话题内相关文本的内容计算

刘冬明[1]，杨尔弘[2]

（1.中北大学，山西省太原市 030051；2. 北京语言大学，北京市 100083）

摘要：信息的暴涨给文本处理带来了更多的挑战。话题检测能够把大量的信息以文本为单位有效的组织起来，然而最终用户并不需要涉及某一话题的所有文本，仅仅关心该话题的具体内容。在我们能够实现根据相关文本智能表达话题内容推送给用户之前，自动从相关文本中挑选符合用户需求的文本是一个非常有意义的工作。本文致力于相同话题的文本之间的内容比较计算，目的便是能够有效的选出满足需求的文本。我们通过对话题进行重新定义，并根据此定义设定了话题和文本的表示方法，给出了基于该表示方法的话题和文本之间的内容比较计算方法。最后，通过实验说明了这一系列方法的有效性。

关键词：话题定义；文本表示； 话题检测；文本内容计算

中图分类号：TP391　　**文献标识码**：A

The Text Content Computing within an Topic

Abstract: Because of Skyrocketing information, text processing is encountering more challenges. Topic detection can effectively organize a lot of information with the text as an unit, but the end user does not need all the texts on a topic. They may just concern the specific content of the topic. Before we can automatically push content to the user with intelligent expression in accordance with the relevant texts in the topic, it is a very meaningful work that selecting the text in line with the needs of users from the associated texts. This paper will compare the content between the texts in a topic and effectively select the text which meets the needs of the user. We redefine the topic and set topic and text representation pattern according to this definition. Then we give calculation method between the texts and topic based on the representation pattern. Finally, the experiment illustrates the effectiveness of this series approach.

Key words: Topic Definition of Topic; Textual Representation; Topic Detection; Text Content Computing

1 引言

随着互联网的迅速发展，信息量规模巨大，然而相同或相近内容的信息，特别是新闻话题，在位置上分散，在形式上多样，导致信息难以高效的利用。话题检测与跟踪[1]、话题演化等相关技术正是为了将信息根据内容以有效合理的形式组织起来。然而对于最终用户来讲，需要的不是关于某一话题成堆的文本，而是关于该话题的具体内容或某一方面的内容。要想精准的给予用户所需信息，需要依据内容的自动语言生成技术，然而该技术目前还远未达到应用的要求；另外也可采用多文档自动文摘技术，但在效果上远不如原文流畅易读。

本文研究目标就在于从描述某一话题的大量文本中寻找符合用户需求的文本，这样的文本或者包含了整个话题的来龙去脉，或者包含了用户想要了解的该话题某一方面的较为全面的内容。因此，本研究以话题检测与跟踪应用的结果作为输入信息，采用一定方式表示文本内容，从而对不同的文本进行内容上的比较计算，最后结合用户需求选出特定的文本。

从上述过程可以看到文本之间基于内容的比较计算是解决此问题的关键所在，而比较计算的基础却在于文本的表示方法。文本表示直观来讲就是将可以让人理解的文本在尽量保留目标任务所需信息的前提下，将其转化成方便计算机处理的一种形式。这种“形式”是文本表示方法的关键所在，不同的计算机处理需求会要求有不同的形式。传统的文本表示方法有

基金项目：国家语委“十二五”科研规划项目：媒体教育领域话题检测及话题库建设（No：YB125-43）

作者简介：刘冬明（1972——），男，讲师，自然语言处理，北京语言大学语言学及应用语言学专业在读博士；杨尔弘（1965——），女，教授，自然语言处理，计算语言学。

向量空间模型、概率模型、图模型等。其中向量空间模型是目前话题检测、文本分类以及信息检索中应用最广泛的模型构架，由 Salton 于 1975 年提出[2]。在该模型中，每一个特征项作为向量空间的一维，每个文本就表示成了 n 维空间的一个点。根据特征选取及特征度量的不同，向量空间模型包括许多形式，典型的如词频权重 TF*IDF 模型[3]、概念表达模型[4]、多词表达模型[5]、句子表示模型[6]、语义关系模型[7]、维基百科类别模型[8]等等。概率模型是信息检索领域较为成熟的模型[9]，在许多应用中取得了不错的实用效果，然而其特征选择范围受限、语料稀疏等问题难以解决导致其应用范围有限。这两种表示方式中各个特征项之间要求独立，通常为了提高效率将文本作为“词袋”处理，忽略了文本中的关联信息，导致对文本内容表达的先天缺失，因此无法对文本内容进一步进行刻画。图模型含有其他模型经常忽略的关联特征，用结构图而非集合来表示文本，最能体现新闻报道内容和结构，也最难构造[10]。使用图模型来表示文本通常会导致复杂度的增大，同时由于理论基础不足和知识资源欠缺，致使执行效率低下，并且引入噪音较多，在后续处理中会有放大效应，最终结果未必如简单的模型。

本研究中的文本表示同上述方式的差别在于：本研究中的文本表示是基于话题的，也即在已知该文本所对应话题的情况下，采用一种形式表示出该文本所描述的相关话题的内容。这种表示方式的基础在于话题的内容，因此，有必要首先对话题给出一个明确的可操作性的定义，在此基础上再定义话题和文本的表示方式和计算方式。

下面第二节描述本文所提出的话题的可操作性定义，第三节具体说明话题及相应文本的表示方法和内容比较计算方法，第四节给出这种表示方式和计算方法的实验验证过程和结果，最后是总结和展望。

2 话题定义

话题，虽然在语言表达上看似比较明确，但是要给出一个具体的定义却很困难。历史上，许多学者都对话题给出过定义，至今为多数研究者所接纳的定义如下：

定义 1：话题指一个核心事件或活动以及与之直接相关的事件或活动。

文献[11]中，Cieriet al 详细解释了话题和事件的关系，给出了具体规则，并在 TDT2 和 TDT3 评测[2]中以此作为指导方针，制定了评测语料库。虽然这个定义沿用至今，但是实际上，脱离了 TDT 的评测语料库之后，这个定义依然模糊，难以操作。原因如下：

a) 事件是该话题定义中的核心元素，但是目前没有成熟的技术能够比较精确的识别语料中的事件，因此依据该话题定义难以准确的表达话题；

b) 两个事件是否直接“相关”同样也没有明确的机器可操作的判断方式；

当前的研究仅仅将此定义作为话题检测与跟踪系统的参考，无法真正依照此定义实现相关应用。因此不同的具体实现技术暗含了不同的对话题的理解，其应用结果是不具有可比性的。

本文在详细考察上述定义在话题应用研究中实际状况以及本文的研究目标的基础上，为了能够更好对话题中的文本进行基于内容的合理的比较计算，提出了一个关于话题的可操作性定义，该定义建立在对于话题中的元素——事件，进行内部分解，从而使计算机能够在不失效率的前提下，有效的表示和计算话题的内容。

为了引入本文的话题定义，我们首先需要定义抽象事件和抽象脚本：

定义 2：抽象事件是指具体事件中移除所有实体之后的描述，其表征就是描述事件的词即事件词。

定义 3：抽象脚本是指连续发生的一系列具有前因后果关系抽象事件。

---

[2] 指美国国家标准技术研究所 (NIST)举办的话题检测与跟踪的国际会议和相应的系统评测。

例如：在关于故宫被盗的话题中，“盗窃”、“抓捕”等就是抽象事件，而将所有抽象事件有机的联系在一起就构成了抽象脚本：展览活动中展品被盗、立案、侦查、抓捕、起诉、结案的过程，其形式化表示为一组相关的事件词的集合，即：｛“被盗”、“立案”、“侦查”、“抓捕”、“起诉”、“结案”｝。

定义 4：话题是一个或多个抽象脚本和具体的时间、地点、人物等实体相结合的描述。

以上定义可以看到，本文将话题分成了具有关联关系的两个部分：抽象脚本集和实体集。实体集中的每一个实体都和话题中的某一个抽象脚本相关，其实质作用即将抽象事件转化成了具体的话题中的事件。例如：｛“被盗”、“立案”、“侦查”、“抓捕”、“起诉”、“结案”｝这个抽象脚本，如果结合了实体集｛“故宫”、 “展品”、 “公安机关”、 “嫌疑人”｝就代表了故宫被盗这个话题。

对比定义 4 和定义 1，可以得出如下几点：

a) 两个定义对话题组成部分的描述角度不同；

定义 1 将话题表示为一个集合，其元素是事件或活动，在这个集合中仅有一个特定元素即核心事件，其余元素处于同等地位，同时隐含表达了包含核心事件同其他元素之间具有“相关”关系的集合。而定义 4 是对话题分成两个不同性质的部分：抽象脚本集和实体集，同时隐含表达了包含抽象脚本同实体具有的关系集合。

b) 两个定义对话题组成的描述粒度不同；

定义 1 中关于话题组成部分中事件或活动是其不可拆分的原子成分，在定义 4 中实质上将具体事件拆成了抽象事件和涉及的实体两个部分，并且这两个部分以当前的应用技术来讲都是可识别的。

c) 两个定义对相关性的表达不同；

定义 1 中关于事件“相关”没有明确的说明，定义 4 中将这种“相关”表达成了抽象事件同属于一个抽象脚本的关系，而抽象脚本本质上是基于一定场景的，也就是说应该是一种比较稳定的结构，可以通过人工方式或机器学习方式来构造，因此可以说通过这一定义，计算机可以学习和表示出抽象事件的相关性，而事件的相关性其本质就是其所对应的抽象事件的相关性。

d) 抛开相关评论部分之后，二者对于话题定义的外延在本质上是相同的。

从以上几点可以看到，仅从定义的内涵出发，本文提出的定义 4 是对话题的更加结构化、细致化的描述，结合其组成成分的计算机可识别性，在外延本质相同的情况下，应该更具有可操作性。

3 话题和文本的表示和计算

3．1 话题和文本的表示

根据上一节给出的话题的可操作性定义，可以将话题表示成如图 3-1 所示的形式：

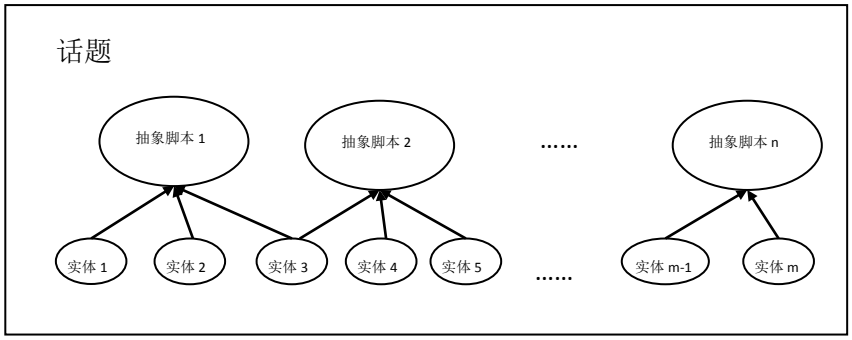

图 3-1 话题表示示意图

这里需要说明如下几点：

a) 抽象脚本在表现形式上就是抽象事件的集合，而抽象事件表现为表示事件的词，一般来说是动词或具有动词意义的名词，我们统称之为事件词。因此，这里抽象脚本就是相关事件词的集合。

b) 实体在表现形式上可能是话题中提到的命名实体或者和事件相关的普通名词。

c) 每一个实体都对应一个抽象脚本，不能对应抽象脚本的实体对于话题内容来讲是没有意义的。同时，一个实体可能对应多个抽象脚本。

从这种表示方式可以看出，抽象脚本是话题的核心，而实体只有依附于抽象脚本才在话题表达中具有意义，这和我们对话题的定义是一致的。因此这种话题表示方式的关键条件是必须事先具备关于要表示的话题的抽象脚本，并且抽象脚本的质量决定了这种表示方式对话题内容的表达能力。

依据图 3-1 和上述分析，可以将话题采用如下三元组的形式化表示：

$$T=<J,E,R>$$

$$J=\{A_1, A_2,\ldots A_i, \ldots A_N\}\ \text{其中}A_i=\{w_{i1}, w_{i2},\ldots w_{ik}, \ldots w_{iM}\}$$

$$E=\{e_1,e_2,\ldots e_i, \ldots e_K\}$$

$$R=\{<A_i,U_i>|\ A_i \in J,\ U_i \subseteq E\}$$

其中，T 表示话题内容，J 表示抽象脚本集合，A_i表示抽象脚本，w_{ik}代表抽象脚本 A_i中的事件词，E 为话题所包含的所有实体构成的集合，e_i 为话题中的某一实体，U_i 为和抽象脚本 A_i关联的实体的集合，是 E 的子集。

对于话题中文本的表示方式，根据本文前面的论述，我们主要研究的是基于话题内容的文本表示，因此可以设定文本的表示方式同话题表示方式一样，差别仅仅在于具体的文本在内容上的是话题的一部分。

3．2 基于话题内容的文本计算

本文的目标在于考察文本表达了话题的多少内容，或者表达了话题某一方面的多少内容，并给出相应的数值，作为推送最终用户的依据。这里所说的话题的某一方面，指的就是话题中的某一个抽象脚本。我们把获取的数值称之为文本对话题的内容覆盖度。

因此，如果将话题用上一小节三元组的形式化方式表达出来以后，那么需要针对每一个话题内的文本，将其表示为该话题三元组的形式，然后以基于集合的方式进行内容覆盖度的计算。其中关键之处有三个：

1) 抽象脚本和文本的关系

词同文本的关系是明确的，因为我们通常将文本作为词袋处理，即词的集合，那么具体一个词要么属于文本要么不属于文本。而抽象脚本本身是事件词的集合，它同文本的关系就成了两个集合之间的关系，难以简单的确定文本是否包含抽象脚本。因此，这

里我们采用阈值过滤的方法将其转化为布尔值，这里设 x 表示文本，A_i表示抽象脚本，S_x表示文本 x 中的所有事件词构成的集合，w 表示事件词，那么文本对抽象脚本的包含程度：

$$\mu(A_i, x)=\frac{\sum_{w\in S_x\cap A_i} f(w)}{\sum_{w\in A_i} f(w)} \qquad \text{公式 3-1}$$

其中$f(w)$是事件词w的权值，这里采用领域特征值[12]来计算，表示w的领域特性。

2) 文本关于话题的某个抽象脚本的内容覆盖度

如果根据上述判别方法判断出文本 x 包含抽象脚本 A_i的程度，然后综合考虑同这个抽象脚本相关的实体，采用如下公式计算 x 对于 A_i的内容覆盖度：

$$\text{Coverage}(A_i, x) = \mu(A_i, x) * (\lambda_A + \lambda_E \times \left(\frac{E_{xA_i}}{E_{TA_i}}\right) + \lambda_G \times (\frac{G_{xA_i}}{G_{TA_i}})) \qquad \text{公式 3-2}$$

式中λ_A、λ_E、λ_G分别为抽象脚本本身、该抽象脚本所对应的命名实体、该抽象脚本所对应的普通名词的权重，这里要求 $\lambda_A + \lambda_E + \lambda_G = 1$。$E_{xA_i}$为文本 x 中抽象脚本$A_i$所对应的命名实体数量，$E_{TA_i}$为话题中抽象脚本$A_i$所对应的命名实体数量，同样$G_{xA_i}$为文本 x 中抽象脚本$A_i$所对应的普通名词数量，$G_{TA_i}$为话题中抽象脚本$A_i$所对应的普通名词数量。因此如果抽象脚本 A_i包含了话题中关于抽象脚本A_i的全部内容，则 x 对于 A_i的内容覆盖度为 1。

通过公式 3-2 还可以看出，x 对于 A_i的内容覆盖度计算是建立在$\mu(A_i, x)$即文本 x 对抽象脚本的A_i包含程度基础上的，如果包含程度很低，那么再多的命名实体和普通名词也不会提高多少内容覆盖度，因此这个公式旨在使实体隶属于抽象脚本，并且通过抽象脚本发生作用，同本文对话题的定义是一致的，更加有效清楚的表达了话题的内涵。

3) 文本关于话题的整个内容覆盖度

$$\text{Coverage}(x) = \frac{\sum_{i=1}^{|T_A|} \alpha_{A_i} \times \text{Coverage}(A_i, x)}{\sum_{k=1}^{|T_A|} \alpha_{A_k}} \qquad \text{公式 3-3}$$

其中，T_A表示话题 T 的抽象脚本集合，α_{A_i}为抽象脚本 A_i中事件词的平均领域特征值，用以刻画抽象脚本的重要程度。这样，$\frac{\alpha_{A_i}}{\sum_{k=1}^{|T_A|} \alpha_{A_k}}$就表征该抽象脚本在话题中的权重，$\alpha_{A_i}$计算公式如下：

$$\alpha_{A_i} = \frac{\sum_{i=1}^{|A_i|} f(w_i)}{|A_i|} \qquad \text{公式 3-4}$$

这个公式的意义就在于给话题中每一个抽象脚本的内容覆盖度加权平均后作为整个文本的内容覆盖度。

4 实验及结果分析

4．1 实验设计

本实验的目标在于验证话题的定义和话题及文本的表示是否有效。以本文引言中提出的应用需求作为验证方法，即通过采用上一节提出的话题及文本表示方式和内容覆盖度的计算方法，在某一个话题相关的文本集合中找到满足用户需求的文本。

本实验的前提是必须有该话题的抽象脚本。由于本文的目标所在，我们采用人工的方式从话题中提取抽象脚本，这样能够保证抽象脚本的质量，使实验能够尽可能排除其它干扰因素，真实的反应本文所提出话题和文本的表示和计算方式的能力。

实验步骤如下：

1) 从新浪网站专题“故宫被盗”中下载了 100 篇文本作为实验语料，并对其进行分词标注。

2) 从所有文本中提取事件词（动词或动名词），人工整理并分类，确定了六个抽象脚

本：分别代表展览过程、防范过程、侦查过程、抓捕过程、盗窃过程、司法过程的事件词集合。

3) 人工从 100 篇文本中挑选出对每个抽象脚本描述最详细的文本，以及整体上描述最详细的文本和描述内容最少的文本各自 5 篇。

4) 将所有的 100 篇文本的合集作为话题，从文本中以句子作为语义范围提取每个抽象脚本对应的实体集合，然后以上一节公式分别计算每个文本相对话题中每个抽象脚本的内容覆盖度以及整体的内容覆盖度，针对每个抽象脚本找到内容覆盖度最大的 5 篇文本，整体内容覆盖度最大的 5 篇文本以及内容覆盖度最少的 5 篇文本。

5) 将计算结果同人工挑选的结果进行对比。

计算内容覆盖度时采用的参数取值为 $\lambda_A = 0.61$，$\lambda_E = 0.27$，$\lambda_G = 0.12$，是根据经验以及重复实验来确定的。

4．2 结果分析

实验结果如下：

表 4-1：文本覆盖度计算结果

抽象脚本名称	抽象脚本部分内容	正确率
展览过程	参观 陈列 展示 展览 …	1
防范过程	警惕 站岗 值勤 夜巡 …	0.8
侦查过程	查清 核查 勘查 拘留 …	1
抓捕过程	抓捕 擒获 追赃 落网 …	0.8
盗窃过程	窥视 藏匿 混入 逃掉 …	1
司法过程	处罚 惩治 惩罚 招供 …	0.6
整体覆盖度最大		0.8
整体覆盖度最小		1

这里的正确率采用如下方式计算：

$$\text{正确率}=\frac{\text{人工方式和内容覆盖度计算方式选定的相同文本的数量}}{5}$$

从结果可以看出，这种计算方法完全能够符合我们的应用要求。实验结果是根据符合条件的前 5 名得出的，如果每一种类只选一个文本的话，那么除了最后一条“整体覆盖率最小”和“司法过程”之外都和人工方式是一致的。原因是整体覆盖率最小的计算结果有 5 篇都是为零的，如：“单士兵：故宫，你丢了最宝贵的文化钥匙”、“民办博物馆将纳入免费开放范围”、“有憾于故宫的接连失守”等等。这些文本的内容大部分只是提及了故宫被盗，没有描述过程，或者转述了其他相关话题如“锦旗错别字”和“故宫内建会所”，或者是全部评论。对于抽象脚本“司法过程”的正确率较低是因为我们选取的描述该抽象脚本的事件词同测试中人工选取的一篇文本中的描述偏差较大，由此也可以看出这种选取和计算方法的关键在于抽象脚本的质量，不论是人工确定抽象脚本还是采用自动生成抽象脚本的方式，抽象脚本对于内容覆盖的全面性和准确性是会相互制约的。

虽然实验结果非常好，但是这是在人工针对待测语料专门制定的抽象脚本的基础上产生的，在实际的应用中不可能针对每一个话题人工制定抽象脚本，因此本实验也只能说明话题表示方法和计算方法的有效性，实际应用还必须依赖合理的抽象脚本自动生成。

5 总结和展望

本文基于话题检测之后的需求，为话题赋予了一个可操作性定义，在此基础上提出了话题及话题中文本的表示方式，文本和话题内容的计算方法，并通过实验进行了验证。在这一系列概念中，关键之处在于本文所提出的抽象脚本的概念，通过它可以清晰的鉴定话题的内容范围，可以量化话题和文本的内容，特别是抽象脚本具有一定的稳定度，一次生成可以多

次使用，这样可以有效的提升计算结果的效率和效果。

我们在实验中为了说明文本和话题表示及计算方法的有效性，采用了人工生成抽象脚本的方法。在实际的应用中，由于话题种类繁多，这种方式较为费时费力，因此在下一步的研究中，我们将考虑将待测话题结合语义知识来自动生成抽象脚本，并且通过实际应用不断提高抽象脚本的质量；同时由于抽象脚本的稳定性，我们将考虑如何将抽象脚本作为知识库应用于话题检测与跟踪中，以提升当前的话题检测与跟踪应用系统。

参考文献：

[1] 洪宇, 张宇, 刘挺, & 李生. (2007). 话题检测与跟踪的评测及研究综述. *中文信息学报*, *21*(6), 71-87.

[2] Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." Communications of the ACM 18.11 (1975): 613-620.

[3] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

[4] Zeng, C., Lu, Z., & Gu, J. (2008, December). A new approach to Email classification using Concept Vector Space Model. In *Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on* (Vol. 3, pp. 162-166). IEEE.

[5] Liddy, E. D. (1998). Enhanced text retrieval using natural language processing.*Bulletin of the American Society for Information Science and Technology*,*24*(4), 14-16.

[6] Keikha, M., Khonsari, A., & Oroumchian, F. (2009). Rich document representation and classification: An analysis. Knowledge-Based Systems, 22(1), 67-71.

[7] Scott, S., & Matwin, S. (1998, August). Text classification using WordNet hypernyms. In *Use of WordNet in natural language processing systems: Proceedings of the conference* (pp. 38-44).

[8] 王锦, 王会珍, & 张俐. (2011). 基于维基百科类别的文本特征表示. *中文信息学报*,*25*(2), 27-31.

[9] Jones, K. S., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments: Part 1.*Information Processing & Management*, *36*(6), 779-808.

[10] Schenker, A., Last, M., Bunke, H., & Kandel, A. (2004). Classification of web documents using graph matching. *International Journal of Pattern Recognition and Artificial Intelligence*, *18*(03), 475-496.

[11] Cieri, C., Strassel, S., Graff, D., Martey, N., Rennert, K., & Liberman, M. (2002). Corpora for topic detection and tracking. In *Topic detection and tracking*(pp. 33-66). Springer US.

[12] 刘冬明，杨尔弘.(2014) 量化词语的领域特征[J]. 中文信息学报, 28(5): 46-50.