

基于主题本体树的文本流层次主题检测技术

张 琪¹, 陈 千², 郭 鑫³

(1 神华集团 神华和利时信息技术有限公司, 北京 100011; 2 山西大学 计算机与信息学院, 山西 太原 030006;
3 同济大学 电子与信息工程学院, 上海 201804)

摘 要: 本文首先提出主题本体树层次结构的主题结构模型, 并研究了基于主题本体树的层次主题检测技术. 在真实新闻数据上的仿真实验表明, 该方法能从文本流中有效检测层次主题, 实验表明三种测度中混合对称测度效果更好.

关键词: 文本流; 层次主题检测; 主题本体树

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2013)07-0060-04

Topic Ontology Tree Based Hierarchical Topic Detection in Text Stream

ZHANG Qi¹, CHEN Qian², GUO Xin³

(1 Shenhua Heshili Information Technology Ltd., Beijing 100011, China;

2 School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China;

3 College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: This paper puts forward the hierarchical structure tree theme ontology theme structure model, and based on the theme ontology tree level theme detection technology, and finally presents the simulation experiment. In the real news of data simulation experiment results show that this method can detect effectively flow from text hierarchical theme, in addition, three measure mixed symmetrical measure the effect is better. Finally, the research field is discussed.

Key words: text stream; hierarchical topic detection; topic ontology tree

1 引言

随着互联网的出现, 尤其是以维基百科、博客、BBS 等为特征的 Web2.0 时代的到来, 互联网已经成为人类最大的信息资源来源^[1]. 在信息海洋中, 人们感兴趣的话题往往分散在不同的时间和不同地点, 人们很难对这些信息进行有效的管理和利用. 话题检测和跟踪技术 (Topic Detection and Tracking) 就是为了解决这种类型的问题而诞生的^[2]. 本文对主题检测研究是传统 TDT 的一个扩展和延伸, 主要对象是一般文本流.

主题检测主要面临两个难点: ①对不同粒度的事件的识别; ②对同一粒度的事件的聚类 and 分组. 传统的做法是首先对当前活动文档进行主题和事件的抽取、其次需要将抽取出来的主题和事件同之前的事件进行相似度比较, 并将事件归为不同的主题. 这种方法对主题事件的层次性体现不够, 粒度上不够灵活.

Bettina 在事件检测中引入了拓扑结构的方法, 并论证了在高维情况下, 该方法的稳健性和准确性得到了较好的体现^[3]. 文献^[4]提出了基于单路径聚类的方法进行话题检测, 但该算法未从主题和事件

收稿日期: 2012-10-15; **修回日期:** 2012-11-11

基金项目: 国家自然科学基金项目 (71171148, 61103069); 国家“八六三”计划项目 (2012BAD35B01); 上海市信息委专项基金项目 (11DZ1501703)

等更细粒度进行研究.另外,主题检测已有的方法还包括基于命名实体方法、增量式 TFIDF 模型、基于语义链方法等^[5-6].

本文提出一种基于主题本体树的层次文本流主题检测方法,从而能够方便用户对不同粒度层次的主题进行抽取.

2 文本流

定义 1 文本流(text stream).文本流是一系列具有动态性、周期性、潜在无限性,以时间顺序连续到达的文本数据记录,它是主题和事件的载体.

同一个文本流中的文本数据属于同一类型题材,比如 Reuters 新闻和 NIPS 历年会议论文集是两个文本流,它们的类型是不同的,不能随意将他们合并为一个文本流,一个是新闻报道类型的,一个是科技文献类型的,具有一定的领域性.在某一个很小的时间段,可能有多篇文档同时到达.若当前 t 时刻到达的文档集为 C_t ,则文本流 Stream 其形式化定义如下:

$$\text{Stream} = \{C_1, C_2, C_3, \dots, C_t, \dots\} \quad (1)$$

其中 C_i 表示第 i 个时间段到来的文本集, C_i 表示为

$$C_i = \{D_1, D_2, \dots, D_j, \dots\}$$

其中 D_j 表示 i 时刻的文档集 C_i 中的第 j 篇文本文档.

定义 2 主题检测(Topic Detection).主题检测的目的是识别不断到达的文本流中蕴含的主题.主题检测过程可以描述为:

Input: D over time

$$\text{Output: } \forall d \in D, \quad t = f_{TD}(d) \quad (2)$$

$$f_{TD}: D \rightarrow T_{incr}, \quad t \in T_{incr} \quad (3)$$

其中, D 为文本流集合, T 为主题集合,它随时间不断增长, f_{TD} 表示从文本流到主题的映射.检测识别出来的主题可能是系统已知的,也可能是系统未知的.

3 基于主题本体树的层次文本流主题检测

3.1 主题本体树

本体是“共享概念模型明确的形式化的规范说明^[7]”,它将某个应用领域抽象概括成一组概念以及概念之间的关系,以及利用这些术语和关系构成的规定这些词汇外延的规则的定义”.本体理论从一个哲学概念到用于计算机人工智能领域,经历过一个漫长的过程.一般来讲,本体可以表示为一个五元组

集合: $O = \{C, R, P, I, A\}$, 其中 C 为概念集合; R 为概念之间关系的集合; P 为概念的属性集合,包括属性所属的概念,属性名和属性值; I 为概念相关的实例集合; A 为本体中的公理集.

定义 3 主题本体树(Topic Ontology).主题本体树是一种用本体理论指导主题结构建模的层次结构.

同本体类似,主题本体树也是一个集合,其中,主题相当于本体中的概念,主题属性对应着构成一颗倒立的树状结构.根据主题是否是最上层或最下层,主题分为根主题、叶子主题和中间主题.根据主题的父子层关系又分为父主题和子主题.

为了简单起见,研究一个四层结构的主题本体树,如图 1 所示.其中带阴影的圆圈代表主题,椭圆代表事件,而不同的灰度小圆圈分代表该主题的三种不同属性.该主题本体树一共分为四层:文档层、根主题层、叶子主题层、概念层.文档层只有一个根节点,根主题层是一系列根主题节点,即对于每一篇文档而言,可以看成是根主题的按照一定比例的混合;叶子主题层由一系列的叶子主题构成,依此类推,每一个根主题可以表示成叶子主题上的多项分布.所有的叶子主题均是概念层中所有概念上的多项分布.另外,只有叶子主题层才能具有主题属性,并且具有事件实例.需要注意的是,主题本体树可以在根主题层和叶子主题层加入任意多层中间主题,其中每个父主题层中的任意节点是相邻下一层的子主题上的多项分布.因此从根主题层到叶子主题层,每一层逐层都可以用一个 dirichlet 分布来表示.

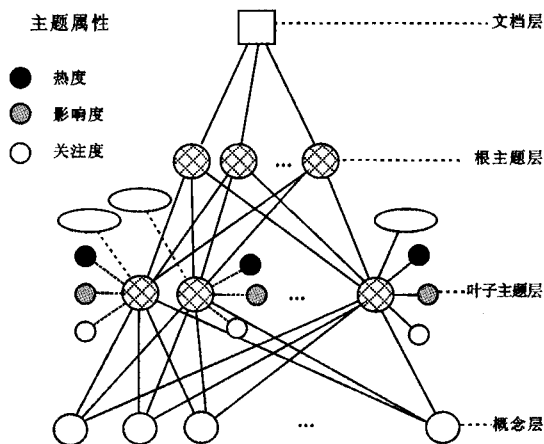


图1 一种四层树状层次结构的主题本体树

3.2 基于主题本体树的主题相似度计算方法

为了计算主题之间的相似度,一般采用信息论的 KL 散度.由于主题是采用分布来计算的,主题 p 到 q 的相似度度量方法如下

$$KL(p \parallel q) = \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)} \quad (4)$$

也可以将主题看成向量,通过计算向量的夹角余弦得到两个主题的相似度.主题间的相似度量方法如下,

$$\cos(\vec{p}, \vec{q}) = \frac{\sum_i p_i \cdot q_i}{\|\vec{p}\| \cdot \|\vec{q}\|} \quad (5)$$

其中, $\|\cdot\|$ 代表向量的模,一般取 $\|\vec{p}\| = \sqrt{\sum_i p_i^2}$. 这两种度量方式不同的是,前者是非对称的度量,即从主题 T_p 到 T_q 的度量并不等于从主题 T_q 到 T_p 的度量,也即 $KL(p \parallel q) \neq KL(q \parallel p)$. 可以采用这种度量方法来度量主题或事件的转移度.而夹角余弦则不同,它描述的是空间两个向量之间的距离,是对称的,因此是一个标准的测度.

还有一些改进的度量方法,如下 sym-KL,该度量方法将 KL 散度进行了修正,得到一个对称的测度.

$$\text{sim}(p, q) = \frac{1}{2}(KL(p \parallel q) + KL(q \parallel p)) \quad (6)$$

当文本流处理某个特定的当前文档时,我们执行在线主题检测算法,在线主题检测算法大致描述如下:

算法 1 在线主题检测任务算法

```

Input: Text stream over time, a certain document being processing
Output: topics and events to be added in topic Ontology
1 for t = 1, 2, ..., T
2   for each arriving document Dt
3     An event with a timestamp was generated,
4     if topic is newly detected,
5       Add this topic to a Topic List,
6       Compute topic correlation
7       Add this topic to it.
8       Attached newly established topic
9     else if topic is existed
10      | Do nothing.
13    end if
14  end for
15 end for
  
```

注意:算法 1 的任务必须以增量的形式完成.该任务的第一个输出是带有时间标签和地点、对象的时间,同时分析抽取该事件所属的叶子主题.该任务的第二个输出是更新后的主题本体树.该树状结构将随着时间的推移而不断的学习.

4 实验结果

本小节在真实数据上设计在线主题检测实验.为了获得真实的新闻报道数据集,从而便于开展在线主题检测研究,我们采用开源网络爬虫工具(Heritrix 1.14.0)网络爬虫工具,自行设计的抽取器和过滤器从雅虎英语官方网址(<http://news.yahoo.com/>)中抽取新闻报道 html 文档.我们通过该工具获得 2657 篇有效的页面文档.并将内容和时间戳、源 url 地址以及题目等抽取出来存入 txt 文本文件中.当文本流处理某个特定的当前文档时,我们执行在线主题检测算法^[8-9].

为了模拟真实的文本流环境,我们一边用 Heritrix 从网络上抓取和抽取过滤文件,一方面执行在线主题检测算法.设置了抓取的时间戳从 2011 年三月份到 2011 年四月份最新的新闻报道.实验语料包含的主要根事件主题如表 1 所示. TDT 使用误报率和漏报率作为算法的评测标准,本文采用 F-score 来评测主要的性能.

表 1 实验语料包含的主要根主题事件

ID	事件名	起始时间	结束时间
E1	311 日本大地震	2011. 3. 11	2011. 6. 14
E2	312 日本大海啸	2011. 3. 11	2011. 6. 14
E3	朝鲜反核游行	2011. 3. 22	2011. 3. 25

F-score 在文献[5]中被用来很好的评测模型的性能.其公式如下,

$$F\text{-score} = \frac{\text{precision} \times \text{recall}}{\lambda \text{precision} + (1 - \lambda) \text{recall}} \quad (7)$$

其中, precision 表示准确度, recall 表示召回率,这里设 $\lambda = 0.5$.

另外,由于本文提出的层次主题模型,能对各种粗细粒度的主题进行抽取,提取前十个概念词,为了方便显示,只显示 3~4 个概念词,效果如图 2 所示.

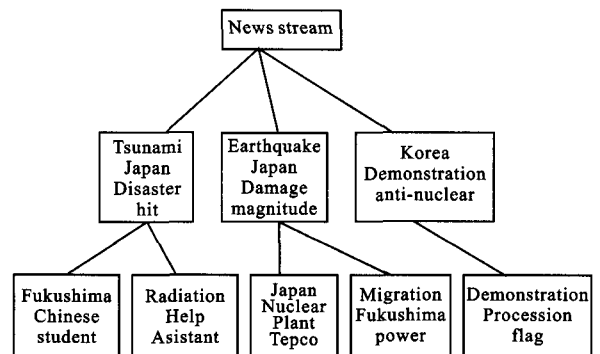


图 2 针对新闻流抽取的主题本体树层次结构

通过仿真实验,得到该算法在三种测度情况下 F-score 的对比,如表 2 所示.

表 2 余弦夹角、KL 散度和对称散度的 F-score 对比

id	cosine	KL-Convergence	sym-KL
E1	0.6795	0.8122	0.8692
E2	0.5819	0.7416	0.6825
E3	0.8017	0.7532	0.6882

由表 2 数据可知,sym-KL 散度的测度方法在实际中的效果平均高于夹角余弦和 KL 散度.另外,由于 E1 主题本身具有更大的辨识度,因此,其 F-score 值相比较于其他两个根主题而言略显优势.

5 结束语

本文提出一种基于主题本体树的层次主题检测方法,能从不同的主题抽象粒度对文本流进行挖掘.由于初次尝试本方法用于真实语料库的规模不大,因此该算法有待改进并用于更大规模的文本流语料库中.

随着主题检测的不断深入研究和发 展,针对不同的问题和不同的应用领域和应用方面,主题检测又衍生出新的分支,包括突发事件检测、热点话题分析、主题演化分析、短文本流的分析、主题源识别(也叫话题发起者检测)、多元异构文本流主题分析、主题情感倾向分析均成为热点研究分支,这些分支在政府的舆情监测、互联网安全、网络舆论犯罪、科技文献的研究趋势分析、新闻档案的整理等等多个领域有着极为重要的研究意义,有待进一步深入研究.

参考文献:

[1] Kleinberg J. Temporal dynamics of on-line information streams[M]. Data Stream Management: Processing High-Speed Data Streams. Springer. Boston: 2006.

[2] 楚克明,李芳. 基于 LDA 话题关联的话题演化[J]. 上海交通大学学报,2010,44(11):1496-1500.

[3] 于满泉,骆卫华,许洪波,等. 话题识别与跟踪中的层次化话题识别技术研究[J]. 计算机研究与发展,2006,43

(3):489-495.

[4] Berendt B, Subasic L. Measuring graph topology for interactive temporal event detection[J]. Kunstliche Intelligenz, 2009, 23(2): 11-17.

[5] Allan J, Papka R, Lavrenko V. Online new event detection and tracking [C] // 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. USA:New York,1998.

[6] Harrington T, Tausz A. Topological classification of datasets without an explicit metric[R]. in CS229: Machine Learning Final Projects, Stanford University, 2008.

[7] Wallach H M. Topic modeling: beyond bag-of-words[C]// Proceedings of the 23rd international conference on Machine learning. USA: Pittsburgh, 2006: 977-984.

[8] Tom Gruber. A translation approach to portable ontology specifications[J]. Knowledge Acquisition, 1993,5 (2): 199-199.

[9] 陈千,向阳,魏韩. 概念匹配方法综述[J]. 计算机应用研究,2010, 27(4):1201-1206.

作者简介:

张 琪 男,(1964-),高级工程师. 研究方向信息管理、决策支持系统.



陈 千 男,(1983-),博士. 研究方向为机器学习、文本挖掘、主题检测、文本流挖掘.(通信作者)



郭 鑫 女,(1982-),博士研究生. 研究方向为数据降维、ERP、数据仓库与数据挖掘.