文章编号：1003-0077（2011）00-0000-00

基于事件元素无向图的查询扩展方法[1]

叶雷，高盛祥，余正涛*，秦广顺，洪旭东

（昆明理工大学 信息工程与自动化学院，昆明 650500）

通讯作者：余正涛 ztyu@hotmail.com

摘要：借助新闻事件元素之间的关联特性，提出了基于事件元素无向图的查询扩展方法，利用新闻事件元素之间的关联关系进行查询扩展提升新闻事件检索效果。首先分析候选事件文档与查询项的关系，确定待扩展的元素；然后利用事件元素之间的关联关系构建无向图，通过事件向量空间计算边的权重；最后，利用无向图节点权重模型计算事件元素权重，依据权重进行事件元素扩展。在新闻事件查询扩展方面进行了对比试验，结果表明提出的查询扩展方法取得了较好的效果。

关键词：新闻事件；查询扩展；事件元素；事件元素无向图

中图分类号：TP391　　**文献标识码：**A

A Query Expansion Method Based on Undirected Graph of Event Elements

Lei Ye, Shengxiang Gao, Zhengtao Yu*, Guangshun Qin, Xudong Hong

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500)

Corresponding Author: Zhengtao Yu, ztyu@hotmail.com

Abstract: Taking the relationship between event elements into account, we propose a query expansion method based on undirected graph of event elements, which utilizes the relevance between news event elements to conduct query expansion to improve news event retrieval. Firstly, analyzing the relationship between candidate events and queries, we select out the elements to be extended. Then, we construct an undirected graph using the extracted event elements and the relationship between them, and compute the edge weights through event vector space. Lastly, we compute the weight of event elements by the undirected graph model of node weight, and extend event elements according to the weights computed. By compared experiments on new event query expansion, it is proved that the proposed query expansion method has a good effect on news event retrieval.

Keywords: news event; query expansion; event elements; undirected graph

1 引言

在信息检索领域，查询扩展技术是指在用户输入的原查询项的基础上，加入一些与查询相关的词语，组合成新的查询，用来解决查询信息不全的问题，有助于提高检索的性能。目前，在信息检索领域，常见的查询扩展有基于语义知识词典、基于全局文档集分析和基于局部文档集分析三种方法。基于语义知识词典的方法[1]通常利用 WordNet、HowNet 或同义词词林等语义知识词典中提供的同义关系和上下位关系选取新词；基于全局分析的技术[2,3]以词关联假设为基础，分析文档集中的全部语词，计算所有词语对的关联强度，利用词语之间的关联实现查询扩展；基于局部分析的技术[4-8]首先进行初始检索，把与查询最相关的 N 篇文档当做扩展词的来源，在这些文档中寻找与查询项相关的词语实现查询扩展，经典的局部分析方法叫做局部反馈（Local Feedback），也称伪相关反馈（Pseudo Feedback）。

近几年，新闻事件的查询扩展方法成为专家学者研究的热点，如仲兆满等人研究了基于

基金项目：国家自然科学基金(61472168、61175068)；云南省自然科学基金重点项目（No.2013FA130），云南省科技创新人才基金项目(No.2014HE001)资助。

局部分析面向事件的查询扩展方法[9]（LA-EO），该方法针对事件的特点将查询项分成事件项和限制项，重点研究了扩展事件的选取以及查询项与文本相似度的计算，并与其他查询扩展方法进行了实验对比，针对事件类信息检索，LA-EO 具有更优的检索性能；如文献[10]提出的基于事件本体的查询扩展（EO-QE）,该方法在查询扩展中引入了事件本体，重点探讨了事件的四元组概念以及在不同事件元素下的扩展策略。

上述面向事件的查询扩展方法，通过引入事件属性和本体的思想，在事件类信息检索中有了一定的提高。在新闻事件检索中，检索的对象是事件，事件由事件元素组成，包括事件的主体、事件发生的时间与地点等，而且同一事件或相似事件的事件元素之间具有关联关系。因此，针对新闻事件的检索，查询扩展的层面应该是事件元素而不是一般的词语。如果查询项包含了目标事件的更多元素，那么检索效果就会更好。例如，在检索关于“5•12 汶川地震”事件的过程中，原查询项是“汶川县，地震”，经过扩展后查询项变成了“汶川县，地震，2008 年 5 月 12 日，救援，死亡……”，查询项中包含了更多目标事件的元素词语，那么检索的精度会得到提高。因此，本文提出了基于事件元素的查询扩展方法，结合无向图的思想，进行新闻事件信息的查询扩展。

2 事件表示

事件可以定义为在一个特定时间和环境发生的、由若干对象参与的、表现出若干动作特征的一件事情。事件的表示模型有很多种，本文对于事件的定义来自于 ACE[11]。根据该定义，事件是一个由事件触发词（Trigger）和描述事件属性的元素组成的集合。图 1 表述了一个事件的构成。其中“出生”是该事件的触发词，“习近平”是对象要素，“1953 年”是对应的时间，“陕西富平”则是地点。

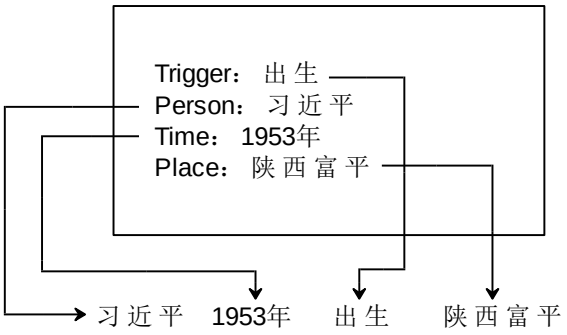

图 1 “出生”事件的表示模型

3 基于事件元素无向图的查询扩展

基于事件元素无向图的查询扩展方法流程如图 2 所示，首先采用伪相关反馈技术从待检索文档集中获取 m 篇与查询最相关文档，用作待扩展的事件元素词的来源。然后分析文档与查询的关系，把文档标记成相关文档或者相似文档，相关文档提取事件的所有元素，相似文档只提取事件的触发词元素。抽取文档中的事件元素，构造以事件元素为节点的无向图，利用事件元素无向图节点权重模型计算图中节点的权重，结合文档与查询的分析，进行事件元素扩展。

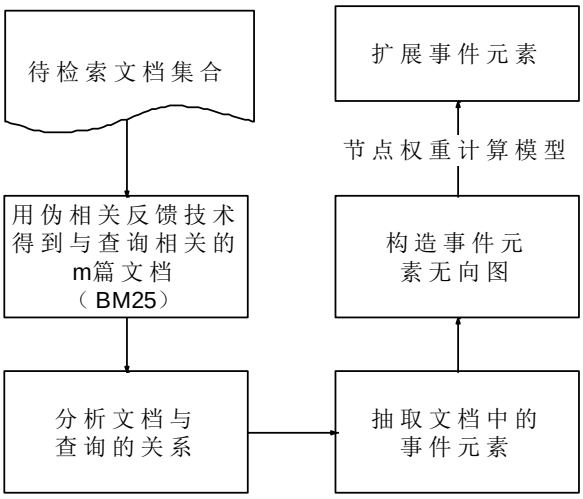

图 2 基于事件元素无向图的查询扩展方法流程图

3.1 用伪相关反馈技术获取文档

伪相关反馈技术首先通过普通检索算法从初始文档中寻找到一个初始结果，然后假定其中排名最前面的 m 篇文档是相关的，最后在这个假设条件下进行相关反馈。本文采用基于 BM25 算法作为伪相关反馈技术的检索算法从待检索文档集中获取文档，作为扩展词的来源。BM25 检索算法是一种经典的概率统计检索算法[12]，用以计算文档 D 和查询 Q 的相似性。给定一个查询 Q，包含关键词 $q_1,\ldots,q_t$，查询和文档的 BM25 分值计算方法如下：

$$BM25(A,Q)=\sum_{i:f(q_i,A)>0}\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5}\cdot\frac{f(q_i,A)\cdot(k_1+1)}{f(q_i,A)+k_1\cdot(1-b+\mathrm{b}\cdot\frac{|A|}{avgal})}\cdot\frac{(k_3+1)f(q_i,Q)}{k_3+f(q_i,Q)} \qquad (1)$$

其中，N 代表候选文档的总数；$n(q_i)$ 为含有查询关键词 q_i 的文档数量；$f(q_i,A)$ 是查询词 q_i 在 A 中出现的频率；$f(q_i,Q)$ 表示查询词 q_i 在 Q 中出现的频率；$|A|$ 表示文档长度，$avgal$ 是所有候选文档的平均长度；k_1，k_3 和 b 是调节因子。

3.2 文档和查询的关系分析

事件元素包括事件触发词和其他元素词（时间、地点和人物等），多个事件可能会出现相同的事件触发词，但是事件的其他元素词不会相同，例如事件 1“2008 年 5 月 12 日，汶川县发生 8.0 级地震”和事件 2“2015 年 4 月 25 日，尼泊尔发生 8.1 级地震”的触发词都是“地震”，其他属性词却不一样。因此，我们认为触发词为事件的共性元素，其他元素词为事件的特性元素。

如果文档中仅出现查询的事件触发词，没有出现其他元素词，那么把文档标记为查询相似文档；如果文档中既出现查询的事件触发词，又包含了其他元素词，那么把文档标记为查询相关文档。对于查询相似文档，把文档中抽取到的事件共性元素（触发词）作词为待扩展词语；对于查询相关文档，把从文档中抽取到的所有事件元素词作为待扩展词语。

3.3 事件元素的抽取

在构造无向图之前，需要抽取事件元素作为待扩展词，关于事件元素的抽取方法可参看文献[13]，本文采用的抽取方法如下：对新闻事件语料进行命名实体识别，通过触发词词典匹配语料中的触发词，在每一个触发词的上下文中搜索相邻的命名实体，匹配到的命名实体包括人名、机构名、地点和时间等，作为事件发生的对象、地点和时间要素。触发词及其相关的事件元素即结构化地表示了一个事件。

3.4 事件元素无向图的构建

为了有效地表示事件元素之间的关联关系，采用无向图的方法。无向图的构建，具体可

以分为两个部分，第一个部分是节点以及边的生成，第二个部分是边权重的计算。下面详细说明事件元素的无向图构建过程。

3.4.1 无向图节点和边的生成

将抽取到的每一个事件元素对应无向图中的一个节点，在这个过程中需要注意两点，一是多个事件可能会出现相同的事件元素，这些共现的事件元素在图中只用一个节点表示；二是对时间表达式做统一处理，比如“2015 年 3 月 9 日”和“2015 年 3 月 9 日 17 时 59 分”，将这个时间统一标记为“2015 年 3 月 9 日 17 时 59 分”。每个事件内部的元素节点两两相连，形成无向图的边。

例如，事件 1“2008 年 5 月 12 日 14 时 28 分，汶川县发生 8.0 级地震”，事件 2“2008 年 5 月 12 日，大量师生死于坍塌校舍之下”和事件 3“2015 年 4 月 25 日，尼泊尔发生 8.1 级地震”构成的无向图如图 3 所示。

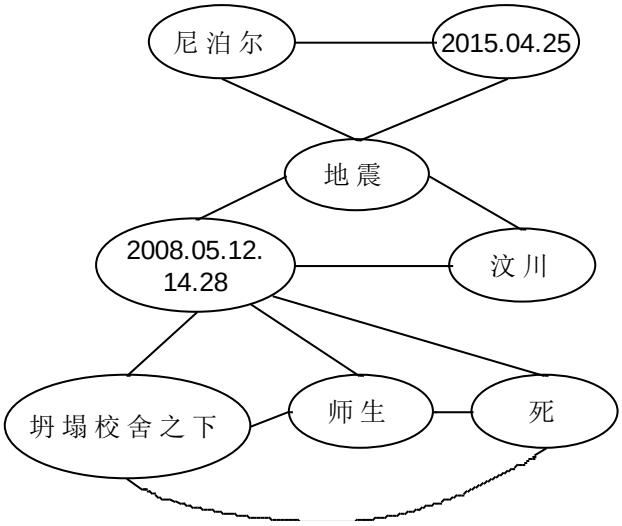

图 3 事件元素无向图

3.4.2 无向图边权重的计算

事件元素无向图边的权重反映的是事件元素之间的关联程度。借鉴向量空间模型的思想，以新闻事件为特征，构建向量空间，来表征事件元素，将事件元素间的关联程度计算问题转换为事件特征空间上向量之间的相似度计算问题。

（1）事件向量空间的构造

假设从伪相关反馈的 N 篇文档中抽取到 M 个事件，形成子事件集合 $K=\{W_1,W_2,\cdots,W_m\}$，根据集合 K 里面的每一个子事件作为向量空间中的一维，构建一个 M 维的向量空间。

（2）事件元素到向量空间的映射

利用对相应维度赋布尔值的方法，实现事件元素到向量空间的映射。以事件元素 e_1 映射到 M 维的向量空间 $K=\{W_1,W_2,\cdots,W_m\}$ 为例，具体步骤如下：

1）对事件元素 e_1，判断其在第一维度 W_1 上面的取值。如果 e_1 在事件 W_1 里面出现，则该维度值为 1，反之为 0。

2）循环第一步，计算 e_1 在每一维度上的取值。

3）计算出 e_1 在每一个维度上的取值之后，就得到了 e_1 在向量空间中的特征向量。

利用上面的方法，实现事件元素到 M 维向量空间上的映射。

（3）事件元素特征向量相似度计算

使用向量之间的余弦夹角对事件元素进行相似度计算，即：

$$\mathrm{Sim}(\mathrm{S}_i,\mathrm{S}_j)=\cos\theta=\frac{\sum_{k=1}^{m} w_{ik}\times w_{jk}}{\sqrt{(\sum_{k=1}^{m} w_{ik}^2)(\sum_{k=1}^{m} w_{jk}^2)}} \quad (2)$$

其中，θ 表示 S_i，S_j 在向量空间中的向量夹角，W_{ik} 为 S_{i} 在第 k 维上面的取值，W_{jk} 是 S_j 在第 k 维上面的取值，m 表示事件向量空间的维数。

事件元素两两之间计算相似度，得到事件元素的相似度矩阵，如下所示：

$$M=\begin{bmatrix} \mathrm{Sim}_{1,1} & \mathrm{Sim}_{1,2} & \cdots & Sim_{1,i} & \cdots & \mathrm{Sim}_{1,n} \\ Sim_{2,1} & Sim_{2,2} & \cdots & Sim_{2,i} & \cdots & Sim_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ Sim_{i,1} & Sim_{i,2} & \cdots & Sim_{i,i} & \cdots & Sim_{i,n} \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ Sim_{n,1} & Sim_{n,2} & \cdots & Sim_{n,i} & \cdots & Sim_{n,n} \end{bmatrix} \quad (3)$$

将事件元素之间的相似度作为无向图中两个节点的边的权重。

3.5 事件元素无向图节点权重计算模型

如果两个事件元素节点之间相连，就说明这两个元素节点存在联系。当计算某节点的权重时，如果相连接的节点越多，那么其权重也应该越大。同时也要考虑原查询事件元素的影响，如果某个节点与原查询事件元素相连接，那么该事件元素节点应该更加重要，更有可能作为原查询事件的扩展元素。

基于上述思想，新闻事件元素的无向图节点重要度模型可以表示为：

$$SW(i)=\begin{cases} (1-d)+d\times\sum_{j=1,j\neq i}^{n} M_{i,j}\times\dfrac{SW(j)}{\sum_{k=1,k\neq j}^{n} M_{j,k}}, & j\notin Q; \\ (1-d)+d\times\sum_{j=1,j\neq i}^{n} M_{i,j}\times\dfrac{SW(j)}{\sum_{k=1,k\neq j}^{n} M_{j,k}}\times\lambda, & j\in Q; \end{cases} \quad (4)$$

其中，$SW(\mathrm{i})$ 表示节点 i 的权重（重要度），d 表示阻尼系数，一般设置为 0.85，$M_{i,j}$ 是相似度矩阵 M 中的值，表示无向图中第 i 个事件元素和第 j 个事件元素之间的相似度，Q 是原查询事件元素集合，λ 是大于 1 的可调参数。

3.6 事件元素的扩展和权重设置

计算出事件元素无向图节点的权重，即得到事件元素的评分值。按照权重从高到低的顺序扩展前 N 个元素，同时考虑到 3.2 节关于文档和查询的关系分析，对于查询相似文档，只扩展触发词元素，相关文档扩展所有元素。

一般来说，不同的扩展词对应的重要性也不同，扩展得到的查询词权重的设置是否合理对检索性能有很大的影响。考虑到扩展词的评分越高，它在查询中所占的权重应该越高，我们的权重设置使用如下公式[6]：

$$W(q\mid Q_{\mathrm{exp}})=\alpha\times W(q\mid Q)+\beta\times\frac{Score(q)}{MaxScore} \quad (5)$$

其中，$W(q\mid Q_{\mathrm{exp}})$ 表示查询词 q 在扩展后查询 Q_{exp} 中的权重，$W(q\mid Q)$ 表示查询词 q 在

初始查询 Q 中的权重，通常使用 q 在 Q 中的频度表示， $Score(q)$ 代表扩展词 q 的得分值，即公式（4）所得结果，$MaxScore$ 表示所有扩展词评分的最大值。α 和 β 为两个大于 0 的可调参数，通常情况下都设置成 1。

4 实验及分析

4.1 检索性能评价指标

信息检索技术常用指标有平均准确率(Mean Average Precision，MAP)，n 位置的准确率（precision at position n，P@n）等，下面将分别进行介绍。

4.1.1 Mean average precision (MAP)

MAP（Mean Average Precision）把所有查询项的 AP 放在一起求宏平均，作为衡量系统对多个查询的平均检索质量[14]，反映系统在全部相关文档上性能的单值指标。检索到的相关文档位置越靠前，那么 MAP 值便会越高。假如没有检索出相关文档，则 MAP 值是 0。计算方法如下表示：

给定一个查询,其平均 AP 的计算公式如下：

$$AP = \frac{\sum_{n=1}^{N}(p@n * rel(n))}{\#total\ relevant\ docs\ \ for\ this\ query} \tag{6}$$

其中，N 为文档的个数， $rel(n)$ 是关于第 n 个文档相关性的一个二元函数。

$$rel(n) = \begin{cases} 1, if\ \ the\ n^{th} doc\ is\ relevant \\ 0, otherwise \end{cases}$$

针对多个查询，我们通过对所有查询的 AP 值求平均得到 MAP：

$$MAP = \frac{\sum_{i} AvgP_i}{number\ of\ \ queries} \tag{7}$$

4.1.2 Precision at position n (P@n)

P@n 是指对一个排序结果，返回前 n 个结果的准确率。有时用户使用搜索引擎搜索想要的结果可能只对返回的前 n 个页面感兴趣，P@n 就是从这样的角度对检索性能进行衡量的评价标准。其公式如下：

$$P@n = \frac{\#relevant\ docs\ in\ top\ n\ results}{n} \tag{8}$$

例如，对于查询返回的前 5 个文档是{irrelevant（不相关），relevant（相关），relevant，relevant，irrelevant}，那么 P@1 到 P@5 的值分别是{0，1/2，2/3，3/4，3/5}。对于查询集合，通过对所有查询的 P@n 值求平均得到 P@n 值。

4.2 实验数据说明

目前已有的信息检索语料多用于通用搜索评测，设置的查询项并不面向新闻事件。所以，我们通过互联网爬虫技术从中国新闻网、新华网、新浪新闻等新闻网站收集了 3000 个面向新闻事件的页面，并对这些页面做一些处理，每个页面保留其标题及正文，把这些文本作为实验语料。爬取的新闻主题包括汶川地震、MH370 失联、昆明火车站恐怖袭击等一系列社会热点话题。查询项的设置采用与用户使用搜索引擎相一致的方式，即输入若干个词语当做查询。针对爬取的新闻文本，我们人工设计了 8 个面向事件的查询项，新闻文本的分布及相应查询项的设置如表 1 所示。

表 1 实验数据分布及查询项设置

新闻主题	文本数量	查询项
512 汶川地震	500	汶川　地震
昆明火车站恐怖袭击	400	昆明火车站　恐怖袭击
香港占中事件	300	香港　占中
上海外滩踩踏事故	300	上海外滩　踩踏
马航 MH370 失联	600	马航　失联
425 尼泊尔地震	500	尼泊尔　地震
深圳山体滑坡事故	200	深圳　山体滑坡
银川公交车纵火事件	200	银川公交车　纵火

4.3 实验设计与结果分析

为了验证所提方法的有效性，设计了两个试验。实验 1 比较了事件元素扩展词的个数对检索性能的影响，在得到扩展词之后，逐渐增加扩展词词数，比较检索结果；实验 2 比较不同扩展方法的检索性能，采用经典的查询扩展算法 Rocchio 方法[4]作为对照。

4.3.1 事件元素扩展词个数的影响

事件元素扩展词数量的选择会影响检索结果精度，扩展词语的数量太少的话，达不到查询扩展的目的，而扩展词数量过多的话，又会引起查询噪音问题。对查询事件元素扩展的个数从 0~18 之间做了对比实验，针对 8 个事件查询主题，所提方法在不同扩展词个数下得到的平均 MAP 和 P@5 如图 4 所示：

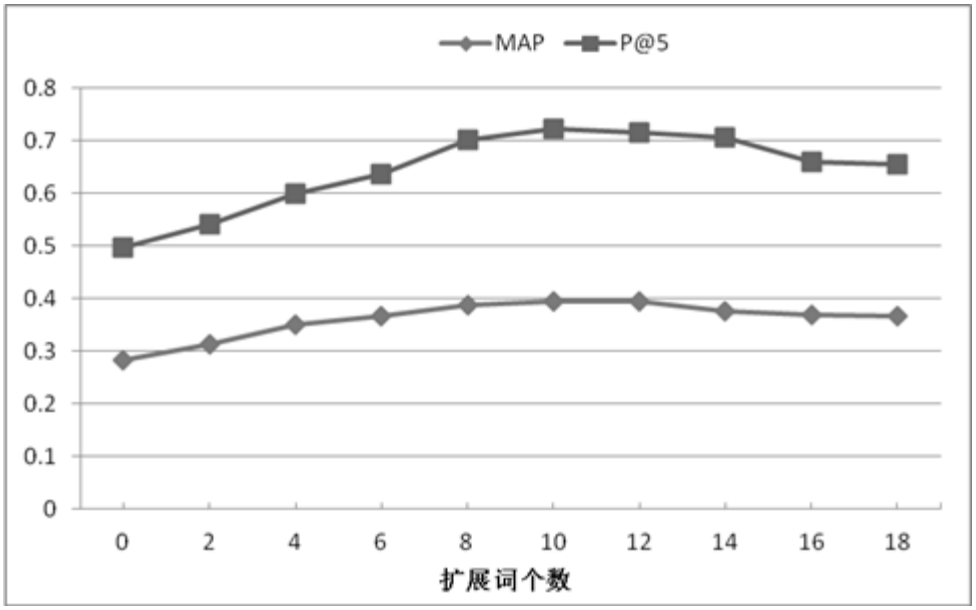

图 4 不同扩展词个数的实验结果

从图 4 可以看出，扩展词数量在 0 到 10 之间时，随着查询事件元素扩展词数量的增加，检索性能 MAP 和 P@5 都有一定幅度的提高。当扩展词个数达到 10 的时候，检索性能达到最优，平均 MAP 为 0.3943，P@5 为 0.7232。而扩展词个数达到 10 之后，增加扩展词个数

并没有继续提高检索性能，反而性能有一定的下降。从此可见，查询扩展词数目应该选择10 的时候检索性能最好，太多的话会引入噪音。

4.3.2 不同查询扩展方法的性能比较

为验证所提查询扩展方法的有效性，在相同扩展词个数（10 个）的条件下，我们与其它扩展方法进行了对比实验。其中，不进行查询扩展的方法记为 M_1，Rocchio 方法记为 M_2，本文所提方法记为 M_3。实验采用的评价指标为 MAP 和 P@5，实验结果如图 5 所示：

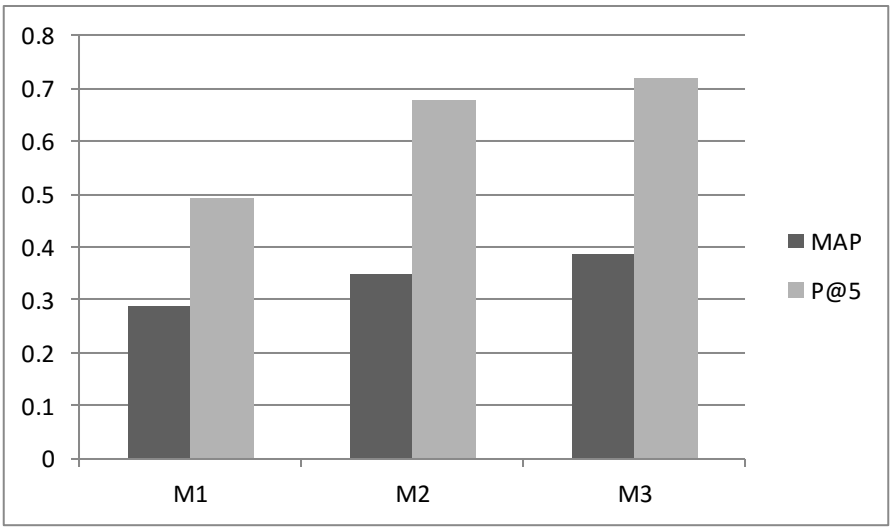

图 5 不同扩展方法的实验结果

由图 5 可见，在相同的条件下，所提出的扩展方法的检索性能相较其他方法都有一定的提高。这主要是因为考虑到面向事件的查询扩展，应该扩展的是事件元素词，而不是一般的词语。同时，考虑到了事件元素之间存在的关联以及原查询项的事件元素对节点权重的影响，通过构建事件元素无向图节点权重模型进行查询扩展，取得了很好的效果。

5 总结

本文提出了一种基于事件元素无向图的查询扩展方法。利用新闻事件元素构成无向图，计算出事件元素之间的相似性，通过节点权重模型计算事件元素的重要程度并得到扩展词。实验证明所提方法在面向事件的查询扩展中取得了较好的效果，能够提高事件检索的性能。进一步将深入考虑新闻事件的主题对事件元素的影响，以及不同类型主题的事件元素的共性，研究新闻事件的查询扩展方法。

参考文献：

[1] Richardson R, Smeaton A. Using WordNet in a Knowledge-Based Approach to Information Retrieval[J]. Working paper CA-0395, School of Computer Applications, Trinity College Dublin. 1995.

[2] Wei J, Bressan S, Ooi B C. Mining term association rules for automatic global query expansion: methodology and preliminary results[C]//Web Information Systems Engineering, 2000. Proceedings of the First International Conference on. IEEE, 2000, 1: 366-373.

[3] Zhang CQ, Qin ZX, Yan XW. Association-Based segmentation for Chinese-crossed query expansion[J]. IEEE Intelligent Informatics Bulletin, 2005,5(1):18−25.

[4] C. Buckley, G. Salton, J. Allan, et al. Automatic query expansion using SMART[C]// Overview of Text Retrieval Conference. 1994:69-80.

[5] Song M, Song I Y, Hu X, et al. Integration of Association Rules and Ontology For Semantic Query Expansion[J]. Data & Knowledge Engineering, 2007, 63(1): 63-75.

[6] 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.

[7] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, 2008, 243-250.

[8] 黄名选,严小卫,张师超.基于矩阵加权关联规则挖掘的伪相关反馈查询扩展[J].软件学报,2009, 20(7):1854-1865.

[9] 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[J].情报学报,2012, 31(2):151-159.

[10] Zhong ZM, Li CH, Guan Y, Liu ZT. A method of query expansion based on event ontology[J]. Journal of Convergence Information Technology,2012, 7(9):364−371.

[11] ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Events. National Institute of Standards and Technology. 2005.

[12] S. E. Robertson, S. Walker, S Jones, et al. Okapi in TREC3[C]//Proceedings of Text Retrieval Conference, Gaithersburg, USA. U.S. National Institute of Standards and Technology, NIST Special Publication 500-225: 1994. 109-126.

[13] 赵妍妍, 秦兵, 车万翔,等. 中文事件抽取技术研究[J]. 中文信息学报, 2008, 22(1):3-8.

[14] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval[M]. New York: ACM press, 1999.

叶雷（1992—），男，硕士生，主要研究领域为信息检索、自然语言处理。
Email:yelei0128@gmail.com

高盛祥（1977—），女，博士生，主要研究领域为信息检索、机器翻译。
Email:gaoshengxiang.yn@foxmail.com；

余正涛（通讯作者）（1970—），男，教授/博导，博士，主要研究领域为自然语言处理、信息检索、机器翻译。
Email:ztyu@hotmail.com