华南理工大学学报(自然科学版)
第32卷增刊 Journal of South China University of Technology Vol. 32 Suppl.
2004年11月 (Natural Science Edition) November 2004

文章编号:1000-565X(2004)S-0042-05

在线新闻主题检测系统的设计与应用

万小军 杨建武

(北京大学 计算机科学技术研究所,北京 100871)

摘 要:利用主题检测技术可以从海量新闻信息中实时检测到主题信息,从而将新闻信息按照主题组织并加以利用.文中通过改进加窗策略,采用自适应倒排文档频率,设计了一个中文新闻主题检测系统并进行了实验.结果表明了该系统的有效性.该系统在新华网数据中心的成功应用进一步表明系统达到了实用需求.

关键词:主题检测;增量式聚类算法;加窗策略;自适应倒排文档频率

中图分类号:TP391 **文献标识码**:A

因特网上文本信息的爆炸式增长给信息检索技术带来了巨大的挑战,新闻报道则是其中的主要信息类型之一.人们越来越难以快速准确地从网上检索到高质量的有用新闻信息.传统的检索方式都建立在用户对自己查询需求的深刻理解之上,也就是说,用户在检索新闻信息时需要准确地把自己的查询需求表达成查询表达式,这种查询需求到查询表达式的转换上的任何偏差都会严重影响检索结果.然而在现实生活中,人们通常会有一种最基本的检索需求,比如"今天发生了些什么大事"、"上个月什么事件最热门"等等,在这类检索需求中,用户无法精确定义自己的需求,只能抽象地加以描述,这类检索需求是现今基于关键词检索的搜索引擎难以满足的.新闻主题的自动检测技术则将海量新闻信息按照主题进行组织,并以一定的方式展现给用户,满足用户的需要.利用主题检测技术,我们可以将文本信息按照其表达的主题进行层次式的归类和组织,方便用户的检索浏览和选择使用,也可以主动将主题信息推送给用户,实现个性化的服务,这种应用主要体现在以海量文本的处理为核心的内容管理系统中.在各大门户网站如新浪、搜狐、雅虎等的新闻频道上,主题检测技术可以取代人工完成自动专题生成、热点新闻生成等任务.同时,通过主题检测技术,可以对主题之间相互演化、发展趋势进行分析,为用户提供更高层次的服务.

图1是通过主题检测技术得到从2003年3月1日至25日与伊拉克相关的两个主题演化图,从图1看出"联合国对伊武器核查"和"美英对伊动武"这两个主题在时间上的联系及每个主题的演化过程.

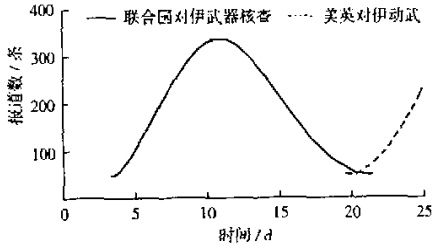

图1 与伊拉克相关的主题演化图

Fig. 1 The evolution diagram of Iraq-related topics

文章内容安排如下:第1节分别介绍了相关概念和技术;第2节阐述了本文提出的在线主题检测系统采用的技术;第3节给出了实验和结果;第4节介绍了相关应用;最后一节是结论和展望.

1 相关概念和技术

1.1 相关概念

主题检测技术和主题追踪技术合称为主题检测与追踪(TDT)。国际TDT会议不仅包括主题检测技术和主题追踪技术,还包括文档分割,主题首篇文

收稿日期:2004-08-27

作者简介:万小军(1979-),男,博士生,主要从事信息检索与自然语言处理方面的研究.E-mail:wanxiaojun@icst.pku.edu.cn

档检测，文档相关性检测等技术.

TDT 会议对主题的定义如下[1,2]：

主题　一个种子事件或活动以及与它直接相关的事件和活动.

事件　在特定时间特定地点发生的事情.

一个主题的例子如下：

标题　"法轮功"组织非法信号攻击卫星

种子事件　覆盖全国的鑫诺卫星 6 月 23 日至 30 日，陆续遭到"法轮功"非法电视信号攻击，卫星转发器传输的"村村通"广播电视工程中的中央电视台 9 套节目和 10 个省级电视台的节目受到了严重干扰.

相关事件　包括造成的影响和后果，政府各部门的各种反应，相关部门的补救措施等，但不包括法轮功组织的其他有关事件以及政府对法轮功组织的取缔等事件.

新闻主题检测就是从新闻文本集合中自动地检测出各个主题，将各个文档划归到相应的主题，并且能够实时地针对新来的文本信息检测新的主题. 图 2 表示了主题检测的意义.

图 2　主题检测的意义

Fig. 2　The meaning of topic detection

1.2　主题检测技术概述

与主题检测技术相关的技术包括聚类、分类、检索、主题追踪等，其中最相关的技术是聚类技术，主题检测技术是聚类技术在特定领域的一个应用.

主题检测的目的就是要按照主题将文档进行聚类，可以分为回溯检测与在线检测两种. 回溯检测可以看作是传统的文本聚类，代表性的算法有基于平均分组的层次聚类法[2-4]（GAC）. 而在线检测则是以在线的方式从实时文档流中标识新主题事件的开始位置，要求系统在表达新主题事件的文档出现时标识出该主题事件.

在线检测的算法有著名的在线增量式聚类算法[2-4]（INCR）及其改进，包括加窗策略[3,4]，基于时间的选择模型[5]，基于时间的阈值模型[6,7]. 此外，在文档的表示模型上，分别有人提出了自适应倒排文档频率[2-4]（Adaptive IDF），主题相关的倒排文档频率[8]，语言模型[9,10]，词链[5]等改进方法.

在线检测具有更实际的应用意义，是本文讨论的重点，如没有特殊说明，下文中出现的主题检测均指在线主题检测.

1.3　评价技术

主题检测与追踪会议采用统一的语料库和评价准则评测各个参与者的 TDT 系统. 除了采用与传统的信息检索和文本分类类似的查准率、查全率、F-measure 来评价结果之外，TDT 主要采用系统错误率来评价结果，主要包括错检率 F 和漏检率 M，一个好的 TDT 系统应该让 F 和 M 同时减小.

表 1　主题检测结果

Table 1　Results of topic detection　　个

类　别	相关文档数	不相关文档数
检测到的文档数目	A	B
没有检测到的文档数目	C	D

表 1 表示某个主题的检测结果，定义漏检率 $M = C/(A + C)$，错检率 $F = B/(B + D)$，为了同时考虑漏检率和错检率，类似 F-measure，TDT 引入了耗费函数[1,7]（Cost Function）

$$C_{\mathrm{Det}} = c_{\mathrm{F}}F(1 - P_{\mathrm{topic}}) + c_{\mathrm{M}}MP_{\mathrm{topic}} \quad (1)$$

式中：P_{topic} 为某个文档属于某个主题的先验概率；c_{F} 和 c_{M} 分别为错检和漏检的代价常量. 在第二届主题检测与追踪大会中，$P_{\mathrm{topic}} = 0.02$，$c_{\mathrm{F}} = c_{\mathrm{M}} = 1.0$. 显然，耗费函数值越小，表明 TDT 系统效果越好.

此外，还可以采用检测错误权衡图[1,7]（DET）来可视化表现错检率和漏检率之间的权衡关系.

2　在线新闻主题检测系统的技术设计

我们设计在线新闻主题检测系统的目的是为了实时地对新闻报道流进行主题检测，从而将新闻报道按照新闻主题有效地组织起来. 新闻报道流可以通过网络爬虫从新闻网站自动收集获得.

我们的系统采用了著名的在线增量式聚类算法，这种算法是对基本的增量聚类算法的改进. 基本的增量聚类算法可简单描述为：

预设一个聚类阈值（Clustering Threshold）t_{c}. 算法顺序处理输入的每篇文档，初始以第一篇文档为种子创建第一个类簇. 对于每一篇输入的新文档，与以前生成的所有类簇进行相似性比较，如果该文档与之前的某个类簇的相似度值大于聚类阈值 t_{c}，那么该文档将属于该类簇；否则，将以该文档为种子创建一个新的主题类簇.

基本的增量聚类算法适合回溯检测,通过调整聚类阈值,算法最后可得到不同粒度层次上的主题类簇,因此合理地选择聚类阈值对主题检测很重要.

对于在线主题检测,系统需要灵活地调整新主题的敏感度,合理界定新主题,因此在线增量式聚类算法预设了另一个创新阈值(Novelty Threshold)t_n,通常$t_c \geq t_n$.算法改进如下:

设s_{max}表示当前文档x与以前主题类簇之间的最大相似度值,那么,(1)如果$s_{max} > t_c$那么这篇文档被标识为"OLD",将该文档加入到最相似的主题类簇中,更新该类簇中心点;(2)如果$t_n \leq s_{max} \leq t_c$,那么这篇文档被标识为"OLD",不作进一步处理;(3)如果$s_{max} < t_n$,那么这篇文档被标识为"NEW",意味着该文档是表达一个新主题的第一篇文档.

由于新闻信息具有很强的时效性,时间是新闻报道的一个关键性特征,而且时间信息比人物、地点等特征更容易获得,考虑到表达同一主题的文档之间的时间相近性,增量式聚类算法还可以通过加窗策略做进一步的改进.

我们采用的加窗策略分为两部分:一是对新闻文档加窗.在计算当前文档和以前某个主题的相似度时,如果某个主题所关联的最新文档在时间上与当前文档相邻,那么对该主题赋予较高的相似度值;二是对主题加窗。对主题加窗的目的有两个,一个是为了提高计算的效率,因为对主题加窗后,当前文档只需和窗口中的已有主题比较相似性,不需要跟所有已发生的主题比较.一个跟对文档加窗的目的类似,在计算当前文档和主题的相似度时,结合主题的时间特性来加强时间相邻性对相似度计算的影响.我们对创建时间较早的主题和当前文档比较时赋予较低的相似度,而创建时间较晚的主题和当前文档比较则赋予较高的相似度,主题的创建的时间可用其产生的顺序来表示.结合两个加窗因素对相似度计算的影响,我们可以得到下面的计算公式:

$$S_{(x,c)'} = \begin{cases} \left(1-\frac{i}{m}\times\alpha\right)\times\left(1-\frac{j}{n}\times\beta\right)\times S_{(x,c)}, \\ \quad \text{如果 } c \text{ 中有文档在文档窗口中} \\ 0, \text{否则} \end{cases} \tag{2}$$

式中:m表示文档窗口的大小;n表示主题窗口的大小;i表示当前文档x和主题c中与x时间最相近的文档相隔的文档个数;j表示主题c和主题窗口中最后一个主题相隔的主题个数;α,β为调节因子.

由于绝大多数主题的生命周期都在一定范围内,因此只要主题窗口大小选择合适,就能避免检测到已经检测出的的主题.而且,主题窗口的存在,对实现也有很大的好处.当主题数量非常多时,无法将所有主题放置在内存,系统不得不多次操作磁盘,效率上产生了瓶颈.对主题加上了窗口,通常情况下可以将窗口中的主题全部放在内存中处理,效率得到很大提高.文档窗口和主题窗口的大小应根据待处理的新闻报道所涉及的新闻范围大小而定.一般说来,新闻范围越大,新闻报道就越多,主题也会越多,窗口就应该加大.

通常情况下,文档的特征表示用向量空间模型,主题类簇的特征则用该簇文档均值点的向量空间模型表示,相似度的计算方法为余弦度量法.

特征的表示通常是选取词,词的权重为$f_{TF} \times w_{IDF}$,其中w_{IDF}为静态IDF,是根据一定的文档集合预先计算出来的,不同的主题事件检测将使用这个固定的IDF,而且以后不会有变化.但是,由于不同的主题事件内容不一样,强调的重点不一样,而且不断在演化发展,使用的词汇也在变化,因此采用动态变化的与主题相关的IDF值将会改进在线主题检测的结果.自适应IDF的计算公式如下

$$w_{IDF(W,p)} = \log_2(N_p/n_{W,p}) \tag{3}$$

式中:p表示当前时间点;N_p表示到当前时间为止已有的文档数量(包含背景文档集合);$n_{W,p}$则表示在时间p时已有文档中包含词W的文档频率.

此外,对于f_{TF}的计算,为了考虑了文档长度的影响,我们采用Okapi系统中的经验值,最终得到公式(4).

$$f_{TF} = \frac{f_{TF}^2}{f_{TF} + 0.5 + 1.5 l_d / l_{avg}} \tag{4}$$

式中:f_{TF}为原始词频;l_d为当前文档长度;l_{avg}为文档平均长度.最后计算每个词的$f_{TF} \times w_{IDF}$权重都要进行规一化.

3 实验及结果

由于我们目前没有参加NIST组织的TDT项目,得不到TDT的用于正式评测的新闻数据,所以我们自己从新闻网站下载了新闻文本建立了训练数据集和测试数据集.目前训练数据集包括50个主题,约500篇新闻报道,测试数据集则包括30个主题,约435篇新闻报道,将这些新闻报道按时间排好序.新闻数据来源为国内门户网站新浪(www.sina.com)、搜狐(www.sohu.com)和网易(www.163.com),类型为国际新闻.新闻报道所属主题信息全

部由手工标注，因此可能由于主观原因存在着不准确性.

目前，我们的原型系统主要处理中文新闻报道，采用了第三方的分词系统，由于该分词系统不提供词性标注功能，而且停用词表不完善，所以我们选择双字词以及多于双字的词作为文本的候选特征.

在用训练集对 TDT 原型系统进行训练和调整之后，系统的主题检测部分的几个参数取值见表 2.

表 2　主题检测参数取值

Table 2　The values of parameters for topic detection

项　目	参数值
文档窗口大小	1 000
主题窗口大小	25
特征词个数	15
聚类阈值	0.10
创新阈值	0.095

我们用原型系统对测试集进行主题检测，$P_{topic}=0.02$，$c_F=c_M=1.0$，并且采用 TDT 2002 所采用的宏平均方法，先对每个主题分别进行统计计算，然后取平均值，得到错检率为 2.97%，漏检率为 32%，耗费函数值为 0.036.

实验表明，如果不采用加窗策略，那么耗费函数值将上升 0.02. 如果采用静态 IDF，耗费函数值将稍微上升 0.005.

此外，还发现特征词的个数对检测结果有较大影响，特征词个数越多，结果反而越差. 我们对特征词个数为 100，50，20，15 的情况分别进行了实验，对应的耗费函数数值分别为0.03，0.023，0.016，0.013.

4　应用实例

目前，已成功将主题检测技术应用于新华网数据中心，进行专题生成、热点新闻生成，取得了较好的效果. 新华网数据中心的基本目标是充分利用分布于全国甚至全球的分社，整合各个地方站点的内容，实现信息的充分共享，消除信息孤岛，并利用数据挖掘等智能化的技术手段达到降低重复劳动，发挥信息利用率. 也就是说，总网的数据和分网的数据，通过数据中心整合来形成完整的不重复的内容. 在这个基础上，进行二次深度利用，如：深入报道、专题、相关等等功能。从而产生新的价值.

新华网数据中心的总体结构如图 3 所示.

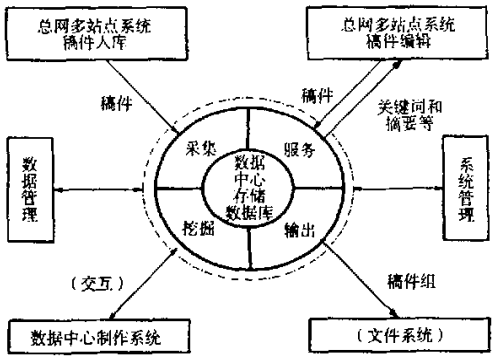

图 3　新华网数据中心总体结构图

Fig. 3　The structural diagram of XHNet's data center

其中数据挖掘是核心模块，指利用数据中心已存贮的信息进行分析挖掘，为编辑人员进行专题制作、新闻追踪、热点新闻分析、热字管理等提供辅助支持. 其中的专题制作和热点新闻分析部分均利用到了主题检测技术. 图 4 是热点新闻分析的界面.

图 4　新华网数据中心界面

Fig. 4　The interface of XHNet's data center

目前，美国的 Candor 公司、IBM 公司等都已经在自己的产品中加入了主题检测功能。

5　结语

本文介绍了新闻主题检测的相关概念和技术，设计了一个在线新闻主题检测系统，给出了实验结果. 该系统在新华网数据中心的应用效果好.

下一步，我们将在系统设计中尝试结合更多的文本的语法、语义特征，例如，词性，基本短语等，以期望提高系统的性能.

检测出主题之后，需要采用一种简洁友好的方

式将该主题的内容展现给用户,这就要求我们对主题簇的内容进行归纳,要涉及到多文档的自动摘要技术,这也是当前研究热点之一.

在应用上,主题检测的功能也将融合进入我们自有产权的文本挖掘产品——方正智思系统.

参考文献:

[1] National Institute of Standards and Technology. The 2002 topic detection and tracking (TDT2002) task definition and evaluation plan, version 1.1[EB/OL]. http://www.nist.gov/speech/tests/tdt, 2002-05-13.

[2] Allan J, Carbonell J, Doddington G, et al. Topic detection and tracking pilot study: final report [A]. In Proceedings of DARPA Broadcast News Transcription and Understanding Workshop [C]. Virginia: Lansdowne Conference Resort Lansdowne, 1998. 194-218.

[3] Yang Y, Pierce T, Carbonell J. A study on retrospective and on-line event detection [A]. In the Proceedings of ACM SIGIR 1998 [C]. Melbourne: Association for Computing Machinery Press, 1998. 28-36.

[4] Yang Y, Carbonell J, Brown R, et al. Learning approaches for detecting and tracking news events [J]. IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval, 1999, 14(4): 32-43.

[5] Hatch P, Stokes N, Carthy J. Topic detection, a new application for lexical chaining? [A]. British Computer Society IRSG 2000 [C]. Cambridge: British Computer Society, 2000. 94-103.

[6] Papka R. On-line new event detection, clustering and tracking[D]. Amherst: Department of Computer Science, University of Massachusetts Amherst, 1999.

[7] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking [A]. In the Proceedings of ACM SIGIR 1998 [C]. Melbourne: Association for Computing Machinery Press, 1998. 37-45.

[8] Dharanipragada S, Franz M, McCarley J S, et al. Story segmentation and topic detection in the broadcast news domain [A]. In Proceedings of the DARPA Broadcast News Workshop [C]. Herndon: National Institute of Standar and Technology, 1999.

[9] Yamron J. P, Knecht S, van Mulbregt P. Dragon's tracking and detection systems for the TDT2000 evaluation [A]. In Proceedings of Topic Detection and Tracking Workshop [C]. U S A: National Institute of Standar and Technology, 2000. 75-80.

[10] Allan J, Lavrenko V, Frey D, et al. Umass at TDT 2000 [A]. In Proceedings of Topic Detection and Tracking Workshop [C]. U S A: National Institute of Standar and Technology, 2000.

Design and Application of an On-line News Topic Detection System

Wan Xiao-jun Yang Jian-wu

(Institute of Computer Science and Tech., Peking Univ., Beijing 100871, China)

Abstract: Topic detection technique can be used to detect news topics from a great amount of news stream, and help us organize and utilize news information according to their topics. In this paper, a Chinese news topic detection system is designed and tested by using the improved time window strategy and adopting the self-adaptive inverse document frequency. It is then indicated that the proposed system is effective. The application of the system in the data center of XinHua Net further illustrates that the system meet the application demands well.

Key words: topic detection; incremental clustering algorithm; time window strategy; self-adaptive inverse document frequency

责任编辑:唐民英