一种利用链接分析的 Web 话题跟踪方法

宋丹，林鸿飞，杨志豪

（大连理工大学计算机科学与工程系 大连 116024）

摘　要： 话题跟踪是一种基于事件的信息组织技术，实现对新闻信息中已有话题的动态跟踪。虽然传统的基于内容计算的话题跟踪方法也可以应用于 Web 话题跟踪，但它并没有利用 Web 的页面特征。文章提出了一种利用内容计算和链接分析相结合来进行 Web 话题跟踪的方法。实验证明这种方法是有效的。

关键词： 题跟踪；链接分析；内容计算

A Web NewsTracking Algorithm with Hyperlink analysis

Song Dan, Lin Hong-fei,Yang Zhi-hao

(1Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024)

Abstract: Topic Tracking is an event-based information organization task that links documents with previously detected events. Although the traditional Topic Tracking approaches based on Content comuting can be used in Web Topic Tracking, it did not use the valuable trait of Web page. This paper presents an approach that combine the hyperlink analysis with content computing. The approach is motivated by experiment.

Keywords: topic tracking; Hyperlink analysis; Content computing

1 引言

随着互联网的不断发展，新闻网页已经成为能够和报纸、电视以及广播齐名的四大媒体之一。与传统媒体相比，互联网新闻在时效性上有着不言而喻的优势。一个TDT系统的功能与一位信息工作者的工作相似，对于一个新的报道能够将其汇总到已识别到的话题中或者将这篇报道视为一个新的话题，它可以把分散的信息有效地汇集并组织起[1]。

话题识别与跟踪研究集中于五个子任务[2]展开：对新闻报道的切分、新事件的检测、报道关系检测、话题识别、话题跟踪。话题跟踪作为话题识别与跟踪(TDT)的一个子任务，是指根据给出的某一话题的一则或多则报道，把与该话题相关的报道联系起来。话题追踪任务是TDT评测中最普通的任务，参加该项评测的单位比参加其它任务单位要多一些。

在过去几年的评测中，进行话题追踪大致有两类方法:一类是基于信息检索的方法，包括向量检索和概率检索;另一类是基于文本分类的方法，例如最近邻分类、神经网、Boosting Bayes分类器、决策树[3]、动态聚类和

基金资助：国家自然科学基金 60373095

作者简介：宋丹（1980-），女，辽宁，硕士，dandong1007@sina.com

支持向量机等[4]。

虽然传统的基于内容计算的话题跟踪方法也可以应用于Web话题跟踪，但考虑到Web页面的如下特点：(1)Web页面之间的超链接是文本文档和Web页面之间最主要的区别，它对聚类一个具有相关性的页面群体提供了非常有价值的信息；(2)一部分Web新闻网页以新闻图片和相关链接为主，而只有极少的文字内容，这使得传统的基于内容计算的话题跟踪方法很难发挥好的效果。

因此，针对上述Web页面的两个重要特点，本文主要研究了链接分析在Web话题跟踪中的应用，提出了一种基于内容计算和链接分析相结合的话题跟踪方法，并针对新闻网页的结构特征对传统的权重和相似度计算方法进行了改进。实验证明该方法是有效的。

2 实现话题跟踪需要思考的问题

2.1 基本概念

话题(Topic)是话题识别与跟踪研究中的一个最基本的概念，它包括一个核心事件或活动以及所有与之直接相关的事件和活动[5]。与话题相对应的一个概念便是报道(Story)，它是指一个与话题紧密相关的、包含两个或多个独立陈述某个事件的子句的新闻片断[5]。如果一篇报道讨论了与某个话题的核心事件直接相关的事件或活动，那么就认为该报道与此话题相关。比如：爆炸现场消防情况、死难人数等都被看作是与某个爆炸话题直接相关。

2.2 话题/报道模型

对于话题跟踪任务，首先要解决用什么模型表示报道和话题的问题[6]。不论是话题还是报道，都要表示成计算机所能识别的形式。本文采用了向量空间模型来表示话题/报道。传统的向量空间模型只依靠词频求权重，从而忽略了文本的结构。但新闻信息有其较强的结构特征：一般新闻报道的题目一定比正文内容更重要，而且开头部分要比后面的部分重要，本文针对上述传统的向量空间模型的不足，采用了一种等级得分的权重计算方法。

2.3 针对新闻网页的等级得分权重计算方法

利用特征词出现的等级，来衡量这个特征词的重要性。在等级计分中，一篇报道中的title和content以及文本部分的第一行(即报道题目)被认为是第一个等级，句子在文本部分出现的次序即为它的等级。句子越后出现，句子中的词的权值也越小。一个特征 t 在报道中出现 m 次，则 t 的等级得分计算公式为：

$$rs(t) = \sum_{k=1}^{m} \frac{1}{2^{\ln t_k}} \tag{1}$$

2.4 与等级得分权重相应的内容相似度计算

本文用报道 d 与话题模型交集部分的权重与两者权重之和的比计算内容相似度。当然每个特征词自身的价值是不同的，因而遵照传统IR的做法，将等级得分乘以倒置文档频率IDF。例如，X，Y 为两个特征词集，则权重的相似度（RWS）为：

$$RWS(X,Y) = \frac{\sum_{K=1}^{|X \cap Y|} rs(t_k) * IDF(t_k)}{\sum_{j=1}^{|X|} rs(t_j) * IDF(t_j) + \sum_{l=1}^{|Y|} rs(t_l) * IDF(t_l)} \tag{2}$$

3 链接分析在Web话题跟踪中的应用

3.1 引入链接分析的原因

虽然基于内容计算的方法至今仍是话题跟踪核心技术的基础。但它根本没有利用网页的特性，可以说是前网络时代的技术。某则报道与话题是否相似，完全依赖于它所拥有的查询项的数目，这使它对那些以新闻图片和相关链接为主只含有少量文本内容的新闻网页很难发挥好的效果。

此外，随着事态的发展，话题往往会发生迁移和分化，以吉林石化公司爆炸这一话题为例，开始主要是有关爆炸现场的报道，接下来是有关爆炸对松花江造成的污染、有关下游哈尔滨停水等问题的报道，关注的焦点经常有所变化。只根据几篇种子报道，依靠与跟踪查询项比较的方法很难进行整个话题相关报道的检索，甚至会漏掉某些方面的相关报道。

综上所述，仅利用内容计算方法来进一步提高Web跟踪系统的性能是很困难的，要想突破必须借助更多的Web分析技术，因此我们在内容计算的基础上引入链接分析技术。

3.2 链接分析方法

链接分析也称结构分析，它的基本思想源于引文排名方法，它是基于这样一个假设[7]：某个页面1通过超链接指向页面2，则页面1与页面2是主题相关的，页面2对于页面1来讲是值得关注的页面。将这种假设应用于话题跟踪则是：某一相关报道通过链接指向报道d，则报道d在很大程度上也将是给定话题的相关报道；那些被种子报道指向、或被多个相关报道指向的网页与给定话题的相似度较高。

3.3 链接加分的计算方法

本文中链接加分是指：根据链接关系计算得出的网页A应得到的加分，算法的思想源于PageRank算法。当网页A指向网页B时，PageRank就认为“网页A投了网页B一票”。可根据网页的得票数评定其重要性。除了考虑网页得票数(即链接)的纯数量之外，还要分析为其投票的网页。“重要”网页所投之票自然份量较重，有助于增强其他网页的“重要性”。

将PageRank算法的这种思想应用于话题跟踪则是：当网页A指向网页B时，A就投了“B是相关报道”的一票；而且A与给定话题的相似度越高，则所投的票分量越重。PageRank最初的基本算法如公式(3)所示，式中：$PR(A)$表示网页A的PageRank值；$PR(Ti)$代表链接到A页的网页Ti的PageRank值；$C(Ti)$表示网页Ti的出度链接数量；d是阻尼系数，0<d<1。PageRank概率$PR(A)$，反映了A的重要程度。

$$PR(A) = (1-d) + d(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} ..... + \frac{PR(T_n)}{C(T_n)}) \quad (3)$$

本文所采用的链接加分的计算方法如公式(4)所示，其中$PS(A)$表示网页A应获得的加分；$RWS(Ti)$代表指向A页的网页Ti与给定话题的内容相似度值；d为相关因子。

$$PS(A) = d(RWS(T_1) + RWS(T_2) + ... + RWS(T_n)) \quad (4)$$

4 利用链接分析的 Web 话题跟踪算法

4.1 算法的基本思想

利用链接分析的跟踪方法的基本思想是：认为那些被种子报道或已确定为相关报道的网页所指向的新闻网页在很大程度上将是给定话题的相关报道，在内容计算的基础上根据链接分析给这些网页加分。例：如果报道A是种子或者它的内容相似度大于阈值α，A通过相关链接指向网页B，则A就投了“B是相关报道”的一票，即A可以利用自己的内容相似度得分RWS依据公式(4)给网页B的PS值加分ΔPS，ΔPS由公式(5)求出。如果种子报道A通过相关链接指向网页B_1，B_2，…，B_m，那么B_1，B_2，…，B_m这m个网页的PS值均能获得ΔPS的加分。

$$\Delta PS(B) = d * RSW(A) \quad (5)$$

4.2 算法的流程

判断某个新报道是否是相关话题，通常的做法是把新报道和已知话题进行比较，如果相似度高于某个阈值，则把新报道标识为话题的相关报道。本文采用的是增量聚类算法，算法流程如图 1 所示.

顺序处理报道，一次处理一则：(1)内容相似度计算；(2)比较阈值α，如果大于则认为它一定是相关报道，执行(3)，小于阈值α则开始处理下一则报道；(3)根据锚文本提取相关链接，删除不在网页集的链接，为它所指向的网页的PS值加分ΔPS。当所有的报道均按上述步骤被处理完一遍，则加分的操作也全部完成了，这时$PS(d)$就等于报道d的按公式(4)求得的链接加分值。最后对每则报道的内容相似度和链接加分进行求和，判定所有最终得分大于阈值α的为相关报道，组织并输出相关报道。

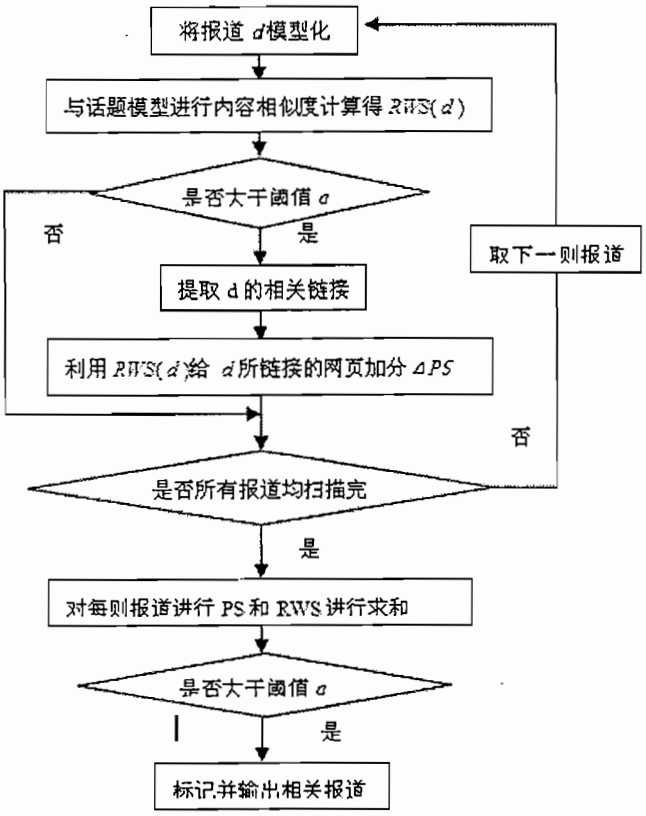

图 1 跟踪算法流程图

Fig.1 Flow chart of the tracking alogrithm

5 实验结果

5.1 系统性能评价

话题跟踪系统的性能的主要指标为查准率和查全率：查准率是指系统识别出的相关报道数与系统找出的全部报道数的比率；查全率是指系统识别出的相关报道数与所有相关报道总数的比率，也可以使用综合评价指标F来进行评价，其计算公式如下：

$$F = \frac{precision \times recall \times 2}{precision + recall} \tag{6}$$

5.2 实验及其结果分析

在实验中，我们从雅虎、新浪、新华网、北方网、NEN新闻中心这5大网站手工挑出600篇新闻网页，其中150篇作为训练集，450篇作为测试集。我们对吉林石化公司爆炸这一话题进行跟踪，相关报道应该有：爆炸现场、伤亡人员数量及抢救情况、疏散群众及抢险行动，爆炸后政府和石化公司及社会各界的态度、爆炸对松花江造成的污染、松花江下游哈尔滨停水、哈尔滨停水后采取的应急措施、松花江污染的整治等问题的报道。

为了比较利用链接分析前后的变化，我们采取了完全基于内容计算的跟踪算法、基于内容计算与链接分析相结合的跟踪算法两种方法。每种方法分别取种子报道数为1和4(4个种子均来自不同网站且报道话题的不同侧面)做了两次实验。结果如表1和表2所示：

表 1 种子报道数为 1 的跟踪结果

Tab.1 The result of tracking when N_t =1

门槛值	查全率(%)		查准率(%)		F(%)	
	内容计算	引入链接分析	内容计算	引入链接分析	内容计算	引入链接分析
0.1	84	98	74.05	76.91	78.49	86.18
0.2	64.29	74.57	86.87	88.47	73.89	80.93

0.3	56.57	68	93.83	94.82	70.58	79.2
0.4	32.57	51.71	96.61	97.83	48.72	67.66
0.5	9.71	24.29	97.14	98.83	17.66	39

由于实验跟踪的话题特征比较鲜明，所以查准率和查全率从总体上均偏高。由上表可知：加入链接分析后查全率和查准率都有所提高。原因在于：一些以新闻图片和相关链接为主只含有少量文本内容的相关报道，由于其被多个相关报道指向则获得较多加分，最终得分大于阈值 α 被正确识别出来。这样的网页如：
http://news.enorth.com.cn/system/2005/11/14/001164464.shtml和
http://news.enorth.com.cn/system/2005/11/14/001164465.shtml。另一个原因在于：一些话题焦点迁移后的相关报道由于获得较多链接加分而被正确识别出来。如：因为我们取的种子报道是http://news.enorth.com.cn /system /2005 /11/14/001163476.shtml，则http://news.enorth.com.cn/system/2005/11/22/00117 0401.shtml target=_blank便是话题焦点迁移后的相关报道，由于它的内容相似度不到0.1，在仅用内容计算的跟踪方法时它被认为是不相关，使用链接分析之后，由于它被相似度较高的网页指向（如：http:// news.enorth.com.cn/system/2005 /11/22/00116 9976.shtml target=），获得较多链接加分最终被正确地识别为相关报道。由于以上两个原因系统的查全率大大提高了。由于新识别出的相关报道大部分都是正确的，这使系统的查准率也有所提高。综合评价指标F值也明显高于引入链接加分之前，在种子数为1时表现更为突出。

表 2 种子报道数为 4 的跟踪结果

Tab.2 The result of tracking when N_t =4

门槛值	查全率(%)		查准率(%)		F(%)	
	内容计算	引入链接分析	内容计算	引入链接分析	内容计算	引入链接分析
0.1	93.16	98.42	64.72	65.96	76.38	78.99
0.2	78.42	91.32	76.21	78.86	77.30	84.63
0.3	58.84	68.42	93.91	94.89	72.35	79.51
0.4	48.95	61.32	98.94	99.15	65.50	75.78
0.5	35.89	58.16	99.27	99.55	52.72	73.42

6 结束语

话题跟踪是指根据给出的某一话题的一则或多则报道，把与该话题相关的报道联系起来，是一个直接面向应用的研究方向。由于web新闻信息的某些特殊性，决定了仅仅利用内容计算方法来进一步提高Web跟踪系统的性能是很困难的，要想突破必须要借助更多的Web分析技术。本文提出了一种借助链接分析技术的Web新闻信息话题跟踪方法，该方法对Web新闻网页针对性更强，并有效地提高了Web话题跟踪的系统性能。但实验没有使用公共的测试语料，因此很难与相关研究的结果进行比较，这将是下一阶段工作的主要任务。

参考文献：

[1] James Allan. Introduction to Topic Detection and Tracking. In James Allan, editor,Topic Detection and Tracking: Event-based Information Organization [M], Kluwer Academic Publishers, 2002.

[2] J Allan, J Carbonell, G Doddington et al. Topic Detection and Tracking Pilot Study Final Report[A] Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop[C], San Francisco: Morgan Kaufmann Pubilishers,Inc, 1998: 194-218.

[3] Yiming Yang, Tom Ault, Thomas Pierce, and Charles W. Lattimer, Improving text categorization methods for event tracking, SIGIR 2000, pp. 65-72.

[4] P. van Mulbregt, I. Carp, L. Gillick, S. Lowe and J. Yamron, Text Segmentation and Topic Tracking on Broadcast News via a Hidden Markov Model Approach[A], Proceedings of the DARPA Broadcast News Workshop[C], February 1999.

[5] 金珠，林鸿飞. 基于 HowNet 的话题跟踪及倾向性分类研究[J] .情报学报，2005，24(5) : 555-561.

[6] 李保利，俞士汶. 计算机识别与跟踪研究[J] . 计算机工程与应用, 2003 , 39(17) : 7-10.

[7] 刘悦，杨志峰，程学旗. 利用链接分析技术提高搜索引擎查找质量的研究[J] .微电子学与计算机, 5(2002): 18-21