软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(10):2366−2378 [doi: 10.3724/SP.J.1001.2013.04382]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向 Web 新闻的事件多要素检索方法*

仲兆满[1], 李存华[1], 刘宗田[2], 戴红伟[1]

[1](淮海工学院 计算机工程学院,江苏 连云港 222005)

[2](上海大学 计算机工程与科学学院,上海 200072)

通讯作者: 仲兆满, E-mail: zhongzhaoman@163.com, http://www.hhit.edu.cn

摘 要: 针对用户获取事件类信息的需求,在分析 Web 新闻特征、事件多要素检索特点的基础上,研究了面向 Web 新闻的事件多要素检索方法.首先,提出了面向 Web 新闻的事件多要素检索模型;然后,使用 BNF(Backus-Naur form)形式化定义了事件多要素查询项;最后,结合事件的动作要素、Web 新闻标题的重要性及事件项与约束项之间的距离,提出了事件查询项与文档相关性的计算方法.设置了 16 个事件多要素查询项,基于 Baidu 搜索引擎对 *P@n* 指标进行了实验分析,所提方法得到的平均 *P*@10 结果为 0.87,平均 *P*@20 结果为 0.83.对 16 个事件查询主题,通过人工标注语料的方法对 *F-measure* 指标进行了实验分析,所提方法得到的平均 *F-measure* 为 0.74.结果表明,所提方法对事件多要素的检索较为有效.

关键词: 事件多要素检索;Web 新闻;事件检索模型;相关性计算

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 仲兆满,李存华,刘宗田,戴红伟.面向 Web 新闻的事件多要素检索方法.软件学报,2013,24(10):2366−2378. http://www.jos.org.cn/1000-9825/4382.htm

英文引用格式: Zhong ZM, Li CH, Liu ZT, Dai HW. Web news oriented event multi-elements retrieval. Ruan Jian Xue Bao/ Journal of Software, 2013,24(10):2366−2378 (in Chinese). http://www.jos.org.cn/1000-9825/4382.htm

Web News Oriented Event Multi-Elements Retrieval

ZHONG Zhao-Man[1], LI Cun-Hua[1], LIU Zong-Tian[2], DAI Hong-Wei[1]

[1](School of Computer, Huaihai Institute of Technology, Lianyungang 222005, China)

[2](School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)

Corresponding author: ZHONG Zhao-Man, E-mail: zhongzhaoman@163.com, http://www.hhit.edu.cn

Abstract: To meet the demand of effectively acquiring event information, a method of Web news-oriented event multi-elements retrieval is studied through analyzing characteristics of Web news and event multi-elements retrieval process. Firstly, a model of Web news-oriented event multi-elements retrieval is proposed. Secondly, event multi-elements query terms are formally defined by using the BNF (Backus-Naur form). Finally, incorporating the importance of event action element, Web news title and the distance between event terms and constrained terms, a method of computing the relevance between query terms and the document is proposed. Sixteen event query topics are created to implement the experiments. With the proposed method, this paper evaluates the index *P@n* based on the Baidu search engine, getting average *P*@10 of 0.85 and average *P*@20 of 0.83. This paper also evaluates the index *F-measure* through manually labeling the corpus with same method, obtaining average *F-measure* of 0.74. The results show that the proposed method offers more effective performances.

Key words: multi-event elements retrieval; Web news; event retrieval model; relevance computing

由于现实中的事件在互联网上都有明显的反映,Web 上存在着大量面向事件的新闻报道.借助搜索引擎从

* 基金项目: 国家自然科学基金(60975033)
收稿时间: 2012-08-21; 修改时间: 2012-10-19; 定稿时间: 2013-02-04

互联网上获取事件信息已经是用户的迫切需求.但是,由于互联网上的信息急剧膨胀,通用搜索引擎返回的结果往往量很大且查询不准确.用户在输入某些关键字后,得到的有用信息并不多,对事件类信息的检索更是如此.

例 1:用户想获取“汶川地震死亡”的相关信息,在 Baidu 搜索引擎中输入关键字“汶川 地震 死亡”检索,点击“新闻”类别,获取的前 10 条信息见表 1(2012 年 8 月 10 执行了此查询).

Table 1 Top ten pieces of information for query term “汶川 地震 死亡” by Baidu

表 1 查询项“汶川 地震 死亡”在百度获取的前 10 条信息

序号	标题	Url
1	裁员潮来袭?	http://www.p5w.net/news/gncj/201208/t4412901.htm
2	地震救援亲历者:北京“玩主”的环球探险	http://news.qq.com/a/20120731/000771.htm
3	汶川大地震死亡数字公布细节	http://news.sohu.com/20120726/n349118371.shtml
4	地震 DNA 检测确认遇难者身份需一两年	http://news.sohu.com/20120726/n349118409.shtml
5	阮次山:北京发布因灾死亡名单是一种进步	http://news.ifeng.com/mainland/special/beijingdayu/content-1/detail_2012_07/28/16373285_0.shtml
6	高邮宝应交界处地震导致 1 死 2 伤	http://nj.house.sina.com.cn/esf/2012-07-22/66421.shtml
7	地震预警系统安徽滁州开建 能为京沪高铁提供 ...	http://www.cet.com.cn/dfpd/bwdqzg/565717.shtml
8	丽江发生 5.7 级地震,教育机构也须“未雨绸缪”	http://jrsc.china.com.cn/content/2012-7/20/40_47700.htm
9	新汶川 新思路 新格局	http://roll.sohu.com/20120716/n348212085.shtml
10	成都开建中国首个城市地震预警系统	http://www.chinadaily.com.cn/dfpd/sc/bwzg/2012-07/13/content_15576453.htm

表 1 中的数据经人工分析后发现,只有 2 条信息与查询项相关,分别是第 2 条、第 3 条信息,查准率仅为 20%.可见,已有的搜索引擎技术对事件信息检索的准确率还是比较低的.在百度新闻搜索时,如果仅仅限制在新闻标题中出现搜索关键字,有些情况下返回的信息非常少,比如查询项“2012 北京洪水死亡”,获取的信息条数为 0.

本文的研究目标是输入事件的多个要素,从 Web 新闻中检索出与事件查询项密切相关的信息,适用于获取事件信息的用户群,包括常规事件以及突发事件等.事件是关联了时间、地点、对象等要素,比概念更大的知识单元.用户输入的事件多要素间不是简单的与/或关系,而是有着紧密的约束关系,比如“汶川 地震”中,地点要素“汶川”约束事件动作要素“地震”.又如“地震 死亡”中,动作要素“地震”约束动作要素“死亡”.Web 文档通常会包含多个事件,有的事件是 Web 文档的核心内容,有的事件只是顺便提及,这需要研究事件查询项与文档相似度的计算方法.本文的研究目标与话题检测与跟踪(TDT)的研究目标有较大的不同,TDT 通常以某事件的几篇新闻报道为跟踪条件,从信息源中将与该事件相关的文档识别出来.近期,一些学者围绕话题的演化做了大量的研究工作,目的是构建出话题的不同子话题之间的信息演化趋势.本文的研究成果一方面是直接获取与用户输入的事件检索项密切相关的文档;另一方面可以被 TDT 所借鉴,用于从话题的不同侧面检索获取信息.比如,想构建地震话题的子话题信息趋势图,可以定时地分别用“地震 死亡”、“地震 视察”、“地震 重建”、“地震 反贪调查”等多个事件查询项检索获取信息,进而构建地震子话题的信息趋势图.

围绕上述研究目标,本文介绍了国内外相关工作的研究状况,分析了 Web 新闻报道的特点,提出了面向 Web 新闻的事件多要素检索模型,研究了事件多要素检索的组合特征,提出了事件查询项与文档相关性的计算方法,并进行了实验分析与比较.

1 相关工作

1.1 事件表示

世界是运动的,运动的世界是由事件组成的.很多认知科学家们认为,“事件”关联了参与者、时间和地点等概念,是比“概念”粒度更大的知识单元,“事件”是人类认识、记忆的基本单位.“事件”这一概念在哲学、认知科学、语言学、人工智能等领域的文献中都有涉及.

在计算机信息学领域,知网(HowNet)中将事件定义为“事情”,并将它分为静态和行动两大类.WordNet 中给出的“事件”定义为“在特定地点和时间发生的事情”.在信息检索领域,“事件”被认为是“细化了的用于检索的主

题”,话题识别与跟踪(TDT)评测会议定义事件为“特定时间特定地点发生的事情”,认为事件是小于话题的概念,多个事件组成一个话题.推动事件提取领域发展的 ACE 评测会议将事件定义为包含参与者的特殊的事情,事件通常可以描述为一种状态的改变.在自动文摘领域,Filatova 等人定义了“原子事件(atomic events)”的概念[1],它是动词(或者动名词)及其连接起来的行为的主要组成部分(如参与者、地点、时间等).Vanderwende 等人[2]使用了逻辑形式三元组的方法,围绕动词描述文本中人物、地点、时间等要素与动词(事件)之间的关系,构建了文本的语义图模型.Zhong 等人[3,4]使用了五元组表示事件 $e=\langle A,S,O,T,L\rangle$,$A$ 表示动作,S 表示主体,O 表示客体,T 表示时间,L 表示地点.

总之,事件的定义都是围绕事件的动作要素展开的,虽然包含的具体要素有所差异,但对象、时间、地点是普遍认可的要素.事件之间的关系是一个多格的结构,各个要素都可以建立事件之间的关联.

1.2 事件检索

事件检索是指针对用户输入的查询事件关键字或者事件问句,获取相关文档或者精准的答案.话题的识别和跟踪(TDT)与事件检索有着一定的联系,其主旨是基于事件对信息流进行组织和利用的研究.其研究方向主要是未知话题的识别及已知话题的跟踪,话题跟踪经常提供若干篇新闻报道为种子,利用相关算法自动地将后续相关新闻报道检测加入到已知话题中.

国内外关于事件检索的研究成果不多,与事件检索密切相关的工作主要有:Metzler 等人[5]提出了微博上的结构事件检索方法,对于一个事件查询,返回的结果是历史事件的摘要排序,主要包括查询扩展及摘要生成两个核心步骤.Steven 等人[6]针对历史事件检索的时间约束,使用了简单的启发式技术支持从 Web 文档中获取事件的时间信息,提出使用模糊时间推理算法改善抽取时间的可靠性.吴平博等人[7]以某事件的几篇报道为检索条件,在聚类的基础上手工地对事件框架的侧面词进行整理,并将事件框架的知识用到事件相关文档的检索中.冯礼等人[8]构建的事件框架既包含了事件的各个侧面,也包含了事件抽取的模式,并将其应用到了事件信息的抽取中.李林等人[9]设计了一个基于领域本体的民航突发事件应急案例语义检索系统,并采用基于 SWRL 规则的推理过程实现了民航突发事件应急案例的语义检索.樊孝忠等人[10]提出一种融合事件信息的复杂问句分析方法,将事件视为由多个要素构成的复杂数据对象,利用事件抽取技术获取复杂问句中若干事件,用多个事件语义模型实例表征整个复杂问句的语义信息.

此外,也有一些面向事件的查询扩展的研究成果.扩展方法一方面基于已有的语义资源,比如本体、事件框架等;另一方面基于局部文档集.Lin 等人[11]于 2005 年提出了一种称为“事件本体”的检索技术,该本体的顶层概念为事件的要素(如地点、时间等),将事件的构成要素作为该本体中的主要分类,在检索时可按事件要素对查询词进行扩展.Zhong 等人[12]提出了基于事件本体的查询扩展方法,在事件本体的基础上研究了事件类各要素、事件类与事件类之间的联想扩展.文献[13]提出了一种基于局部分析面向事件的查询扩展方法,该文重点讨论了面向事件的查询项分析、事件项的扩展以及查询项与文本相似度的计算等问题.Hsu 等人[14]建立了事件结构框架信息映射,将其作为一种知识表示模式,对于某查询对象,在事件结构框架信息映射中会给出它的相关的行为,比如查询“汽车”,会联想到“停车”、“加油”、“维修”等行为.

2 Web 新闻特征分析

2.1 新闻报道

新闻报道是对指定时间、特定地点发生事件的报道,具有很强的时效性.在新闻的写法上,通常要求以规范的语言、正确的句法、合理的修辞手法交代清楚新闻事件的重要信息.这些信息被称为新闻的“六何”,包括 who(何人)、where(何地)、what(何事)、when(何时)、why(何故)和 how(如何).

新闻报道在书写习惯上为了吸引读者,多采用倒叙的方式,首先介绍事件结果以引起读者关注.因此,从新闻的结构上看,新闻的标题和首段往往包含了事件的最重要的信息[15].新闻要告诉人们何时、何地、何人(物)、发生了何事、事件的进展如何.尽管新闻标题只有短短的一句话,却是对新闻内容的浓缩和概括.从总体上看,新

闻标题的揭示和阐明功能,使绝大部分标题都能精确地描述何时、何人(或物)、何地、发生了何事这几个方面的内容.

2.2 Web新闻的特点

Web 页面的 HTML 标签包含了很多有价值的信息,有效地利用这些标签信息,有助于提高 Web 文档过滤、检索的准确率[16,17].HTML 标签中,⟨Title⟩,⟨Bold⟩,⟨Header⟩,⟨Italic⟩,⟨Strong⟩,⟨em⟩等标签具有突出页面内容的作用.

此外,Web 新闻通常都包含 META 标签,即 Meta Description,是页面的描述性语句,作用是为了让搜索引擎清楚地了解页面的主要内容.META 标签中,最重要的是标题标签(⟨Title⟩)、关键词标签(⟨Keyword⟩)和描述标签(⟨Description⟩).

⟨Title⟩标签标识文章的标题,⟨Keywords⟩标签包含了概述页面内容的若干关键字,⟨Description⟩标签是对页面内容的概括,相当于页面的摘要.

戴伊克(Van Dijk)在著作《作为话语的新闻》中提出的假设性话语结构图[18],包括了新闻报道的概述及故事情节等内容,而新闻的概述经常体现在 Web 新闻的标题、关键字及描述中,故事情节基本上囊括了事件的信息,而 Web 新闻正文的首段经常是对核心事件的起因、过程或结果的简介.

3 面向 Web 新闻的事件多要素检索方法

3.1 面向Web新闻的事件多要素检索模型

事件多要素检索模型涉及的相关概念解释如下:

定义 1(事件). 事件是指在某个特定的时间和地点发生的、由若干对象参与的、表现出若干动作特征的一件事情.事件的表示模型是多样的,根据事件检索关注的要素特点,本文使用的事件模型是四元组结构:$e=\{t,l,o,a\}$,t 表示时间(何时),l 表示地点(何地),o 表示对象(何人,也可以是其他参与对象,比如“组织机构”),a 表示动作(何事).

新闻报道“六何”提及的“何故”、“如何”两要素在本文中认为是另外一个事件,它们之间是因果、时序或者并发的关系.

定义 2(事件检索). 事件检索是指将输入事件的若干要素作为查询条件,获取与事件查询要素相关的信息.对事件检索而言,事件的时间、地点、对象及动作要素中,动作要素是不可缺少的,否则就不能表征事件,其他要素根据不同的事件类型组合出现.

定义 3(事件项). 在事件多要素查询项中,动作要素能够表征待查询事件的类别,称为事件项.动作要素又称为事件触发词,是事件最为核心的要素.

定义 4(约束项). 在事件多要素查询项中,时间、地点、对象要素用来约束查询事件项的范围,有时动作要素也可以约束查询事件项的范围,这些要素称为约束项.

面向 Web 新闻的事件多要素检索的模型定义为一个四元组:$(D,Q,F,R(Q,d_i))$,其中,D 是 Web 新闻的集合,d_i 是 D 的某一篇文档,Q 是查询事件要素的集合,F 是查询项及文档的表示框架,$R(Q,d_i)$是查询项与文档相关性的计算方法.

在表示 Web 文档时,考虑到 Web 新闻的特点、新闻特征的维数、标签的易提取性,选取的 Web 新闻特征主要包括标题 T(⟨Title⟩)、关键字 K(⟨Keywords⟩)、描述 D(⟨Description⟩)和首段 F(⟨Firstparagraph⟩)这 4 类特征.因此,一篇 Web 新闻可以表示为 $d_i=\{T,K,D,F\}$.

事件多要素检索的查询项表示为 $Q=\{Q_e,Q_c\}$,其中,Q_e 是事件项,Q_c 是约束项.$Q_e=\{e_1,e_2,\ldots,e_m\}$,一般的查询项都包含一个事件项,即 $m=1$;$Q_c=\{t,l,o,e_1,e_2,\ldots,e_n\}$,约束项可以是 t,l 或者 o,也可以是其他事件 e,一般的查询项都包含 1~2 个事件约束项.比如,查询项 Q={“2008 汶川地震死亡”},则 Q_c={“2008”,“汶川”,“地震”},Q_e={“死亡”}.其中,“2008”是时间约束项,“汶川”是地点约束项,“地震”是事件约束项,“死亡”是查询事件项.用户关注的是“2008 汶川地震”事件中的“死亡”事件.

F 可以采用布尔、概率、向量空间等模型来表示 Web 文档及查询项.

$R(Q,d_i)$的计算依赖于所使用的文档表示模型.

面向 Web 新闻的事件多要素检索模型如图 1 所示.

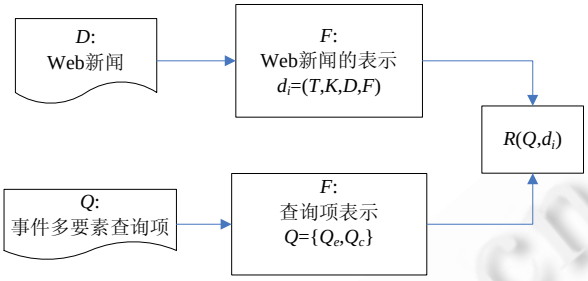

Fig.1 Model of event multi-elements retrieval towards Web news

图 1 面向 Web 新闻的事件多要素检索的模型

3.2 事件多要素检索的组合分析

事件的各个要素在查询项中的作用是不同的,有着很强的约束关系.比如查询项:“8.10 周克华抢劫”,其中,“抢劫”是事件动作要素,标识所要检索的事件类型,称为事件项;“8.10”和“周克华”分别是事件的时间和对象要素,约束查询事件的范围,称为事件的约束项.事件的约束项除了由事件要素充当以外,还有可能由其他事件充当.比如查询项“地震死亡”,用户关心的是“地震”中的“死亡”事件,“地震”充当了约束项.

下面,采用巴科斯范式(BNF 范式)形式化定义事件查询项.

事件查询项::=〈约束项〉〈事件项〉

〈约束项〉::=〈时间〉{〈动作〉}|〈地点〉{〈动作〉}|〈对象〉{〈动作〉}|〈时间〉〈地点〉{〈动作〉}|〈时间〉〈对象〉{〈动作〉}|〈地点〉〈对象〉{〈动作〉}|〈时间〉〈地点〉〈对象〉{〈动作〉}

〈事件项〉::=〈动作〉

〈时间〉::=时间表示格式

〈地点〉::=地点表示格式

〈动作〉::=动作表示格式

〈对象〉::=对象表示格式

约束项中的动作可以重复出现(一般重复 1~2 次),比如查询项“3.10 交通事故死亡赔偿”,“3.10”是时间约束项,“交通事故”和“死亡”都是动作约束项,动作重复了 2 次,“赔偿”是事件项.

对于不同的事件,用户关注的事件要素也不相同.比如,自然灾害、军事冲突、金融这类事件对于时间、地点要素比较敏感,而有关选举、犯罪、科学研究等事件对于时间、人名比较敏感.

根据约束项的组合对事件项的约束作用,从查询项包含的信息量的角度将查询项划分为 3 个级别:

假设某类事件的信息总量为 sum,

级别 1——查询项能明确表明具体事件,不会与其他事件混淆,获取的信息量为 sum_1;

级别 2——模糊的表明具体事件,存在一些混淆事件,获取的信息量为 sum_2;

级别 3——不能明确表明具体事件,存在大量混淆事件,获取的信息量为 sum_3.

对 3 个级别而言,$sum_1<sum_2<sum_2<sum$.级别 1 是最理想的状态,查询项能明确表明某一具体事件;级别 2 虽然是模糊的表明事件,但很多情况下用户是经常使用的,在事件检索时,用户很少会把一个事件涉及的所有要素全部输入;级别 3 包含的混淆事件太多,实际检索的价值不大.

事件多要素检索的组合分析见表 2,数字“1”表示存在该项,数字“0”表示不存在该项.

Table 2 Combinatory analysis of event multi-elements retrieval

表 2 事件多要素检索的组合分析

事件项	约束项				级别	例子	解释
动作	动作	时间	地点	对象			
1	0	0	0	0	3	持枪抢劫	Q_e={持枪抢劫},Q_c={}
1	0	0	0	1	3	周克华持枪抢劫	Q_e={持枪抢劫},Q_c={周克华}.不能确定是哪一次抢劫
1	0	0	1	0	3	重庆持枪抢劫	Q_e={持枪抢劫},Q_c={重庆}.不能确定是哪一次抢劫
1	0	0	1	1	2	周克华重庆持枪抢劫	Q_e={持枪抢劫},Q_c={周克华,重庆}.模糊表明具体事件,但可能同一对象于不同时间在同一地点发生多次同类事件
1	0	1	0	0	2	2012 年 8 月 10 日 持枪抢劫	Q_e={持枪抢劫},Q_c={2012 年 8 月 10 日}.模糊表明具体事件,但可能于同一时间在多地发生同类事件
1	0	1	0	1	1	2012 年 8 月 10 日 周克华持枪抢劫	Q_e={持枪抢劫}, Q_c={2012 年 8 月 10 日,周克华}
1	0	1	1	0	1	2012 年 8 月 10 日 重庆持枪抢劫	Q_e={持枪抢劫}, Q_c={2012 年 8 月 10 日,周克华}
1	0	1	1	1	1	2012 年 8 月 10 日 周克华重庆持枪抢劫	Q_e={持枪抢劫}, Q_c={2012 年 8 月 10 日,周克华,重庆}
1	1	0	0	0	3	持枪抢劫伤亡	Q_e={伤亡},Q_c={持枪抢劫}
1	1	0	0	1	3	周克华持枪抢劫伤亡	Q_e={伤亡},Q_c={周克华,持枪抢劫}
1	1	0	1	0	3	重庆持枪抢劫伤亡	Q_e={伤亡},Q_c={重庆,持枪抢劫}.时间不确定
1	1	0	1	1	2	周克华重庆 持枪抢劫伤亡	Q_e={伤亡}, Q_c={周克华,重庆,持枪抢劫}
1	1	1	0	0	2	2012 年 8 月 10 日 持枪抢劫伤亡	Q_e={伤亡}, Q_c={2012 年 8 月 10 日,持枪抢劫}
1	1	1	0	1	1	2012 年 8 月 10 日 周克华持枪抢劫伤亡	Q_e={伤亡}, Q_c={2012 年 8 月 10 日,周克华,持枪抢劫}
1	1	1	1	0	1	2012 年 8 月 10 日 重庆持枪抢劫伤亡	Q_e={伤亡}, Q_c={2012 年 8 月 10 日,重庆,持枪抢劫}
1	1	1	1	1	1	2012 年 8 月 10 日 周克华重庆持枪抢劫伤亡	Q_e={伤亡}, Q_c={2012 年 8 月 10 日,周克华,重庆,持枪抢劫}

由于对象、地点、时间、动作等要素都有一定的粒度,在刻画不同的事件时体现的粒度也不相同.比如,“2008 年中国地震”就不能表征一个具体的事件,而“2008 年汶川地震”能够表征一个具体的事件,因为地点“汶川”比“中国”的粒度更细.但“2008 中国奥运会”就能表征具体的事件,地点要素“中国”已经不需要细化.同样地,“主席出生”也不能表征一个具体的事件,而“毛泽东出生”能够表明一个具体的事件,因为对象“毛泽东”比“主席”的粒度更细,已经具体到了一个人.本文在对事件诸要素的约束讨论中,假设对象、地点、时间、动作等要素的粒度比较细,已经能够明确地表明确切的对象、地点、时间及动作要素,不会引起歧义.结合新闻学的规律及对表 2 的分析后可见:

(1) 对事件的检索,事件项是不可缺少的,即 Q_e 不能为空集,否则不能表明是事件检索;但如果 Q_c 为空,查询项就代表了某一类事件,事件检索的意义不大,比如,“持枪抢劫”范围太广;

(2) 对约束项 Q_c 而言,在时间要素 t 确定的前提下,地点要素 l、对象要素 o 中任意一个要素出现,即能明确地表征具体的事件,因为在同一时间、同一地点很少会发生两件同样的事件,在同一时间,同一对象也很少会参与两件相同的事件;

(3) 对约束项 Q_c 而言,在时间要素 t 确定的前提下,无其他要素的出现,能够模糊地表征具体的事件,即仅在时间上约束了查询事件的范围.比如,“2008 年春运”涉及的是 2008 年全球的与春运有关的事件;

(4) 对约束项 Q_c 而言,在无时间要素 t 的前提下,地点要素 l、对象要素 o 组合出现,能够模糊地表征具体的事件,因为不同的时间在不同的地点可能发生同样的事件;不同的时间,不同的对象也可能参与同样的事件;但不同的时间,同一对象于同一地点参与同样事件的概率虽然有,但相对小些;

(5) 3 个约束项 t,l,o 都出现的情况下,必能明确地表征具体事件,即同一时间、同一地点、同一对象只能参与一件事情.

3.3 事件要素与文档相关性的计算

3.3.1 事件查询项及 Web 文档的表示

在几类文档表示模型中,向量空间模型使用的最广,且在几个大型的系统中都得到验证,比如著名的 SMART 文本检索系统.本文使用了向量空间模型表示查询项及 Web 文档.

Web 文档 d_i={*T*,*K*,*D*,*F*}的各个部分特征项的权重取词的频度 *TF*(term frequency);查询项 Q={Q_e,Q_c}中,各个项的权值设为 1.

假设查询项 Q 的事件项 Q_e={a_2},约束项 Q_c={*t*,*l*,*o*,a_1},则 Web 文档 d_i 可以表示为公式(1):

$$d_i=\begin{cases}T=\{\langle t^T,w_t^T\rangle,\langle l^T,w_l^T\rangle,\langle o^T,w_o^T\rangle,\langle a_1^T,w_{a_1}^T\rangle,\langle a_2^T,w_{a_2}^T\rangle,...\}\\ K=\{\langle t^K,w_t^K\rangle,\langle l^K,w_l^K\rangle,\langle o^K,w_o^K\rangle,\langle a_1^K,w_{a_1}^K\rangle,\langle a_2^K,w_{a_2}^K\rangle,...\}\\ D=\{\langle t^D,w_t^D\rangle,\langle l^D,w_l^D\rangle,\langle o^D,w_o^D\rangle,\langle a_1^D,w_{a_1}^D\rangle,\langle a_2^D,w_{a_2}^D\rangle,...\}\\ F=\{\langle t^F,w_t^F\rangle,\langle l^F,w_l^F\rangle,\langle o^F,w_o^F\rangle,\langle a_1^F,w_{a_1}^F\rangle,\langle a_2^F,w_{a_2}^F\rangle,...\}\end{cases} \tag{1}$$

公式(1)中,省略号表示文档 d_i 中的其他项,即非事件查询项.

查询项 Q 可以表示为公式(2):

$$Q=\{\langle t,1\rangle,\langle l,1\rangle,\langle o,1\rangle,\langle a_1,1\rangle,\langle a_2,1\rangle\} \tag{2}$$

3.3.2 特征项权值的调整

在事件查询项 Q={Q_e,Q_c}中:动作要素是必须存在的,否则无法表征具体的查询事件;其他几个要素因不同的事件而有不同的差异.给动作要素乘以权重系数λ(λ>1),得到公式(3):

$$d_i=\begin{cases}T=\{\langle t^T,w_t^T\rangle,\langle l^T,w_l^T\rangle,\langle o^T,w_o^T\rangle,\langle a_1^T,w_{a_1}^T\times\lambda\rangle,\langle a_2^T,w_{a_2}^T\times\lambda\rangle,...\}\\ K=\{\langle t^K,w_t^K\rangle,\langle l^K,w_l^K\rangle,\langle o^K,w_o^K\rangle,\langle a_1^K,w_{a_1}^K\times\lambda\rangle,\langle a_2^K,w_{a_2}^K\times\lambda\rangle,...\}\\ D=\{\langle t^D,w_t^D\rangle,\langle l^D,w_l^D\rangle,\langle o^D,w_o^D\rangle,\langle a_1^D,w_{a_1}^D\times\lambda\rangle,\langle a_2^D,w_{a_2}^D\times\lambda\rangle,...\}\\ F=\{\langle t^F,w_t^F\rangle,\langle l^F,w_l^F\rangle,\langle o^F,w_o^F\rangle,\langle a_1^F,w_{a_1}^F\times\lambda\rangle,\langle a_2^F,w_{a_2}^F\times\lambda\rangle,...\}\end{cases} \tag{3}$$

在新闻报道的 *T*,*K*,*D*,*F* 这 4 项中,标题 T 中的特征词更为重要,其他项中的特征词的重要性可以认为相同.给标题 T 中出现的特征词乘以权重系数α(α>1),得到公式(4):

$$d_i=\begin{cases}T=\{\langle t^T,w_t^T\times\alpha\rangle,\langle l^T,w_l^T\times\alpha\rangle,\langle o^T,w_o^T\times\alpha\rangle,\langle a_1^T,w_{a_1}^T\times\lambda\times\alpha\rangle,\langle a_2^T,w_{a_2}^T\times\lambda\times\alpha\rangle,...\}\\ K=\{\langle t^K,w_t^K\rangle,\langle l^K,w_l^K\rangle,\langle o^K,w_o^K\rangle,\langle a_1^K,w_{a_1}^K\times\lambda\rangle,\langle a_2^K,w_{a_2}^K\times\lambda\rangle,...\}\\ D=\{\langle t^D,w_t^D\rangle,\langle l^D,w_l^D\rangle,\langle o^D,w_o^D\rangle,\langle a_1^D,w_{a_1}^D\times\lambda\rangle,\langle a_2^D,w_{a_2}^D\times\lambda\rangle,...\}\\ F=\{\langle t^F,w_t^F\rangle,\langle l^F,w_l^F\rangle,\langle o^F,w_o^F\rangle,\langle a_1^F,w_{a_1}^F\times\lambda\rangle,\langle a_2^F,w_{a_2}^F\times\lambda\rangle,...\}\end{cases} \tag{4}$$

3.3.3 事件查询项间距离的计算

在新闻写作中,为交代清楚一件事情,在文本中较近的距离内,围绕事件动作要素往往需要介绍事件的其他要素,比如时间、地点及对象等.可见,距离事件动作要素间的距离越近,各个事件要素之间的关联就越密切,越有可能在表述同一件事情.相应地,文档的相关度与检索项就越高.

下面以 Web 文档 d_i 的 F 项为例,介绍查询项中约束项与事件项的距离的计算方法.将 F 分词后,Q_c={*t*,*l*,*o*,a_1}和 Q_e={a_2}在 F 中出现的位置依次记为 *pos*(*t*,*F*),*pos*(*l*,*F*),*pos*(*o*,*F*),*pos*(a_1,*F*)和 *pos*(a_2,*F*).

- *t* 与 a_1 之间的距离记为 *dis*(*t*,a_1)=|*pos*(*t*,*F*)−*pos*(a_1,*F*)|;
- *t*,*l*,*o* 与 a_1 的距离之和为 $\sum_{x\in\{t,l,o\}}|pos(x,F)-pos(a_1,F)|$;
- a_1 与 a_2 的距离为 *dis*(a_1,a_2)=|*pos*(a_1,*F*)−*pos*(a_2,*F*)|.

那么,F 中 *t*,*l*,*o*,a_1 与 a_2 的距离之和见公式(5):

$$Dis(F)=\sum_{x\in\{t,l,o\}}|pos(x,F)-pos(a_1,F)|+|pos(a_1,F)-pos(a_2,F)| \tag{5}$$

例 2:设 F 中的文本为:“8 月 10 日重庆周克华持枪抢劫造成了 3 人伤亡”,查询项为 Q=“8 月 10 日重庆持枪

抢劫伤亡”,Q_e={“伤亡”},Q_c={“8月10日”,“重庆”,“持枪抢劫”},则查询项在文本中的距离计算步骤如下:

(1) 对F使用ICTCLAS分词得到:F_1=“8月/t 10日/t 重庆/ns 周克华/nr 持枪/vd 抢劫/v 造成/v 了/u 3/m 人/m 伤亡/vn”;

(2) 依据查询项对分词结果F_1的部分特征词进行合并,得到F_2=“8月10日/t 重庆/ns 周克华/nr 持枪抢劫/v 造成/v 了/u 3/m 人/m 伤亡/vn”;

(3) 统计事件查询项在F_2中的位置,$pos(t,F)=1$,$pos(l,F)=2$,$pos(a_1,F)=4$,$pos(a_2,F)=9$;

(4) 计算查询项在F中的距离,$dis(t,a_1)=|1-4|=3$,$dis(l,a_1)=|2-4|=2$,$dis(a_1,a_2)=|4-9|=5$,则$Dis(F)=3+2+5=10$.

如果同一个事件动作要素a在文本中多次出现,其他要素与其距离取最小值.比如,同一事件动作要素a_i,a_j在文本中出现了2次,一个对象约束要素o在文本中出现了1次,位置依次记为$pos(o,F),pos(a_i,F),pos(a_j,F)$,则$o$与$a$的距离为$dis(o,a)=\min\{dis(o,a_i),dis(o,a_j)\}$.这种情况在Web文档的$T$中一般不会发生,大多出现在$K,D,F$中.

3.3.4 事件查询项与文档间相似性的计算

查询项与文档之间的相似度计算采用经典的余弦向量度量法,同时考虑到相似度与事件要素之间的距离成反比,最后得到查询项Q与文档d_i的在F上的相似度计算方法如公式(6)所示:

$$R(Q,d_i^F)=\frac{\sum_{x\in Q\cap d_i^F}W(x|Q)W(x|d_i^F)}{\sqrt{\sum_{x\in Q}W(x|Q)^2\sum_{x\in d_i^F}W(x|d_i^F)^2}}\times\frac{1}{\log_2 Dis(F)} \tag{6}$$

$Dis(F)$取对数是为了减少距离对相似度平滑性的影响.查询项Q与文档d_i在T,K,D上的相似度计算方法类似于公式(6).

综合考虑文档d_i的T,K,D,F这4个特征项后,一个查询项Q与一篇Web文档d_i的相似度计算方法如公式(7)所示:

$$R(Q,d_i)=R(Q,d_i^T)+R(Q,d_i^K)+R(Q,d_i^D)+R(Q,d_i^F) \tag{7}$$

4 实验设计与结果分析

4.1 实验流程

本文进行实验时的流程如下:

(1) 制定查询事件要素,搜集实验的Web文档,制定评价指标;

(2) 分析查询项中的事件项和约束项;

(3) 查询项及文档中的时间要素的规范化;

(4) 查询项与文档相似度的计算;

(5) 按照相关度排序输出检索结果.

步骤(1):本文参考突发事件的分类及ACE2005标注的事件制定了16个事件查询项,这些查询项基本覆盖了突发事件及ACE2005标注事件的大类别,每个查询项由限定项和事件项组成,详见表3.文献[19]论述的突发事件的分类体系包括3个层次:一级4个大类,二级33个子类,三级94个小类.这个分类体系的第1层分类如下:自然灾害类、事故灾难类、公共卫生事件及社会安全事件.ACE2005对633篇语料涉及的8类事件进行了标注,分别是生命、运动、交易、商务、冲突、联系、人事、司法类别,标注的事件以常规事件为主.两者有一定的重合,比如,突发事件的社会安全事件也属于ACE2005的冲突事件.

实验语料分为两类:1) 一类是基于百度搜索引擎,输入各个查询事件,选百度新闻类别后,对每个查询事件项取返回的20 000条信息作为实验语料,如果返回的信息少于20 000条则取实际返回的信息条数.每条获取的信息包括标题、关键字、描述及正文这4个部分.标题、关键字及描述部分从Html文件中根据标签容易获取,正文的获取采用了课题组在文献[20]中介绍的方法,该方法已经在舆情、企业情报等多个系统中得到应用.对此实验语料,使用的评价指标是$P@n$,$P@n$指标模拟了常用搜索引擎返回的结果,是一个拟人化的指标,目前的搜

索评测中用的较多.*P@n* 指标只关心检索到的结果与查询项是否相关,不考虑返回的文本与查询项相关性的次序,评测起来容易实现.由于很多用户对搜索引擎返回的结果只关注前几页(假设每一页包含 10 条信息),故本文选用了 *P*@10 和 *P*@20 两个指标由人工对返回的结果进行评判;2) 第 2 类语料是基于百度、必应、即刻等搜索引擎,输入各个查询事件,选新闻(或资讯)类别后,对返回的信息人工挑选出约 200 篇相关的信息(每个事件查询项对应的信息条数见表 5),对 16 个查询项而言,将获取的 3 174 篇信息混杂在一起作为实验语料.对此实验语料,使用的评价指标是 *F-measure*,既考虑到信息的查准率 Precision(记为 *P*),又考虑到信息的查全率 Recall (记为 *R*),*F-measure*=(2×*P*×*R*)/(*P*+*R*).

Table 3 Sixteen event query terms

表 3 16 个事件查询项

编号	约束项				事件项
	时间	地点	对象	动作	动作
1		汶川		地震	死亡
2		汶川		地震	反贪调查
3	2012		马英九		竞选
4	911			恐怖袭击	救援
5		重庆	周克华	持枪抢劫	伤亡
6		温州		动车事故	视察
7			三鹿	奶粉污染	问责
8	2010	钓鱼岛	中日		撞船
9	2012	北京		洪水	死亡
10	2012	北京			洪水
11			毛泽东	出生	
12	2008			春运	
13		河南		转账门	
14			三鹿	破产	
15			叙利亚 土耳其	冲突	
16			萨达姆	审判	

步骤(2):本文采用了在指定框中输入相关检索要素的方法,类似于百度、谷歌的高级搜索功能,没有采用自动的事件项和约束项的分析方法.

步骤(3):新闻文本中的时间相对于人名、机构名、地点来说是比较规范的,可以采用规则方法来对它们进行识别.采用规则的方法识别实体的过程类似于使用正则表达式对字符串的匹配过程.Web 文档中的时间可以分为绝对时间和相对时间两种:绝对时间是指明确地说明事件发生的具体时间,比如“2012 年 8 月 10 日”、“8 月 10 日”、“8.10”、“8-10”等,在报道中的表现格式比较单一;而相对时间是指文档中出现的时间是相对于其他事件发生的时间或者新闻报道发表的时间,比如“两天后”、“地震发生之前”、“昨日”、“截止报道发表时”等等,表现的格式较为复杂.在本文中,仅仅使用了事件的绝对时间,没有处理相对时间.识别出时间后,统一规范化为“*年*月*日*时*分*秒”的形式.

步骤(4):使用第 3.3 节介绍的方法计算查询项与文档的相似度.

步骤(5):按照相似度对返回的文档降序排序,输出检索结果.

4.2 *P@n*指标的实验结果

本文使用了两种方法对 *P@n* 指标进行实验比较:

- 一是输入事件查询项,直接获取 Baidu 返回的结果,该方法记作 M_1;
- 二是使用本文提出的方法对 Baidu 返回的结果重新计算事件查询与文档的相似度,然后降序排序输出结果,该方法记作 M_2.

经多次实验后,事件动作要素权值调整参数在[2~3]范围内对结果影响不大,本文中λ=2.文档标题权值调整参数参考的是经验值,文献[7]使用的标题权重参数为 2,文献[21]的标题权重参数为 3,本文中α=2.5.

对表 3 设计的 16 个事件查询项,得到的最终结果见表 4.

Table 4 *P@n* results for sixteen event query terms

表 4 16 个事件查询项的 *P@n* 结果

事件查询项	M_1		M_2	
	P@10	*P*@20	*P*@10	*P*@20
1	0.2	0.1	0.8	0.8
2	0.3	0.2	0.9	0.7
3	0.4	0.35	0.8	0.8
4	0.3	0.2	1.0	0.9
5	0.3	0.25	0.9	0.9
6	0.2	0.2	0.8	0.85
7	0.4	0.4	0.7	0.8
8	0.2	0.2	0.7	0.7
9	0.1	0.1	0.9	0.85
10	0.3	0.25	0.8	0.8
11	0.4	0.25	0.8	0.8
12	0.3	0.35	0.9	0.85
13	0.4	0.4	0.9	0.9
14	0.3	0.25	0.8	0.75
15	0.4	0.35	1.0	0.95
16	0.4	0.35	0.9	0.85
平均值	0.31	0.26	0.85	0.83

由表 4 可见,对于 16 个事件查询项,方法 M_1 得到的平均 *P*@10=0.31,平均 *P*@20=0.26,获取的检索结果是非常不理想的,不能满足用户获取事件类信息的需求.方法 M_2 得到的平均 *P*@10=0.85,平均 *P*@20=0.83,获取的检索结果比较理想.方法 M_1 与方法 M_2 比较,信息检索的准确率有了大幅度提升,平均 *P*@10 指标提高了 0.54,平均 *P*@20 指标提高了 0.57.

对 16 个事件查询项,不带约束动作要素的有 9 项,分别是第 3 号、第 8 号、第 10 号~第 16 号查询项,使用方法 M_1 得到这 9 项的平均 *P*@10=0.34,平均 *P*@20=0.31.带约束动作要素的有 7 项,分别是第 1 号、第 2 号、第 4 号~第 7 号、第 9 号查询项,使用方法 M_1 得到的这 7 项的平均 *P*@10=0.24,平均 *P*@20=0.19.可见,不带约束动作查询项的效果好于带约束动作的查询项,主要原因是带了约束动作的查询项之间的关系更为复杂,并不是简单的关键字的罗列.

对检索结果的分析还可以发现,一些 Web 文档中对于地点、对象等要素的表述经常是多样的.比如,“汶川地震”可以表述为“北川地震”、“四川地震”、“5.12 大地震”、“映秀地震”;又如,“救援人员”可以表述为“救援队”、“医生”、“护士”、“医护人员”、“救援团队”、“消防队”、“消防人员”等.缺乏对这些要素的深入理解,也影响了本文的实验结果.

新闻的标题虽然具有明显的指示作用,但标题经常使用隐喻、引用、对偶、夸张等修辞手法以引起人们的兴趣,这对查询项与文档相似性的计算有很大的影响,比如标题“小巴热闹大把寂寥”、“摩托罗拉发‘最后通牒’”、“唐王大桥下巨野河河水变红”、“河水变成‘黄汤汤’”等.过分强调标题的重要性也会影响信息的查全率.

此外,Web 文档的⟨Keywords⟩部分虽然是关键字的列表,但很多情况下都是短语的形式,为了对⟨Keywords⟩部分进行理解分析,同样需要进行分词及词性标注,比如关键词为“事故调查”、“严重损坏”、“重庆警察”、“810 持枪抢劫”等.

4.3 *F-measure*指标的实验结果

对于 *F-measure* 指标,本文使用了 3 种方法进行实验比较:

- 一是采用经典的文本相似度计算方法,将新闻看作是一个整体,不区分标题、关键字、首段等信息,对文本分词后计算特征词的 *TF*IDF* 作为权值,查询项的权值统一设为 1,使用余弦向量计算查询项与文本间的相似度,该方法记作 S_1;
- 二是区分新闻的标题、关键字、首段信息,对文本分词后计算特征词的 *TF*IDF* 作为权值,查询项的权值统一设为 1,使用余弦向量计算文本间的相似度,该方法记作 S_2;
- 三是使用本文提出的查询事件分析、文本特征权值计算、查询项与文本的相似度计算方法,该方法记

作 S_3.

事件动作要素权值调整参数及文档标题权值调整参数与第 4.2 节的取值相同.

每个事件查询项,按照相似度的大小降序排序,取前 200 篇信息作为返回结果的评价集.对表 3 设计的 16 个事件查询项,得到的 *F-measure* 结果见表 5.

Table 5 *F-Measure* results for sixteen event query terms

表 5 16 个事件查询项的 *F-measure* 结果

事件查询项	信息数(3 174 篇)	S_1			S_2			S_3		
		P	*R*	*F-measure*	*P*	*R*	*F-measure*	*P*	*R*	*F-measure*
1	198	0.46	0.46	0.46	0.54	0.55	0.54	0.73	0.74	0.73
2	201	0.47	0.47	0.47	0.50	0.49	0.49	0.68	0.67	0.67
3	210	0.50	0.47	0.48	0.53	0.50	0.51	0.58	0.55	0.57
4	204	0.40	0.39	0.40	0.50	0.49	0.49	0.88	0.86	0.87
5	196	0.42	0.43	0.42	0.48	0.49	0.48	0.67	0.68	0.67
6	200	0.49	0.49	0.49	0.50	0.50	0.50	0.84	0.84	0.84
7	187	0.51	0.54	0.52	0.65	0.69	0.67	0.89	0.95	0.92
8	200	0.50	0.50	0.50	0.65	0.65	0.65	0.63	0.63	0.63
9	208	0.48	0.46	0.47	0.66	0.63	0.65	0.78	0.75	0.76
10	190	0.47	0.49	0.48	0.56	0.59	0.57	0.79	0.83	0.81
11	192	0.47	0.48	0.47	0.61	0.63	0.62	0.69	0.71	0.70
12	193	0.49	0.50	0.49	0.55	0.56	0.55	0.70	0.72	0.71
13	198	0.49	0.49	0.49	0.55	0.55	0.55	0.70	0.71	0.70
14	200	0.52	0.52	0.52	0.58	0.58	0.58	0.81	0.81	0.81
15	202	0.50	0.49	0.49	0.60	0.59	0.59	0.77	0.76	0.77
16	195	0.47	0.48	0.47	0.58	0.59	0.59	0.73	0.75	0.74
平均值		0.48	0.48	0.48	0.56	0.57	0.56	0.74	0.75	0.74

由表 5 可见,3 种方法中,S_3 的 *F-measure* 为 0.74,比方法 S_1 提高了 0.26,比方法 S_2 提高了 0.18,方法 S_3 相比方法 S_1 和 S_2,有了较大幅度的提升.方法 S_2 与方法 S_1 相比,*F-measure* 也有一定的提升,提高了 0.08,这说明对 Web 信息的检索,利用标题、关键字、描述及首段的内容已经较好地描述了 Web 文档,考虑新闻的全文不仅信息量大,而且存在干扰特征,反而会降低信息的查全率及查准率.方法 S_3 与方法 S_2 相比,考虑事件不同要素的特点,增加了特征项的权值调整,考虑了事件多要素之间的距离约束,*F-measure* 才会有了较大幅度的提升.

5 结束语

本文提出了一种面向 Web 新闻的事件多要素检索方法,该方法一方面结合了 Web 文档的特征,另一方面使用了事件检索的组合分析技术,将两者结合取得了较好的实验结果.本文的主要贡献有:

(1) 提出了一种面向 Web 新闻的事件多要素检索模型,该模型对事件类信息的检索有重要的参考价值;

(2) 分析了事件检索时多要素组合的特征,理清了多要素之间的关系,与传统的不区分查询项的作用、采用等同视之的信息检索观点有较大的不同;

(3) 提出了事件查询项与 Web 文档相似度的计算方法,该方法使用了 Web 文档的结构特征,结合新闻写作的特点提升了事件动作要素、文档标题等特征项的权重,考虑了事件要素之间的距离对相似度计算的影响.

在研究中发现,以下内容还值得进一步探讨:

(1) 事件多要素检索项的自动分析,最好能对用户任意输入的短语、句子或者是问句都能达到自动理解分析的程度;

(2) 对文档中的相对时间的规范化处理需要进一步加强,研究新的处理策略;

(3) 为提高信息检索的效率,面向事件的 Web 新闻索引技术有待进一步研究;

(4) 基于语义资源对事件的动作、时间、地点、对象等要素进行深入的语义分析,这将有助于进一步提高系统的查全率;

(5) 由于语言的差异,本文提出的一些思想可以借鉴应用到其他语言的处理上,但对于时间、地点、对象

等要素的处理还需根据不同的语言采取不同的策略.

致谢 在此,我们向对本文提出中肯修改建议的审稿人表示感谢.

References:

[1] Filatova E, Hatzivassiloglou V. Domain-Independent detection, extraction, and labeling of atomic events. In: Proc. of the RANLP 2003. Borovetz: RANLP Organising Committee, 2003. 145−152. http://www.cs.columbia.edu/nlp/papers/2003/filatova_hatzivassiloglou_03.pdf

[2] Vanderwende L, Banko M, Menezes A. Event-Centric summary generation. In: Proc. of the DUC 2004 Workshop. Boston: NIST Press, 2004. 127−132. http://duc.nist.gov/pubs/2004papers/microsoft.banko.pdf

[3] Zhong ZM, Liu ZT, Liu W, Guan Y, Shan JF. Event ontology and its evaluation. Journal of Information and Computational Science, 2010, 7(1):95−101.

[4] Zhong ZM, Liu ZT, Li CH, Guan Y. Event ontology reasoning based on event class influence factors. Int'l Journal of Machine Learning and Cybernetics, 2012,3(2):133−139. [doi: 10.1007/s13042-011-0046-8]

[5] Metzler D, Cai CX, Hovy E. Structured event retrieval over microblog archives. In: Proc. of the 2012 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Montreal: Association for Computational Linguistics Press, 2012. 646−655. http://www.aclweb.org/anthology-new/N/N12/N12-1083.pdf

[6] Steven S, De Martine C, Kerre EE. Reasoning about fuzzy temporal information from the Web: Towards retrieval of historical events. Soft Computing, 2010,14(8):869−886. [doi: 10.1007/s00500-009-0471-8]

[7] Wu PB, Chen QX, Ma L. Study on intelligent retrieval of event relevant documents based on event frame. Journal of Chinese Information Processing, 2003,17(6):25−30 (in Chinese with English abstract).

[8] Feng L. Breaking events' information extraction based on event frame [MS. Thesis]. Shanghai: Shanghai Jiaotong University, 2008 (in Chinese with English abstract).

[9] Li L, Wang H, Fu Y, Yang X, Wang J. Research on semantic retrieval for civil aviation emergency cases. Computer Engineering and Design, 2011,32(3):1130−1133 (in Chinese with English abstract).

[10] Liu XM, Fan XZ, Liu L. Analysis method of complex questions integrating event information. Journal of South China University of Technology (Natural Science Edition), 2011,39(7):140−145 (in Chinese with English abstract).

[11] Lin HF, Liang JM. Event-Based ontology design for retrieving digital archives on human religious self-help consulting. In: Proc. of the 2005 IEEE Int'l Conf. on e-Technology, e-Commerce and e-Service. Hong Kong: IEEE Press, 2005. 522−527. http://dl.acm.org/citation.cfm?id=1049566 [doi: 10.1109/EEE.2005.70]

[12] Zhong ZM, Li CH, Guan Y, Liu ZT. A method of query expansion based on event ontology. Journal of Convergence Information Technology, 2012,7(9):364−371. [doi: 10.4156/jcit.vol7.issue9.43]

[13] Zhong ZM, Zhu P, Li CH, Guan Y, Liu ZT. Research on event-oriented query expansion based on local analysis. Journal of the China Society for Scientific and Technical Information, 2012,31(2):151−159 (in Chinese with English abstract).

[14] Hsu WL, Wu SH, Chen YS. Event identification based on the information map-INFOMAP. In: Proc. of the IEEE Int'l Conf. on Systems, Man, and Cybernetics. Arizona: IEEE Press, 2001. 1661−1666. [doi: 10.1109/ICSMC.2001.973523]

[15] Wang W, Zhao DY, Zhao W. Identification of topic sentence about key event in Chinese news. Acta Scientiarum Naturalium Universitatis Pekinensis, 2011,47(5):789−795 (in Chinese with English abstract).

[16] Chau M, Chen HC. A machine learning approach to Web page filtering using content and structure analysis. Decision Support Systems, 2008,44(2):482−494. [doi: 10.1016/j.dss.2007.06.002]

[17] Kim S, Zhang BT. Genetic mining of HTML structures for effective Web-document retrieval. Applied Intelligence, 2003,18(3): 243−256. [doi: 10.1023/A:1023293820057]

[18] Yang EH, Zeng QQ, Li TT. Analysis of event information structure in text. Journal of Chinese Information Processing, 2012,26(3): 92−97 (in Chinese with English abstract).

[19] Yang LY, Zhang HJ, Zhang YK. Research on emergency news corpus classification system. In: Proc. of the Chinese Information Society Conf. on Chinese Information Processing. Beijing: Tsinghua University Press, 2006. 403−409 (in Chinese with English abstract). http://cpfd.cnki.com.cn/Article/CPFDTOTAL-ZGZR200611002048.htm

[20] Wang L, Liu ZT, Wang YH, Liao T. Web page main text extraction based on content similarity. Computer Engineering, 2010,36(6): 102−104 (in Chinese with English abstract).

[21] Lei Z. Research on event-based news story analysis technology [Ph.D. Thesis]. Changsha: National University of Defense Technology, 2006 (in Chinese with English abstract).

附中文参考文献:

[7] 吴平博,陈群秀,马亮.基于事件框架的事件相关文档的智能检索研究.中文信息学报,2003,17(6):25−30.

[8] 冯礼.基于事件框架的突发事件信息抽取[硕士学位论文].上海:上海交通大学,2008.

[9] 李林,王红,付宇,杨璇,王静.民航突发事件应急案例语义检索方法研究.计算机工程与设计,2011,32(3):1130−1133.

[10] 刘小明,樊孝忠,刘里.融合事件信息的复杂问句分析方法.华南理工大学学报(自然科学版),2011,39(7):140−145.

[13] 仲兆满,朱平,李存华,管燕,刘宗田.一种基于局部分析面向事件的查询扩展方法.情报学报,2012,31(2):151−159.

[15] 王伟,赵东岩,赵伟.中文新闻关键事件的主题句识别.北京大学学报(自然科学版),2011,47(5):789−795.

[18] 杨尔弘,曾青青,李婷婷.事件信息结构分析.中文信息学报,2012,26(3):92−97.

[19] 杨丽英,李红娟,张永奎.突发事件新闻语料分类体系研究.见:中国中文信息学会学术年会论文集.北京:清华大学出版社,2006. 403−309. http://cpfd.cnki.com.cn/Article/CPFDTOTAL-ZGZR200611002048.htm

[20] 王利,刘宗田,王燕华,廖涛.基于内容相似度的网页正文提取.计算机工程,2010,36(6):102−104.

[21] 雷震.基于事件的新闻报道分析技术研究[博士学位论文].长沙:国防科学技术大学,2006.

仲兆满(1977－),男,江苏赣榆人,博士,副教授,主要研究领域为智能信息检索.
E-mail: zhongzhaoman@163.com

刘宗田(1946－),男,教授,博士生导师,主要研究领域为人工智能,软件工程.
E-mail: ztliu@shu.edu.cn

李存华(1963－),男,博士,教授,主要研究领域为数据挖掘,人工智能.
E-mail: cli@hhit.edu.cn

戴红伟(1975－),男,副教授,CCF 会员,主要研究领域为人工智能.
E-mail: hongweidai@hotmail.com