

基于特征加权的事件要素识别

付剑锋 刘宗田 刘 炜 单建芳

(上海大学计算机工程与科学学院 上海 200027)

摘 要 事件抽取是自动内容抽取(Automatic Content Extraction, ACE)会议评测的任务之一,事件要素识别是事件抽取的一个子任务。分析了事件抽取和事件要素识别的研究现状,提出了一种基于特征加权的事件要素识别算法(Feature Weighting Based Event Argument Identification, FWEAI)。该算法首先对分类算法中的 ReliefF 特征选择算法进行改进,将其应用于聚类算法中。改进的 ReliefF 算法(FWA)根据各个特征对聚类的不同贡献分配不同的权值,然后采用 KMeans 算法对事件要素进行聚类。实验结果表明, FWEAI 算法可以提高事件要素识别的准确率。

关键词 特征加权, ReliefF 算法, 事件要素识别, 事件抽取

中图分类号 TP391 **文献标识码** A

Feature Weighting Based Event Argument Identification

FU Jian-feng LIU Zong-tian LIU Wei SHAN Jian-fang

(School of Computer Engineering & Science, Shanghai University, Shanghai 200072, China)

Abstract Event extraction is a task of Automatic Content Extraction (ACE) program. Event Argument Identification is sub-task of Event extraction. The state-of-the-art of event extraction and event argument identification was given. An algorithm named Feature Weighting Based Event Argument Identification (FWEAI) was proposed, which improved ReliefF, a feature selection algorithm in classification algorithm, and employed it in clustering algorithm. The improved ReliefF (FWA) can assign different weights to different features according to their contributions to clustering, then Event arguments were clustered by using KMeans clustering algorithm. Experimental results demonstrate that FWEAI algorithm improves the precision on Event Argument Identification.

Keywords Feature weighting, ReliefF, Event argument identification, Event extraction

1 研究背景

事件抽取是自动内容抽取^[1] (Automatic Content Extraction, ACE)评测会议的任务之一。ACE 评测会议由美国国家标准技术研究所(National Institute of Standards and Technology, NIST)主办,旨在推动和发展信息抽取的相关技术。2005 年,ACE 的评测内容中加入了事件抽取任务。ACE 中定义事件由事件触发词(Trigger)和事件要素(Arguments)组成,其事件抽取任务包括事件识别和事件要素识别两部分。事件识别要判断一个包含事件触发词的句子是否是现实世界中发生的事件,事件要素识别是从已识别的事件中识别事件的时间(Time)、地点(Place)、参与者(Participants)等要素。表 1 是 ACE 中文事件标注指南(ACE Chinese Annotation Guidelines for Events)中的一个关于出生事件要素的例子。

2006 年, Ahn^[2] 提出把事件要素识别当作多元分类问题,采用基于分类学习的方法在 ACE 英文语料上实现事件识别和事件要素识别。Ahn 的方法中存在正反例不平衡的问题以及为每一类事件要素单独构造多元分类器在语料规模较小时

存在数据稀疏的问题。赵妍妍^[3]对此进行了改进,提出对触发词进行扩展,并且采用将不同类别中相同的事件要素合并构造多元分类模型的方法进行事件要素的识别,在 ACE 中文语料上的实验证明了事件要素识别效果的有效提高。Tan^[4]先采用局部特征选择和正反特征融合的方法识别事件,然后使用多层模式匹配的方法在 ACE 中文语料上识别事件要素。由于所采用的规则比较有限,导致事件要素识别效果不太好。

目前基于机器学习的事件要素识别主要采用监督(分类)学习的方法。这种学习需要大规模人工标注的熟语料库作为训练集,以获取事件要素的相关知识,学习的效果依赖于语料的质量和规模。如果语料不够充分,往往使得识别效果不理想。因此,一方面要加速事件标注语料库的建设,另一方面要探索利用无监督学习的方法在生语料中识别事件要素。

聚类分析是模式识别中一种重要的无监督分类方法,聚类算法可以在没有任何先验信息下对数据进行聚类分析。典型的聚类算法包括 KMeans 算法、OPTICS 算法、EM 算法等。然而无论是哪种算法,都隐含假设待分析数据集中对象的各

到稿日期:2009-04-17 返修日期:2009-06-26 本文受国家自然科学基金(60575035)和上海市重点学科建设项目(J50103)资助。

付剑锋(1978—),男,博士生,主要从事自然语言处理的研究, E-mail: feng93017@tom.com; 刘宗田(1946—),男,教授,博士生导师,主要从事人工智能、软件工程的研究; 刘 炜(1978—),男,博士,讲师,主要从事 Agent、自然语言处理的研究; 单建芳(1979—),女,博士生,主要从事自然语言处理的研究。

个特征对聚类的贡献是均匀的,不考虑不同特征对聚类效果的不同贡献。而实际应用中,不同的特征具备不同的区分能力,根据特征的区分能力赋予不同的权值可以改善聚类效果^[5,6]。

表1 出生事件要素示例

Person-Arg	PER	The person who is born	[李傻傻],原名蒲荔子,生于1981年11月,湖南隆回人。
Time-Arg	TIME-within	When the birth takes place	李傻傻,原名蒲荔子,生于[1981年11月],湖南隆回人。
Place-Arg	GPE/LOC/FAC	Where the birth takes place	譬如出生在[北京],叫“京生”;出生在[台湾],叫“台生”

本文提出一种基于特征加权的事件要素识别算法(Feature Weighting Based Event Argument Identification, FWEAI)。考虑到不同特征的区分能力,首先对分类算法中的ReliefF特征选择算法进行改进,将其应用于聚类算法中的特征加权,根据各个特征对聚类的不同贡献分配不同的权值。然后用聚类算法对事件要素进行聚类。实验表明,FWEAI算法可以提高事件要素识别的准确率。

2 FWEAI 算法

2.1 特征加权

采用机器学习的方法识别事件要素,就是借鉴分类的思想,将给定的某个事件要素映射到 k 个事件要素类型中的一个。特征选择是分类问题的一个重要步骤。为了有效区分事件要素,综合考虑了词性特征、句法特征、位置特征以及上下文特征来描述词语的特征。一般来说,可以将中心词 w_i 的左右各 d 个词看作是 w_i 的上下文,由于与 w_i 关系最密切的是 w_i 的左右两个词 w_{i-1} 和 w_{i+1} ,因此取 $d=1$ 。所选特征如下:

- 1) w_i, w_i 的词性以及 w_i 在句子中的句法特征;
- 2) w_{i-1}, w_{i-1} 的词性以及 w_{i-1} 在句子中的句法特征;
- 3) w_{i+1}, w_{i+1} 的词性以及 w_{i+1} 在句子中的句法特征;
- 4) 触发词 t, t 所属的事件类别、 t 的词性以及 t 在句子中的句法特征;
- 5) w_i 与 t 之间的位置关系(w_i before or after t)。

如果将上述特征直接用于事件要素识别的聚类分析,则隐含假设各个特征对聚类贡献是均匀的,不考虑各个特征对聚类的不同影响。直觉上,中心词 w_i 的特征对聚类的贡献应该最大,削弱 w_i 的贡献会降低聚类性能。为了区别不同特征对聚类的不同贡献度,采用特征加权的方法为每一类特征赋予一定的权重来提高聚类效果。

设数据集 $X = \{x_1, x_2, \dots, x_n\}$, 其中 n 是对象数, x_i 是 X 中的第 i 个对象。 $F = \{f_1, f_2, \dots, f_m\}$ 是 $1 \times m$ 的特征矩阵, f_j 表示 F 中的第 j 个特征, $value(f_j, x_i)$ 表示对象 x_i 在特征 f_j 上的取值。 $W = \{w_1, w_2, \dots, w_m\}$ 是 $1 \times m$ 的权值矩阵, $w_j \in \mathbb{N}$, 表示分配给特征 f_j 的权重。任意给定两个对象 x_a 和 x_b , 它们之间的相似度 $Sim(x_a, x_b)$ 计算公式为:

$$Sim(x_a, x_b) = \frac{\sum_{j=1}^m w_j E(value(f_j, x_a), value(f_j, x_b))}{\sum_{j=1}^m w_j} \quad (1)$$

其中,

$$E(value(f_j, x_a), value(f_j, x_b)) = \begin{cases} 1, & \text{if } value(f_j, x_a) = (f_j, x_b) \\ 0, & \text{otherwise} \end{cases}$$

w_j 用来调节特征 f_j 所占的权重, 分母 $\sum_{j=1}^m w_j$ 对相似度进行标准化, 保证 $Sim(x_a, x_b)$ 的取值在 $[0, 1]$ 之间。

ReliefF 是分类算法中常用的一种特征选择方法^[7,8], 其基本思想是: 对于数据集 X 中任意给定的一个对象 x , 首先在同类中分别找出 x 的 l 个最近邻 $p_r, r=1, \dots, l, p_r \in class(x)$, 然后在各个不同的类中分别找到 x 的 l 个最近邻 $q_r, r=1, \dots, l, q_r \notin class(x)$ 。特征矩阵 F 中任意特征 f 所对应的权值 w 的计算公式为:

$$w = w - \sum_{r=1}^l diff(f, x, p_r) + \sum_{q_r \notin class(x)} \frac{P(class(q_r))}{1 - P(class(x))} \sum_{r=1}^l diff(f, x, q_r) \quad (2)$$

其中, $P(class(x))$ 表示对象 x 所属类出现的概率, 可用 $class(x)$ 中的对象总数比上数据集 X 中的对象总数得到。 $diff(f, x, p_r)$ 为对象 x 和对象 p_r 在特征 f 上的差异。对于离散型特征, 定义差异函数:

$$diff(f, x, p_r) = \begin{cases} 0, & \text{if } value(f, x) = value(f, p_r) \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

用式(2)对特征矩阵中的每一个特征进行计算, 重复若干次, 便可得到特征矩阵中每一个特征的权重。

ReliefF 算法需要预先知道对象的类别, 而在聚类分析中, 对象的类别都是未知的。因此, 首先用 KMeans 算法对数据集 X 进行聚类, 得到 k 个类簇。此外还对对象的选择进行约束, 保证尽量选择隶属度较高的对象。约束方法为: 随机选择 k 个类簇中的某个类簇 C , 从中选择与聚类中心相似度最高的对象 x 。然后分别找出与 x 同类簇的 l 个最近邻以及每一个不同类簇中的 l 个最近邻来计算权值矩阵 W 。算法描述如下:

算法1 特征加权算法 (Feature Weighting Algorithm, FWA)

输入: 数据集 X ; 类簇数 k ;

输出: 权值矩阵 W ;

- 1) 初始化权值矩阵 $W = \{w_1, w_2, \dots, w_m\}$, set $w_j = 0, j = 1, \dots, m$;
- 2) 用 KMeans 算法进行聚类;
- 3) for(int $i = 0; i < n; i++$) {
- 4) 随机选择一个类簇 C , 从中选择与聚类中心相似度最高的对象 x ;
- 5) 找出与 x 同类簇的 l 个最近邻 $p_r, r = 1, \dots, l, p_r \in C$;
- 6) for each cluster $C' \neq C$ {
- 7) 从类簇 C' 中找到 l 个最近邻 $q_r, r = 1, \dots, l$;
- 8) }
- 9) for(int $j = 1; j \leq m; j++$) {
- 10) 用式(2)计算 w_j ;
- 11) }
- 12) }

算法第3行中的 n 是迭代次数, 由用户指定。

2.2 算法框架

FWEAI 算法首先对输入的生语料进行分词、词性标注、句法分析等预处理。预处理之后的数据经过特征选择, 用 KMeans 进行聚类, 然后使用改进的 ReliefF 算法计算权值矩阵。权值矩阵根据各个特征对聚类的不同贡献, 为其分配相应的权值。之后, 再次使用 KMeans 算法对其进行聚类, 将相

同类型的事件要素聚集在同一个类簇中。其中,从特征选择到输出特征权值矩阵属于特征加权算法(FWA)。整个 FWEAI 算法的框架如图 1 所示。

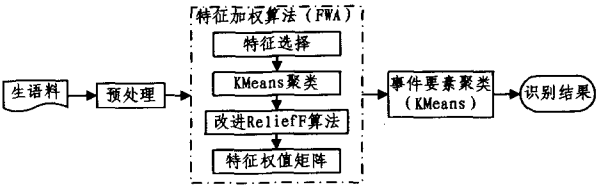


图1 FWEAI 算法框架

3 实验结果及分析

为了评价 FWEAI 算法的性能,我们使用 java 语言实现了该算法。实验中分词和句法分析工具采用了哈工大 LTP 语言技术平台提供的模块。由于 ACE 语料不对外公开评测,本文实验语料收集了来自于中国政府网以及新浪、搜狐、新华网等关于突发事件的报道,共 8 类事件,每类事件各 100 篇,共 800 篇文档。不同事件类中的相同事件要素(比如时间要素和地点要素)视为一类要素,共包含 16 类事件要素,如表 2 所列。

表2 实验采用的突发事件语料

事件名称	事件要素个数	语料篇数
台风	3	100
地震	3	100
交通事故	5	100
爆炸	3	100
食物中毒	5	100
禽流感	4	100
骚乱	3	100
空袭	4	100

图 2 是 FWA 算法迭代 20 次后得到的各个特征所对应的权值,其中特征 1,2 和 3 分别表示中心词 w_i 词本身、词性及句法特征;特征 4,5 和 6 分别表示 w_{i-1} 词本身、词性及句法特征;特征 7,8 和 9 分别表示 w_{i+1} 词本身、词性及句法特征;特征 10,11,12 和 13 分别表示触发词 t 本身、触发词 t 所属事件类别、 t 的词性及句法特征;特征 14 表示中心词 w_i 与触发词 t 的位置关系。可以看出,中心词 w_i 本身所具备的特征以及 w_i 与 t 的位置关系对聚类贡献比较大。

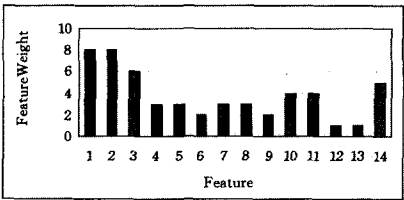


图2 FWA 算法得到的各个特征所对应的权重

我们将 FWA 算法计算得到的权值矩阵用于 FWEAI 算法的后续事件要素聚类,并且将 FWEAI 算法与没有经过特

征加权的 KMeans 聚类算法识别事件要素进行比较。分别抽取语料的 20%,40%,60%,80%和 100%进行试验,实验结果采用准确率 P (Precision)评价。实验结果如表 3 所列。

表3 FWEAI 算法与 KMeans 算法识别事件要素结果比较

语料比例 (%)	Precision (%)	
	KMeans Only	FWEAI
20	50.32	67.05
40	51.16	66.90
60	52.39	67.29
80	50.55	67.65
100	52.69	67.55
平均	51.42	67.28

从表 3 可以看出,仅采用 KMeans 算法不对特征进行加权,在不同比例语料上的平均准确率为 51.42%,而 FWEAI 算法在不同比例语料上的平均准确率提高到 67.28%。因此,采用特征加权算法根据不同特征对聚类的不同贡献分配相应的权重,可以明显提高事件要素识别的准确率。

结束语 本文分析了事件抽取和事件要素识别的研究现状。针对目前研究中的不足,提出了一种基于特征加权的事件要素识别算法(FWEAI)。算法采用 FWA 算法计算权值矩阵,根据不同特征对聚类的不同贡献,分配相应的权值,然后采用 KMeans 算法对事件要素进行聚类。在生语料上的实验表明,FWEAI 算法可以提高事件要素识别的准确率。由于 FWEAI 算法采用的是无监督的聚类学习,聚类的结果需要人工加以判断。用半监督的聚类算法进一步提高聚类准确率,并且通过标记对象自动推导聚类的结果,是下一步的研究方向。

参考文献

[1] Doddington G, Mitchell A, Przybocki M, et al. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation[C]//Proc. LREC. 2004;837-840

[2] Ahn D. The stages of event extraction[C]//Proceedings of the COLING-ACL 2006 Workshop on Annotating and Reasoning About Time and Events. 2006;1-8

[3] 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

[4] Tan H, Zhao T, Zheng J. Identification of Chinese Event and Their Argument Roles[C]//Computer and Information Technology Workshops, 2008 CIT Workshops 2008 IEEE 8th International Conference on. 2008;14-19

[5] Wettscchereck D, Aha D. Weighting features [M]. Springer, 1995; 347-358

[6] 李洁,高新波,焦李成. 基于特征加权的模糊聚类新算法[J]. 电子学报,2006(1)

[7] Kononenko I. Estimating attributes; analysis and extensions of RELIEF[M]. New York; Springer-Verlag, 1994; 171-182

[8] Robnik-Sikonja M, Kononenko I. Theoretical and Empirical Analysis of Relief and RRelief[J]. Machine Learning, 2003, 53 (1/2):23-69

(上接第 229 页)

[7] Xie Gang, Zhang Jinlong, Lai K K. A group decision-making model of risk evasion in software project bidding based on VPRS [C]//Lecture notes in computer science. 2005,3642:738-742

[8] 毕文杰,陈晓红. 一种基于可变精度粗糙集的群体分类决策方法[J]. 系统工程,2007,25(8):94-97

[9] Zhang Mei, Wu Wei-zhi. Knowledge reductions in information systems with fuzzy decisions [J]. 工程数学学报,2003,20(2):

53-58

[10] 袁修久,何华灿. 优势关系下模糊目标信息系统约简的辨识矩阵[J]. 空军工程大学学报:自然科学版,2006,7(2):81-84

[11] 林琳. 计算机审计风险分析及防范对策[J]. 中国管理信息化:综合版,2005(7):44-46

[12] 蒋锡元. 会计电算化条件下的审计风险与防范措施[J]. 中国管理信息化,2006(5)