一种基于形式概念分析的事件演化分析算法①

刘　艺[1]，张召乾[2]

[1](河南省农业科学院 图书馆，郑州 450002)
[2](解放军信息工程大学 信息工程学院，郑州 450002)

摘　要：针对传统基于内容相似度的事件关系计算方法不能分析出事件间的潜在关系的问题，提出了基于 FCA 的事件关系计算方法。该方法利用根据话题的三层结构模型，对话题中的事件进行属性提取，并依据特征频率因子进行属性选择。利用两个事件之间的属性关系建立形式背景，以此为基础形成概念格。用基于概念格的相似度分析发现事件之间潜在。实验证明了这种方法的有效性。

关键词：话题；事件；概念格；事件演化；形式概念

Event Evolution Algorithm Based on FCA

LIU Yi[1], ZHANG Zhao-Qian[2]

[1](Library of Henan Academy of Agricultural Science, Zhengzhou 450002,China)
[2](Information Engineering University, Zhengzhou 450002, China)

Abstract: Traditional event evolution analysis method can't find out the relationship between potential connected events in a topic. We proposed a new method based on FCA. This method uses the idea of topic three layer model to extract the properties in the event and filter the useless properties in the light of PFF(Property Frequency Factor). It constructs the Formal Context with property relation between events and analyzes the event evolution relation with similarities algorithm based on concept lattice. Practical examples show our method is effective.

Key words: topic; event; concept lattice; event evolution; formal concept

1　引言

互联网信息的不断膨胀使得用户关注的某些方面的信息往往孤立地分散在很多不同的地方，并且出现在不同的时间，用户难以获取其感兴趣的信息，话题识别与跟踪(TDT)技术就是在这种情况下应运而生的。话题追踪是 TDT 的一个子任务，能够帮助人们把分散的信息有效地汇集并组织起来，从大体上了解一个话题的全部信息[1]。

话题是由在某个时间，某个地点，由某些人或组织参与的某项事件及所有与该事件有关系的事件组成的[2]。因此，只要找出话题中各个事件的发展演化关系，就能清楚地了解事件的来龙去脉。针对传统基于内容相似度的事件关系计算方法不能分析出事件间的潜在关系的问题，本文提出一种基于 FCA 的事件关系计算方法，从话题的新闻报道中进行属性抽取，根据特征选择算法进行过滤，构成了事件－属性的对应关系，并根据概念格的层次关系分析事件间存在的潜在关系。实验证明该方法是有效的。

2　相关概念

话题：

话题是对一个核心事件或活动以及所有与之直接相关的事件和活动的概要描述，比如北京奥运会。

事件：

就是发生在某个时间，某个地点，由某个人参与的一项活动。它可能有一定的条件，也有可能产生一定的后果。同样以汶川地震为例，汶川发生地震就是一个事件。在这里，要同话题的定义区分开，话题包含一个事件及其相关的所有事件。汶川发生地震只是一个事件，它连同解放军进驻灾区，党和国家领导人

① 收稿时间:2010-11-02;收到修改稿时间:2010-12-06

亲赴灾区，国际社会提供救援等一系列事件构成汶川地震话题。

子话题：

子话题是指话题的某个方面，以伊拉克战争为例，对战争的影响的报道，伤亡情况的报道，以及美英局势的报道都是伊拉克战争这个话题的子话题。话题 T 在 i 时刻的状态 T^i 是由若干个子话题组成，记为 $T^i=\{T_1^i,T_2^i,\cdots,T_n^i\}$，每个子话题 T_j^i 代表话题在 i 时刻的某一个方面。

3 话题的三层结构模型

图 1 描述的是抽象出来的话题三层结构模型图。图顶层的是事件层；事件层下面是子话题层，最下面是新闻报道层。这三类节点之间的层内与层间之间的关系可以形式化描述如下：

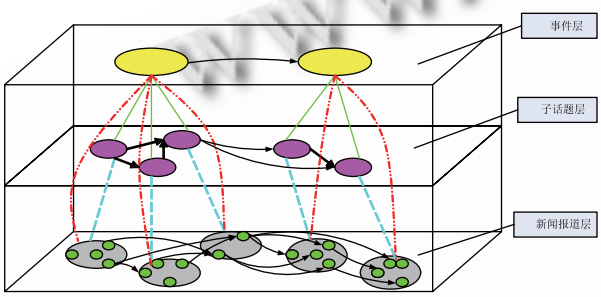

图 1 话题三层结构模型

以 $G(E,T,S,E_e,E_t,E_s)$ 表示这个三层结构模型，其中：

3.1 抽象节点

$E=\{E_1,E_2,\cdots,E_n\}$ 表示的是最顶层的新闻事件集合，集合中的每一个 E_i 都属于且只属于话题 T。每一个事件节点同时又可以划分为多个子话题节点，也就是说，每个事件中包含了多个子话题节点，图 1 中，深色的节点表示了这些子话题节点。

$T=\{T_1^1,T_2^1,\cdots,T_j^i\}$ 表示的是子话题集合，T_j^i 表示第 i 个时间段中的第 j 个子话题。任意的 T_j^i 属于且只属于一个 E_i。相邻时间段的，描述同一个事件的若干子话题构成一个事件 $E_i=\{T_i^1,T_j^2,\cdots,T_k^n\}$，图 1 中用灰色节点来表示这些子话题。

$S=\{S_1,S_2,\cdots,S_m\}$ 表示新闻报道集合，每个新闻报道属于且仅属于一个子话题。反过来，每一个子话题由一个时间段中讨论同一事件的新闻报道组成。图 1 中用最底层深色节点来表示新闻报道节点。

3.2 层内关系的层次化描述

E_e 表示事件节点内部关系形成的边集，用一个大小为 $n\times n$ 的矩阵 W_{ee} 来表述该关系集合。n 代表话题内的事件的个数，矩阵内任意元素 W_{ij} 表示第 i 个事件与第 j 个事件之间的关联关系。W_{ij} 值越大，关系越密切。

E_t 代表子话题节点内部形成的边集。用一个 $m\times m$ 的矩阵 W_{mm} 来描述该关系集合，m 代表子话题节点的个数，在本文中，我们只考虑相邻时间片中子话题的关系。同样的，矩阵内任意元素 W_{ij} 表示第 i 个子话题与第 j 个子话题之间的关联关系。W_{ij} 值越大，关系越密切。

E_s 表示新闻报道节点内部关系形成的边集，用一个大小为 $k\times k$ 的矩阵 W_{ss} 来表述该关系集合。k 代表话题内的报道的个数，矩阵内任意元素 W_{ij} 表示第 i 个报道与第 j 个报道之间的关联关系。W_{ij} 值越大，关系越密切。新闻报道之节点间的关系的计算可以通过余弦相似度求得。

3.3 层间关系的形式化描述

E_{E-T} 代表新闻事件层与子话题层节点之间关系形成的边集。该集合的每一个元素对应于某个新闻事件节点与某个子话题之间的从属关系。用一个大小为 $m\times n$ 的矩阵 W_{ET} 来描述该关系集合，m 代表子话题节点的个数，n 代表事件节点的个数，矩阵内任意元素 a_{ij} 表示第 i 个事件与第 j 个子话题之间的从属关系。a_{ij} 为 1 表示第 j 个子话题属于第 i 个事件，也就是说它是事件 i 在某个时间段内的状态。

E_{T-S} 代表新闻报道层与子话题层节点之间关系形成的边集。该集合的每一个元素对应于某个新闻报道节点与某个子话题之间的从属关系。用一个大小为 $m\times n$ 的矩阵 W_{TS} 来描述该关系集合，m 代表子话题节点的个数，n 代表新闻报道的个数，矩阵内任意元素 a_{ij} 表示第 i 个新闻报道与第 j 个子话题之间的从属关系。a_{ij} 为 1 表示第 j 个报道属于第 i 个子话题，也就是说它是的主要内容是关于话题的相关侧面的。

4 事件模型的建立

事件是发生在某时，某地，由某些人参与的一项活动。因此可以将事件表示为图 2 所示的多向量模型。

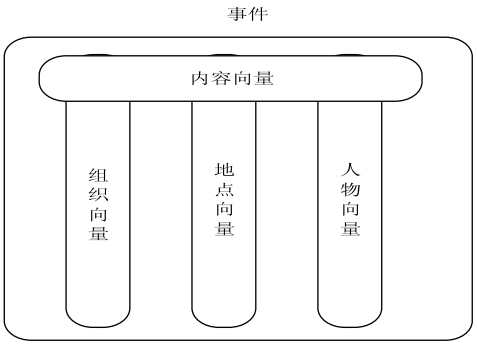

图 2　话题的多向量模型示意图

下面举例说明事件的多向量表示。对于发生在 2003 年 3 月 19 日的巴勒斯坦民族权力机构主席阿拉法特卸任，阿巴斯继任的事件，可以将其表示为

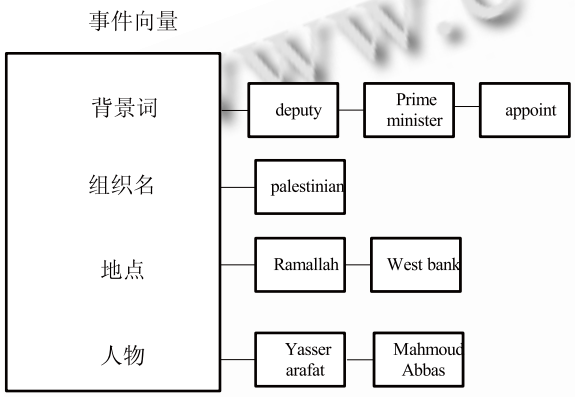

图 3　阿巴斯继任事件的多向量表示

现在，我们来形式化的描述事件与子话题间的关系。例如，现有一个话题 T_i 的子话题集合 $T_i=\{T_i^1,T_i^2,\cdots T_i^n\}$，那么话题中的事件集合 $E=\{E_1,E_2,\cdots E_m\}$ 必有如下约束：

$$\forall i\ E_i\in 2^C \tag{1}$$

$$\forall i,j\ s.t.i\neq j, E_i\cap E_j=\varnothing \tag{2}$$

$$\forall C_i\ \exists\ E_k\in E\ s.t. C_i\in E_k \tag{3}$$

第一条约束规定了每个事件是子话题的幂集中的一个元素。第二条约束规定了每个子话题 T_i 只能属于一个事件 E_j。最后一个则是说每个子话题属于事件集合 $E=\{E_1,E_2,\cdots E_m\}$ 中的一个事件。这样，我们可以定义一个从子话题到事件的映射：

$$f(C_i)=E_k\ iff\ C_i\in E_k \tag{4}$$

定义好了事件，下一步就要对表示事件之间关系的边进行定义，把事件之间的有向边集定义为 $\varepsilon=\{(E_i,E_j)\}$，两个事件之间的边表示它们之间存在关系。边的方向表示事件的因果或者是时序关系，因果关系是指 B 事件的发生是 A 事件发生的结果。而时序关系则是指 B 事件发生在 A 事件之后，但并不一定是 A 事件的必然结果。例如，在一个讨论飞机失事事故的话题中，飞机失事（事件 A）和对失事事故的调查（事件 B），显然事故调查是飞机失事的结果。因此，可以在 A 与 B 之间连一条从 A 指向 B 的有向边代表他们之间的因果关系。在考虑下面两个事件，温家宝访问英国（事件 A）与温家宝在剑桥演讲（事件 B），在这两个事件中，B 事件发生在 A 事件的后面，并且有一定的关系，但却并不是因果关系。这就是时序关系，区分这两种关系需要对事件高度的理解与丰富的经验，因此，本文中不去区分这两种关系，把他们统一称之为关系。

下面举例来说明事件之间的关系。以 TDT3 中的话题 30005，美国对 bin-Laden 的控告话题作为实例。这个话题中共有 23 篇新闻报道，他们分别来自 5 个事件，我们在图中表示他们之间的关系。图中每个节点代表对 bin-Laden 的控告话题中的一个事件，事件 2 对 bin-Laden 的审讯和控告的发生是事件 1 中央调查局搜集证据的结果。类似的，事件 3 伊斯兰国家的威胁，事件 4 来自穆斯林世界对此审判的反映和事件 5 中央调查局宣布对 bin-Laden 的悬赏这三个事件都是事件 2 的结果。

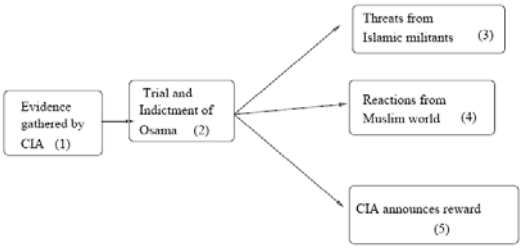

图 4　bin-Laden 控告话题内的事件关系

在一个新闻事件中，事件的地点、人物、时间等是描述事件内容的重要元素，事件演化关系的分析就要提取出事件的关键要素，展现事件的基本面貌。为了能够合理表示事件和事件之间的关系，我们将事件的属性简化为以下信息：

1) 人物集合：$P=\{p_1,p_2,\cdots,p_i,\cdots,p_m\}$；

2) 组织集合：$O=\{o_1,o_2,\cdots,o_i,\cdots,o_s\}$；

(3) 地点集合：$L=\{l_1,l_2,\cdots,l_i,\cdots,l_t\}$；

(4) 事件集合：$E=\{e_1,e_2,\cdots,e_i,\cdots,e_n\}$；

设$W=P\cup O\cup L$，则单个事件E_i的初始属性集合为W_i，$W_i\subseteq W$，事件e_i表示为ω维向量$e_i(p_{i1},\cdots,p_{im},o_{i1},\cdots,o_{is},l_{i1},\cdots l_{it}),\omega=|W_i|$，由此构成了事件的向量空间模型$V_E$：

$$V_E=\begin{pmatrix}e_1\\ \cdots\\ e_i\\ \cdots\\ e_n\end{pmatrix}=\begin{pmatrix}p_{11} & \cdots & p_{1m} & o_{11} & \cdots & o_{1s} & l_{11} & \cdots & l_{1t}\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ p_{i1} & \cdots & p_{im} & o_{i1} & \cdots & o_{is} & l_{i1} & \cdots & l_{it}\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ p_{n1} & \cdots & p_{nm} & o_{n1} & \cdots & o_{ns} & l_{n1} & \cdots & l_{nt}\end{pmatrix} \quad (5)$$

目前，命名实体识别技术[3,4]已经取得了很好的研究成果，为以上新闻要素的提取提供了技术保证，我们选择 Stanford 自然语言处理中心的研究成果[6]作为本文新闻要素提取的工具，从其发布的自然语言提取结果来看，英文命名实体识别任务的准确率和召回率均达到 92%以上。

5 属性的提取与选择

要想分析多个事件的关联程度，首先要获得获得单个事件以及与该事件相关的人物、地点和组织。

根据事件中所有的子话题和新闻报道，仍以“中美关系”话题为例，在话题中假定中美关系出现了一些新情况，比如“中美经济战略对话”、“中国导弹击毁气象卫星”、“台海局势”等，这些事件以及事件相关的一些内容是我们所关注的，因此信息资源的扩充是关键。

数据准备过程由以下步骤组成：

根据上一章得到的新闻报道，利用命名实体识别技术提取报道中的人名、组织名和地名，即得到与事件E_i相关的人物集合P_i、组织集合O_i和地点集合L_i，其并集$W_i=P_i\cup O_i\cup L_i$为所有 N 个返回文档中包含的人物、组织和地点的并集。

考虑到提取出得元素存在噪音的可能性，并不是集合P_i,O_i,L_i中所有元素都与事件相关，而这些不相关元素造成了计算开销的增大和结果的准确性降低，因此集合元素的选取需要考虑到无关信息的过滤问题。如果将集合P_i,O_i,L_i中的每个元素看作事件的特征，那么无关元素的过滤问题即转化为传统的特征选择问题。

特征选择用于去除某些不相关或不重要的特征，由此，引入“特征频率因子”度量特征的权重，特征频率因子表示每个集合中的元素在所有查询结果中出现的频率，数学描述为：

$$f(p_j)=\frac{N_{p_j}}{\sum_{i=1}^{m}N_{p_i}} \quad (6)$$

其中N_{p_j}表示人物p_j在返回结果中出现的次数。组织频率因子：

$$f(o_j)=\frac{N_{o_j}}{\sum_{k=1}^{s}N_{o_k}} \quad (7)$$

其中N_{o_j}表示组织o_j在返回结果中出现的次数。地点频率因子：

$$f(l_j)=\frac{N_{l_j}}{\sum_{k=1}^{t}N_{l_k}} \quad (8)$$

其中N_{l_j}表示地点l_j在返回结果中出现的次数。

引入特征频率因子后，集合W_i中的元素描述为(w_i,g_i)，其中w_i代表某个标识词，g_i代表该词对应的特征频率因子。

最后由特征选择阈值$\varphi_p,\varphi_o,\varphi_l$决定某标识词是否作为事件$E_i$的特征，去掉无关信息后得到初始事件要素集合$W_i$的子集$W_i'$，以此作为事件$E_i$的属性集合。

新闻要素提取后，对于事件E_i，其描述特征为提取后的新闻要素，由人物、地点、组织构成，表示为N_i维向量$E_i(p_{i1},\cdots,p_{im},o_{i1},\cdots,o_{is},l_{i1},\cdots l_{it}),N_i=|W_i'|$，如果将向量中$N_i$的个参数看作属性，则构成了事件－属性的对应关系，以“中美经济战略对话”为例，通过上述信息检索和要素提取过程，我们得到了与该事件相关的事件－属性集：

E_i(胡锦涛,布什,保尔森,温家宝,…,吴仪,中美商贸联委会,中美经济联委会,…,中美科技联委会,中国,美国,…,北京)

6 事件关系分析与实验

对于事件的相关关系，利用概念格的特性可以找到其中隐含的事件相关性。以中美军事问题为例，近年来，与中美军事问题相关的事件有很多，如“中国威胁论”、“朝核危机”、“中国导弹击毁气象卫星”、“中国潜艇跟踪美航母”、“台海局势突变”、“美国新联盟战略”等相关事件。这些事件在关键词上几乎不存在

交集，但并不意味着事件间不存在隐含的关系和含义。

在形式概念分析中，我们从事件的角度分析上述事件与地域范围间隐含的关系，以地点作为 FCA 中的对象集合，以事件作为 FCA 中的属性集合，则数据获取阶段得到的对象与属性的关系如图 5。

	中国潜艇跟踪美航母	中国威胁论	中国导弹击毁气象卫星	朝核危机	台海局势突变	美国新联盟战略
中国	×	×	×	×	×	×
美国	×	×	×	×	×	×
台湾	×	×			×	×
加拿大			×			
澳洲			×			
日本		×	×	×	×	×
英国			×			×
韩国				×		
朝鲜				×	×	

图 5　中美军事相关事件的对象－属性关系

生成概念格如图 6 所示：

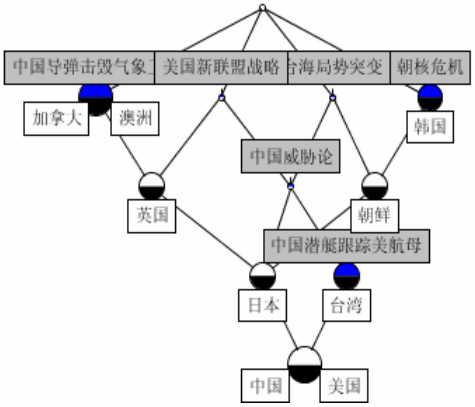

图 6　中美军事相关事件概念格

在中美军事相关事件的概念格中，不同的事件以属性的形式描述其发生的地域范围，概念格表达了现实情况中的很多隐含的关系：

1）从概念格分析结果来看，很多看似不相关的事件实际上都涉及到台湾这个地点，进一步我们可以猜测出这些事件存在与“台湾问题”的相关性。

2）实验中选择的多个事件在话题中涉及的范围有所不同，比如“中国潜艇跟踪美航目”是一个具体的事件，而“美国新联盟战略”则属于一个时间跨度大、地域范围广的事件，事件与事件之间存在潜在的包含关系，从概念格的分析结果来看，这种潜在的包含关系可以通过概念格的层次关系表达出来。

3）从表面上看“朝核危机”与“台海局势突变”似乎没有什么关系，而从概念格中分析可以得出其在事件发生的地域范围上都涵盖了朝鲜，实际上“台海局势”和“朝核危机”都属于美国为突变积极准备的内容。

7　结语

到目前为止，话题中的事件关系计算方法都是基于报道内容的文本相似度计算。在发现事件间存在的潜在关系方面表现比较差。本文提出一种基于 FCA 的事件关系计算方法，从话题的新闻报道中进行属性抽取，根据特征选择算法进行过滤，构成了事件－属性的对应关系，并根据概念格的层次关系分析事件间存在的潜在关系。随着形式概念及事件演化分析技术的迅速发展，事件关联挖掘将会有更大的发展空间。

参考文献

1 Allan J. Topic Detection and Tracking: Event-Based Information Organization. USA: Kluwer Academic Publishers, 2002.1–16.

2 Ault TG, Yang YM. Information Filtering in TREC-9 and TDT-3: A Comparative Analysis Information Retrieval, 2002,(5):159–187.

3 Zhou GD, Su J. Named entity recognition using an HMM-based chunk tagger. Annual Meeting of the ACL. Proc. of the 40th Annual Meeting on Association for Computation Linguistics. Philadelphia, Pennsylvania, 2001. 473–480.

4 McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and Web. Proc. of the 7th Conference on Natural Language Learning, 2003,4:188–191.

5 Dasarathy BV. Nearest Neighbor (NN) Norms: NN Patern Classification Techniques, IEEE Computer Society Press: Las Alamitos, California, 1991.

6 The Stanford Natural Language Processing Group. 2006-08, http://nlp.stanford.edu/ ner/index.shtml