基于触发词指导的自相似度聚类事件检测

张先飞[1] 郭志刚[1] 刘 嵩[1] 程 磊[2] 田雨暄[1]
(解放军信息工程大学信息工程学院 郑州 450002)[1] (中国人民解放军61081部队 北京 100094)[2]

摘 要 传统方法将事件检测任务看作分类问题,将词作为实例来训练分类器,容易导致训练正反例不平衡,同时,在语料库规模较小时存在一定的数据稀疏问题。首先避开以词为实例进行分类,在事件类别判断上引入聚类思想,在事件触发词的指导下,采用自相似度对K-means聚类算法中的K值进行自收敛,优化了聚类算法。然后结合命名实体及其位置信息,对事件类别进行详细定位,很好地解决了传统事件检测对类别模板的依赖性,所检测的事件在文本摘要、检索和主题检测与追踪上得到了很好的应用。

关键词 事件检测,触发词,自相似度,命名实体,聚类
中图法分类号 TP391 **文献标识码** A

Self-similarity Clustering Event Detection Based on Triggers Guidance

ZHANG Xian-fei[1] GUO Zhi-gang[1] LIU Song[1] CHENG Lei[2] TIAN Yu-xuan[1]
(Information Engineering Institute, PLA Information Engineering University, Zhengzhou 450002, China)[1]
(PLA 61081 Unit, Beijing 100094, China)[2]

Abstract Traditional method of Event Detection and Characterization (EDC) regards event detection task as classification problem. It makes words as samples to train classifier, which can lead to positive and negative samples of classifier imbalance. Meanwhile, there is data sparseness problem of this method when the corpus is small. This paper didn't classify event using word as samples, but clustered event in judging event types. It adapted self-similarity to convergence the value of K in K-means algorithm by the guidance of event triggers, and optimized clustering algorithm. Then, combining with named entity and its comparative position information, the new method further ensures the pinpoint type of event. The new method avoids depending on template of event in tradition methods, and its result of event detection can well be used in automatic text summarization, text retrieval, and topic detection and tracking.

Keywords Event detection, Trigger, Self-similarity, Named entity, Clustering

1 引言

事件检测与描述(Event Detection and Characterization, EDC)源自ACE(Automatic Content Extraction)会议[1]。ACE会议主要研究从新闻语料中自动检测实体(Entity)、关系(Relation)、事件(Events)等内容。目前,事件检测已成为ACE会议中除实体识别(Entity Detection and Recognition, EDR)和实体关系识别(Relation Detection and Recognition, RDC)之外的又一研究热点。本文事件检测的定义来自于ACE会议,事件由事件触发词(Trigger)和描述事件结构的元素(Argument)构成[1]。一个典型的例子:"张三于1981年出生在山东青岛",其中,"出生"是该事件的触发词,"张三"、"1981年"、"山东青岛"是构成这个事件的3个元素。

当前事件检测主要有两种方法:模式匹配方法和机器学习方法。模式匹配法采用模式匹配算法将待检测句子与固定模板匹配[2],代表有Surdeanu和harabagiu针对开放域的事件检测系统——FSA系统[3]。这种方法准确率高,但对具体领域的依赖性强,可移植性差。机器学习方法将事件检测看作分类问题,主要包括事件类别识别和事件元素识别两个步骤,事件类别由事先定义的事件模板来决定,ACE会议定义了8种事件类别及33种子类别,每种事件类别对应唯一的模板。上述例子中,事件类别为ACE会议定义的Life,子类别为Be-Born,而3个构成元素分别对应Be-Born子类别模板中的3个元素标签,即:Person, Time及Place。分类学习是目前事件检测研究多数采用的方法。2002年,Hai Leong Chieu和Hwee Tou Ng首次引入最大熵分类器[4],用于事件元素的识别;David Ahn于2006年结合MegaM和Timbl两种机器学习方法分别实现了事件检测中的事件类别和元素识别[5];国内事件检测代表性的算法有2008年赵妍妍的基于触发词扩展和二元分类相结合的事件检测方法[6]。然而,分类法将每个词作为一个实例来训练分类器,容易引入大量的反例,导致正反例严重不平衡。此外,事件类别的多元分类及每类事

到稿日期:2009-04-17 返修日期:2009-07-18 本文受863国家重点基金项目(2007AA01Z439)资助。
张先飞(1981—),男,博士生,主要研究方向为中文信息处理、信息抽取及数据挖掘,E-mail:zhangxianfei2003@126.com;**郭志刚**(1973—),男,讲师,主要研究方向为Web文本挖掘;**刘 嵩**(1985—),男,硕士生,主要研究方向为自动内容抽取、事件检测;**程 磊**(1981—),男,工程师,主要研究方向为通信信号处理;**田雨暄**(1988—),女,硕士生,主要研究方向为文本信息抽取。

件元素单独构造多元分类器在语料库规模较小时存在一定的数据稀疏问题。

针对以上问题，本文避开以词为实例进行分类，尝试在事件类别判断上引入聚类思想，提出基于触发词指导的自相似度聚类事件检测算法。首先利用触发词来确定 K-means 聚类初始质心，同时结合自相似度策略来确定 K 值，以解决聚类算法中 K 值及初始质心选取的问题，实现对文本事件进行初步类别判断；然后，结合人名、地名、组织机构名、时间等命名实体及相对位置信息，进一步确定事件类别，完成事件检测。实验结果表明，本方法突破事件检测对事件类别模板的限制，所检测事件的召回率较分类法检测事件有明显提高，在文本摘要、分类与检索和主题检测与追踪中得到了很好的应用。

2 K-means 聚类算法

给定 d 维数据集 $X=\{x_i \mid x_i \in R^d, i=1,2,\cdots,N\}$，将其聚成 K 个类 $\omega_1,\omega_2,\cdots,\omega_K$，质心为 $c_1,c_2,\cdots,c_K$，其中 $c_i=(1/n_i)\sum_{x\in\omega_i}x$，$n_i$ 是类 ω_i 中数据点的个数。聚类目标函数为：$J=\sum_{i=1}^{k}\sum_{j=1}^{n_i}d_{ij}(x_j,c_i)$，其中 $d_{ij}(x_j,c_i)$ 是 x_j 与 c_i 之间的欧氏距离。

K-means 聚类步骤如下：

1)从 X 中随机选择 K 个初始参照点 $c_1,c_2,\cdots,c_K$；

2)以 $c_1,c_2,\cdots,c_K$ 为参照点，对 X 进行划分。若满足 $d_{ij}(x_i,c_j)=\min\limits_{m=1,2,\cdots,k}d_{im}(x_i,c_m)$，其中，$j=1,2,\cdots,K$，$i=1,2,\cdots,N$，则将 x_i 划分到类 ω_j 中；

3)根据式 $c_i=(1/n_i)\sum_{x\in\omega_i}x$，重新计算类的质心 $c_1^*,c_2^*,\cdots,c_K^*$；

4)若对于任意 $i\in\{1,2,\cdots,K\}$，$c_i^*=c_i$ 都成立，则算法结束，当前的 $c_1^*,c_2^*,\cdots,c_K^*$ 代表最终的聚类结果；否则，令 $c_i=c_i^*$，重新执行步骤 2)。

为了防止步骤 4)中出现无限循环的情况，通常设置一个固定的阈值 th，当对于所有的 c_i，都有 $|c_i^*-c_i|<th$ 时，算法结束。

基于 K-means 聚类的事件检测需要解决以下问题：

1)聚类类别数 K 的确定。文档中事件的个数事先不知道，所以需要确定 K 的值。

2)初始质心的选择。K-means 是以质心为参照点进行聚类的，质心的选取决定着最终所聚类别事件的核心内容，因此，如何选取聚类初始质心在基于聚类的事件检测中尤为重要。

3 基于自相似度的最大最小聚类

为了解决 K-means 聚类中存在的问题，这里引入了基于自相似度的最大最小原则[7,8]，利用自相似度的自收敛策略来确定 K 值的选取，同时在事件触发词的指导下完成基于 K-means 聚类的事件检测。

假设文本中的一个事件为一个文本单元，任意两个文本单元间的距离用余弦相似度来计算。比如文本单元 i 的特征向量为 $x_i=(a_1,a_2,\cdots,a_n)$，j 的特征向量为 $x_j=(b_1,b_2,\cdots,b_n)$，则两个文本单元的相似度：

$$\mathrm{sim}(x_i,x_j)=(\sum_{m=1}^{n}a_mb_m)/\mathrm{sqrt}(\sum_{m=1}^{n}a_m^2)\times\mathrm{sqrt}(\sum_{m=1}^{n}b_m^2) \quad (1)$$

其中，$a_m,b_m(1\leqslant m\leqslant n)$ 为文本单元的第 m 个特征对应的权值，n 为两个文本单元的特征并集总数。

设聚类样本集合为：$X=\{x_1,x_2,\cdots,x_N\}$，N 为样本个数。计算所有样本间的相似度 $\mathrm{sim}_{ij}=\mathrm{sim}(x_i,x_j)$，当 $i=j$，则 $\mathrm{sim}_{ij}=1$。

定义全局平均相似度：

$$\bar{s}=\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathrm{sim}_{ij}/N(N-1)/2 \quad (2)$$

定义最大最小平均相似度：

$$\bar{s}_{\mathrm{minmax}}=\{\max(\mathrm{sim}_{ij})+\min(\mathrm{sim}_{ij})\}/2 \quad (3)$$

其中，$i=1,2,\cdots,N-1$，$j=i+1,i+2,\cdots,N$。

定义平均自相似度，类别 ω_k 的平均自相似度：

$$\bar{s}_{kk}=\frac{s_{kk}}{\frac{n_k(n_k-1)}{2}},n_k=|\omega_k| \quad (4)$$

其中，

$$s_{kk}=\sum_{i\in\omega_k}\sum_{\substack{j\in\omega_k\\ j\neq i}}\mathrm{sim}_{ij} \quad (5)$$

i 与 j 同为类 ω_k 中的样本，因此称 s_{kk} 为类 ω_k 的自相似度。如果 $i\in\omega_{k_1}$ 且 $j\in\omega_{k_2}$，则 $s_{k_1k_2}=\sum\limits_{i\in\omega_{k_1}}\sum\limits_{j\in\omega_{k_2}}\mathrm{sim}_{ij}$ 为类 ω_{k_1} 和 ω_{k_2} 的互相似度。

定义全局平均自相似度门限：

$$\bar{s}_{Gth}=\max(\bar{s},\bar{s}_{\mathrm{minmax}}) \quad (6)$$

假设选取了 k 个质心，各类的平均自相似度分别为：$\bar{s}_{11},\bar{s}_{22},\cdots,\bar{s}_{kk}$，增加一个质心后，各类的平均自相似度为：$\bar{s}'_{11},\bar{s}'_{22},\cdots,\bar{s}'_{kk},\bar{s}'_{k+1k+1}$。

定义局部自适应的平均自相似度门限 $\bar{s}_{th}$：

$$\bar{s}_{th}=(\bar{s}_{11}+\bar{s}_{22}+\cdots+\bar{s}_{kk}+\bar{s}'_{11}+\bar{s}'_{22}+\cdots+\bar{s}'_{kk})/2k \quad (7)$$

该门限值随每次聚类动态变化。

如果出现

$$\frac{(\bar{s}'_{11}+\bar{s}'_{22}+\cdots+\bar{s}'_{kk})}{(|\bar{s}'_{11}-\bar{s}_{th}|+|\bar{s}'_{22}-\bar{s}_{th}|+\cdots+|\bar{s}'_{kk}-\bar{s}_{th}|)}\geqslant\frac{(\bar{s}_{11}+\bar{s}_{22}+\cdots+\bar{s}_{kk})}{(|\bar{s}_{11}-\bar{s}_{th}|+|\bar{s}_{22}-\bar{s}_{th}|+\cdots+|\bar{s}_{kk}-\bar{s}_{th}|)}$$

且 $\bar{s}_{k+1k+1}\geqslant\bar{s}_{Gth}$ (8)

则继续选取下一个质心聚类。

基于自相似度的最大最小聚类过程如下：

1)将文本中事件触发词作为初始聚类质心。检测触发词所在句级文本单元相似度最小的两个样本 s_1 与 s_2，将其作为初始质心，其他样本按照相似度分别划分到 s_1 与 s_2 所在的类别中，然后分别计算 $\bar{s}_{11}$ 与 $\bar{s}_{22}$，此时聚类类别数 $K=2$。如果出现具有同样最小相似度的其他样本对，则同样重新计算出以此对样本为质心后各类的平均自相似度 $\bar{s}'_{11}\bar{s}'_{22}$。具体方法二者选其一：

(a)计算不同样本对做质心后，各类之间的互相似度 $s_{k_1k_2}$ 和 $s'_{k_1k_2}$。比较 $s_{k_1k_2}$ 与 $s'_{k_1k_2}$，取值小的那对样本作为初始质心。

(b)不计算互相似度，仅利用平均自相似度进行初始质心的选取。比较 $\frac{(\bar{s}_{11}+\bar{s}_{22})}{(|\bar{s}_{11}-\bar{s}_{Gth}|}+|\bar{s}_{22}-\bar{s}_{Gth}|)$ 与 $\frac{(\bar{s}'_{11}+\bar{s}'_{22})}{(|\bar{s}'_{11}-\bar{s}_{Gth}|}+|\bar{s}'_{22}-\bar{s}_{Gth}|)$ 选取比值较大的那对样本为初始质心；如果比值相等，则选择 $|\bar{s}_{11}-\bar{s}_{22}|$ 与 $|\bar{s}'_{11}-\bar{s}'_{22}|$ 较小的那对样本。

这种样本对选择策略有两个好处：既可以尽量保证聚类后各类的平均自相似度不能太小，同时也避免了选取的样本

聚类后两个类的平均自相似度相差太大。

2)用同样的方法选取下一个质心。聚类后计算每个类别的平均自相似度,直到不满足式(8),停止聚类,并确定类别数 K。

4 基于自相似度聚类的事件检测

基于自相似度聚类的事件检测的主要思想是在事件触发词的指导下,抛开预先定义的事件类别的限制,对原始文档集进行初步事件聚类,然后结合命名实体及相对位置信息对事件进行详细定位,准确找出文档中所包含的事件。基于触发词指导的自相似度聚类事件检测流程图如图 1 所示。

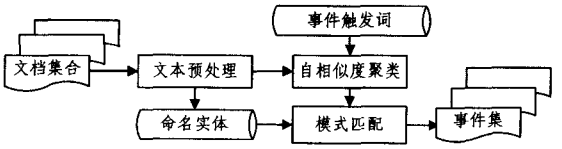

图 1 基于自相似度聚类的事件检测流程图

基于自相似度聚类完成事件检测的具体步骤如下:

(1)对文档集进行文本预处理,主要包括中文分词、词性标注等,完成对自然文本的初步预处理;

(2)根据词性标注的结果过滤出文本集中包含的触发词,并与事先标注的事件触发词进行匹配,匹配时要考虑近义词的识别;

(3)在匹配出的事件触发词的指导下,利用自相似度 K-means 算法对以事件触发词为中心的文本单元进行自相似度聚类,根据所聚类别初步确定所检测事件;

(4)根据第(1)步对文档集预处理的结果,提取其中的人名、地名、组织机构名及时间等命名实体,同时按照在文本中出现的顺序记录命名实体的相对位置信息;

(5)在命名实体位置信息的指导下,将所提取命名实体与第(3)步中初步聚类的结果进行模式匹配,最终完成文档中事件的准确检测。

5 实验及性能分析

本文采用网络采集数据作为实验数据,经过网页文本内容提取处理后,取其中 500 篇文本进行文本预处理及标注工作,标注内容包括人名、地名、组织机构名、时间等实体信息和事件触发词、元素及其属性,并对事件触发词进行同义词归类,这样做有利于算法的实验验证。本文采用信息抽取中的评价标准:准确率 P(precision)、召回率 R(Recall)和 F 值来进行性能评价,各评价指标定义如下:

$$F=\frac{2PR}{P+R} \tag{9}$$

$$P=\frac{\text{抽取准确的事件个数}}{\text{抽取事件总数}} \tag{10}$$

$$R=\frac{\text{抽取准确的事件个数}}{\text{标注事件总数}} \tag{11}$$

5.1 不同聚类算法的事件检测结果

本实验比较几种聚类算法事件检测的结果,所验证的聚类算法包括:经典 K-means 算法(算法 1)、自相似度指导的 K-means 算法(算法 2)及基于触发词指导的自相似度 K-means 聚类(算法 3)。其中自相似度指导的 K-means 算法中的初始质心不是采取触发词指导,而是采取每次计算最小相似度迭代的方法进行聚类。各算法的实验结果如表 1 所列。

表 1 3 种聚类算法指导下的事件检测结果

	召回率(R)	准确率(P)	F1-值
算法 1	38.91%	50.36%	43.91%
算法 2	56.18%	61.24%	58.61%
算法 3	61.31%	65.47%	63.32%

由于 K-means 算法的类别数是个经验值,并且聚类初始质心是随机选取的,因此所检测事件的性能十分不理想。基于自相似度的 K-means 算法在初始质心的选取上采用自相似度比较,同时通过自收敛可以确定所聚类别数,因此所检测事件的效果要比传统 K-means 算法好很多;然而由于其初始质心只依靠自相似度比较来确定,如果第一个质心选取错误,则会导致错误的延续,因此事件检测结果仍不理想。基于触发词指导的自相似度聚类事件检测则解决了这个问题,在触发词的指导下选取初始质心,可以避免初始质心选取错误而导致的错误延续,同时加入命名实体及相对位置信息对事件进行精确定位,可使所检测事件效果最好。

5.2 与分类法事件检测结果对比

将本文基于触发词指导的自相似度聚类方法(本文方法)与 Ahn 的基于机器学习法对事件类别和元素进行识别的方法[5](简称机器学习法)以及基于触发词及二元分类的事件检测方法[6](简称二元分类法)结果进行比较。为了使比较结果具有可信性,在实验中对文献[5,6]中的方法进行了实现,并在同一个环境下对本文实验数据进行对比,结果如表 2 所列。

表 2 3 种方法对比试验结果

	召回率(R)	准确率(P)	F1-值
机器学习法	43.18%	57.93%	49.48%
二元分类法	53.27%	68.87%	65.47%
本文方法	61.31%	65.47%	63.32%

利用机器学习法对事件类别和元素类别进行识别,首先对文档中的事件进行了类别划分的限制,同时由于分类将词作为实例来训练分类器,引入了大量反例,这些都是直接导致其分类效果差的主要原因。基于触发词和二元分类事件检测方法虽然引入了触发词来构造候选事件,减少了分类训练中的反例个数,使得事件检测准确率有所提高,但是此方法仍然局限于预先定义类别的事件检测,而没有定义的事件则无法检测,因此导致了其召回率不高。本文方法则突破预先定义事件类别的局限,根据事件触发词对文档中的事件进行自动聚类,使得所检测结果的召回率有显著提高。然而,聚类由于是无指导的类别区分方法,因此检测结果中有部分重复的事件,导致了检测准确率有所下降,今后可考虑采取一定的修剪策略来提高准确率。

结束语 本文抛开传统事件检测根据预先定义事件类别进行事件分类的思路,在触发词的指导下,研究了如何将 K-means 聚类算法应用于事件检测,利用自相似度策略,结合命名实体及相对位置信息对事件进行检测,提升了事件检测的性能。但是,事件检测仍处于起步阶段,不是仅仅在有限个触发词的指导下就能很好地检测出事件,还需要一个崭新的模型并充分利用触发词、命名实体信息及语义模型,才能更好地进行事件检测。

参 考 文 献

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

(下转第 220 页)

代表处理的任务和任务的处理量，如 0(22)表示此批次处理任务 0，处理量为 22。图 3 中的任务 0～5 之间的关系由图 4 所示的 STN 表达(注意这里的序号都减 1)。再调度要求为在原调度中插入序号为 3 的任务，其订单处理量为 30。从图 4 可以看出，需要插入的 1 和 4 两个任务(批次)，在图中体现为 0 和 3。

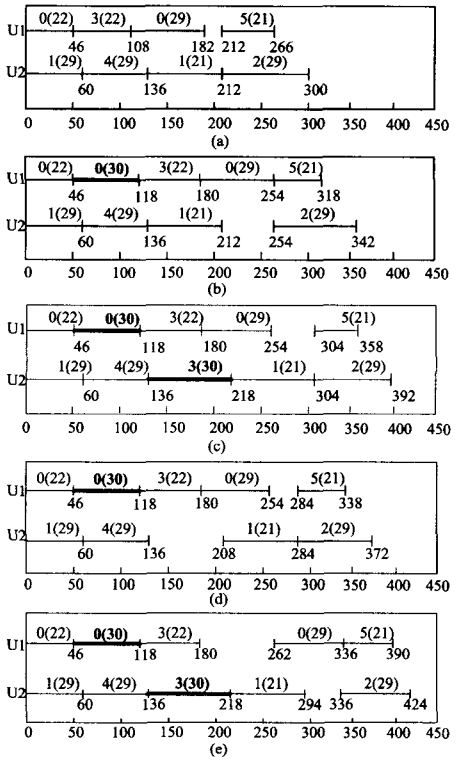

图 3　ABR 和 RSR 调度实例

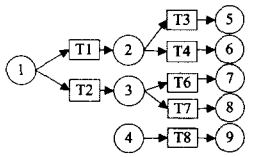

图 4　STN 表达的多目的批量过程

在图 3 中，(a)为原调度，(b)和(c)为应用 ABR 法分别插入第一个任务和插入两个任务后获得的再调度结果，(d)和(e)为应用 RSR 法分别插入第一个任务和插入两个任务后获得的再调度结果。从图 3 可以看出，应用 ABR 再调度法，插入第一个任务后其新调度的完成时间为 342，开始时间为 228，插入二个任务后其完成时间为 392，开始时间延迟为 410。应用 RSR 再调度法，插入第一个任务后其新调度的完成时间为 372，开始时间为 380，插入二个任务后其完成时间为 424，开始时间延迟为 552。可以看出，不论在完成时间指标上，还是在开始时间延迟指标上，ABR 法均优于 RSR 法。

结束语　本文研究了复杂多目的批量过程在受到延迟扰动时的再调度问题。在深入分析多目的批量过程调度问题和 Job Shop 调度问题异同的基础上，借鉴针对 Job Shop 的受影响操作再调度 AOR 算法，提出了多目的批量过程的受影响批次再调度 ABR 算法。仿真算例结果显示，提出的 ABR 算法在新调度完成时间和原调度开始时间延迟两种评价指标上均优于现有的右移再调度算法。后续工作将研究在更改作业处理顺序和资源的条件下如何提高再调度在新调度质量和原调度稳定性上的综合性能。

参 考 文 献

[1] Cott B J, Macchietto S. Minimizing the Effects of Batch Process Variability Using Online Schedule Modification[J]. Computers and Chemical Engineering, 1989, 13: 105-113

[2] Kim M, Lee I B. Rule-based Reactive Rescheduling System for Multi-purpose Batch Processes[J]. Computers and Chemical Engineering, 1997, 21: 1197-1201

[3] Honkomp S J, Mockus L, Reklaitis G V. A Framework for Schedule Evaluation with Processing Uncertainty[J]. Computers and Chemical Engineering, 1999, 23: 595-609

[4] Vin J P, Ierapetritou M G. A New Approach for Efficient Rescheduling of Multiproduct Batch Plants[J]. Industrial and Engineering Chemistry Research, 2000, 39: 4228-4238

[5] Roslof J, Harjunkoski I, Bjorkqvist J, et al. An MILP-based Reordering Algorithm for Complex Industrial Scheduling and Rescheduling[J]. Computers and Chemical Engineering, 2001, 25: 821-828

[6] Mendez C A, Cerda J. Dynamic Scheduling in Multiproduct Batch Plants[J]. Computers and Chemical Engineering, 2003, 27: 1247-1259

[7] Mendez C A, Cerda J. An MILP Framework for Batch Reactive Scheduling with Limited Discrete Resources[J]. Computers and Chemical Engineering, 2004, 28: 1059-1068

[8] Abdullah C, Amarnath B. Cycle time reduction in batch processing by upstream rescheduling[C]// IIE Annual Conference and Expo 2007-Industrial Engineering's Critical Role in a Flat World. 2007: 1471-1476

[9] Floudas C A, Lin X. Continuous-time versus Discrete-time Approaches for Scheduling of Chemical Processes: A Review[J]. Computers and Chemical Engineering, 2004, 28: 2109-2129

[10] Abumaizar R J, Svestka J A. Rescheduling job shops under disruptions[J]. International Journal of Production Research, 1997, 35, 2065-2082

(上接第 214 页)

[2] Surdeanu M, Harabagiu S, Williams J, et al. Using Predicate-Argument Structures for Information Extraction[C]// Proceedings of ACL. 2003: 8-15

[3] Surdeanu M, Harabagiu S. Infrastructure for open-domain information extraction[C]// Proceedings of the Human Language Technology Conference. 2002: 325-330

[4] Chieu Hai Leong, Ng Hwee Tou. A Maximum entropy Approach to Information Extraction from Semi-Structured and Free Text[C]// Proceedings of the 18th National Conference on Artificial Intelligence. 2002: 786-791

[5] Ahn D. The Stages of Event Extraction[C]// Proceedings of the Workshop on Annotations and Reasoning about Time and Events. 2006: 1-8

[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] Ding C, He Xiaofeng. Cluster Merging and Splitting in Hierarchical Clustering Algorithms[A]// Proceedings of the 2002 IEEE International Conference on Data Mining[C]. Maebashi City, Japan: Maebashi TERRSA, 2002: 139-146

[8] Ding C, He X, Zha H, et al. A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering[A]// Proceedings of the IEEE Internationl Conference[C]. San Jose, California, USA: Data Mining, 2001: 107-114