情 报 学 报 ISSN 1000 - 0135
第 31 卷 第 2 期 151 - 159, 2012 年 2 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 31 No. 2 , 151 - 159 February 2012

doi:10.3772/j.issn.1000-0135.2012.02.005

一种基于局部分析面向事件的查询扩展方法[1)]

仲兆满[1] 朱 平[3] 李存华[1] 管 燕[1] 刘宗田[2]

(1. 淮海工学院 计算机工程学院, 连云港 222005; 2. 上海大学 计算机工程与科学学院, 上海 200072;
3. 国际介藤网络中心, 北京 100102)

摘要 针对用户获取事件类信息的需求,提出了一种基于局部分析面向事件 LA-EO(local analysis-event-oriented)的查询扩展方法,该方法将查询项区分为事件项和限定项两类分别处理。文章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。围绕突发事件领域,使用搜索引擎和定点采集相结合的方法收集了 4011 篇文本,设置了 10 个查询项对本文提出的方法进行了实验比较。结果表明:LA-EO 与 Rocchio 机制(记作 LA-Rocchio)和局部上下文分析(记作 LA-LCA)扩展方法相比,对事件类信息的检索,LA-EO 具有更优的检索性能。

关键词 信息检索 查询扩展 局部分析 面向事件 Rocchio 局部上下文分析

Research on Event-Oriented Query Expansion Based on Local Analysis

Zhong Zhaoman[1], Zhu Ping[3], Li Cunhua[1], Guan Yan[1], Liu Zongtian[2]

(1. *School of Computer Engineering, Huaihai Institute of Technology, Lianyungang*, 222005;
2. *School of Computer Engineering and Science, Shanghai University, Shanghai*, 200072;
3. *International Centre for Bamboo Rattan, Beijing*, 100102)

Abstract Aiming at the demand for event information by users, we propose a method——local analysis-based event-oriented (LA-EO) query expansion, which divides query terms into two categories: event terms and qualifying terms. The paper emphatically discusses the analysis of event-oriented query terms, the expansion of event terms and the computing between query terms and documents. We have collected 4011 documents and set ten query topics centering on emergency domain to evaluate our proposed methods. The results show that LA-EO query expansion offers more effective performances for retrieving events, compared with LA-Rocchio and LA-LCA.

Keywords Information retrieval, query expansion, local analysis, event-oriented, Rocchio, LCA

1 引言

查询扩展是指在原查询词的基础上加入相关的词,从而组成新的、更准确的查询词集。它利用计算机语言学、信息学等多种技术,以用户原查询为基础,把与原查询相关的词添加到原查询,以便更完整地描述原查询所隐含的语义或主题,帮助信息检索

收稿日期: 2011 年 4 月 7 日

作者简介: **仲兆满**,男,1977 年生,博士,副教授,主要研究方向:信息检索、事件本体。E-mail:zhong zhao man@163.com。**朱平**,男,1982 年生,工程师,主要研究方向:知识管理、信息检索。**李存华**,男,1963 年,博士,教授,主要研究方向:模式识别与人工智能。**管燕**,女,1976 年生,硕士,讲师,主要研究方向:模式识别。**刘宗田**,男,1946 年生,上海大学博士生导师、教授,中国人工智能学会粗糙集与软计算专业委员会常委,主要研究方向:智能信息处理、软件工程。

1) 基金项目:国家自然科学基金项目(60975033)、国家科技部项目(2009GJC10043)。

系统提供更多有利于判断文档相关性的信息,是弥补用户查询信息不足,改善信息检索的查全率和查准率的有效手段。其核心问题是如何设计和利用扩展词的来源。查询扩展方法大致分为两类:基于语义知识辞典的方法和基于语料库的方法[1]。基于语义知识辞典的方法在进行查询扩展时通过已有的语义知识辞典来进行扩展词的选取(常用的是辞典或本体)。基于语料库的方法又可细分为两种:全局分析方法和局部分析方法。局部分析方法假设初始检索结果的前若干篇文档是相关的,然后根据一定的策略选取局部文档集的关键字扩展到查询项中。最有代表性的扩展项选取策略是 Rocchio 机制(简称 Rocchio)[2~4] 和局部上下文分析(记作 LCA)[5],Rocchio 根据关键字在局部文档集中出现的频率选取扩展项,LCA 通过计算查询项与局部文档集中关键字的关联强度选取扩展项。多次 TREC 评测会议表明,局部分析方法是一种计算量小且不依赖于外部资源,但十分有效的查询扩展方法[6]。

与事件检索相关的研究主要是围绕事件的识别和跟踪展开的。由美国国防高级研究计划委员会主办的话题识别与跟踪(TDT)评测会议定义事件为"特定时间特定地点发生的事情",认为多个事件组成一个话题。较多的学者采用聚类的方法进行事件识别,如文献[7]。微软的 Li 等在文献[8]中利用概率方法完成了事件回顾识别任务。Yang 等[9]提出了将信息检索和机器学习技术应用于事件探测和追溯中。事件的识别和跟踪针对的是新事件,是对未知事件的发现与追溯。本文所研究的是面向事件的查询扩展,是从文本集合中检索出与用户输入的已知事件最为相关的文档。

近几年,有些学者开始在基于本体的查询扩展中引入了事件的思想。Lin 等[10]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,将事件的要素作为该本体中的主要分类,在检索的时候可以按事件要素对查询词进行扩展。Han[11]提出了一种基于事件的人物本体模型,他认为可以根据人物之间的关系构造本体,同时人物会关联一些特定的事件。

事件框架是事件类知识表示的另外一种形式。Hsu 等[12]建立了事件结构框架,对于某查询对象,在事件结构框架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。吴平博等[13]利用手工确定的句型模板构造了提取规则,用于从处理后的文本中提取事件信息填充到事件框架的侧面中,并将事件框架的知识用到事件相关文档的检索中。这类方法的实现需要事先构造本体或事件框架。

随着各类突发事件的频繁发生,获取网络上事件类的信息已经变得非常迫切。在很多情况下,用户借助搜索引擎获取事件类信息,对事件类信息的查询和其他查询有许多不同,主要表现在:

(1) 事件是由事件触发词标识,关联了参与者、时间和地点等要素,比"概念"粒度更大的语义单元。事件的各个要素在查询内容中作用是不同的。比如,查询项:"汶川地震"和"孟买 恐怖袭击",其中,"地震"、"恐怖袭击"是事件触发词,标识所要检索的事件类型,在本文中称为事件项。"汶川"和"孟买"是事件的地点要素,限定查询事件的范围,在本文中称为事件的限定项。

(2) 限定项除了由事件要素充当以外,还有可能是由其他事件充当的。比如查询项"地震 救援",用户关心的是"地震"中的"救援"事件,"地震"充当了限定项。由于事件项和限定项在查询项中的作用不同,所以应该采取不同的处理策略。

(3) 事件之间有着紧密的联系,一个话题总是关联了一些特定的事件。提及话题"地震",人们自然的就联想到"死亡"、"救援"、"重建"等事件;看到话题"竞选","演讲"、"辩论"、"投票"等事件自然浮现到脑海中,而不需要关心具体的事件要素。据此,可以进行事件到事件之间的联想扩展。

但已有的查询扩展方法没有分析查询项的不同作用,没能使用面向事件的查询扩展技术,因此不能很好地解决一些面向事件的信息检索问题。本文提出了一种基于局部分析面向事件的查询扩展方法(LA-EO),重点研究了面向事件的查询项分析、事件到事件的联想扩展等关键技术,并与 Rocchio 机制(LA-Rocchio)和局部上下文分析方法(LA-LCA)进行了实验比较,对检索结果进行了对比分析。

2 基于局部分析面向事件的查询扩展

2.1 查询扩展的模型

本文提出的基于局部分析面向事件的查询扩展模型如图 1 所示。

在图 1 所示的模型中,统计所有查询项在文本中出现的频次,按照频次的高低选择排在最前面的 n 篇文本组成第一次检索获取的局部文档集 N。局部文档集 N 对查询扩展效果有较大影响,为了获取

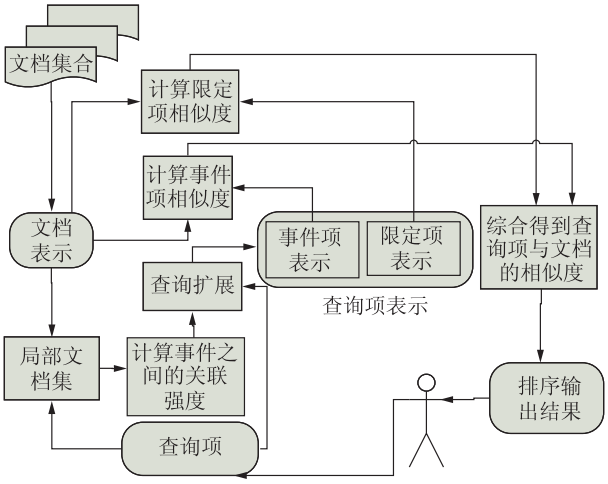

图 1　面向事件的查询扩展模型

好的局部文档集，一些研究者提出了对返回文档集重排的算法[14,15]。对局部文档集的重排不是本文研究的重点。明显的，如果局部文档集与用户的查询相关性强，则面向事件的查询扩展的效果将会更好，即本文介绍的方法同样适用于查询扩展的全局分析。

对局部文档集 N 进行面向事件的分析，识别出文本中的事件，计算查询事件与局部文档集 N 中事件的关联强度，作为选择扩展事件的依据。

对查询项进行面向事件的分析，区分为事件项与限定项两类，对事件项进行事件间的查询扩展。分别计算查询项的事件项与文本的相似度、查询项的限定项与文本的相似度，综合得到整个查询项与一篇文档的相似度。最后根据相似度的大小返回若干篇文本作为最终的检索结果。

文本及查询项的表示方法与所采用的信息检索模型有关。由于布尔模型是一种精确检索，不易处理模糊的检索，没有考虑查询项的不同权值，向量空间模型 VSM 得到了广泛的研究与应用。VSM 采用了部分匹配的检索策略，即包含了部分查询项的文本也可以作为检索结果。本文的文本及查询项的表示都使用了向量空间模型。

下面讨论本模型的几个关键技术。

2.2　文本中事件的识别

面向事件的查询扩展首先需要识别出文本中的事件，这些事件是计算关联强度、选择扩展的依据。

事件的触发词一般是由动词、动名词和名词充当的。很多动词都是实义动词，有明确的意义，如“睡觉”、“赌博”、“救援”等。但有些动词抽象意义较高，本身不能表达一件事情，常和其他动词或名词连用，这类动词不能充当事件触发词，我们将此类抽象意义高的动词称之为停用动词，如“存在”、“发生”、“开始”、“让”等。

根据停用动词的抽象程度，我们将停用动词分为两大类：第一类动词，不作为事件触发词，直接舍弃，如常见的是存现动词、能愿动词、判断动词、使令动词、主观感觉、猜想、阐述等动词；第二类动词，不作为事件触发词，但与其一起出现的动词、名动词或者名词可以作为事件触发词，如“发生火灾”、“开始演讲”、“禁止吸烟”等，“发生”、“开始”和“禁止”不是事件触发词，而名词“火灾”、动词“演讲”和“吸烟”是事件触发词。这两类停用动词的情况汇总如表 1 所示。

根据北大计算语言所给出的《汉语文本词性标注标记集》，文本经 ICTCLAS 分词工具分词、滤除停用词后，事件的识别步骤如下：

步骤 1：提取出所有的动词(/v)和名动词(/vn)作为候选事件集 E1。

步骤 2：从 E1 中直接删除第一类停用动词，从而得到候选事件集 E2。

步骤 3：从 E2 中选取第二类停用动词，假设当前选取的第二类的某个停用动词为 v_1，在文本中从 v_1 的位置开始先向后查找 v_1 后面出现的名词(之所以选名词，因为在步骤 1 中已经将动词和名动词作为事件触发词了)，如果查找到了某个名词为 n_1，则将 n_1 添加到 E2 中，从 E2 中删除 v_1；如果 v_1 的后面没有出现名词，则从 v_1 的位置开始向前查找 v_1 前面出现的名词，查找到的名词同样添加到 E2 中，从 E2 中删除 v_1；如果向前和向后都没发现名词，则直接将 v_1 从 E2 中删除。无论是向后还是向前查找，查找的范围都局限于一个子句的 k 距离个词(所谓子句是指由标点符号“。”、“，”、“：”、“？”、“！”分开的一个汉字序列，k 距离指一个词到 v_1 所间隔的词的个数)。最终得到候选事件集 E。

2.3　查询项中限定项和事件项的判别

查询项中限定项和事件项的判别有两种可用的方法：① 用户进行事件检索时，在不同的文本框中输入相关内容，即用户指定了查询项中的事件项和限定项；② 让计算机借鉴语义资源、局部文档集等自动判别查询项中的限定项和事件项。方法一类似于常用搜索引擎的高级搜索功能，虽然增加了用户的工作量，但提高了查询项判别的准确率。在本文中，使用了第一种方法，第二种方法在本文中不做讨论。

表1 停用动词表

两类抽象动词	停用动词	
第一类停用动词	存现动词	属于、有、包括、显现、存在、无、没有、具有、出现、还有、成为
	能愿动词	会、愿意、可以、能够、能、可能、愿、不能、不得
	趋向动词	来、去、上、下、起、到、回、回到、出、入、进、进入、起来、靠近、到来、向前、走、离、前来、往来、下来、回去、前往、下去
	判断动词	是、为、乃,系,即
	使令动词	使、让、令、叫、宵禁、禁止、勒令
	主观感觉、猜想、阐述等动词	感觉、猜、猜想、想、认为、相信、说、说道、称、宣称、介绍、宣布、提出、暗示、明示、表明、表示、指出、坚持、主张、强调、重申、介绍、呼吁、希望、感到、表达、要求、告诉、关注、得知、说明
	心理动词	喜欢、恨、气愤、觉得、思考、厌恶、想、支持、反对、赞同、同意、勉励、盼望、爱、反感
	其他动词	要、应该、应、不可、可、得到、继续、产生、有利于、需要、处理、像、成、请、问、意味着、予以、值得、看得见
第二类停用动词	发生、发动、开始、造成、加强、强化、削弱、提高、提升、降低、获得、展开、提供、准备、充满、受到、进行、取得、推动、制定、达成、加快、加大、举行、赢得、争取、进行、取得、参加、保持、面临、推进、准备、开创、恢复、筹办、开展、召开	

2.4 扩展事件的选取

2.4.1 事件间关联强度的计算

LA-Rocchio方法选取扩展词的标准是计算词在局部文档集中出现的频次。LA-LCA方法考虑了词之间的关联,以词在局部文档集中的共现度作为选取扩展词的标准。文本集合中一个词与一个查询项的共现度的计算方法如下[5]:

$$co\_\mathrm{deg}ree(c,q_i)=\log_{10}(co(c,q_i)+1)^{idf(c)}/\log_{10}(n) \quad (1)$$

$$co(c,q_i)=\sum_{d\mathrm{in}N} tf(c,d)tf(q_i,d) \quad (2)$$

公式(1)和公式(2)中,q_i指某个查询项,c指文本集中的一个词,N是局部文档集,n是文本集合的篇数。其中,公式(2)是计算共现度的核心,$tf(c,d)$是词c在文档d中出现的次数,$tf(q_i,d)$是查询项q_i在文档d中出现的次数。由公式(1)和公式(2)可知,$co(c,q_i)$值的大小主要取决于$tf(c,d)$和$tf(q_i,d)$的乘积。

例1:假设有一查询项q_i为"地震",局部文档集N中包含两篇文本d1和d2,待扩展的词是c1"救援"、c2"通话"。q_i和c1、c2在d1和d2中出现的次数如表2所示。

表2 q_i和c1、c2在d1和d2中出现的次数

	q_i	c1	c2
d1	10	8	2
d2	30	10	12

通过表2可见,在文档d1中,"地震"与"救援"的关联强度明显大于"地震"与"通话"的关联强度;在文档d2中,"地震"与"救援"、"通话"的关联强度非常接近,关联强度都比较弱,没有什么区别。据此,由文档d1和d2,根据人工的判断,"地震"与"救援"的关联强度应该大于"地震"与"通话"的关联强度。

使用词与一个查询项的共现度的计算公式(1)和公式(2)计算"地震"与"救援"、"通话"的关联强度。得到"地震"与"救援"、"通话"的关联强度相等,都是1.714 729。这与人工的判断结果不相符合,可见,使用共现度的方法计算事件之间的关联强度并不合理。原因是:动态的"事件"不同于静态的"概念","事件"是随着时间的变化而动态迁移,即一个文本集合中,往往是一部分文本在描述一个子话题,而另外一部分文本在描述另外一个子话题。

采用共现度的方法计算事件之间的关联强度，由于是用两个事件出现的次数进行了相乘，故一个事件的出现次数多少的变化很容易引起共现度的变化。比如，例 1 中所举的文档 d2，就是因为 q_i 在 d2 中出现的次数较多，q_i 与 c1 相乘后累加导致共现度计算的不合理。

事件之间的关联更多地表现出一个事件发生后，另一个事件发生的概率大小，是条件概率的关系。公式(2)不适合描述事件之间的关联强度。

对局部文档集合 N 中的一篇文本 d，事件 e_i 对 e_j 的关联强度定义为：

$$w_{ij}^{d} = \frac{F_{e_j}}{F_{e_i}} \tag{3}$$

其中，F_{e_i} 表示事件 e_i 在文档 d 中出现的次数，F_{e_j} 表示事件 e_j 在文档 d 中出现的次数。对文档 d 而言，如果 $F_{e_i} = 0$，则无须计算 F_{e_j}，直接有 $w_{ij}^{d} = 0$。

如果 $w_{ij}^{d} > 1$，则进行归一化处理，令 $w_{ij}^{d} = 1$。

对整个局部文档集 N，事件 e_i 对 e_j 的关联强度定义为：

$$w_{ij} = \frac{\sum_{d \in M} w_{ij}^{d}}{|M|} \tag{4}$$

其中，M 表示局部文档集 N 中出现过事件 e_i 的文档的集合。

接着例 1，使用公式(3)和公式(4)计算“地震”对“救援”、“通话”的关联强度。“地震”对“救援”的关联强度为 0.5665；“地震”对“通话”的关联强度为 0.3，得到的结果与人工的判断比较吻合，可见，条件概率的思想更适合描述事件之间的关联强弱。

局部分析方法的三种扩展项选取策略：Rocchio 方法(LA-Rocchio)、局部上下文分析(LA-LCA)和面向事件(LA-EO)的比较见表 3 所示。

表 3　三种扩展项选取方法的比较

扩展方法	LA-Rocchio	LA-LCA	LA-EO
扩展项选取的指标	扩展项在文档集中出现次数	扩展项与整个查询串共现度	扩展项与整个事件项关联强度
核心计算方法	次数相加	次数相乘，再相加	次数相除，再相加
语义关联	否	是	是
适用范围	扩展项与查询项无语义关联	扩展项与查询项有语义关联	扩展项与查询项有语义关联，且表现为条件概率关系

表 3 说明，对事件类信息的检索，LCA-EO 方法更适合。所以，在本文中，使用了条件概念的思想衡量事件之间的关联强度，并进一步作为选取扩展事件的依据。

2.4.2　扩展事件的选取

依据公式(3)和公式(4)，可选择 *top-k* 个事件作为扩展事件。对 k 的确定，按阈值选择或指定个数是研究中常用的方法[16]，而且 k 的变化对检索结果有很大的影响。

假设查询项 Q 包含 m 个事件 $Q_e = \{e_1, e_2, \cdots, e_m\}$，文本集中待扩展的事件记作 e_x，则 e_x 与 $Q_e = \{e_1, e_2, \cdots, e_m\}$ 的关联强度计算方法如式(5)所示：

$$f(e_x, Q_e) = \sum_{e_i \in Q_e} w_{ix} \tag{5}$$

其中，w_{ix} 是事件 e_i 对事件 e_x 的关联强度。

2.5　查询项的权值设置

查询项是由事件项和限定项组成的。原查询项中的事件项集记作 Q_e，各个事件项 $e_i \in Q_e$，扩展后的查询事件项集记作 $Q_e^{\exp}$；原查询项的限定项集记作 $Q_{e'}$，各个限定项 $e'_i \in Q_{e'}$，扩展后的限定项集记作 $Q_{e'}^{\exp}$。查询项 x($x \in Q_e$ 或 $x \in Q_{e'}$)在初始查询 Q 中的权重记做 $W(x \mid Q)$，查询项 x 在扩展后的查询 $Q^{\exp}$ 中的权重记做 $W(x \mid Q^{\exp})$。

查询项权重的设置是否合理对检索的效果有较大的影响，但其权值的设置一直是一个难以解决的问题。最常用的权值设置是采用 Rocchio 方法，Rocchio 公式见式(6)所示：

$$W(x \mid Q^{\exp}) = \alpha \times W(x \mid Q) + \beta \times \frac{\sum_{d \in N} W(x \mid d)}{n} \tag{6}$$

其中，$W(x \mid Q)$ 是查询项 x 的初始查询 Q 中权重，$W(x \mid Q^{\exp})$ 是查询项 x 在扩展后查询 $Q^{\exp}$ 中的权重，$W(x \mid d)$ 为查询项 x 在文档 d 中的权重，α 和 β 为两个大于 0 的可调参数，通常情况下，设定 $\alpha = \beta = 1$。$W(x \mid Q)$ 的计算是根据 x 在查询项 Q 中出现的频次，一般有 $W(x \mid Q) = 1$。根据式(6)可知，$W(x \mid Q^{\exp})$ 的值主要取决于 x 在局部文档集 N 中的平均权重。

LCA 根据文档集中词与查询词关联的强度，指定了计算查询项权重的计算方法如式(7)所示：

$$wt_i = 1.0 - 0.9 \times i/s \tag{7}$$

其中，wt_i 是第 i 个查询项的权重，s 是查询项的个数。

Rocchio方法可用于无关系的项的权值量化，LCA方法更适合于有关系的项的权值量化。因此，本文对查询项中限定项的权值设置采用了Rocchio方法，查询项中事件项的权值设置采用了LCA中的量化方法。

2.6 查询项与文档相似度的计算

文档 d 的特征项的权值取该特征项在文档 d 中出现的频次。

扩展的查询事件项向量 $Q_e^{\exp}$ 与文档向量 d 的相似度计算公式如式(8)所示：

$$S(Q_e^{\exp},d)=\frac{\sum_{e_i\in Q_e^{\exp}\cap D_e} W(e_i\mid Q_e^{\exp})W(e_i\mid d)}{\sqrt{\sum_{i=1}^{|Q_e^{\exp}|} W(e_i\mid Q_e^{\exp})^2\sum_{i=1}^{|d|} W(e_i\mid d)^2\,|Q_e^{\exp}|}} \quad (8)$$

其中，$e_i\in {}^{\exp}_{e}\cap d$ 表示 e_i 是扩展的查询事件项向量 $Q_e^{\exp}$ 和文档 d 的共同事件项，$|Q_e^{\exp}|$ 和 $|d|$ 分别是向量 $Q_e^{\exp}$ 和 d 的大小。

设定权值后的查询限定项向量 $Q_{e'}^{\exp}$ 与文档向量 d 的相似度计算公式如式(9)所示：

$$S(Q_{e'}^{\exp},d)=\frac{\sum_{e'_i\in Q_{e'}^{\exp}\cap D_{e'}} W(e'_i\mid Q_{e'}^{\exp})W(e'_i\mid d)}{\sqrt{\sum_{i=1}^{|Q_{e'}^{\exp}|} W(e'_i\mid Q_{e'}^{\exp})^2\sum_{i=1}^{|d|} W(e'_i\mid d)^2\,|Q_{e'}^{\exp}|}} \quad (9)$$

其中，$e'_i\in Q_{e'}^{\exp}\cap d$ 表示限定项 e'_i 是查询限定项向量 $Q_{e'}^{\exp}$ 和文档向量 d 的共同限定项，$|Q_{e'}^{\exp}|$ 是向量 $Q_{e'}^{\exp}$ 的大小。

公式(8)和公式(9)之所以要除以 $|Q_e^{\exp}|$、$|Q_{e'}^{\exp}|$，是因为向量 $Q_e^{\exp}$ 的事件个数远大于 $Q_{e'}^{\exp}$ 中限定项的个数，导致 $S(Q_e^{\exp},d)$ 的值远大于 $S(Q_{e'}^{\exp},d)$ 的值，这样就大大地弱化了限定项的限定作用，容易引起“查询漂移”。

整个查询(包括事件项和限定项)与文档 d 的相似度计算公式如式(10)所示：

$$S(Q^{\exp},d)=S(Q_e^{\exp},d)+S(Q_{e'}^{\exp},d) \quad (10)$$

3 实验设计及结果分析

3.1 实验目的

实验目的是验证面向事件的查询扩展的有效性，所以在实验时没有使用文本的标题、首段、首句等启发式信息。对于本文提出的面向事件的查询扩展方法，从以下两个角度进行了实验比较：① 区分查询项的类别，扩展的事件项个数对检索性能的影响；② LA-EO与LA-Rocchio、LA-LCA扩展方法在检索性能上的比较。

3.2 实验语料及评测方法

已经有一些中文搜索的评测语料，如2005年863信息检索，它使用的语料集是CWT100G，提供了50个查询项；SEWM系列会议的信息检索评测，它更侧重于Web的信息检索的评测，语料库有CWT100G、CWT200G等几个版本；CIRB030评测，它的语料是纯文本格式的，用XML做了一些标记，提供了42个查询项；搜狗实验室在2008年也推出了3.0版本的互联网搜索语料。

已有的搜索语料面向的是通用的搜索评测，收集的面很广，而且设置的查询项只有很少一部分是面向事件的。所以，我们围绕突发事件领域收集评测语料，突发事件一级分为4个大类，二级分为33个子类[17]，在搜集突发事件语料的时候涉及到上述33个子类中的9个子类，即围绕9个子类制定了一些查询关键词，其中重点是“地震”、“火灾”、“食物中毒”、“交通事故”和“恐怖袭击”5个子类。借助Google搜索引擎，输入一些查询关键字，收集了1639篇文本；使用爬虫工具，从指定的一些站点上下载了2435篇文本。然后对所有的文本按照标题进行了排重，最后剩下4011篇文本作为本文实验的语料。

查询项的设置采用了与用户使用搜索引擎最为一致的方式：输入若干个关键字。人工设置了10个查询项，对于每个查询项，使用P@10和P@20作为评价指标。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。

使用了Pooling技术确定每个查询项的标准答案。对于P@n，一个查询项的标准答案的确定具体步骤是：

(1) 取4种方法返回的前 n 篇文本合并得到一个集合S；

(2) 人工从这个文本集合S中选取相关的文档作为一个主题的标准答案。

表4列出了本文中使用的10个查询项。每个

查询项由限定项和事件项组成。

表 4　10 个查询项

编号	限定项	事件项
1	汶川	重建
2	汶川;地震	死亡
3	地震	救援
4	2008;汶川	地震
5	新疆	恐怖袭击
6	印度	恐怖袭击
7	河北	交通事故
8	河北;交通事故	死亡
9	克拉玛依;12.8	火灾
10	全球	灾难

3.3　扩展的事件项的个数对检索性能的影响

LA-EO 查询扩展的算法步骤如下:

(1) 计算初始查询 Q 中所有查询项在每篇文档 d 出现的次数之和 $Cou(d)$;

(2) 按照 $Cou(d)$ 大小降序排序,选择前面的 n =30 篇文本作为局部文档集 N;

(3) 使用面向事件的查询扩展方法,得到扩展事件项向量 $Q_e^{\exp}$ 和限定项向量 $Q_{e'}^{\exp}$;

(4) 将文档向量化 d;

(5) 分别计算 $S(Q_e^{\exp},d)$ 和 $S(Q_{e'}^{\exp},d)$,最终得到 $S(Q^{\exp},d)$,按照 $S(Q^{\exp},d)$ 大小降序排序输出最终检索结果。

步骤(2)选取局部文档数目的依据是文献[18]的建议值 10 ~ 50。

在局部文档集数目为 30,对于查询项“汶川 重建”,图 2 是 LA-EO 扩展方法在扩展事件的个数为 0 时得到的查询结果,图 3 是 LA-EO 扩展方法在扩展事件的个数为 6 时得到的查询结果。图 2 和图 3 列出的是 top-10 个文档的名称及其与查询项的相似度。相似度在输出时保留了小数点后 6 位。

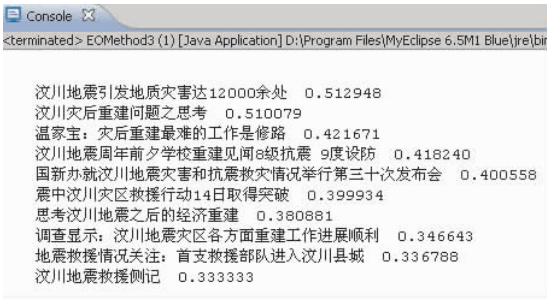

图 2　扩展项个数为 0 时的查询结果

图 3　扩展项个数为 6 时的查询结果

对查询事件扩展的个数从 0 ~ 20 之间做了实验比较。表 5 列出了 10 个查询项的平均结果。

不同的扩展事件项个数获取的 10 个查询项的平均 P@10 和 P@20 如图 4 所示。

从表 5 和图 4 可见,在区分查询项类别的情况下,对初始查询不进行扩展,P@10 为 0.57、P@20 为 0.49,检索的结果已经不是很差。随着查询事件

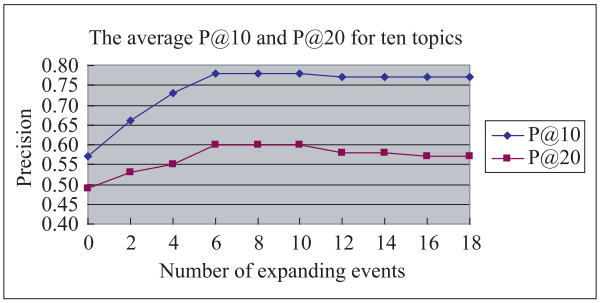

图 4　10 个查询项的平均 P@10 和 P@20

表 5　不同的扩展事件项的个数对检索性能的影响

扩展方法	指标	事件项扩展的个数									
		0	2	4	6	8	10	12	14	16	18
LA-EO	P@10	0.57	0.66	0.73	0.78	0.78	0.78	0.77	0.77	0.77	0.77
	P@20	0.49	0.53	0.55	0.60	0.60	0.60	0.58	0.58	0.57	0.57

扩展个数的增多,P@10 和 P@20 都有一定幅度的提高。在扩展的事件个数达到 6 的时候,检索性能已经非常理想, P@10 为 0.78,P@20 为 0.60。随后,增加扩展事件的个数并没有改善检索性能,反而有所下降。可见,查询扩展事件的个数较多时,容易带来噪声,干扰查询效果。

3.4 LA-EO 与 LA-Rocchio、LA-LCA 的性能比较

LA-Rocchio 方法比较简单,在此不做进一步阐述,有兴趣的读者可以阅读文献[2]。LA-LCA 的算法步骤如下:

步骤(1)、(2)同于 2.3 小节介绍的 LA-EO 查询扩展的算法步骤(1)、(2);

(3) 使用 LCA 方法,得到扩展向量 $Q^{\exp}$;

(4) 计算 $Q_e^{\exp}$ 与 d 的相似度 $S(Q^{\exp},d)$,按照 $S(Q^{\exp},d)$ 大小降序排序输出最终检索结果。

其中第(3)步是 LCA 方法的核心,扩展词选择的标准采用如下函数进行

$$f(c,Q) = \prod_{w_i \text{ in } Q} (\delta + co\_\deg ree(c,w_i))^{idf(w_i)} \quad (10)$$

公式(10)中, $co\_\deg ree(c,w_i)$ 是词与整个查询项的共现度。计算词 c 与整个查询项 Q 的共现度时,既可以侧重于词与整个查询串的共现度,也可以侧重于词与某个查询项的共现度,δ 是起到此作用的调节参数。δ 值大,侧重于词与单个查询项的共现度;反之,δ 值小,侧重于词与整个查询串的共现度。在本文的实验中, $\delta = 0.01$。$idf(w_i)$ 考虑到不同的查询项的重要度不同。

例如,对于查询项“汶川 重建”,使用三种扩展方法得到的前十个扩展项如表 6 所示。

表 6 三种扩展方法获取的前 10 个扩展项

扩展方法	扩展项
LA-EO	恢复 地震 支援 建设 施工 援建 审计 受灾 规划 完成
LA-Rocchio	地震 灾区 资金 规划 工作 恢复 四川 灾害 情况 问题
LA-LCA	地震 规划 恢复 工作 灾区 资金 建设 四川 项目 灾害

从表 6 可见,不同的扩展方法得到的扩展项有较大的不同,LA-EO 和 LA-Rocchio 有 70% 是不同的,LA-EO 与 LA-LCA 有 60% 是不同的。而且,三种方法得到的查询项的排序都有些不同,即使是得到相同的查询项,但不同的排序影响了查询项的权值,对计算查询项与文本的相似度也会有较大的影响。

对 LA-Rocchio 和 LA-LCA 的扩展词的个数从 0~40 做了实验,三种方法取 10 个查询项的平均结果的最优值进行了对比。表 7 列出了对比结果。

表 7 三种扩展方法获取的最优的检索性能的比较

Query Expansion Method	扩展词或事件的个数	P@10	P@20
LA-Rocchio	12	0.59	0.52
LA-LCA	16	0.63	0.54
LA-EO	6	0.78	0.65

从表 7 可见,三种不同的查询扩展方法,LA-EO 是最好的,LA-Rocchio 是最差的。对评价指标 P@10 和 P@20,LA-EO 比 LA-Rocchio 分别提高了 0.19 和 0.13。主要原因:一方面 LA-Rocchio 和 LA-LCA 没有区分查询项的不同类型,分别处理;另一方面,LA-Rocchio 和 LA-LCA 没有采用面向事件的联想扩展策略。另外,实验结果还表明,对于事件类信息的查询项,查询扩展的个数在较少的情况下已经可以取得很好的查询结果,对于 LA-Rocchio 和 LA-LCA 方法扩展项的个数建议为 10~16,而对于 LA-EO 扩展方法扩展事件的个数为 6 个左右。这与针对通用信息的查询扩展方法所建议的扩展项的个数有些不同,如文献[5]通过实验表明使用 30 个扩展项获取的检索性能只是稍逊于使用 70 个扩展项的检索性能,文献[18]建议的扩展项的个数以 30~100 为宜。这也说明,对不同领域的查询问题,获取较好的检索性能所使用的扩展项的个数是有差别的。

4 结论

本文提出了一种基于局部分析面向事件的查询扩展方法,该方法对查询内容采用了面向事件的分析技术,将查询内容区分为事件项和限定项两种类型,分别采用不同的策略处理。与经典的扩展策略 Rocchio 和 LCA 方法相比,在实验结果上体现出了面向事件的查询扩展在实际应用中的优势。在研究中发现,以下内容还值得进一步探讨:

(1) 从部分文档中统计获得的事件之间的关联并不非常准确,对事件项的扩展,可以考虑借鉴语义

词典或本体进行补充扩展。

（2）事件的各个要素在查询内容和文本中都有不同的表现形式，如文本介绍“汶川地震”，而查询项是“中国 地震”，需要借助地理本体才能更好地实现地点要素的匹配；又如，文本介绍“今年 5 月的地震”，而查询项是“2008 年 地震”，需要提取文本报道的时间，对文本和查询项中的时间要素进行规整才能够做到时间的准确匹配。

（3）采用索引技术是提高信息检索效率的有效途径，基于事件的文本索引技术有待进一步研究。

参 考 文 献

[1] 王瑞琴，孔繁胜. 基于查询扩展和词义消歧的语义检索[J]. 情报学报，2010，29(1)：16-21.

[2] Buckley C, Salton G, Alan J, et al. Automatic query expansion using SMART[C]//Harman D. Proceedings of the 3rd text retrieval conference (TREC-3). National Institute of Standards and Technology, Gaithersburg, MD, 1995：69-80.

[3] Ko Y, An H, Seo H. Pseudo-relevance feedback and statistical query expansion for web snippet generation[J]. Information Processing Letters, 109, 2008：18-22.

[4] 吴丹，何大庆，王惠临. 基于伪相关反馈的跨语言查询扩展[J]. 情报学报，2010，29(2)：232-239.

[5] Xu J, Croft B W. Improving the effectiveness of informational retrieval with local context analysis [J]. ACM Transactions on information systems, 2000, 18(1)：79-112.

[6] Voorhees E, Harman D. Overview of the Sixth Text Retrieval Conference (TREC-6) [C]//Voorhees E. Proceedings of the 6th text retrieval conference (TREC-6). NIST Special Publication, 1998：240-500.

[7] Yang Y, Pierce T, Carbonell J. A study of retrospective and on-line event detection[C]//Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, 1998：28-36.

[8] Li Z, Wang B, Li M, et al. A probabilistic model for retrospective news event detection[C]//The 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005：106-113.

[9] Yang H, Chua T S, Wang S, et al. Structured use of external knowledge for event-based open domain question answering [C]//The 26th annual international ACM SIGIR conference on Research and development in information retrieval Toronto. Canada: ACM Press, 2003：33-40.

[10] Lin H F, Liang J M. Event-based Ontology design for retrieving digital archives on human religious self-help consulting [C]//The 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service. Hong Kong, China, 2005：522-527.

[11] Han Y. Reconstruction of People Information based on an Event Ontology [C]//The 2007 IEEE International conference on natural language processing and knowledge engineering. Beijing, China, 2007：446-451.

[12] Hsu W L, Wu S H, Chen Y S. Event identification based on the information map-INFOMAP [C]//The IEEE International Conference on Systems, Man, and Cybernetics, Tucson, Arizona, USA, 2001：1661-1666.

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

[14] Hearst M A. Improving full-text precision on short queries using simple constraints[C]//The symposium on document analysis and information retrieval. Las Vegas, NV, 1996：237-267.

[15] Mitra M, Singhal A, Buckley C. Improving automatic query expansion [C]//The 21st annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, 1998：206-214.

[16] Chang Y, Qunis I, Kim M. Query reformulation using automatically generated query concepts from a document space [J]. Information Processing and Management, 2006, 42(2)：453-468.

[17] 杨丽英. 突发事件新闻语料分类体系研究[C]//中文信息处理前沿进展(中国中文信息学会二十五周年学术年会论文集). 北京：清华大学出版社，2006.

[18] 丁国栋，白硕，王斌. 一种基于局部共现的查询扩展方法[J]. 中文信息学报，2006，20(3)：84-91.

（责任编辑 马 兰）