doi:10.11887/j.cn.201404029 http://journal.nudt.edu.cn

词网络的新闻事件关联建模*

张 辉[1,3],李国辉[1],徐新文[2],贾 立[1],孙博良[1]
(1. 国防科技大学 信息系统与管理学院,湖南 长沙 410073;
2. 国防科技大学 指挥军官基础教育学院,湖南 长沙 410073;3. 61226部队,北京 100079)

摘 要:互联网上每天都会报道许多新闻事件,为了挖掘各事件间的关系,提出一种新闻事件关联建模方法。该方法首先利用TF-IEF和相邻词合并策略对事件的相关报道集提取关键词,然后综合多种词共现度量窗口对事件关键词的关联关系建模,构建事件关键词关联网络,最后依据事件间共有关键词的程度建立事件关联模型,从而建立事件关联网络。实验表明该方法能够较准确地提取报道集的关键词,较好地发现事件间的关联关系。

关键词:词网络;新闻事件关联;词共现
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2014)04-0169-08

Modeling association of news events on term network

ZHANG Hui[1,3], *LI Guohui*[1], *XU Xinwen*[2], *JIA Li*[1], *SUN Boliang*[1]
(1. College of Information System and Management, National University of Defense Technology, Changsha 410073, China;
2. College of Basic Education for Officers, National University of Defense Technology, Changsha 410073, China; 3. Unit 61226, Beijing 100079)

Abstract: There are many news events reported daily on the Internet. An innovative method is proposed to mine event-relations between news. Following an adjacent term-combining strategy, this method primarily utilized a so-called term frequency & inverse event frequency (TF-IEF) model to extract key phrases from the corresponding reports set as to a particular event. Then term co-occurrence windows were employed to calculate the associating degree of every single term pair. This degree is indicative in building event key phrase-networks. Further, two matters were correlated to shape the event relation-network model: (I) common key phrases as mediators within event key phrase-network, and (II) the degree of commonness of key phrases within different observed events. An experiment was conducted to examine the performance of proposed method. The results show that the method can accurately extract key phrases and comprehensively mine associations between events.

Key words: term network; news event association; term co-occurrence

由于互联网技术的快速发展,网络媒体都会对每天发生的新闻事件进行实时报道,形成海量新闻文本。新闻事件是发生在某个具体时间、某个具体地点的一项活动,可能由某主体引起或组织。通过新闻事件探测技术[1]能够将事件相关的报道进行聚类,形成大量的事件相关报道集,每个报道集描述一个新闻事件。不同新闻事件间或多或少地存在各种联系,例如日本地震与日本农产品出口、中东政治事件与全球石油价格变动等。通过对不同的事件报道集进行关联挖掘,获取用户感兴趣的内在联系或知识,一直都是情报挖掘研究的核心内容。如果用人工方式对海量的新闻报道进行整理、分析,查找事件间的关联知识,将是一件极其费时、费力的工作。本文提出一种新闻事件关联建模及可视化方法,该方法以词网络技术为基础,挖掘事件间的关联,自动构建事件关联词网络,辅助开源情报[2]获取。

词网络分析[3-5]是获取各种关联信息、知识的一种非常有效的方法,词网络的节点是词语,词语是最基本的文本语义单元。词语间的关联组合构成语句、段落、篇章等语义单元,从而对事件涉及的主题进行描述和表达。新闻事件与词语的语义层级关系如图1所示,从图中可以看出:(1)事件主题表达是由多个词集合共同完成,词是事件主题描述的最小语义单元;(2)要充分发现事件间的关联,仅从新闻事件的报道、段落、语句等语义层较难实现,需要从事件的词语语义层出发,构建事件的词关联网络,并在不同事件的共有词基

* **收稿日期**:2013-09-05
基金项目:国家自然科学基金项目(61170158);国家部委资助项目;湖南省自然科学基金项目(12JJ5028)
作者简介:张辉(1983—),男,湖南湘潭人,博士研究生,Email:zhanghui@nudt.edu.cn;
李国辉(通信作者),男,教授,博士,博士生导师,Email:guohli@nudt.edu.cn

础上,构建出事件间的关联网络。因此,将词网络分析技术应用到事件关联分析中,能够较充分地挖掘出事件间的关联关系。

新闻事件之间存在的关联是具体的,很少有共性。为了简化研究,本文忽略事件关联的具体意义,将所有的关联进行抽象:如果两个事件中包含相同的关键词,并且这个共有的关键词分别与两个事件中的其他关键词的词共现度超过预先设置的阈值,则认为两个事件间存在关联关系。词共现度是指两个词语同时在同一新闻事件中共现的频繁程度,如果两个词存在共现,则认为这两个词语之间存在着一定的关联。词共现度是词语之间关联度的度量,是用定量的方法描述事件中两个词语关联的紧密程度。在事件的词网络中,关键词是网络节点,词关联关系是节点间的边,边的权重表征词共现度。因此,可以通过计算事件中关键词共现度,直接建立事件词网络,同时对不同事件词网络中的共有关键词节点进行分析,构建事件关联网络。

本文以词网络分析技术为基础,首先通过 TF-IEF(Term Frequency & Inverse Event Frequency)[1] 和相邻词合并策略提取事件报道集中关键词集,其中 TF-IEF 是一种计算词语表征事件权重的模型,然后通过建立事件词共现模型,构建事件关键词关联网络,最后根据不同事件间的共有词程度,建立事件间的关联网络。

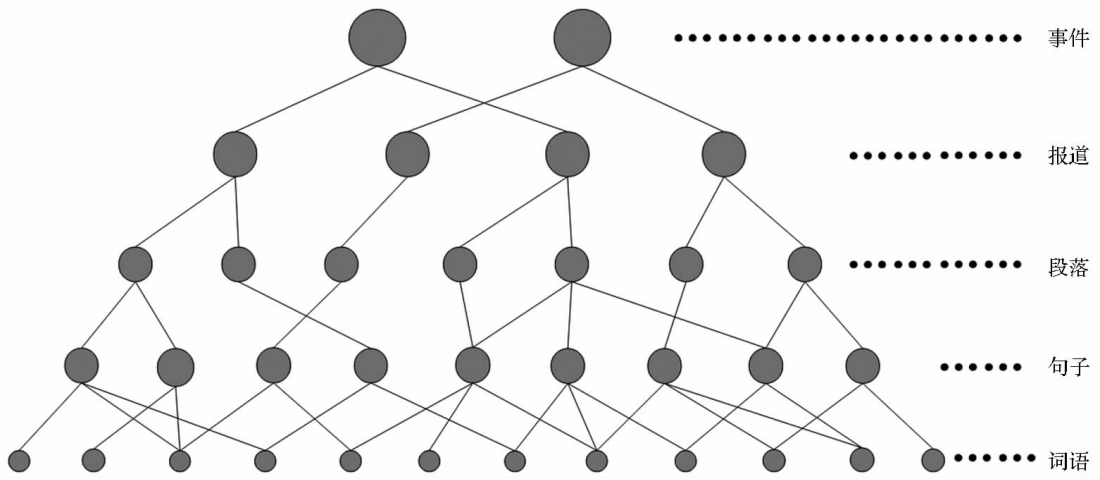

图 1 新闻事件与报道、段落、句子、词语间的语义层级关系

Fig. 1 Hiberarchy relationship between events and documents, paragraphs, sentences, terms

1 相关研究

近年关于词网络的研究主要还是利用统计方法,对文本集中不同共现窗口中的词对进行统计分析,计算共现词间的关系权重,通过共现词构建文本集的词网络。在语句共现窗口方面,Tseng[4,6]、Ho[5]等提出一种基于频繁字词项合并的关键词提取方法,并通过分析关键词对在单篇新闻报道的语句窗口中共现关系进行建模,进而获取关键词对在整个文本集中的关联关系,从而构建词关系网络,并将其应用到犯罪和民众文化素养调查;Perrin[7]结合日本语言的特点,提出利用词缀模式和类模式两种方法提取文本集的关键词集,然后比较了三种根据关键词对提取对应语句的方法,依据词对共现语句的多少判断词对关联的程度,最后利用关键词网络对关键词关联关系进行可视化。在段落共现窗口方面,Kawai[8]使用词缀模式对关键词进行提取,通过将词对的共现窗口设置为段落,构建词对关联网络,同时针对复杂的词网络结构,提出从网络结构和语义结构两个角度对网络进行简化。在篇章共现窗口方面,Lim[9]等对文本集中的词建立其文本向量,当词出现在文本中时,则向量中的对应元素值为非零,相反则为零,然后通过计算两个词的文本向量相似性,判断该词对是否存在关联关系,根据词对关系构建语义词网络。

2 事件关联建模

事件关联分析过程主要包括四个部分:(1)事件报道集预处理;(2)事件报道集关键词提取;(3)事件关键词共现度计算;(4)事件共有词关联度计算。构建事件关联分析流程如图 2 所示。

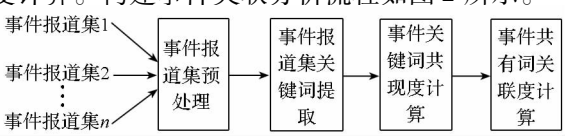

图 2 事件关联建模流程

Fig. 2 Modeling process of event association

2.1　事件关键词提取

事件关联分析首先提取事件报道集中的关键词。提取关键词首先要对事件中的文本进行分词(中文文本)、相邻词合并、去停用词处理,得到事件包含的词语集;然后通过计算词语权重提取关键词。

首先利用中科院开发的 ICTCLAS[10] 软件进行中文分词。分词结果中每个词 t 都会出现前后相邻字词(句首和句尾除外),即 t 之前的一个词和之后的一个词。例如对"陆委会主委蔡英文指示三项管理原则来规划开放两岸小三通"分词后得到"陆/m　委/ng　会/v　主/n　委/vg　蔡/nr　英文/nz　指示/n　三/m　项/q　管理/vn　原则/n　来/f　规划/v　开放/v　两岸/n　小/a　三/m　通/q",那么"英文"的相邻词即为:"蔡"(前相邻词)和"指示"(后相邻词)。分词系统经常将具有完整语义的词组分成多个词(字),例如句中的"陆委会""蔡英文""小三通"等,这样使得词的语义丢失、不完整或引发歧义,不利于提取到具备完整语义的关键词。因此,对于分词系统产生的分词结果,利用文献[4-6]中提出的频繁相邻词(字)合并算法对结果进行优化,尽量使每个词均具备较完整的语义。对优化后的分词结果利用文献[11]的停用词表进行停用词过滤。

关于单篇文本关键词提取的方法比较多,但针对文本集提取关键词的方法不多见。本文利用 TF-IEF 模型[1] 计算事件相关报道集的词语权重,并将词语按权重值大小进行排序,选取前 Top-N 个词语(不足 N 个词的取实际数目,N 大小一般不超过 20)作为事件的关键词。TF-IEF 的核心思想:当一个词在一个事件报道集中频繁出现,而在其他事件报道集中很少出现,则认为该词具有很好的事件区分能力。TF-IEF 模型将事件作为词语权重计算的基本单元,直接计算词语表征事件主题的能力大小。

设 $E=\{E_i \mid i=1,2,\cdots,l\}$表示一个事件集合,$\omega(t_k,E_i)$表示第 i 个事件报道集中第 k 个词的权重值,$E_i=\{(t_k,\omega(t_k,E_i)) \mid k=1,2,\cdots,m\}$表示事件词向量 E_i 的 m 个词及其权重,TF-IEF 计算事件词权重如公式(1)、(2)所示:

$$\omega(t_k,E_i)=\frac{[1+\log_2 TF(t_k,E_i)]\cdot IEF(t_k)}{\sqrt{\sum_{k=1}^{m}\{[1+\log_2 TF(t_k,E_i)]\cdot IEF(t_k)\}^2}} \tag{1}$$

$$IEF(t_k)=\log_2\frac{|E|+1}{|EF(t_k)|+0.5} \tag{2}$$

其中,$TF(t_k,E_i)$是词 t_k 在事件 E_i 包含的各报道中出现总次数,$|EF(t_k)|$是出现词 t_k 的事件数;$|E|$是总的事件数。

2.2　事件关键词关联建模

事件关键词关联可以通过词共现模型进行度量、分析。词共现模型认为:若某两个词经常在同一窗口单元中共同出现,则它们在一定程度上共同表达某种语义信息,并由此进一步构成某一个主题思想。目前词共现模型基本只采用单一共现窗口的统计量进行评估,精度较低,抽取结果还需要人工检查。

从图 1 的事件与词语知识层级关系可以看出,事件的关键词共现主要有四种窗口单元:(1)句子共现;(2)段落共现;(3)报道共现;(4)事件共现。单一研究句子共现模型[4-7]和文本共现模型[8]比较多,而综合多种词共现窗口的共现模型不多见,因此本文综合考虑多种共现窗口,提出一种新的词共现模型,构建事件关键词关联关系。

本文组合句共现、段落共现及报道共现三种窗口对报道中的词对进行关联分析,构建单篇报道中关键词对关联模型如下:

$$w_{d_i}(t_{ij},t_{ik})=Aw_{sen}(t_{ij},t_{ik})+Bw_{par}(t_{ij},t_{ik})+Cw_{doc}(t_{ij},t_{ik}) \tag{3}$$

其中,A、B、C 均为系数,值域为$\{0,1\}$,$w_{d_i}(t_{ij},t_{ik})$表示词语 t_{ij}和 t_{ik}在报道 d_i 中的关联权重;$w_{sen}(t_{ij},t_{ik})$、$w_{par}(t_{ij},t_{ik})$、$w_{doc}(t_{ij},t_{ik})$分别表示词语 t_{ij}和 t_{ik}在句子、段落、报道中的共现权重;

这种共现模型充分考虑了三种共现情况,但不重复考虑共现。即词对在句中共现,则 A 等于 1,B 和 C 均等于 0;词对只在段落和报道共现时,则 B 等于 1,A 和 C 等于 0;词对在句中和段落中均不共现,考虑报道共现,相应 C 等于 1,A 和 B 等于 0。这样定义模型主要是考虑:句中共现的语义关联最紧密,段落共现语义关联相对次之,篇章共现的语义关联最弱。

三种共现的词对计算公式分别为:

$$w_{sen}(t_{ij},t_{ik})=\frac{2\times f_{p\_tit}(t_{ij},t_{ik})\times f_s(t_{ij},t_{ik})}{f_s(t_{ij})+f_s(t_{ik})}\times\ln(1.72+|S_i|) \tag{4}$$

$$w_{par}(t_{ij},t_{ik})=\frac{2\times f_p(t_{ij},t_{ik})}{f_p(t_{ij})+f_p(t_{ik})}\times\ln(1.72+|P_i|) \tag{5}$$

$$w_{doc}(t_{ij},t_{ik})=\frac{f_{d_i}(t_{ij})}{f_E(t_j)}\times\frac{f_{d_i}(t_{ik})}{f_E(t_k)} \tag{6}$$

$$f_{p\_tit}(t_{ij},t_{ik}) = \begin{cases} 1+f_{tit}(t_{ij},t_{ik}) & t_{ij}\text{和}t_{ik}\text{在标题中共现} \\ 1 & t_{ij}\text{和}t_{ik}\text{不在标题中共现} \end{cases} \tag{7}$$

$$f_E(t_j) = \sum_{d_i \in E} f_{d_i}(t_j) \tag{8}$$

$$f_E(t_k) = \sum_{d_i \in E} f_{d_i}(t_k) \tag{9}$$

其中，$f_{tit}(t_{ij},t_{ik})$表示词语 t_{ij}和 t_{ik}在标题中的共现次数；$f_s(t_{ij},t_{ik})$表示词语 t_{ij}和 t_{ik}共现的句子数；$f_s(t_{ij})$、$f_s(t_{ik})$表示词语 t_{ij}、t_{ik}分别出现的句子数；$f_p(t_{ij},t_{ik})$表示词语 t_{ij}和 t_{ik}共现的段落数；$f_p(t_{ij})$、$f_p(t_{ik})$表示词语 t_{ij}、t_{ik}分别出现的段落数；$f_{d_i}(t_{ij})$、$f_{d_i}(t_{ik})$表示词语 t_{ij}、t_{ik}在报道 d_i 中分别出现的次数；$f_E(t_j)$、$f_E(t_k)$表示词语 t_j、t_k 在事件 E 中分别出现的次数；$|S_i|$表示报道 d_i 中包含的句子数；$|P_i|$表示报道 d_i 中包含的段落数；$\ln(1.72+|S_i|)$和 $\ln(1.72+|P_i|)$是对长报道中的词关联权重进行补偿[4]，因为相对于短报道，长报道将弱化词关联权重，所以需 $\ln(1.72+|S_i|)\geqslant 1$、$\ln(1.72+|P_i|)\geqslant 1$，因此真数部分加 1.72。公式(4)对标题中的词对共现赋予较高的权重，因为新闻标题信息含金量高，标题能够概略反映报道的主题。

在计算完单篇报道中的关键词共现权重后，对事件的整个报道集中所有关键词对的关联权重进行计算，将关键词对在事件每篇文本中的关联权重进行求和，并用“词单位数”和词对的逆文本频率对其进行加权，公式如下：

$$w_E(t_j,t_k) = \frac{\log_2\left(l_{t_j}\times l_{t_k}\times \frac{n}{df_{t_{jk}}}\right)}{\log_2 n} \times \sum_{i=1}^{n} w_{d_i}(t_{ij},t_{ik}) \tag{10}$$

其中，$df_{t_{jk}}$为事件 E 中同时出现词 t_j 和词 t_k 的报道数(共现文本频率)，l_{t_j}为词 t_j 的词单位数，l_{t_k}为词 t_k 的词单位数，n 为事件 E 中包含的新闻报道数。

式(10)中的“词单位数”是指一个词所包含的最小的完整词义单位的个数[12]。例如：“国民党主席”的词单位数是 2(包含的词单位为“国民党”和“主席”)。对于大多数关键词来说，较长的词所陈述的语义更具体、清晰，专指性较好。同时，关键词对共现的文本频率越大，说明词对越普遍平常，关联意义不大。所以文中利用“词单位数”和词对的逆文本频率对其进行加权，词越长且文本频率越小，则词对关联权重越大。

计算完词对共现权重后，建立事件关键词关联矩阵 A，矩阵的大小是 $n\times n$，A 是一个对称矩阵；A 中元素 a_{ij}的值为非零与零，值为零表示词 t_i 与 t_j 不存在共现，否则存在；n 表示事件包含的关键词数，n 等于上节 $Top-N$ 中 N 的值。

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \tag{11}$$

其中，

$$a_{ij} = \begin{cases} w_E(t_i,t_j) & i\neq j \\ 1 & i=j \end{cases} \tag{12}$$

新闻文本中段落数小于句子数，根据式(3)、(10)、(11)可知关联矩阵 A 的计算复杂度为 $O(n^2 m_{avg} s_{avg})$，其中 n 为关键词数，m_{avg}为事件包含的平均文本数，s_{avg}为文本包含的平均句子数。因为本文只针对新闻报道，报道的长度一般都比较小，且事件报道经过去重处理后包含的数量不会很多，所以 m_{avg}和 s_{avg}均不会很大，所以关联矩阵的计算复杂度主要看 n 的大小，为保证计算效率，Top－N 中 N 取值不宜过大。

2.3 事件关联建模

本文建立事件联系主要是通过关联词来实现的。当两个事件中存在相同的词(关联词)时，则认为两个事件存在关联，并且认为关联词越多以及关联词与事件关联权重越大，则事件关联度越大。因此通过关联词建立事件关联模型：

$$w(E_i,E_j) = \sum_{t_l \in T_l} w(t_l,E_i) \times w(t_l,E_j) \tag{13}$$

$$w(t_l,E_i) = \sum_{t_k \in E_i, l\neq k} w_{E_i}(t_l,t_k) = \sum_{j=1,j\neq l}^{n} a_{lj} \tag{14}$$

其中，$w(E_i,E_j)$表示事件 E_i 与 E_j 的关联权重，T_l 是事件 E_i 与 E_j 之间的关联词集。$w(t_l,E_i)$表示关联词 t_l 与事件 E_i 的关联权重。

根据公式(13)，构建事件关联矩阵 R，矩阵大小是 $m\times m$。R 中元素 r_{ij}表示事件 E_i 与 E_j 间的关联权重；当 r_{ij}为零时，表示事件 E_i 与 E_j 间不存在关联，否则存在关联；m 表示总的事件数。

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix} \tag{15}$$

其中，

$$r_{ij} = w(E_i,E_j) \tag{16}$$

根据公式(13)和(15)，可知矩阵 R 的计算复杂度为 $O(|T_l|m^2 n)$，其中 $|T_l|$为事件 E_i 与 E_j 之

间的关联词数,因为 $|T_l|$ 一般比较小,而 n 大小由 Top – N 决定,所以矩阵 R 的计算复杂度主要看 m 的大小。为了提高效率,可以对事件时间进行加窗,只计算窗内事件间的关联度;同时也可以通过 $|T_l|$ 控制计算量,即设置一个阈值,只有当 $|T_l|$ 大于该阈值时才计算事件间关联度。

3　事件关联可视化

3.1　事件关键词关联网络

每个事件关键词集代表一个主题,通过事件关键词关联模型计算得到词关联矩阵,利用图模型对关联矩阵进行展示,构建事件关键词关联网络。

定义 1(事件关键词关联网络):事件关键词关联网络 G 表示为二元组:$\langle T, A\rangle$,其中 $T = \{t_i\}_n$ 为事件中的 n 个关键词集合,t_i 表示对应于节点 i 的一个关键词;$A = \{a_{ij}\}_{n\times n}$ 为事件关键词关联矩阵,由 2.2 节计算得到,如果 $a_{ij} \neq 0$,则表示节点 i 与 j 存在邻接边;如果 $a_{ij} = 0$,则表明 t_i 与 t_j 之间缺乏足够的联系。

事件关键词关联网络示例如图 3 所示,每个实心圆点代表一个事件关键词,点与点之间的连线表示两个关键词之间的关联关系,边的权重即为 a_{ij} 对应的值。示例中展示的所有关键词都建立了直接或间接的关联,而在实际应用中,一个事件的关键词关联网络也可能存在相互不连通的几个子图。

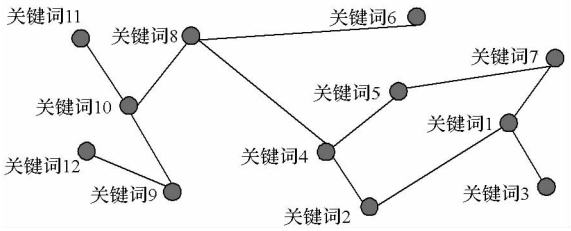

图 3　新闻事件关键词网络示例

Fig. 3　Sample of keyphrase network for news event

3.2　事件关联网络

每个关键词集对应一个事件主题,而不同事件间存在共有关键词,即事件关联词,这些词具有连接不同事件的重要作用,所以利用关联刻画不同事件间的关联特征,为不同事件的关键词网络建立连接,从而形成事件关联网络。

定义 2(事件关联网络):事件关联网络 W 表示为三元组:$\langle G_{all}, R, L\rangle$,其中 $G_{all} = \{G_i\}_m$ 为所有事件关键词网络集合,G_i 对应于一个事件关键词关联网络;$R = \{r_{ij}\}_{m\times m}$ 为事件关联矩阵,如果 $r_{ij} \neq 0$,则表示事件 E_i 与 E_j 存在关联,r_{ij} 的大小表示事件关联权重;如果 $r_{ij} = 0$,则表明事件 E_i 与 E_j 之间不存在联系;$L = \{L_{ij}\}$ 为构建事件关联的关联词集,L_{ij} 表示事件 E_i 与 E_j 之间的共有关键词集,即关联词集。

事件关联网络示例如图 4 所示,其中实心圆点表示事件特有关键词,空心圆点表示事件间共有的关联词。

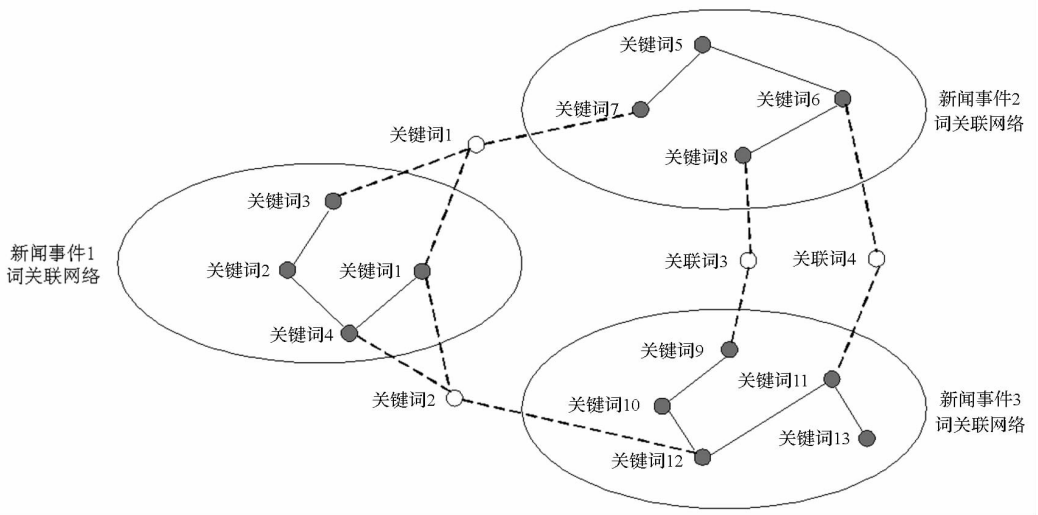

图 4　新闻事件关联网络示例

Fig. 4　Sample of news events relationship network

4　实验结果及分析

4.1　实验评价方法

实验包括两部分。第一部分,使用本文的事件关键词提取方法进行实验;第二部分,使用本文关键词关联模型进行实验,计算词关联关系。

实验采用准确率(Precision Rate,P)、召回率(Recall Rate,R)、F – 度量(F-measure,F)来进行

评价。对于一个事件 E,设答案集合(由标注人员抽取出的关键词集合或共现词对集合)为 K_E,本文方法得到的关键词集合(共现词对集合)为 K'_E,假设关键词集合(共现词对集合)K''_E表示 K_E 与K'_E的公共部分,公式为:

$$K''_E = K_E \cap K'_E \tag{19}$$

则准确率 P 就是本文方法提取的正确的关键词数(共现词对数)与提取到的总的关键词数(共现词对数)的比值,公式为:

$$P_E = |K''_E| / |K'_E| \tag{20}$$

则召回率 R 就是本文方法提取的正确的关键词数(共现词对数)与答案集的关键词数(共现词对数)的比值,公式为:

$$R_E = |K''_E| / |K_E| \tag{21}$$

则 F - 度量为:

$$F_E = \frac{2 \times P \times R}{P + R} = \frac{2K''_E}{K_E + K'_E} \tag{22}$$

4.2 实验数据

实验数据主要有两组,第一组(Data1)是从中国期刊网(http://www.cnki.net)下载的 108 篇学术论文的摘要,这些论文均是利用主题词检索,从检索到的结果中随机选择。论文涉及主题及篇数:"数据流聚类"10 篇,"话题探测与跟踪"5 篇,"命名实体识别"12 篇,"人脸检测、人脸识别"23 篇,"图像分割"25 篇,"视频对象跟踪"18 篇,"数据降维"6 篇,"步态识别"9 篇。它们分别来自《计算机学报》《自动化学报》《软件学报》《电子与信息学报》《中国图象图形学报》《中文信息学报》《模式识别与人工智能》。

第二组实验数据(Data2)是利用网络爬虫工具从中国新闻网国际专栏中采集的 2011 年 3 月 11 日至 2011 年 5 月 11 日部分新闻报道,共计 1982 篇。通过人工方式对采集的报道进行事件识别,选取其中 6 个事件共 23 篇相关报道进行实验,如表 1 所示。

表 1 事件列表

Tab. 1 The list of events

事件	事件概要	报道数
E_1	日本大地震	5
E_2	俄罗斯增加对日本天然气供应	3
E_3	墨西哥现巨型鱼群	2
E_4	切尔诺贝利核泄漏	4
E_5	七成泰国人反对兴建核电站	3
E_6	日本地壳移动	6

4.3 关键词提取实验

关键词提取实验时以 Data1 中论文的关键词集为答案集,以 TF-IDF[13-14] 作为实验的基准方法,使本文方法与之进行对比。实验仅抽取摘要中的关键词,答案集中将摘要中没有出现的关键词去除。

实验时,抽取前 Top - N 个关键词,其中 N 分别从 6 取到 10,共计进行五组基于 TF-IEF 和基于 TF-IDF 方法的比较实验,根据实验结果计算各个主题抽取关键词的准确率(召回率),并求和平均得到平均准确率 P_{av}(平均召回率 R_{av})。用 P_{av}和 R_{av}绘制实验结果对比曲线如图 5 所示。

从图 5 中可以看出,本文的报道集关键词提取方法要明显优于基于 TF-IDF 的基准方法。特别是当关键词数为 6 时,本文方法的 P_{av}和 R_{av}分别为 77.4% 和 60.1%,而 TF-IDF 方法为 41% 和 35.6%,提取效果提升了将近一倍。

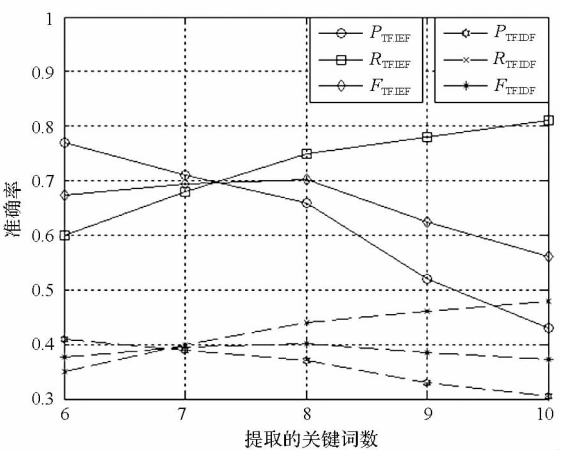

图 5 不同词数的关键词抽取性能比较

Fig. 5 Performance of keyword extraction

4.4 事件关键词关联实验

事件关联研究是基于事件关联词分析进行的,关联词分析的准确性与否直接关系到事件关联的准确性,所以本文对事件关键词关联算法性能进行实验对比,以说明事件关联的准确性。关键词关联实验在 Data2 数据集上进行,首先根据本文的关键词提取方法获取各事件报道集的关键词集,然后依据事件关键词关联模型对关键词关联关系进行计算,进而获得关键词关联矩阵,并统计出不同事件的共有关联词,计算关联词关联事件的权重,依此构建关键词关联网络。文中选择文献[4]中性能最好的 LAR Model + TScore 方法及文献[4]中的方法作为本文关键词共现计算的比较基准。将本文的词共现(关联)方法与文献[15]和[4]方法进行比较实验,比较结果如表 2

所示。通过比较可以看出,本文方法的性能指标均优于参与比照的两种方法。同时,本文方法提取的共现词对数最多,其召回率也最高,说明其能较充分发现词对关联关系。

表 2　词对关联实验比较

Tab. 2　The contrastive of different term co-occurrence method

方法	共现词对数	P_{av}(%)	R_{av}(%)	F
本文	287	84.9	86.3	0.86
文献[15]	258	81.4	82.1	0.82
文献[4]	269	76.2	72.5	0.74

4.5　事件关联网络实例

以数据 Data2 的事件关键词关联实验结果为基础,利用 Netdraw 软件对其部分关键词关联关系进行可视化展示,构建事件关键词关联网络如图 6 所示。图中圆圈外带黑方框的是事件关联词,其他不同形状的节点图标表示不同事件关键词,每一类图标表示一个事件关键词集。如果单从表 1 的事件列表出发,可能很难发现事件间的一些联系,如泰国民众反对兴建核电站与日本地震的关系,墨西哥出现的巨型鱼群与日本地震间的关系。但是通过对底层词语的关联分析,发现这些事件间是存在各种联系的,这也进一步验证本文方法对挖掘事件关联的有效性。

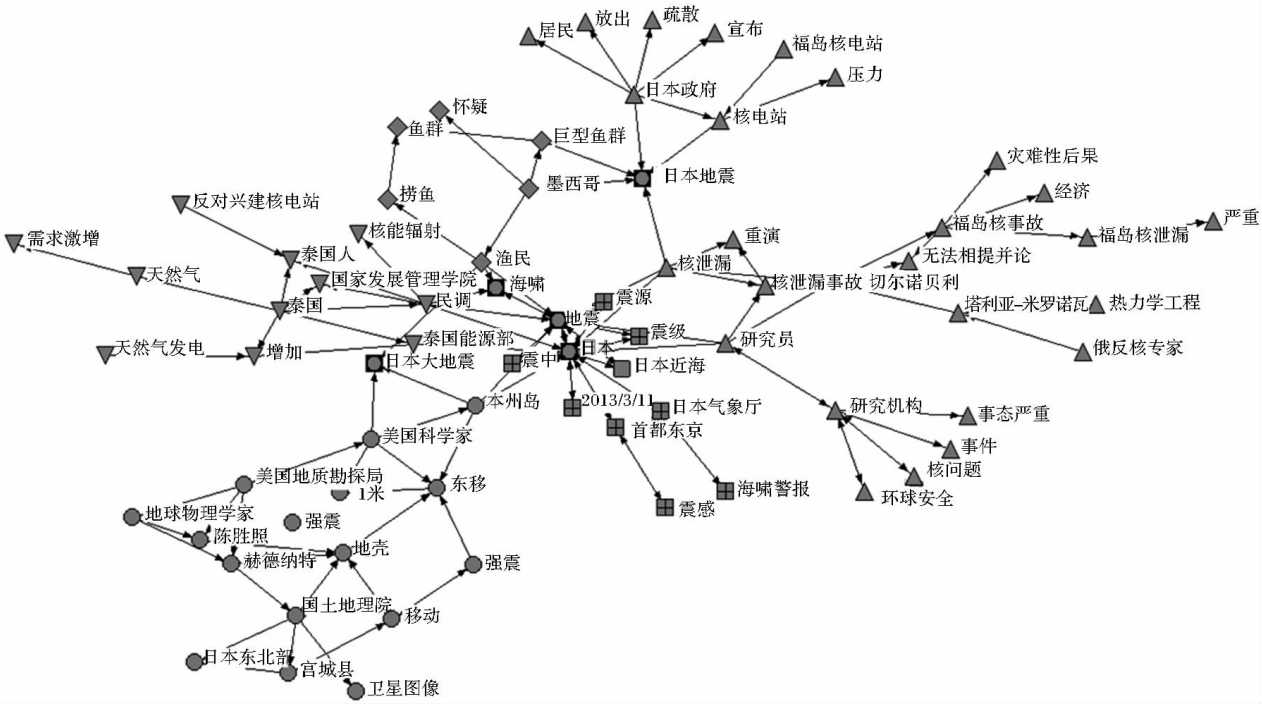

图 6　事件关联网络实例

Fig. 6　Example of news events relationship network

5　结论

网络媒体每天报道了诸多新闻事件,很多可能隶属于不同的主题,从表面上看,貌似发现不了它们之间有任何联系,但实际上它们存在各种形式的关联。为了有效挖掘出新闻事件间的各种关联知识,本文从词语语义层出发,尝试通过挖掘事件各关键词间的语义关联,进而发现事件间的语义关联。本文提出一种事件报道集关键词提取方法,然后基于词共现思想对提取的关键词对建立关联模型,最后在事件关键词关联模型的基础上,通过统计、计算事件关联词与不同事件的关联权重,从而构建事件关联模型。通过实验及示例,本文方法能够有效挖掘出事件间的关联关系,发现潜在的知识、模式。

参考文献(References)

[1] 张辉, 李国辉, 贾立等. 一种基于 TF · IEF 模型的在线新闻事件探测方法[J]. 国防科技大学学报, 2013, 35(3):55-60.
ZHANG Hui, LI Guohui, JIA Li, et al. On-line news event detection based on TF · IEF model[J]. Journal of National University of Defense Technology, 2013, 35(3):55-60. (in Chinese)

[2] Clive B. Web mining for open source intelligence[C]// Proceedings of the 12th International Conference on Information Visualization. Washington: IEEE Computer Society, 2008:321-325.

[3] Jacobs N. Co-term network analysis as a means of describing

the information landscapes of knowledge communities across sectors[J]. Journal of Documentation, 2002, 58(5):548 -562.

[4] Yuen-Hsien Tseng, Zih-Ping Ho, Kai-Sheng Yang, et al. Mining term networks from text collections for crime investigation[J]. Expert Systems with Applications, 2012, 39:10082 -10090.

[5] Zih-Ping Ho, Yuen-Hsien Tseng, Kai-Sheng Yang, et al. Term mining for relation visualization and exploration-some practical applications in crime investigation[C]// Proceeding of The 3rd International Conference on Data Mining and Intelligent Information Technology Applications, 2011:1 -4.

[6] Yuen-Hsien Tseng, Chun-Yen Chang, Shu-Nu Chang Rundgren, et al. Mining concept maps from news stories for measuring civic scientific literacy in media [J]. Computer&Education, 2010, 55:165 -177.

[7] Perrin T, Kawai H, Kunieda K, et al. Global dynamics network construction from the web[C]// Proceeding of The International Workshop on Information-Explosion and Next Generation Search, 2008:69 -76.

[8] Kawai H, Kunieda K, Yamada K, et al. Visualization for statistical term network in newspaper [C]// Proceeding of NTCIR-7 Workshop Meeting, 2008:549 -554.

[9] Lim Cheon Choi, Kyung Ung Choi, Soon Cheol Park. An automatic semantic term-network construction system [C]// Proceeding of International Symposium on Computer Science and its Applications, 2008:48 -51.

[10] Zhang H P, LIU Q. Calculation of the Chinese lexical analysis system ICTCLAS [CP/OL]. Institute of Computing, Chinese Academy of Sciences, 2002. http://sewm. pku. edu. cn/QA/reference/ICTCLAS/FreeICTCLAS/

[11] 张晓燕. 新闻话题表示模型和关联追踪技术研究[D]. 长沙:国防科技大学, 2010.
ZHANG Xiaoyan. Research on the representation model and technologies of link detection and tracking on news topic [D]. Changsha: National University of Defense Technology, 2010. (in Chinese)

[12] 韩客松, 王永成. 一种用于主题提取的非线性加权算法[J]. 情报学报, 2000, 19(6):650 -653.
HAN Kesong, WANG Yongcheng. A non-linear term weighting method used for subject distillation[J]. Journal of The China Society for Scientific anf Technical Information, 2000, 19(6):650 -653. (in Chinese)

[13] LI J Z, FAN Q A, ZHANG K. Keyword extraction based on TF/IDF for Chinese news document[J]. Wuhan University Journal of Natural Sciences, 2007, 12(5):917 -921.

[14] 王灿辉, 张敏, 马少平, 等. 基于相邻词的中文关键词自动抽取[J]. 广西师范大学学报:自然科学版, 2007, 25(2): 161 -164.
WANG Canhui, ZHANG Min, MA Shaoping, et al. Chinese keyword extraction algorithm based on neighbour words[J]. Journal of Guangxi Normal University: Natural Science Edition, 2007, 25(2):161 -164. (in Chinese)

[15] 郭锋, 李绍滋, 周昌乐, 等. 基于词汇吸引与排斥模型的共现词提取[J]. 中文信息学报, 2004, 18(6):16 -22.
GUO Feng, LI Shaozi, ZHOU Changle, et al. Co-occurrence word retrieval based on the lexical attraction and repulsion model[J]. Journal of Chinese Information Processing, 2004, 18(6):16 -22. (in Chinese)