Web新闻流的增量演进分析

邱江涛[1,2] 唐常杰[2] 乔少杰[2] 李太勇[1,2]
(西南财经大学经济信息工程学院 成都 610075)[1] (四川大学计算机学院 成都 610065)[2]

摘 要 将互联网上的新闻事件按照时间顺序和事件依赖关系组织起来呈现给用户,可以帮助用户方便快捷地了解新闻事件演进过程。定义了Web新闻流增量演进任务(IEA)来实现这一需求。与一些类似的工作比较,IEA具有以下特点:适合Web新闻事件的流特征,以图的方式在时间线上增量更新新闻话题的事件演化过程。为了完成IEA任务,定义了一个事件进展图(EEG)数据结构,并相应地提出了EEG构造和整理算法。实验证明,该方法可以有效地实现新闻事件时间线分析的任务。

关键词 Web挖掘,事件时间线分析,事件进展图
中图法分类号 TP311.13

Incremental Evolution Analysis of Web News Stream

QIU Jiang-tao[1,2] TANG Chang-jie[2] QIAO Shao-jie[2] LI Tai-yong[1,2]
(School of Economic Information Engineering, South Western University of Finance and Economics, Chengdu 610075, China)[1]
(School of Computer, Sichuan University, Chengdu 610065, China)[2]

Abstract With a large number of news available on the Internet everyday, it is an interesting work to automatically organize news events by time order and dependencies between events. The work may help user to conveniently and quickly navigate news event evolution. This paper defined an Incremental Evolution Analysis (IEA) task to meet the need. Compared with existed works, IEA present event evolution in graph manner, and incrementally update the process of events evolving so as to better fit feature of stream of news on the Internet. This paper proposed an Event Evolving Graph (EEG) structure and the building and tidying algorithm of EEG. Experiments demonstrate utility and feasibility of the method.

Keywords Web mining, Events timeline analysis, Event evolution graph

1 引言

当今互联网已经成为第四媒体,通过互联网我们可以对世界上发生的新闻进行充分了解。但是大量信息充斥在互联网上,若需要对一个新闻事件的来龙去脉有清晰的了解,就必须花相当多的时间和精力在互联网上搜索。一些搜索引擎可以帮助提供相关新闻的链接,但是没有提供按照新闻事件发生的时间顺序和依赖关系对事件进行组织的功能。需要用户从搜索引擎提供的大量网页链接中自己寻找,这是一项费时的工作。

本文定义了Web新闻流增量演进分析IEA(Incremental Evolution Analysis)来自动化组织新闻事件演化。IEA任务可以:(1)从收集的关于某个话题(Topic)的新闻网页集合中检测属于该话题的新事件,确认事件之间的依赖关系,绘制关于该话题的事件在时间顺序上的演进图。(2)收集新时期,如第二天,关于话题的新网页,检测新事件,确认新事件和演进图中其他事件的依赖关系,增量更新演进图。图1是关于嫦娥一号发射这一新闻话题的事件演进图。

目前已有一些研究与IEA任务相似,如话题检测与跟踪TDT(Topic Detection and Tracking)、比较文本挖掘CTM(Comparative Text Mining)、时态文本挖掘TTM(Temporal Text Mining)和事件线索化(Event Threading)等,我们将在相关工作一节中进行详细讨论。

这些研究工作或者仅以线性的方式描述新闻事件的演进;或者假设一次获得了所有的文档,然后在文档集合上进行事件的检测和分析,但是当需要在Web新闻流上跟踪事件的进展时,这样的方法不能有效更新事件的进展。

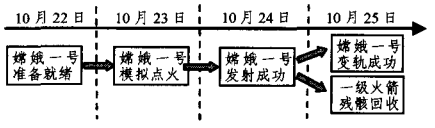

图1 一个事件进展图的例子

本文的主要贡献包括:定义IEA任务和EEG数据结构;提出EEG的构造和整理算法,有效地实现新闻事件的时间线分析任务。

本文余下部分组织如下:第2节介绍了相关工作;第3节形式化描述了IEA任务。第4节对EEG图的构造方法,以

到稿日期:2008-04-08 本文受十一五国家科技支撑计划(编号2006BAI05A01),国家自然科学基金(编号60773169)资助。
邱江涛(1972-),男,博士,讲师,研究方向为数据挖掘;唐常杰(1946-),男,教授,博士生导师,研究方向为数据挖掘。

及对图的整理方法进行了详细的描述。第5节通过详实的实验验证了新方法的有效性。最后是结论。

2 相关工作

CTM[1](Comparative Text Mining)任务在多个文本集合上发现潜在的主题(Theme)。作者提出一个概率混合模型，使用EM算法在整个文本集合上估计模型的参数，从而得到各主题。TTM[2](Temporal Text Mining)任务使用文献[1]中的模型发现潜在主题，然后构造主题进化图并分析主题的生命周期。将文本集合按时间段划分为可以重叠的子集，使用概率混合模型抽取最突出的主题。Mei的方法需要在整个文档集合上产生背景模型，因此不能进行增量的主题进化图的构造。

Event Threading[3]假设一次获得了关于某个话题的所有文本，使用聚类的方法对文本集合进行聚类，每个类即一个事件，一个事件用一系列故事(story)来描述。该方法不能解决增量更新的问题。文献[6-8]可以挖掘新事件，但不能反映出事件的进展路线。

关于话题检测与跟踪已经有很多研究[9]。话题检测把讨论相同话题的stories聚成一类。目的是在文本流中发现新话题的出现。话题跟踪在一个文本流上跟踪关于相同话题的事件。话题跟踪的结果是一个单独的时间线上在某一时刻出现的事件。

3 问题描述

本文假设每篇文档仅包含一个事件，每个事件具有时间戳，时间粒度为天。

定义1(事件) 设d是一篇文档，事件e是d中词的概率分布。它反映了词w在d中出现的概率，符号如表1所列。用一个一元概率语言模型描述则是$\{P(w|d)\}_{w\in d}$，并且$\sum_{w\in d}P(w|d)=1$。

表1 符号表

符号	含义	符号	含义	符号	含义
e	事件	c	文档集合	o	节点
d	文档	t	时间戳	l	层跨度
w	词	v	词集	u	类

定义2(话题) 一个事件集合中，所有事件共同具有的背景，称为话题。话题是比事件粒度更高的新闻内容的描述。

例如，两个事件'嫦娥一号探月卫星发射成功'和'嫦娥一号卫星第一次变轨成功'，它们都是关于嫦娥一号的新闻事件。'嫦娥一号'就是两个事件的背景，也称作话题。

定义3(事件进展图EEG) EEG是一个按层划分的有向无环图。

1)EEG中的一个节点是一个四元组($Child\_List$, $Next$, TC, e)，其中$Child_List$是孩子节点线性表。$Next$指向同层的下一个节点，TC是一个文本分类器，e是节点表示的事件。

2)一个节点与孩子节点之间的联系称为一条边。

3)一个层节点数组$Layer$的每个记录是一个二元组(L, t)。L指向一个层的第一个节点，t是层的时间戳，时间粒度为天。

图2是EEG图的一个例子。同层节点具有相同的时间戳。相同的事件聚成一类组成一个节点。即EEG图中每个节点描述了一个事件。每条边表示两个事件之间的依赖关系。边只存在于不同层的节点之间。

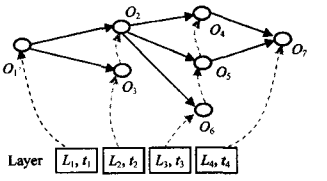

图2 EEG数据结构

定义4(新闻流增量演进分析IEA) 有时间戳为t的文档集合c，由c建立EEG，或者由c以增量方式更新EEG的任务称为Web新闻流增量演进分析。

4 事件进展图构造

观察1 设有两个事件A和B。如果事件A是事件B的延续或进展，则通常满足两个条件：事件A的时间戳比事件B的时间戳更新，事件A和事件B的相似度高。

EGG的构造过程实际上就是寻找事件和事件之间依赖关系的过程。按照观察1我们提出如下构造事件进展图的方法：

1)搜集关于某个话题的网页，按照每个网页的日期属性，即时间戳，将网页划分到不同集合$\{c_1,\cdots,c_k\}$。

2)构造EEG的第一层：选择满足t最新的文档集合c_1，按照话题对c_1进行文本聚类，检测并去除噪声类。每一类产生一个EGG中的节点。

3)构造第$i+1$层：用第$i-l+1\sim i$层的分类器对文档集合c_{i+1}中的文档分类，每个文档形成一个节点，这样在第$i+1$层形成$|c_{i+1}|$个节点。将父节点集合相等的节点合并。

4)对EEG进行整理操作。

我们使用文献[4]中的文档聚类方法OHC将文档集合按照话题聚成不同的类，文本分类采用文献[5]的PARC方法。因此，本文将对EEG构造与更新和EEG整理进行研究。

4.1 产生节点

事件进化图包括多个层。建立第一层节点和其他层节点的方法不同。本文认为每篇文档描述了一个事件。每天报道的新闻事件不计其数，即使针对某一个话题，每天也会有很多的新闻事件产生。当IEA任务将注意力放在重要的事件，就需要对事件的重要性进行判断。

观察2 新闻事件有这样的特点，相比不重要的事件，重要的事件会有更多的新闻源进行报道，因此关于某一个事件的报道越多越说明该事件重要；反之则说明不重要。

由观察2，在建EEG的第一层时，需要将描述相同事件的文档聚为一类。每个类是相同事件文档的集合，不同的类描述了不同的事件。文献[4]提出了一种可以根据主题进行文本聚类的算法OHC，产生的聚类结果包括多个类。每个类代表一个事件。

定义5(孤立类) 设聚类结果S中包含多个类$S=\{s_1, s_2,\ldots,s_n\}$，每个类$s_i\in S$中有$k_i$个数据。如果一个类$s_j\in S$中的数据个数$k_j$与其他类中的数据个数相比明显要少，则称类$s_j$是孤立类。

由定义5，重要事件和不重要事件在类内数据个数上应该存在显著差异。因此可以采用基于偏差的孤立点分析方

法，找出在数据个数上有显著较少的类别，即孤立类；然后删除孤立类。

将文档划分成不同的类且去噪后，需要建立节点。这时对每个类产生一个文本分类器。我们使用文献[5]中的算法PARC来产生文本分类器。

建立其他层的节点时为了更新EEG，需要确定每个事件与其他层的事件的关联。已知EEG中每个节点包含一个分类器。因此对新层的每个文本，用所有分类器对它进行判断，如果一个文本 d 被一个节点 o 的分类器识别，则用 d 形成一个新类，节点 o 与新类 s 之间建一条边 l。如果一个文本 d' 不能被任何一个分类器所识别，则作为噪声删除。当考察完所有 n 个文档 $(d_1,\cdots,d_n)$，形成了 $k(k\leqslant n)$ 个类 $(s_1,\cdots,s_k)$。每个类 s_i 有父节点集合 PO_i。对于同层的两个类 s_h 和 s_j，如果它们的父节点集合 $PO_i=PO_j$，则称两个类为等价类。将所有等价类合并，最后形成 $m(m\leqslant k)$ 个类。对每个类，建立节点，就完成了EEG的更新。具体算法描述见算法1的Build-OtherLayer。

算法1 EEG生成与更新算法

```
输入:文档集合 c
输出:EGG
Procedure buildEEG(c) //建立 EEG
1 us←(OHC (c); // 用 OHC 聚类算法产生聚类结果
2 DelNoise(us);
3 eeg=new EEG(); buildLayer(eeg);
4 For each s∈cs DO
5    BuildNode(s,eeg);
Procedure updateEEG(c) //更新 EEG
7 For each document d in c DO
8    build a cluster si and put it into set cs;
9    For each node o in i−l∼i−1 levels DO
10      If(o. TC(d)=true)Then
11         put o into parent node set of si POi;
12 MergeCluster (us); //合并类
13 buildLayer (eeg);
14 For each cluster s in cs DO
15    BuildNode (s,eeg);
Procedure BuildNode(s,eeg) // 建立节点
18 e←(getEvent(s); tc←(PARC(s);
19 insertEEG(new Node(e,tc));
20 For each node po in PO of s
21       update child-list of po; //更新父节点的孩子节点
```

算法第3和13步的函数 *buildLayerg* 更新层数组 *Layer* 和新层节点的 *next* 指针。在考虑事件的的进展时，一些重要话题事件密集，每天都有新事件发生，因此 l 设为1，只考虑新事件和最后一个层的节点之间的依赖关系。而一些话题，事件稀疏，几天才会有新事件发生，那么为了保持图的连续性，可以把层跨度 l 设得大一些。一个新事件也将检测是否和 l 天前的事件有依赖关系。函数 *DelNoise* 进行孤立类的检测和删除。

算法第18步的 *GetEvent* 函数从类中获得事件。按照定义1，事件是一个一元语言概率模型。当需要在一个类(文档集合)中获得事件时，我们将类中的所有文本作为训练样本，事件作为参数，通过使用下面的概率模型估计参数，来获得事件 e。

$$\hat{p}(w|e)=\frac{\sum_{d\in u}c(w,d)}{\sum_{w'\in v}\sum_{d\in u}c(w',d)}$$

u 表示类，v 表示词集，d 表示文档，$c(w,d)$ 表示词 w 在文档 d 中出现的次数。

4.2 EEG整理

按照算法1产生的EEG，图的结构可能会很复杂。复杂的原因既包括话题进展的复杂，也包括在产生节点的过程中造成了路径的复杂。我们只想保留相对最明确清晰的路径，即保留主干事件的进展路径。枝节和等价节点路径的存在是导致EEG复杂的直接原因。

定义6(枝节) 设EEG包含 n 个层，如果存在路径从第一层的节点出发可以至少到达位于第 $n-l+1$ 层的节点，这样的路径称为树干路径。查找所有的树干路径。如果存在节点 O 不在树干路径上，那么我们称 O 为枝节。

我们认为，如果两个事件的时间戳相差 $l+1$ 天，两个事件是不可能关联的(l 参数有用户设定)。因此从第一层到第 $n-l$ 层的叶子节点没有可能在第 $n+1$ 层上演进出新事件。这些叶子节点就是不在树干上的节点。

定义7(等价节点) 设EEG上存在两个节点 O_1 和 O_2，如果 O_1 的父节点集合和子节点集合 PO_1 和 CO_1，与 O_2 的父节点集合和子节点集合 PO_2 和 CO_2，有 $PO_1=PO_2$，且 $CO_1=CO_2$。则称节点 O_1 和 O_2 是等价节点。

将EEG中的枝节删除和将等价节点合并可以减少EEG的复杂性。图3a是整理前的EEG图，图中的粗线描述树干路径。设参数 l 为2，则节点 O_3 没有在树干路径上因此是枝节。节点 O_4 和 O_5 具有相同的父节点集合{ O_2 }和子节点集合{ O_7 }，因此 O_4 和 O_5 是等价节点。经过整理操作后，删除了枝节，并将等价节点合并，得到如图3b所示的整理后EEG。

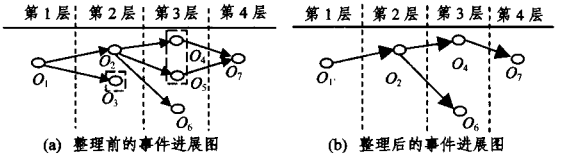

图3

等价节点只存在于一个节点的孩子节点中，因此只需要在一个节点的孩子节点中找等价节点。在EEG整理算法中，我们层次遍历EEG的每个节点，在每个节点的孩子节点中寻找等价节点。EEG的整理包含两种操作，枝节修剪和等价节点合并。具体算法描述见算法2。

算法2 EEG整理算法

```
输入:EEG
输出:整理后的 EEG
1 set all node in EEG as twig node;
2 For each node o in first layer DO
3    DeepFirTraverse(st,o); //st 为一个辅助栈
5 delete all twig;
6 For each node v in LayerTraversal DO //层次遍历中的每个节点
7    For p in v. Child_List DO
8       If ∃q in v. Child_List s. t. p. child=q. child and p. parent=q.
parent Then
9          Merge (p,q); // 合并等价节点
```

(下转第231页)

mbining Image and Linguistic Analysis[A]// Proc. International Symposium on Visual Computing (ISVC), Lecture Notes in Computer Science. Vol. 4292, 2006: 1686-1695

[16] Abdolah C, Golshah N, Alfred M. Sketch-based image matching using angular partitioning. IEEE Trans. on Pattern Systems and Humans, 2005, 35(1): 28-41

[17] Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag, 1995

(上接第 195 页)

Procedure DeepFirTraverse(st, node o)//从一个顶层节点 node 开始一个深度优先遍历

11 st. push(o);

12 If o is leaf in later two layer Then

13 set each o in st as non-twig node;

14 Else If (o is null) Return

15 For each child of node DO

16 deepFirTraverse(st. child)

17 st. pop();

5 实验

我们用 JAVA 实现了上述算法。实验在 P4 2.6G CPU, 512M 内存, Windows XP 计算机上进行。从互联网搜集了 18 天(2007 年 10 月 24 日至 11 月 11 日)关于嫦娥一号发射的共 1840 篇相关事件报道网页组成数据集 D1。

实验在两个数据集上产生事件进化图。首先在'嫦娥一号'数据集上产生事件进展图,如图 4 所示。图 4 中每个节点使用属于该节点的某个文档的标题作为事件描述,事件表如表 2 所列。当进行 EEG 的整理操作时,对等价节点进行了合并。合并后节点的事件描述为合并前多个节点事件描述的并集,如节点 9,17 等是合并后节点。为了便于观看进展图,删除了部分枝节。层跨度参数设为 $l=1$。

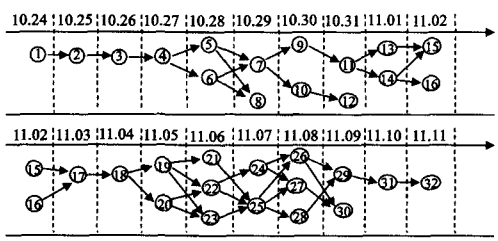

图 4 数据集 D1 的事件进展图

从事件进展图中可以观察到,嫦娥一号发射的主事件进展在图中做了详细的描述,而相对不重要的事件,则被忽略掉了。

表 2 事件表

1	外电:承载中国人希望,嫦娥一号成功发射升空	17	嫦娥一号今离月还有 3 万公里仍需飞行 1 天。嫦娥一号昨天实施首次轨道中途修正。
2	嫦娥一号卫星首次变轨成功	18	嫦娥一号直飞月球捕获点。胡锦涛温家宝致电祝贺嫦娥一号卫星近月制动圆满成功
3	嫦娥一号卫星成功实施第二次变轨。嫦娥一号第一次变轨成功,推进系统工作正常。	19	胡锦涛温家宝致电祝贺嫦娥一号卫星近月制动圆满成功
4	月球探测卫星嫦娥一号运行正常各系统状态良好	20	嫦娥一号将实施首次近月制动将成为月球卫星
5	嫦娥一号卫星 10 月 29 日将实施第二次近地点变轨。	21	胡锦涛温家宝贺嫦娥一号首次近月制动成功
6	嫦娥一号状态良好 深空测控网全面启动	22	嫦娥一号最快可于 7 日宣布"奔月"成功喜讯
7	北京时间 29 日 18 时 01 分 嫦娥一号第三次变轨成功。嫦娥一号卫星今起给地月照相	23	"嫦娥一号"今日 11 时左右进行第二次近月制动
8	嫦娥一号卫星第三次变轨成功未来十天有四大看点	24	嫦娥一号卫星完成第三次近月制动
9	嫦娥一号有关数据将在一年后国际共享。嫦娥一号探月卫星成为我国飞行最远的航天器	25	国家航天局发言人宣布嫦娥一号卫星成功绕月
10	黄江川:嫦娥一号卫星紫外敏感器达国际先进水平	26	嫦娥一号开始环月工作 11 月下旬传回图片语音。嫦娥一号 11 月下旬传回第一段语音
11	嫦娥一号卫星飞离地面高度 11 月 1 日将再创新高。已达远地点 12 万公里嫦娥一号今奔月。	27	嫦娥一号测控开展国际合作 卫星运行面临四风险
12	为什么说嫦娥一号发射环节取得圆满成功	28	专家称日凌现象可能干扰嫦娥一号通信
13	嫦娥一号卫星正式脱离地球怀抱开始奔月旅程	29	图文:嫦娥一号探月卫星进入近月工作轨道
14	嫦娥一号预计 11 月 5 日 11 时 25 分进入月球捕获轨道	30	嫦娥一号升空 VIEWGOOD 应用进入新领域
15	嫦娥一号卫星测控首次实现国际联网。嫦娥一号预计 11 月 5 日 11 时 25 分进入月球轨道	31	嫦娥一号全面体检 首次面对日凌考验
16	嫦娥一号首次轨道修正取消	32	嫦娥一号成功经受"日凌"挑战

结束语 在互联网已经成为第四媒体的今天,互联网用户有这样的需求:将互联网上的新闻事件按照时间顺序和事件依赖关系组织起来呈现给用户,帮助用户方便快捷地对新闻事件演进过程进行了解。从这一需求出发,我们定义了 IEA 任务。并通过定义 EGG 数据结构,以及实现 EGG 的构建和整理来完成 IEA 任务。和其他类似任务相比我们的方法可以在低事件粒度上增量完成对事件发展的描述。实验证明了我们提出方法的有效性。

参 考 文 献

[1] Zhai Chengxiang. A Cross-Collection Mixture Model for Comparative Text Mining// Proceedings of KDD 2004

[2] Mei Qiaozhu. Discovering Evolutionary Theme Pattern from Text -An Exploration of Theme Text Mining// Proceedings of KDD 2005

[3] Event Threading within News Topics// Proceedings of CIKM 2004

[4] Qiu Jiangtao, Tang Changjie. Topic-Oriented Semi-supervised Documents Clustering// Proceedings of SIGMOD 2007 Workshop IDAR

[5] Qiu Jiangtao, et al. A Novel Text Classification Approach based on Enhanced Association Rules. LNAI 4632. ADMA 2007

[6] 吴平博,陈秀群,马亮. 基于时空分析的线索性事件的抽取与集成系统研究. 中文信息学报,2006,20(1)

[7] 赵华,赵铁军,于浩,等. 面向动态演化的话题检测研究. 高技术通讯,2006,16(12)

[8] 贾自艳,何清,张海俊,等. 一种基于动态进化模型的事件探测和追踪算法. 计算机研究与发展,2004,41(7)

[9] Allan J, Carbonell J, Doddington G, et al. Topic detection and tracking pilot study: Final report// Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. 1998: 194-218