基于超图的汉越双语新闻话题要素提取*

涂子令，周 枫，余正涛，严 馨，洪旭东
（昆明理工大学 信息工程与自动化学院，昆明 650500）

摘 要： 针对汉越双语新闻话题文本集合中新闻话题要素提取的问题进行了研究，在超图模型的基础上，运用了 PageRank 随机游走排序方法。首先根据触发词激励的方法提取新闻中的事件要素；然后在此基础上构建话题超图模型，将汉越事件要素作为节点，将文本集合中的句子作为超边，根据概率评估函数计算节点和超边的初始权重；最后采用 PageRank 随机游走方法对汉越事件要素进行评分，最终得到汉越话题要素。实验结果表明，该方法相比只考虑单文本事件要素提取方法的效果有显著提高，说明了基于超图的 PageRank 方法提取新闻话题要素的准确性。

关键词： 汉越双语；事件要素；超图；随机游走；话题要素
中图分类号： TP391.1 **文献标志码：** A **文章编号：** 1001-3695(2017)08-2278-04
doi:10.3969/j.issn.1001-3695.2017.08.008

Extraction of news topic elements for Chinese Vietnamese bilingual based on hypergraph

Tu Ziling, Zhou Feng, Yu Zhengtao, Yan Xin, Hong Xudong
(*School of Information Engineering & Automation, Kunming University of Science & Technology, Kunming* 650500, *China*)

Abstract: This paper studied the problems of news topic elements in the Chinese and Vietnamese bilingual news topic text collections. Based on hypergraph model extracted, it used the PageRank random walk ordering method. First according to the trigger word incentive method, it extracted the news event elements, and then on the basis of this, it constructed topic hypergraph model. It took the Chinese and Vietnamese elements as nodes and the sentences of text collection as a hyper-edge, it calculated the initial weights of nodes and hyperedges according to probability evaluation function. Finally, it used the PageRank random walk method to score the elements of the Chinese-Vietnamese event, and finally obtained the elements of the Chinese-Vietnamese topic. Results show that the proposed method can significantly improve the extraction performance compared to the method only considered single text event feature extraction. It shows the accuracy of extraction of news topicby PageRank method based on hypergraph elements.

Key words: Chinese and Vietnamese; event elements; hypergraph; random walk; topic elements

0 引言

越南与我国的交流日益密切，两国有很多共同的国际和地区热点问题，围绕同一问题会有很多分别用汉语和越南语表达的新闻文本。从汉越双语新闻话题文本集合中提取出与话题紧密相关的不同语言要素，对于了解新闻话题的全貌、对比分析不同国家对话题事件的不同看法等都具有重要作用。

话题是包含种子事件和一系列外延事件的集合，用汉越两种语言分别描述同一话题叫做汉越双语话题。事件要素是以单文本为对象进行提取的，无法对话题的文档集合进行整体的概括表达。但是可以先将单文本中的事件信息抽取出来，然后根据文本之间事件要素的联系在话题文本集合中提取出具有共性的事件要素作为新闻话题要素来精简地描述一个话题文本集合。基于汉越双语词典，可以找到意思相近的汉越双语事件要素，将汉越双语话题关联起来，最终通过超图模型提取出汉越话题要素。本文的主要目的是解决如何从汉越话题文本集合中提取出汉越话题要素。考虑到话题内汉越不同语言文本间存在事件要素互译、句子及文本间词语共现等多种关系，采用超图模型进行建模，提出了基于超图的汉越双语新闻话题要素提取方法。

汉越双语新闻话题要素的提取，主要是解决跨语言的话题表征问题。本文运用了基于超图的随机游走方法，首先通过触发词激励的方法得到汉越事件要素（对象、时间、地点、触发词）；再以汉越事件要素为节点、句子为超边来构建超图模型；然后根据概率评估函数得到节点和超边的初始权重；最后通过基于超图的随机游走迭代算法取排名较高的汉越事件要素为汉越新闻话题要素，汉越新闻话题要素具体包括与话题事件集合相关的人物（who）、时间（when）、地点（where）、触发词（what）等。

收稿日期：2016-05-23；修回日期：2016-07-11 基金项目：国家自然科学基金资助项目（61562049）

作者简介：涂子令（1991-），男，湖北孝感人，硕士研究生，主要研究方向为自然语言处理（514678798@qq.com）；周枫（1958-），男，云南昆明人，副教授，硕导，硕士，主要研究方向为数据挖掘、自然语言处理；余正涛（1970-），男，云南昭通人，教授，博导，博士，主要研究方向为自然语言处理、信息检索、机器翻译；严馨（1969-），女，云南昆明人，副教授，硕导，硕士，主要研究方向为数据挖掘；洪旭东（1989-），男，安徽马鞍山人，博士研究生，主要研究方向为信息检索、机器翻译.

1　相关工作

当前针对双语话题文本集合进行要素提取研究还相对较少,大多数相关研究针对的是单文本级别的要素提取问题。事件抽取是对单文本级别的要素提取,事件抽取主要有模式匹配和机器学习两种方法。模式匹配方法有: 梁晗等人[1] 提出了一种基于框架的信息抽取模式,利用框架的继承—归纳特性来提取事件信息; 冯礼[2] 使用预定义的事件框架进行新闻要素的信息抽取。由于这类方法依赖于具体领域,所以可移植性差。与模式匹配方法相比,机器学习方法的鲁棒性和灵活性比较好,且比较客观,不需要太多的人工干预和领域知识。所以基于机器学习的方法已成为目前的研究热点和主流。机器学习的方法有: 2002 年 Chieu 等人[3] 首次引入最大熵分类器,用于事件元素的识别; 2006 年 Ahn[4] 结合 Mega 和 Timbl 两种方法分别实现了事件抽取中事件类别和元素的识别这两个主要步骤。Ji、赵妍妍、张先飞等人[5~7] 提出的事件抽取方法也属于触发词激励的范畴。触发词激励的方法是目前人们用得比较多的,将每个词作为一个实例进行训练,然后判断是否为触发词,再根据触发词的位置来提取相关事件元素。由于事件抽取方法没有考虑话题文本集合中不同文档间的关联,以及无法表达汉越双语事件要素之间的关系,所以本文将在话题文本集合中,以汉越事件要素为节点、句子为超边,用超图表示汉越话题文本集合,通过 PageRank 随机游走算法得到汉越话题要素。

2　汉越新闻事件要素的建立

本文将已经得到的双语话题文档集合进行预处理,对于汉语新闻文本,用中国科学院 NLPIR 分词工具,对于越南语新闻文本,用昆明理工大学潘清清,构建的越南语分词平台[8] ,分别进行分词、去除停用词。然而文档集合里面的词太多,需要过滤掉一部分词语来简化后面的超图构建。因为话题是相关事件的集合,可以从词语集合中提取出事件要素对词语集合进行过滤。本文对于事件的定义来自于 ACE[9] ,根据定义,新闻事件由事件触发词(trigger) 和描述事件结构的要素(argument) 构成。图 1 表述了一个事件的构成。其中“出生”是该事件的触发词,事件的主要组成部分“阮富仲”“1944 年”“越南河内市”分别对应着该事件的三个要素,即对象、时间、地点。

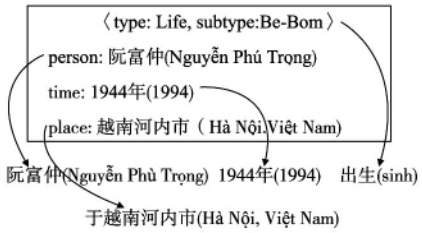

图1　“出生”事件的基本组成要素

在新闻文本中,需要抽取的元素是事件发生的对象、时间、地点和触发词。事件触发词一般是由动词和动名词构成。通过命名实体识别和词性标注,可以抽取到一个新闻页面中出现的动词和动名词作为事件触发词,选取事件触发词前后相邻的命名实体,包括人名、机构名、地点和时间作为事件发生的对象、地点和时间要素。因为本文是对汉越双语的研究,要把汉语话题新闻与越南语话题新闻相关联起来。由于汉越双语描述的是同一种话题,所以文本集合中的双语事件要素有很多意思是相近的,可以通过汉越双语词典标查找出意思相近的事件要素组成对来建立双语的对应关系。

3　基于超图的新闻话题要素提取算法

3.1　构建超图模型

本文定义 $HG(V,E)$ 为一个普通超图,$V=\{v_1,v_2,\cdots,v_n\}$ 代表的是一个节点的集合,$E=\{e_1,e_2,\cdots,e_m\}$ 代表是边的集合。一条超边 e 是 V 的一个子集,在这里可以表示为 $\cup_{e\in E}e=V$。为了用超图建模来表示汉越话题文档集合,用节点来表示汉越新闻事件要素,用超边来表示话题文档集合中的句子。构建的超图模型如图 2 所示。

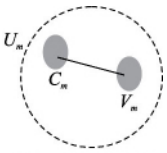

图2　超图节点

在图 2 中,表示的是超图节点 U_m 即为汉越新闻事件要素,其中 C_m 表示汉语事件要素,V_m 表示越南语事件要素。

在图 3 中,以汉越新闻事件要素表示超图的节点,以文本集合中的句子作为超边(e_1、e_2、e_3) ,其中句子 e_1 中包含有与 e_2、e_3 句子中一部分相同的汉越新闻事件要素,这样就将汉越双语的话题新闻文本集合用超图表示出来了。用汉越新闻事件要素来表示节点不仅可以让汉越双语关联起来,还可以增强事件要素的重要性。

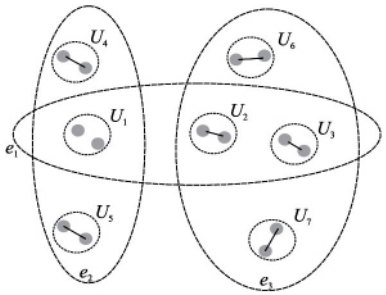

图3　超图模型

为了从图形描述转换为量的准确描述,需要对超图有基本公式的定义。本文对超图量的定义来自于 Wang 等人[10] 在文章中提到的超图模型定义。在考虑集合中超边的权重(即句子的权重) 时,一个有权无向超图可以定义为 $HG(V,E,\omega)$,在这里 ω: $E\rightarrow R^+$ 表示超边的权重。关联矩阵表示为 $H\in R^{|V|\times|E|}$,定义如下:

$$h(v,e)=\begin{cases}1 & v\in e\\ 0 & v\notin e\end{cases} \tag{1}$$

超图当中点的度定义如下:

$$d(v)=\sum_{e\in E}w(e)\,h(v,e) \tag{2}$$

式(1) 表示的关联矩阵并没有考虑节点的权重(即汉越事件要素的权重) ,应将点的权重考虑进来,超图的关联矩阵表示为 $W\in R^{|V|\times|E|}$,定义如下:

$$w(v,e)=\begin{cases}w(v_e) & v\in e\\ 0 & v\notin e\end{cases} \tag{3}$$

在考虑了节点权重的情况下,超图当中超边的度定义如下:

$$\delta(e)=\sum_{v\in e}w(v,e) \tag{4}$$

其中: D_v 和 D_{ve} 分别代表点和超边的度对角矩阵, W_e 代表的是边权重的对角矩。

3.2 节点和超边的权重计算

由于最终提取的是汉越新闻话题要素,计算节点的权重也就是计算汉越新闻话题要素的权重。话题是在一定的时间内发生的,随着时间变化,所以新闻话题要素也是跟时间相关的。这个特性在文档分布上面的表现就是新闻话题要素的频率在文档集合之间的波动变化比较大。样本方差跟检验样本和总体分布的波动程度成正比。根据上述分析可以提出汉越新闻话题要素权重的计算公式为

$$\text{weight}(w) = \frac{tf(w)}{df(w)} \times \sigma = \frac{tf(w)}{df(w)} \times \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(tf_i(w) - \overline{tf(w)})^2} \tag{5}$$

其中: w 表示汉越新闻事件要素; $tf(w)$ 表示 w 在整个语料中出现的频率; $df(w)$ 表示 w 出现的文档频率; $tf_i(w)$ 表示 w 中的事件要素分别在第 i 篇文档中的出现频率; $\overline{tf(w)}$ 表示在各个文档当中出现的平均频率。得到的汉越新闻事件要素权重也就是节点的权重 $w(v,e)$ 。在本文中,认为节点的权重在所有的超边中都是相同的。

对于超边权重计算的问题,可以通过节点的权重来计算得到超边的权重。通过上述节点权重的计算公式(5) ,已经得到了节点的权重 $w(v,e)$ 。Bellaachia 等人[11] 提出的超边计算公式如下:

$$W(e) = \frac{1 + \sum_{v\in e} w(v,e)}{(\sum_{v\in e} h(v,e))^2} \tag{6}$$

其中: $\sum_{v\in e} w(v,e)$ 代表的是句子 e 包含的话题信息之和,由于句子中词语越多,包含的话题信息也就越多,为了消除句子长度对句子权重计算带来的不公平性,用句子 e 包含的话题信息之和除以 $(\sum_{v\in e} h(v,e))^2$ 即句子的长度的平方。

3.3 基于超图的随机游走

本文中,在超图模型里面不仅考虑了超边的权重,而且还添加了节点的权重,是为了得到一个更加普适的随机游走算法。对随机游走过程的描述是这样的: 首先冲浪者从与当前点 u 关联的所有边中以 $w(e)$ 与包含 u 的所有边的权重之和 $\sum_{\hat{e}\in E(u)} w(\hat{e})$ 的比值作为概率选择一条边 e; 然后冲浪者以目标点 v 的权重 $w(v,e)$ 和边 $e \in E(u) \cap E(v)$ 中的所有节点权重之和的比值作为概率选择目标点 v。则从点 u 到点 v 的概率计算方法如下:

$$P(u,v) = \sum_{e\in E} w(e)\frac{h(u,e)}{\sum_{\hat{e}\in E(u)} w(\hat{e})}\frac{w(v,e)}{\sum_{\hat{v}\in e} w(\hat{v},e)} = \sum_{e\in E} w(e)\frac{h(u,e)}{\sum_{\hat{e}\in E(u)} w(\hat{e})}\frac{h(v,e)\times w(v_e)}{\sum_{\hat{v}\in e} w(\hat{v},e)} = \sum_{e\in E(u)\cap E(v)} \frac{w(e)}{\sum_{\hat{e}\in E(u)} w(\hat{e})}\frac{w(v_e)}{\sum_{\hat{v}\in e} w(\hat{v},e)} \tag{7}$$

矩阵表示形式为

$$P = D_v^{-1} H W_e D_{ve}^{-1} W^{\mathrm{T}}$$

其中: D_v 为式(4) 中的点度对角矩阵; H 为式(1) 中的关联矩阵; W_e 为边权重对角矩阵,对角线上的值为对应边的权重; D_{ve}^{-1} 是式(4) 中的超边度对角矩阵; W^{T} 为式(3) 中带有权重的关联矩阵 W 的转置矩阵。为了避免回路,把 P 当中的对角线元素置为 0,然后再把 P 归一化,使每一行元素之和为 1。

采用 PageRank 方法[12] 来进行随机游走, v 为待排序的点权重向量, α 为阻尼系数,具体计算方法如下:

$$v(i+1) = \alpha P^{\mathrm{T}} v(i) + (1-\alpha)\, e/n \tag{8}$$

其中: 阻尼系数 α 一般被赋值为 0.85; n 是超图当中节点的个数,即是文档中汉越新闻事件要素的个数; $e \in \mathrm{R}^{n\times 1}$ 是长度为 n 的单位向量; $\alpha P^{\mathrm{T}} v$ 表示冲浪者从当前点 u 选择一条边跳转到另外一个点, $(1-\alpha)\, e/n$ 表示冲浪者以 $(1-\alpha)/n$ 的概率跳转到任意的其他点。

在 PageRank 随机游走的迭代过程中,当 v 的值发生变化时,点的权重也就发生了变化,所以 D_v、D_{ve}、W_e 和 W 的值均将改变。因此刚开始使用式(5) 得到汉越新闻事件要素权重作为初始值,通过点的权重,可以得到 D_v、D_{ve}、W_e 和 W 的值,从而可以计算出 P 的值; 然后进行第一次迭代,迭代得到的值即汉越新闻话题要素的权重向量,可以重新计算出 D_v、D_{ve}、W_e 和 W 的值,从而得到 P 的值,进行下一次迭代。如此循环往复,直至相邻两次迭代 v_n 和 v_{n-1} 对应位置元素差值的绝对值小于某个固定阈值(设置为 0.0001) ,则停止迭代,同时得到了汉越新闻话题要素的最终得分,取得分最高的 K 个汉越事件要素为新闻话题要素。这里 K 的取值要适中,太大包含的无效词太多,太小则不能涵盖话题的主要信息。

4 实验结果与分析

4.1 实验数据

实验数据来源上选择了中国以及越南的主流媒体。313 篇中文新闻文本主要来源于《人民日报》、新华社、门户网站; 215 篇越南语新闻文本主要来源于越讯社、《越南人民日报》《越南外交报》《越南每日快讯》《越南政府》等这些比较主流的新闻媒体。

4.2 评价方法

实验先通过人工挑选出汉越新闻事件要素集合中的 20 个汉越事件要素作为汉越新闻话题要素提取的标准结果。由于有关新闻话题要素的提取研究比较少,所以用 Ahn[4] 提出的事件要素提取方法进行对比。在横向比较方面用汉越双语新闻话题要素提取的方法跟两种语言单独提取新闻话题要素的方法进行对比。对比所用指标包括平均的准确率、召回率和 F 值,具体如下所示:

准确率 = 正确识别的特征词总数/识别出的特征词总数

召回率 = 正确识别的特征词总数/测试集中存在的特征词总数

F 值 = 正确率 × 召回率 × 2/(准确率 + 召回率)

4.3 结果分析

在实验中,先要确定从已经排名好的汉越新闻事件要素中提取出多少个事件要素作为汉越新闻话题要素(即 K 的取值) ,在这里分别取值为 15、25、35 进行比较,然后取结果最好的那一组为 K 值。选取不同 K 值的实验结果如表 1 所示。

表 1 选取不同 K 值的实验结果

K 值	准确率	召回率	F 值
15	0.600	0.450	0.514
25	0.560	0.70	0.622
35	0.485	0.850	0.618

从表 1 的结果可以看出,当 K 值取 25 时实验结果是最好

的。因为当 K 值太小时,召回率太低; 当 K 值太大时,准确率太低; 当 K 值取 25 时,F 值是最高的,且对应的召回率、准确率都比较高。所以 K 值取 25 时为最佳。

在做对比实验时,用到提取事件要素的方法以及本文提出的方法(proposed) 用于在单语料(Chinese、Vietnam) 上进行,分别作为纵向的比较和横向的比较。

a) Ahn's method[4] ,是一种触发词激励的事件抽取方法,该方法最重要的一个步骤是判断句子中的每个词是否是最能描述某个事件发生的触发词,最终实现了事件抽取中事件类别和元素的识别。

b) proposed(Chinese) ,用汉语新闻事件要素来构建超图模型,然后用 PageRank 随机游走算法进行迭代计算,最终得到汉语新闻话题要素。

c) proposed(Vietnam) ,基于超图的 PageRank 随机游走方法用于相同的越南语话题文本集合当中,最终提取出新闻话题要素。

d) proposed,在汉越双语新闻话题文本集合当中,用汉越事件要素来构建超图,用基于超图的 PageRank 随机游走方法得到汉越新闻话题要素。

表 2 为新闻话题要素提取评价结果。从表 2 的对比实验结果可看出,双语语料下,本文提出的在基于超图的 PageRank 随机游走迭代算法要优于其他方法,证明了此方法的有效性。Ahn 提出的是一种触发词激励的事件抽取方法,该方法只是在单新闻文本中提取出事件要素,对话题的整体性覆盖不足。用越南语做语料来进行实验时,效果明显要比中文做语料进行实验的效果差,因为越南语在文本预处理阶段获取新闻事件要素的时候误差比较大。然而汉越双语做语料时的效果比中文单独做语料时效果要好,是因为汉越双语事件要素在计算重要词语权重时有一个相互加强的作用,比单独提取的效率要高。

表 2　新闻话题要素提取评价结果

方法	准确率	召回率	F 值
Ahn's method	0. 240	0. 300	0. 267
proposed(Vietnam)	0. 480	0. 600	0. 533
proposed(Chinese)	0. 520	0. 650	0. 578
proposed	0. 560	0. 700	0. 622

以 2014 年 5 月 3 日开始的中越"981 钻井平台"爆发的一系列事件为话题,用相关方法抽取到的汉越新闻话题要素如表 3 所示。表中是以中越海洋石油"981"相关事件为话题的文档集合,K 的取值为 25,最终以汉越双语的形式展现新闻话题要素。

5　结束语

针对新闻话题要素提取,本文提出了基于超图的PageRank 随机游走方法,该方法经过迭代计算以后,对汉越新闻事件要素集合给出一个有关话题信息重要性的排序,最终得到汉越新闻话题要素(who、when、where、what) 。实验结果也证明了基于超图的 PageRank 随机游走方法提取汉越话题要素的有效性。下一步的工作将围绕话题演化与追踪展开研究,用新闻话题要素来表达话题整个演化的过程。

表3　汉越新闻话题要素提取实例结果

方法	汉越新闻话题要素
人工提取	中国(đồ sứ)、越南(Việt Nam)、981钻井平台(Nền tảng 981 khoan)、2014年5月2日(02 Tháng 5 2014)、西沙群岛(Quần đảo Hoàng Sa)、干扰(sự can thiệp)、黎海平(Li Haiping)、越南"专属经济区"(Việt Nam "khu vực đặc quyền kinh tế")、华春莹(Hoa Xuân Oánh)、侵犯(Cuộc xâm lược)、中国主权(chủ quyền của Trung Quốc)、2014年5月11日(Th á ng 11 năm 2014)、示威游行(Các cuộc biểu tình)、胡志明市(Thành phố Hồ Chí Minh)、爆发(thoát ra khỏi)、反华罢工游行(Chống Trung Quốc diễu đình công)、烧砸(đốt hit)、2014年5月17日(Tháng 17 năm 2014)、驻越大使馆(Đại sứ quán tại Việt Nam)、3000人回国(3000 trở lại)
Anh's method	中国、越南、2014年5月11日、胡志明市、爆发、烧砸、河内、游行、抗议、警方、群众、驱散、巡逻、声称、维稳、外交法码、军事、媒体、威胁、世界和平、安抚、歌剧院、国际法、聚集、静观态度
proposed (Vietnam)	đồ sứ、Việt Nam、tàu chính thức、tàu vũ trang、Quần đảo Hoàng Sa、sự can thiệp、Li Haiping、Việt Nam "khu vực đặc quyền kinh tế"、Hoa Xuân Oánh、Tháng 11 năm 2014、Thành phố Hồ Chí Minh、thoát ra khỏi、đốt hit、Đại sứ quán tại Việt Nam、tham khảo、hoạt động、luật pháp quốc tế、Chủ quyền của các nước láng giềng、Vùng đất nước ngoài、Hoa Kỳ、khiêu khích、Nhà máy Trung Quốc、bờ biển lục địa、gìn giữ hòa bình、chống Trung Quốc
proposed (Chinese)	中国、越南、981钻井平台、海洋石油、争端、2014年5月2日、公务船、武装船、冲撞、西沙群岛、摩擦、渔船、保卫、华春莹、侵犯、中国主权、灯塔、海洋划界、外交对侍、2014年5月11日、胡志明市、烧砸、驻越大使馆、干扰、不稳定因素
proposed	中国(đồ sứ)、越南(Việt Nam)、981钻井平台(Nền tảng 981 khoan)、2014年5月2日(02 Tháng 5 2014)、公务船(tàu chính thức) 、武装船(tàu vũ trang) 、冲撞(đụng nhau) 、钻探(khoan) 、撤走(rút) 、南海(biển Đông) 、西沙群岛(Quần đảo Hoàng Sa)、干扰(sự can thiệp)、诬蔑(phỉ báng) 、摩擦(xích mích) 、渔船(câu cá) 、保卫(phòng thủ) 、华春莹(Hoa Xuân Oành)、侵犯(Cuộc xâm lược)、中国主权(chủ quyền của Trung Quốc)、2014年5月11日(Th á ng 11 năm 2014)、胡志明市(Thành phố Hồ Chí Minh)、爆发(tho á t ra khỏi)、烧砸(đốt hit)、驻越大使馆(Đại sứ qu á n tại Việt Nam)、不稳定因素(sự bất ổn)

参考文献:

[1] 梁晗,陈群秀,吴平博. 基于事件框架的信息抽取系统 [J] . 中文信息学报,2006,20(2) : 40-46.

[2] 冯礼. 基于事件框架的突发事件信息抽取 [D] . 上海: 上海交通大学,2008.

[3] Chieu H L,Ng H T. A maximum entropy approach to information extraction from semi-structured and free text [C] //Proc of the 18th National Conference on Artificial Intelligence. 2002: 786-791.

[4] Ahn D. The stages of event extraction [C] //Proc of Workshop on Annotating & Reasoning About Time & Events. 2006: 1-8.

[5] Ji Heng,Grishman R. Refining event extraction through unsupervised cross-document inference [C] // Proc of the 46th Annual Meeting of the Association for Computational Linguistics. 2008: 254-262.

[6] 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.

[7] 张先飞,郭志刚,刘嵩,等. 基于触发词指导的自相似度聚类事件检测 [J] . 计算机科学,2010,37(3) : 212-214.

[8] 潘清清,周枫,余正涛,等. 基于条件随机场的越南语命名实体识别方法 [J] . 山东大学学报: 理学版,2014,49(1) : 76-79.

[9] ACE(automatic content exraction) Chinese annotation guidelines for events [R] . [S. l.] : National Institute of Standards and Technology, 2005.

[10] Wang Wei,Li Sujian,Li Jiwei,*et al*. Exploring hypergraph-base semi-supervised ranking for query-oriented summarization [J] . Information Sciences,2013,237(13) : 271-286.

[11] Bellaachia A,Al-Dhelaan M. Multi-document hyperedge-based ranking for text summarization [C] //Proc of ACM International Conference on Information and Knowledge Management. New York: ACM Press,2014: 1919-1922.

[12] Chen P,Xie H,Maslov S,*et al*. Finding scientific gems with Google's PageRank algorithm [J] . Journal of Informetrics,2007,1(1) : 8-15.