于方法中不同因素的贡献的评价

在第二组实验中,我们评价了 LFIC 方法中各种因素的贡献。首先,我们想通过实验证明结合了命名实体和关键词的基本类索引比仅仅使用命名实体或者关键词的索引方式更为合理和有效。

表 1 比较了 3 种使用不同基本类索引方式的实验结果。我们可以明显地发现,由命名实体与关键词相结合的基本类索引在准确率和召回率方面都要优于仅使用关键词作为索引的方法。这说明仅用关键词来对主题进行描述是不够的。

通过表 1 我们还可以发现,仅使用命名实体索引得到的召回率是最高的,但是其准确率却是最低的。这是因为,在实验数据中,多个主题可能和同一个命名实体相关,比如"美国"这个命名实体可能与"美国攻打伊拉克","美国大选","胡锦涛出访美国"等多个主题相关。因此单一使用命名实体进行索引可能会使多个主题被索引到同一个基本类中,从而导致生成许多召回率很高而准确率很低的类。

表 1　LFIC 中不同的基本类索引方式的比较

索引方式	准确率(%)	召回率(%)
命名实体 + 关键词	94.66	84.46
仅使用关键词	88.19	78.08
仅使用命名实体	78.31	86.94

其次,我们通过实验证明基本类合并的必要性。为此,我们将 LFIC 方法的聚类结果和没有进行基本类合并的结果进行了比较。比较结果如表 2 所示。我们注意到,没有进行基本类合并的聚类结果其召回率低于 50 %,这么低的召回率对于大多数的具体应用而言都是不可以接受的。然而,这一结果非常正常,因为关于同一主题的一系列文本可能会被索引到多个基本类中。因此,为了对这些不完整的基本类进一步聚类,合并基本类的操作是不可或缺的。这里,我们可以得出这样的结论,即合并基本类可以显著提高 LFIC 的召回率。

表 2　合并基本类与不合并基本类的比较

	准确率(%)	召回率(%)
合并基本类	94.66	84.46
不合并基本类	96.90	49.63

5　结论

本文提出了一种新的文本聚类方法——LFIC,其主要有以下三个特点: 第一,LFIC 是一种基于主题的文本聚类方法,可以通过基本类索引来实现主题的识别;第二,LFIC 综合使用了命名实体以及词性等语言学特征来构建索引与表征文本。第三,本方法通过对基本类的合并来提高方法的召回率。

实验证明,LIFC 方法在保证一个可以接受的召回率的同时,可以实现很高的聚类准确率。同时,

实验还证明,基本类索引,语言学特征的使用以及基本类合并等技术对于LIFC方法都是有效的。

不可否认,LIFC算法仍存在一些缺陷。首先,LIFC算法依赖于下层模块的性能,尤其是命名实体识别模块的性能。其次,LFIC中需要设定一些阈值。这些阈值设定的合理性也将会影响方法的有效性。

在未来的工作中,我们将尝试把更多的命名实体类型,比如日期、时间等作为语言学特征加入到LFIC中。我们相信这些特征将有利于对主题的精确描述。此外,由于目前应用的基本类合并算法比较简单,因此我们会对该算法进行改进。

参考文献:

[1] Hatzivassiloglou V, Gravano L and Maganti A. An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering [A]. In: Proceedings of the 23rd ACM SIGIR Conference, Athens [C]. 2000. 224-231.

[2] Zamir O and Etzioni O. Web Document Clustering: A Feasibility Demonstration [A]. In: Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. 1998. 46-54.

[3] Gusfield D. Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology [M]. Cambridge, UK: Cambridge University Press, 1997.

[4] Lee D-L, Chuang H and Seamons K. Document Ranking and the Vector-Space Model [J]. IEEE Software, 1997, 14 (2): 67-75.

[5] Kummamuru K, Lotlikar R, Roy S, et al. A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results [A]. In: Proceedings of the 13th International Conference on World Wide Web [C]. 2004. 658-665.

获奖消息

第四届中国科协期刊优秀学术论文评选结果于近日揭晓。由我学报推荐的、发表于《中文信息学报》2002年第5、6期的论文"北京大学现代汉语语料库基本加工规范"(作者:俞士汶、段慧明、朱学峰、孙斌)荣获"第四届中国科协期刊优秀学术论文"。在此特向论文作者表示祝贺!

"中国科协期刊优秀学术论文"评选活动自2003年起,每年一届,旨在进一步推进我国学术繁荣、促进学术交流、提高学术期刊质量,对更多一流学术成果在国内学术期刊上发表起到导向和推动作用。此次评选活动涉及全国学会105个,期刊283种。经中国科协期刊优秀论文评审委员会专家评审、中国科协常委会学术与学会工作专门委员会审定,中国科协网站、科技导报社网站等有关媒体公示无异议,共评选出200篇论文为第四届中国科协期刊优秀学术论文。

希望广大中文信息处理领域的科研工作者将更多的一流学术论文在我学报上发表。《中文信息学报》今后将继续向中国科协推荐所刊的优秀论文。