面向 TDT 的动态多文档文摘研究

刘美玲[1,2], 赵铁军[1], 郑德权[1], 于 洋[1]

(1. 哈尔滨工业大学 计算机科学与技术学院,哈尔滨 150001, mlliu@mtlab.hit.edu.cn;
2. 东北林业大学 信息与计算机工程学院,哈尔滨 150040)

摘 要: 应用话题检测技术中的聚类思想,分析了动态多文档文摘时间特性.利用时间信息的阈值变化得到不同的多文档聚类结果,进而生成基于动态网页信息数据流的多文档文摘.针对不同阈值的多文档文摘比较,了解时间信息在动态的多文档文摘中的重要性.实验证明在 TDT 的话题检测技术中,可通过调整 α 值的方法来平衡 TDT 的输出结果,进而生成质量更好的动态多文档文摘.

关键词: 话题检测;时间信息;动态阈值;多文档文摘
中图分类号: TP391 **文献标志码:** A **文章编号:** 0367-6234(2010)11-1767-04

Research on dynamic multi-document summarization by topic detection and tracking technology

LIU Mei-ling[1,2], ZHAO Tie-jun[1], ZHENG De-quan[1], YU Yang[1]

(1. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, mlliu@mtlab.hit.edu.cn;
2. College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China)

Abstract: In this paper time characteristics of dynamic multi-documents summarization has been analyzed by the clustering idea in topic detection technology. From the change of time information threshold value, different multi-document clustering and multi-documents summarization for the dynamic Web information data stream can be generated. By compared with different threshold values, the importance of time information in a dynamic multi-document summarization is understanded. Experimental result means that you can counterpoise the output of TDT by adjusting α value and generate better quality dynamic multi-document summarization.

Key words: topic detection; temporal properties; dynamic threshold; multi-document summarization

作为一个新兴的自然语言处理研究方向,话题检测与跟踪(Topic Detection and Tracking, TDT)[1]是一项旨在基于事件对信息进行组织和利用的研究,也是为了应对信息过载问题而提出的一项应用研究.话题检测本质上类似于无监督的聚类研究,但通常的聚类基于全局信息实现,而话题检测中的聚类以增量方式实现.

本文研究的动态多文档文摘作为话题检测的重要组成部分和输出结果,生成事件的线索报告.基于话题检测技术的动态中文多文档文摘,结合了 TDT 的时序性及动态特性,针对于巨大的网络信息量,能够在庞大的数据流中按照话题聚类,形成同一主题文档集合,进而生成多文档文摘,便于使用者按照主题查询相关文摘,得到最新的摘要式信息.

1 相关工作与理论方法

本文使用了话题检测技术中的聚类思想,利用时间信息的阈值变化得到不同多文档聚类结果,进而生成基于动态网页信息数据流的多文档文摘,方便用户按照主题的不同进行查询,缩短了网页信息的浏览时间.通过不同阈值的多文档文摘比较,了解时间信息在动态的多文档文摘中的重要性.

收稿日期: 2009-11-03.
基金项目: 国家高技术研究发展计划重点资助项目(2006AA010108);
国家自然科学基金重点资助项目(60736044).
作者简介: 刘美玲(1981—),女,博士,讲师;
赵铁军(1962—),男,教授,博士生导师.

1.1 话题检测技术

话题检测与跟踪研究的最终目标是在多种形式多种语言的新闻报道流中准确的自动检测新话题并对已知话题进行跟踪,所以可以说话题检测任务是对这一终极目标的最佳体现.

目前大多数话题检测算法都是基于某种聚类思想实现,比如:Single-Pass 聚类方法[2-3],K-NN[4],K-means[5]以及层次聚类算法[6]. 话题通常用质心来表示,话题特征项的权值重、权重通常用 TFIDF 计算. 本文所采用的报道和话题的表示模型为:1)报道的基于动态 IDF 的向量空间表示模型;2)话题的中心向量表示模型;3)报道和话题之间的相似度计算. 目前话题检测算法主要是对文本聚类算法的改进和延伸. 在话题检测研究中,通常都是预先设定阈值,并且一旦设定在整个检测过程中就不再改变,这种做法是不合理的. 首先,由于话题的时间集中性[7],随着时间的推移和该话题相关的报道将越来越少,所以应该在检测过程中逐渐加大阈值;其次,不同的话题应该采取不同的阈值.

1.2 多文档文摘技术

多文档文摘系统,根据采用的方法的不同,大致可以分为[8-9]:

1)基于单文档文摘技术的方法. 基于单文档文摘技术生成多文档文摘是指采用单文档文摘中有效的方法计算多文档集合中的每个句子的权值,选出权重最高的若干句子组成文摘[10]. 常利用对句子进行加权的特征有文档的词频统计、句子在原文中的位置和句子长度等信息.

2)基于信息抽取的自动文摘. 信息抽取指的是从大量的原始数据中抽取有价值的信息[11]. 这些信息包括特定的实体、关系或事件. 基于信息抽取的自动文摘包括两个阶段:从文本中获取信息和填充模板生成文摘. 文摘的模板中包含了所处理的对象的各个属性,其中需要从原文中获取的各项信息以空槽的形式表示.

3)基于多文档集合特征的方法. 这种方法主要是指将多文档集合作为一个整体进行研究,通过对多文档集合中的句子按照其表达意思的相近程度重新组合聚类,然后从不同的类别中抽取文摘句[12].

本文使用基于原文抽取的多文档文摘系统,此类型系统流程为:1)根据各种形式特征对句子加权;2)根据句子权重选取句子并去除冗余信息;3)对句子排序以生成一篇连贯的文摘. 综合句子加权、选取、排序技术实现一个中文多文档自动文摘系统.

1.3 基于时间信息的话题检测技术

时间信息在自然语言处理(Natural Language Processing, NLP)是许多自然语言处理任务的基础,如多文档文摘系统中需要按照时间顺序排列相关的信息[13]. 新闻报道含有丰富的时间信息. 基于上述分析,应用基于话题的持续时间的动态阈值模型为

$$\mathrm{Threshold}(T,t) = \theta + \alpha(\mathrm{Time}(S) - \mathrm{Time}(T)). \quad (1)$$

式中:$\mathrm{Threshold}(T,t)$ 为 t 时刻话题 T 的阈值,θ 为话题刚建立时的阈值,且 θ 是一个常数;α 为时间信息在动态阈值中所占的比例,且 α 是一个可调参数. $\mathrm{Time}(S)$,$\mathrm{Time}(T)$ 分别为报道时间和话题的建立时间,d.

1.4 比值法

在话题检测的基本模型中,首先计算报道和所有已存在话题的相似度,然后通过最高相似度与相似度阈值的比较得到报道和话题是否相关的判断结果[14-15]. 但是,在基于动态阈值模型的话题检测算法中,每个话题的相似度阈值都可能不相同. 为了让报道和不同的话题的相似度具有可比性,本文应用了比值法,用来选择与某个报道最相似的话题,基于比值法选择最相似话题的过程为:

1)计算报道和话题的相似度.

2)计算相似度与话题的阈值的比值,最大比值记为 $Ratio_{\max}$,得到最大比值的话题为 $T_{\max}$,如果 $Ratio_{\max} \geqslant 1$,那么报道和话题 $T_{\max}$ 相关;

3)如果 $Ratio_{\max} < 1$,那么新建立一个话题.

实验结果证明,基于话题持续时间的动态阈值模型配合比值法的使用取得了很好的效果.

1.5 多文档文摘系统

本系统的主要功能是:首先对话题相关文档集中的每篇文本进行分句、分词的预处理,应用停用词表进行停用词的过滤,并进行词频统计. 在此基础上进行主题词抽取,并融合词汇信息、位置信息、长度信息来计算句子的权值;然后以权值为依据,不断选择权值大且含新信息多的句子加入摘要,直到达到指定长度;最后通过一定的排序算法对句子进行排序,从而生成一篇具有一定逻辑性的,连贯、流畅的文摘.

2 实验结果和分析

本文应用基于话题持续时间的动态阈值模型结合比值法来选择和话题最相似的报道,针对不同的时间阈值比较它们生成的动态多文档文摘质量.

2.1　汉语语料的话题检测实验

本文实验的汉语语料是从网上获取的，用手工标注的新闻报道，共有 113 篇. 语料中所有的报道按照时间顺序排列组成一个文本文件，其中标注了 12 个话题. 为了更好的说明语料库的内容，汉语语料的标记情况如表 1 所示.

表 1　汉语语料标记信息

编 号	话 题	YES
1	2004 台湾立委选举	9
2	3·19 枪击案	11
3	扁宋会	9
4	策划大游行	5
5	反分裂法	11
6	胡锦涛对台讲话	10
7	“汉光”军事演习	6
8	肯德基苏丹红事件	11
9	李登辉访日	11
10	刘翔跨入小学语文教材	9
11	湖南凤凰大桥坍塌	13
12	山东新泰矿难	8

2.2　话题检测评价标准

本文采用 TDT 评测中所使用的归一化检测开销 $(C_{Det})_{\text{norm}}$ 来评测话题检测系统为

$$(C_{Det})_{\text{norm}} = \frac{C_{\text{miss}}P_{\text{miss}}P_{\text{target}} + C_{fa}P_{fa}P_{\neg\,\text{target}}}{\min(C_{\text{miss}}P_{\text{target}}, C_{fa}P_{\neg\,\text{target}})}. \tag{2}$$

式中：P_{miss} 为系统的漏报率，P_{fa} 为系统的误报率，P_{target} 为在信息流中看到一个新话题的概率，而 $P_{\neg\,\text{target}}$ 是在信息流中看到一个老话题的概率，$P_{\neg\,\text{target}} = 1 - P_{\text{target}}$，$C_{\text{miss}}$ 为漏报一个新事件的代价，C_{fa} 为误报一次的代价.

基于上述提出的评价标准，得到了汉语话题检测基本模型在汉语语料上的评测结果，实验结果如表 2 所示，从表 2 中看到话题检测系统在阈值为 0.40 时取得最好结果.

表 2　汉语话题检测评测结果

阈值	P_{miss}	P_{fa}	$(C_{Det})_{\text{norm}}$
0.25	0.481 6	0.054 1	0.746 7
0.30	0.196 6	0.069 3	0.536 1
0.35	0.119 5	0.044 3	0.336 7
0.40	0.140 4	0.029 0	0.282 3
0.45	0.196 2	0.020 2	0.295 3
0.50	0.207 0	0.017 0	0.290 2
0.55	0.303 9	0.004 1	0.323 8
0.60	0.430 4	0.003 2	0.446 2

采用基于时间信息的动态阈值模型即 $\theta + \alpha$，其中，$\theta = 0.4$，而不是单纯的静态阈值 θ. 对 α 取不同值时的性能进行了验证，结果如表 3 所示.

表 3　时间信息的动态阈值模型实验结果

α	P_{miss}	P_{fa}	$(C_{Det})_{\text{norm}}$
0.001	0.289 3	0.002 1	0.299 7
0.002	0.229 9	0.001 7	0.238 1
0.003	0.246 9	0.001 4	0.253 9
0.004	0.245 1	0.001 1	0.250 5
0.005	0.305 6	0.000 9	0.310 1

表3 可以看出基于时间信息的动态阈值模型取得了很好的效果. 但是，随着 α 值增大，话题的阈值将变得越来越大，结果可能使得许多相关报道不能被检测到. 所以在使用本文提出的动态阈值模型时，α 应该取适当的值.

2.3　基于话题检测的动态多文档文摘

基于不同的 α 值，对 TDT 的聚类输出结果进行了实验. 实验表明，基于话题持续时间的动态阈值模型中，不是所有的阈值输出结果都能生成高质量的多文档文摘. 将 $\alpha = 0.001$ 与 $\alpha = 0.003$ 做了对比实验，图 1 为 $\alpha = 0.001$ 时第 1 组 TDT 输出的同一聚类话题多文档摘要结果，图 2 为 $\alpha = 0.003$ 第 1 组 TDT 输出的同一聚类话题多文档摘要结果.

日本政府不顾中国反对向李登辉发放赴日签证
日本称原则上同意向李登辉核发赴日观光签证
日本政府将向李登辉发签证 要求日媒体不要报道
日本纵容台独势力分裂活动又给李登辉发签证
中新网12月21日电 日本共同社消息，日本政府于21日上午向李登辉发放了入境签证。
李登辉计划九月下旬和家人到日本观光。
日本政府发言人、内阁官房长官细田博之在9月1日首度表态：日本官方将慎重评估判断。
司马辽太郎可说是将李登辉之名传遍日本的关键人物。
日本政府允许李登辉赴日活动，是对"台独"分裂势力的纵容和支持，向"台独"势力再次发出了错误信号，损害了中日关系政治基础。中国政府已通过外交渠道向日方提出严正交涉，强烈要求日本政府立即取消允许李登辉到日本活动的决定。

图 1　第 1 组 TDT0.001 文档摘要结果(307)

日本政府不顾中国反对向李登辉发放赴日签证
李登辉持签证抵达日本 中国政府表示强烈不满
日本称原则上同意向李登辉核发赴日观光签证
日本纵容台独势力分裂活动又给李登辉发签证
但据26日几家日本媒体的报道说，日本外务省的态度还是"面有难色"，对发给签证表示慎重。
李登辉计划九月下旬和家人到日本观光。
刘建超回答说：李登辉是急进"台独"势力的总代表。
日本政府发言人、内阁官房长官细田博之在9月1日首度表态：日本官方将慎重评估判断。
中国政府已通过外交渠道向日方提出严正交涉，强烈要求日本政府立即取消允许李登辉到日本活动的决定。
《朝日新闻》称，日本外务省亚大局局长薮中三十二于15日晚，向中国驻日本大使馆公使程永华通报了此事。

图 2　第 1 组 TDT0.003 文档摘要结果(299)

2.4 实验分析和结论

由于评测语料不同,所以没能将实验结果与国内外同类研究的结果进行比较. 而中文文摘的评测系统尚未建立,所以只能通过与自身的结果大致对比,可以看出这种方法取得了比较理想的效果.

通过直观比较可以看出,表示时间信息在动态阈值中所占的比例的 α 参数变化时,所生成的多文档文摘质量有所差别. 同 TDT 的结果略有区别的地方是,随着 α 值的增大,话题的阈值将变得越来越大,结果可能使得许多相关报道不能被检测到,但是文摘所含的信息量也随着增大,有更强的阅读性和流畅程度,信息覆盖率大. 当 α 值增大到一定程度,文档文摘质量下降. 通过比较生成的多文档文摘,可以看出,时间阈值在等于 0.003 的时候文摘质量最好,当时间信息过少或者过多时,文摘都只能得到一般的结果. 说明在 TDT 的话题检测技术中,可以通过调整 α 值的方法来平衡 TDT 的输出结果,以得到更好质量的多文档文摘,得到对于新闻数据流更全面、覆盖率更高的信息摘要.

3 结 论

1) 基于话题检测技术的动态中文多文档文摘,结合了 TDT 的时序性及动态特性,针对于巨大的网络信息量,能够在庞大的数据流中按照话题聚类.

2) 时间是话题的重要特征,中文文摘技术也正朝着动态发展,动态的形成同一主题文档集合,进而生成多文档文摘. 并应用于大规模数据处理上.

参考文献:

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

[2] ALLAN J, LAVRENKO V, FREY D, *et al.* Umass at TDT 2000 [C]//Proceedings of Topic Detection and Tracking Workshop. USA: National Institute of Stander and Technology, 2000.

[3] WALLS F, JIN H, SISTA S, *et al.* Topic detection in broadcast news[C]//Proceddings of the DARPA Broadcast News Workshop. [S. l.]:The Information Retrieval Series, 2002, 7: 97 - 126.

[4] 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[J]. 计算机研究与发展, 2004, 41(7): 1273 - 1280.

[5] JUHA M, HELENA A M, MARKO S. Applying semantic classes in event detection and tracking[C]//Proceedings of International Conference on Natural Language Processing. Mumbai, India: [s. n.], 2002: 175 - 183.

[6] 刘远超, 王小龙, 刘秉权. 一种改进的 K_means 文档聚类初值选择算法[J]. 高技术通讯, 2006, 16(1): 11 - 15.

[7] 秦 钰, 荆继武, 向 继, 等. 基于优化初始类中心点的 K_means 改进算法[J]. 中国科学院研究生院学报, 2007, 24(6): 771 - 777.

[8] 张其文,李明. 多文档文摘提取方法的研究[J],兰州理工大学学报,2007. 33(1): 96 - 99.

[9] 张姝,赵铁军,郑德权,等. 基于浅层分析的多文档自动文摘技术研究[J]. 哈尔滨工业大学学报,2007, 39(7): 1102 - 1105.

[10] QIN Bing, LIU Ting, ZHANG Yu, *et al.* Research on multi-document summarization based on latent semantic indexing [J]. Journal of Harbin Institute of Technology (New Series). 2005,12(1): 91 - 94.

[11] NENKOVA A, PASSONNEAU R, MCKEOWN K. The pyramid method: incorporating human content selection variation in summarization evaluation[C]//ACM Transactions on Speech and Language Processing (TSLP). New York, NY: ACM, 2007: 4.

[12] 赵华,赵铁军,于浩,等. 基于查询向量的英语话题跟踪研究[J]. 计算机研究与发展,2007, 44(8): 1412 - 1417.

[13] CHEN Y J, CHEN H H. NLP and IR approaches to monolingual and multilingual link detection[C]//Proceedings of the 19th International Conference on Computational Linguistics. Morristown, NJ: Association for Computational Linguistics, 2002: 1 - 7.

[14] RADEV D R, TAM D. Summarization evaluation using relative utility[C]//Proceedings of the Twelfth International Conference on Information and Knowledge Management. New York, NY: ACM, 2003: 508 - 511.

[15] LEVOW G A, OARD D W. Translingual Topic Detection: Applying Lessons from MEI Project[C]//Proceedings of Topic Detection and Tracking Workshop. [S. l.]:[s. n.], 2000.

(编辑 张 红)