面向个人简历的事件抽取和检索框架

李 劲[1,2] 张 华[1] 辜希武[3]
(湖北民族学院信息工程学院 恩施 445000)[1] (华中师范大学信息管理系 武汉 430079)[2]
(华中科技大学计算机学院 武汉 430074)[3]

摘 要 个人简历(Curriculum Vitae,Vita)通常包含了丰富的数据,如个人信息、教育背景以及工作经历等。从大量的个人简历中抽取出有用的信息并提供检索服务,可以提供更加全面和完整的个人资料。个人简历中包含的信息可以看成是按时间排序的事件序列。进一步地,可以从不同的个人简历所包含的事件中挖掘出事件之间的关联关系。提出了一个从个人简历中提取并检索事件的框架,它可以自动地从互联网上搜索并下载个人简历文档,并从中提取出感兴趣的事件保存在数据库里,以进一步查询和检索事件。所完成的工作包括:(1)提出了一个事件表示模型,用于描述事件的基本属性及检索事件;(2)基于条件随机场提出了一个概率模型,用于从个人简历中自动提取事件;(3)通过挖掘事件属性之间的共现性,提出了基于事件的检索方法。

关键词 条件随机场,事件检索,事件抽取,事件表示
中图法分类号 TP311 **文献标识码** A

Framework of Vita Event Extraction and Retrieval

LI Jing[1,2] ZHANG Hua[1] GU Xi-wu[3]
(School of Information Engineering, Hubei University of Nationalities, Enshi 445000, China)[1]
(Information Management Department, Central China Normal University, Wuhan 430079, China)[2]
(Colleage of Computer Sicence, Huazhong University of Science and Technology, Wuhan 430074, China)[3]

Abstract A curriculum vitae (henceforth referred to as a vita) usually contains a wealth of abundant data such as personal information, educational background, publications and work experience. It is significant to search, extract and explore the data from these vita documents which may provide a more comprehensive and integral personal profile. This personal profile can be viewed as a series of events. Moreover, we can take advantage of events from different individual's vita to explore and establish relationships between these events and the people involved. In this paper, we presented a framework extracting and explorating vita event, which can retrieve vita documents from the Internet, extract events from these documents and save the events to a database for further exploration. More concretely, the work introduced in this paper includes: (1) an event presentation model which characterizes the basic attributes of events and is utilized for event exploration; (2) a probabilistic model for extracting events from vita documents automatically; (3) an event exploration approach by exploiting the co-occurrence of the event attributes on the basis of the event presentation model and the event extraction approach.

Keywords Conditional random fields, Event retrieval, Event extraction, Event presentation

1 引言

个人简历以半结构化或非结构化的方式来组织个人资料,会包括个人信息、教育背景、发表的论文以及工作经历等。个人简历中包含的这些信息通常按时间顺序来组织,可以被看作具有5个属性的一系列事件(Event)。这5个事件属性为who(事件的主体)、when(事件发生的时间)、where(事件发生的地点)、what(事件导致的结果)、how(事件的动作)。例如,事件"Tom published a conference paper in 2010 when he studied in university"可以由以下5个属性来描述:who(Tom)、when(2010)、where(universiy)、what(conference paper)和how(publish)。

随着互联网的迅速发展,用户更愿意以不同的文件格式,如HTML、PDF或Word将个人简历放到Web上,甚至在个人简历中嵌入多媒体如图像或视频。如果将分布在互联网上的个人简历所包含的信息看作是按时间排序的事件序列,那

到稿日期:2011-08-29 返修日期:2011-12-19 本文受国家自然科学基金(61040006),湖北省自然科学基金(2010CDZ027),湖北省教育厅科技项目(B20101909)资助。

李 劲(1973—),男,博士,副教授,主要研究方向为基于互联网的数据挖掘和数据管理、面向云计算的Web服务及Web服务组合、计算机网络应用及安全、信息管理,E-mail:lj05921@tom.com;张 华(1978—),男,硕士,讲师,主要研究方向为网络应用;辜希武(1967—),男,博士,主要研究方向为分布式计算、数据挖掘、信息检索等。

么不同的个人简历所包含的事件可以通过挖掘事件属性的共现性(例如事件发生的时间和地点相同)而彼此关联起来,即通过挖掘事件之间的语义关联,可以构建一个事件网络(Event Web)[13]。更重要的是,从单个简历中抽取的事件只能从某一方面反映一个人。如果能够从不同简历中挖掘和发现所抽取事件之间的关联关系,我们就可能会得到关于一个人更加全面的个人资料。考虑这样一个例子:一个学生的个人简历中包含的事件也许并不能全面地反映该学生,但如果能从该学生导师的个人简历所包含的事件中发现和该学生的关联关系,就能得到该学生更加全面和完整的资料。

通常,事件可以以一种结构化的方式描述,即一个事件可以用五元组〈who,when,where,what,how〉来表示。然而,个人简历中包含的事件是被隐藏在个人简历的半结构化甚至非结构化的文本中。因此,从非结构化的文本中自动抽取出结构化的事件,是一个具有挑战性的难题,并在近年来被学术界广泛关注。一旦从简历中抽取出事件,就可以将结构化的事件保存在数据库里,进一步进行查询和检索。另外一个重要问题是,如何自动地从互联网上搜索并下载个人简历文档,显然主题爬虫比通用爬虫更适合解决这个问题。在我们的框架中,主题爬虫的关键组件是文本分类器,其用于自动地识别个人简历文档。

事件提取和检索框架的主要组件如图1所示,它们的主要功能将在下文里简要描述。

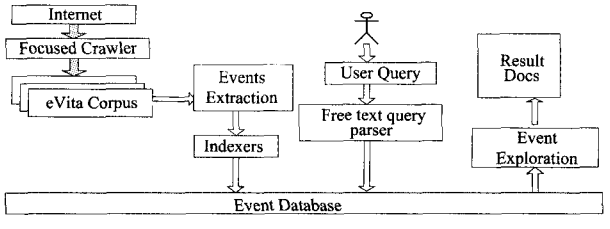

图1 框架总体架构图

(1)主题爬虫负责从互联网搜索和下载个人简历。和通用爬虫不同,主题爬虫只下载指定类型的文档(如个人简历)。因此主题爬虫的关键技术是文本分类。

(2)事件抽取组件负责从下载的个人简历中自动地抽取感兴趣的事件。事件抽取算法的实现基于层次条件随机场(Hierarchical Conditional Random Fields,HCRF)[29],层次条件随机场是条件随机场(Conditional Random Fields,CRF)[17]的扩展。.

(3)事件检索组件利用事件的属性(who,when,where,what,how)挖掘发现事件之间的关联关系,并根据用户的检索条件返回经过排序的相关事件集合。

本研究工作的主要贡献为:

(1)提出了一个统一的事件表示模型,可以描述事件的基本属性,如who,when,where,what,how。能很方便地用于挖掘事件之间的关联关系及事件检索。

(2)基于层次条件随机场,提出了一个概率统计模型,用于自动地从个人简历中提取事件。该模型可以同时标记出事件块(Event Block)和每个事件的属性。

(3)在事件表示模型和事件抽取算法的基础上,提出了事件的检索算法。算法基于事件属性的共现性,如co-when,co-where,co-what,co-who。

本文第2节介绍和本文相关的研究工作;第3节描述事件表示模型及事件抽取算法;第4节介绍事件检索算法;第5节给出了实验方法和实验结果;最后对全文进行总结。

2 相关工作

关于文本分类的研究工作可以分为2类:基于概率统计的方法和基于向量空间模型的方法。基于概率统计的方法,如Naïve Bayes、Bernoulli,利用MAP(Maximum a Posteriori)准则试图找到"最佳"的分类结果;基于向量空间模型的分类方法将文档表示为向量,向量每一维的值为对应term的tf/idf值,然后分类器试图在向量空间里找到一个"最佳"边界对文档向量分类。这类方法中,最有代表性的是支撑向量机(Support Vector Machine,SVM)算法[4]。

信息/事件抽取(Information/Event Extraction)研究领域也有一些不同类型的算法。第一种方法基于分类算法标记文本属性,进而完成信息抽取。例如,文献[12,18]利用SVM确定文本属性的标记。另一种方法则是预先定义网页的模板,然后基于模板进行信息抽取[2,5]。然而,基于分类的信息抽取方法通常是将文档表示为向量空间里的向量,不适合建模文本属性之间的依赖关系;另一方面,基于模板的信息抽取方法存在的问题是要维护大量网页模板,工作量巨大。

一种更有效,因而被更普遍使用的方法是基于概率统计的方法,如隐马尔科夫模型(Hidden Markov Model,HMM)[10]、最大熵模型(Maximum Entropy Model,MEM)[3]、最大熵马尔科夫模型(Maximum Entropy Markov Model,MEMM)[19]和线性链式条件随机场(Linear-chain CRF,LCRF)[17]。这类方法将信息/事件抽取任务看作是文本序列的标记问题。上述模型可以描述输出标记间的依赖关系,标记间的依赖关系可以提高标记结果的精度。但是,上述模型只能建模输出标记间的一阶依赖关系。

作为目前最好的信息抽取的概率模型,CRF以不同方式被改进,以提高信息抽取的精度。例如,HCRF[29]和Tree-structured CRF(TCRF)不仅能建模标记间的线性链式(linear-chain)依赖关系,而且能建模标记间更高阶的依赖关系。因此,CRF模型的扩展模型可以取得比LCRF更好的效果。另外一个CRF的扩展模型将CRF模型扩展到二维,从而可以建模输出标记在二维空间里的依赖关系[28]。Multi-scale CRF[11]可以集成从不同尺度的空间里抽取出的信息。Semi-CRF[21]和CRF的区别是:CRF对单个单词进行标记,Semi-CRF则对一段文本(包含多个单词)进行标记,同时Semi-CRF的特征函数也是对一段文本的特征进行描述。Constrained CRF (CCRF)[15]和Constrained HCRF[27]则利用Viterbi解码算法寻找最佳标记的过程中加入自定义的约束条件,使得计算得到的最佳标记能满足条件约束。Dynamic CRF (DCRF)[24]和Dynamic HCRF (DHCRF)则针对传统的CRF模型的结构必须是静态不变的问题,提出了结构可以动态变化的模型。

另外一些研究则关注事件关联关系的挖掘和事件的检索。例如基于事件的电子纪事[26]和多媒体电子纪事[14],文

献[26]提出了第一个可伸缩的简历系统，其提供了一个集成的多模态交互环境，可以将基于事件的多模态信息集成在一起，供用户检索。文献[1]则提出了基于多媒体的事件检索框架，其允许用户以可视化的方式来检索和浏览共享的多媒体信息。另一方面，文献[23]关注如何利用事件发生时间和地点的共现性来定量计算事件之间的关联度。另外一些基于文本的事件检索系统包括事件检测[7]、层次化事件的构建[9]、事件关联关系推理[22]和事件日历[20]。

3 事件抽取

3.1 事件的表示

个人简历通常将内容分块组织。一份典型的简历通常包括以下内容块：个人基本信息、教育背景、工作经历和论文发表情况等。不同的内容块里，信息项以不同的模式列出，将某个内容块里信息项的列表模式称为 inner-block-pattern。例如，教育背景内容块的 inner-block-pattern 可能是时间、就读大学、所学专业；论文发表情况内容块的 inner-block-pattern 可能是作者、论文题目、所发表的期刊或会议论文集、发表年份。另一方面，不同简历的内容块可能以不同的次序列出，将简历中不同内容块的排列次序称为 inter-block-pattern。例如一个大多数简历的 inter-block-pattern 应该首先是个人基本信息内容块，接下来依次是教育背景内容块、工作经历内容块，最后是发表论文情况。

每个内容块里列出的信息项都是按时间顺序排列的，因此这些信息项可以被看作一系列的事件，这些事件都具有 5 个属性：who(事件的主体)、when(事件发生的时间)、where(事件发生的地点)、what(事件导致的结果)、how(事件的动作)。因此可以将事件 E 表示为一个五元组：

$$E=\langle who, when, where, what, how\rangle$$

例如，一条发表论文的信息项“John Smith, Conditional Random Fields, Proc. of ACM SIGIR, 2010”可以表示为 5 元组〈who (John Smith), what (Conditional Random Fields), where (Proc. of ACM SIGIR), when (2010), how (Publication)〉。其中“who”、“what”、“where”和“when”分别代表作者、文章题目、论文发表的期刊或会议论文集和论文发表时间，而“how”代表事件的动作：publication。

事件的动作属性(how)由包含该事件的内容块决定。在上面的例子中，事件的动作为“publication”，因为该事件位于内容块“publication”中。需要注意的是，“what”和“where”属性在不同的内容块里有不同的意义。例如，一个动作属性为“publication”的事件，其“what”和“where”属性分别代表论文的题目和论文发表的期刊或会议论文集；而对于一个位于工作经历内容块的事件，其“what”和“where”属性分别代表工作职位和工作单位。

基于上述分析，一份个人简历由一系列事件组成，而相同类型的事件被组织在同一个事件块(Event Block)里。受文献[6]的研究启发，用层次化的事件图来表示一份个人简历。图 2 显示了一个二层事件图。事件图高层包含了 3 个事件块，分别为 education、experience 和 publication。每个事件块里包含的事件由事件图低层表示。在图 2 中，不同的事件块包含了不同类型的事件，因此用不同的形状表示(椭圆、矩形和三角形)。

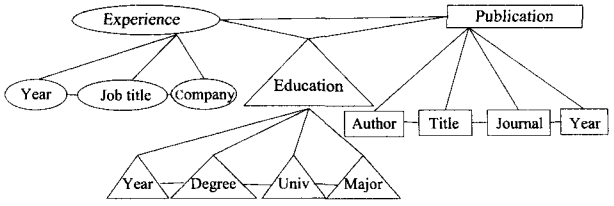

图 2 个人简历 2 层事件图

3.2 事件抽取模型

3.2.1 问题定义

因为个人简历由事件块组成，每个事件块包含不同类型的事件，因此事件抽取需要解决 2 个问题：1)如何自动识别事件块；2)如何自动识别事件块所包含事件的属性(who、when、where、what 和 how)。

如果把个人简历看作是可被观察的单词输入序列，事件抽取问题就可以被描述成单词输入序列的标记问题。即给定一个单词输入序列，如果自动标记出每个单词属于哪个事件块以及每个单词属于事件的哪个属性，单词输入序列的标记问题就可以被形式化定义为给定输入数据序列 $X=\{x_1, x_2, \cdots, x_n\}$，并给定标记字母表 $Y=\{y_1, y_2, \cdots, y_n\}$(即每个单词的标记只能从集合 Y 中选取)，事件抽取的目标为找到最佳的标记序列 Y^*，并满足

$$Y^* = \arg\max_Y P(Y|X) \tag{1}$$

3.2.2 标记字母表

有 2 种类型标记用作事件抽取：事件块的标记和事件属性的标记。由于事件的动作属性(how)由事件块决定，而事件其它属性的意义随着包含该事件块的不同而变化，因此使用更有具体意义的单词来标记不同事件块里的事件属性。使用的标记字母表如表 1 所列。

表 1 标记字母表

Type	Label	Meaning
Event Block	Education	Education block
	Experience	Work experience block
	Publication	Publication block
Event Attribute	EDU_TIME	‘When’ of education
	WORK_TIME	‘When’ of experience
	PUB_TIME	‘When’ of publication
	DEGREE	‘What’ of education
	MAJOR	
	COMPANY	‘Where’ of experience
	SCHOOL	‘Where’ of education
	JOBTITLE	‘What’ of experience
	AUTHOR	‘who’ of publication
	PAPERTITLE	‘what’ of publication
	JOURNAL	
	PROCEEDINGS	‘where’ of publication
	PAGES	

表 1 中没有列出教育背景和工作经历类型事件的事件属性“who”，因为这 2 种类型事件的“who”属性是通过抽取个人基本信息块中的姓名得到的。同时，所有类型事件的“how”属性由包含该事件的事件块决定，因此表 1 中也没有列出。

3.2.3 低层事件属性标记模型

低层事件属性标记的任务是对事件块里的事件项自动标

记出每个单词的事件属性。这里采用的标记模型是 LCRF。下面给出 CRF 的形式化定义。

CRF 形式化定义：假设图 $G=(V,E)$，X 为每个节点 $v\in V$ 上可观察的输入数据的集合，$X=(x_v)_{v\in V}$，Y 为每个节点 $v\in V$ 上的标记结果的集合，$Y=(y_v)_{v\in V}$，集合 X 和 Y 都通过图 G 的节点来索引，则(X,Y)为条件随机场。当满足给定输入数据集合 X，随机变量 y_v 具有马尔科夫性质，即 $P(y_v|X,y_w,w\neq v)=P(y_v|X,y_w,w\sim v)$。其中，$w\sim v$ 表示在图中的节点 w 和 v 为邻居。马尔科夫性质指，节点 v 的标记为 y 的条件概率仅由输入数据 X 和节点 v 的邻居节点的标记决定。

LCRF 假设节点的输出标记间仅满足一阶马尔科夫性质，并且节点的输出标记间彼此通过无向边链接而形成线性链。输出标记的条件概率由一些特征函数及其相应权重来决定：

$$P(Y|X)\propto\exp\sum_{t}\left(\sum_{k=1}^{K}\lambda_k f_k(y_t,y_{t-1},X)+\sum_{h=1}^{H}u_h f_h(y_t,X)\right) \tag{2}$$

在 LCRF 中，有 2 类特征函数：转移函数 $f_k(y_t,y_{t-1},X)$ 定义了相邻位置 t 和 $t-1$ 的输出标记的依赖关系；状态函数 $f_h(y_t,X)$ 定义了位置 t 的输出标记和输入 X 之间的依赖关系。

为了利用 LCRF 来标记低层事件属性，需要定义合适的特征函数，同时通过机器训练来估计特征函数的权重。特征函数的定义将在 3.3 节介绍。

3.2.4 高层事件块标记模型

个人简历中事件由事件块组成，这样的层次关系可以利用 HCRF 来建模事件块之间、事件属性之间的依赖关系。对于 LCRF，特征函数仅仅被定义在一阶 clique(节点)和二阶 clique(边)上。但对于 HCRF，其特征函数可定义在更高阶的 clique 上。因此对于 HCRF，输出标记的条件概率定义为：

$$P(Y|X)=\frac{1}{z(x)}\exp\sum_{c\in C}\sum_{k=1}^{K}\lambda_k f_k(Y_c,X) \tag{3}$$

式中，C 是所有 cliques 的集合，$f_k(Y_c,X)$是定义在 clique c 上的特征函数，λ_k 是特征函数的权重。

为了建模事件属性标记之间和事件块标记之间的依赖关系，用图 2 表示输出标记之间的依赖关系。每一对低层节点之间的边表示了事件属性标记之间的依赖关系，其特征函数由事件属性特征函数定义；从高层的事件块节点到低层每个事件属性节点的边表示了事件块标记和低层事件属性标记间的依赖关系，其特征函数由事件属性特征函数和事件块特征函数共同决定。

3.3 特征函数

本节将定义用于标记事件块和事件属性的特征函数。特征函数是 CRF 及其扩展模型的关键，它决定了利用模型进行事件抽取的精度。特征函数包括事件属性特征函数和事件块特征函数 2 类。

3.3.1 事件属性特征函数

本文定义了 4 种类型事件特征函数。

(1)字典。字典是最简单的特征函数。我们已经建立了关于大学名称、英文个人名字、期刊和会议论文集名称这 4 种字典。表 2 给出了每种字典包含的项数。

表 2 4 种字典包含的项数

Dictionary	Number of items
University	2378
Personal Name	3354
Journal	10604
Proceedings	4627

一个基于字典的特征函数的定义示例为：

$$f_h(y_t,x_t)=\begin{cases}1, & \text{if } x_t\in D_{name}\wedge y_t=\text{AUTHOR}\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

该特征函数的含义是：如果位置 t 的输入单词属于个人姓名字典(D_{name})中的一项，则该单词最有可能被标记为 AUTHOR。

(2)单词特征。大写开头的单词通常是命名实体的名字(如人名、大学名、期刊或会议论文集名)；数字字符串通常代表时间或发表论文的页数。我们采用正则表达式来定义这样的单词特征。表 3 列出了 3 类正则表达式所定义的单词特征。

表 3 单词特征

Feature name	Regular expression
Time	(19|20)\\d{2}\\s*(\\-|to)\\s*(19|20)\\d{2}
Name	[\\p{Lu}]\\.(\\s*[\\p{Lu}]\\.)?[\\s]*[\\p{L}]+
Page number	(page|pages)\\s*[.|:|,|\\s]\\s*\\d+\\s*\\-\\s*\\d+

一个基于单词特征的特征函数的定义示例为：

$$f_h(y_t,x_t)=\begin{cases}1, & \text{if } x_t\in Pattern_{time}\wedge y_t=\text{TIME}\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

(3)单词之间的相邻关系特征。单词之间的相邻关系往往也包含了重要的特征。例如，一个论文发表项通常从作者开始，同时论文发表项的作者、文章题目、论文发表期刊或会议论文集和发表年份之间通常用标点符号分隔开。一个基于单词之间的相邻关系特征的特征函数定义示例为：

$$f_k(y_t,y_{t-1},x_t)=\begin{cases}1, & \text{if } x_t=\text{'.'}\wedge y_{t+1}=\\ & \quad \text{PARERTITLE}\wedge y_{t-1}=\text{AUTHOR}\\ 0, & \text{otherwise}\end{cases} \tag{6}$$

该特征函数的含义是：如果位置 $t-1$ 的单词标记为 AUTHOR，位置 $t+1$ 的单词标记为 PAPERTITLE，位置 t 为标点符号“.”，则特征函数值为 1。即作者和文章名字之间更可能由标点符号“.”分隔开。

3.3.2 事件块特征函数

事件块定义了如下 3 种特征：

(1)事件块标题的视觉特征。输入单词序列的视觉特征，如字体类型、字体大小、字体粗细通常包含了重要的线索，以识别事件块。例如位于不同事件块内的事件通常用相对小的字体，但字体大小和字粗细应该一样；而事件块的标题通常用更大的字体，而且通常用粗体。

(2)事件块标题的内容。不同的事件块的标题往往包含可以用来识别事件块类型的关键单词，例如“publication”和“education”。

(3)事件块所包含事件的属性。一个事件块内的事件属性标记也能用来识别事件块的类型。例如发表论文事件块应包含如下事件属性标记：AUTHOR、JOURNAL 或 PROCEEDINGS。

4 事件检索

事件被抽取出来后，保存在事件数据库中，同时被保存的还有事件的相关信息，如包含事件的原始上下文。本节将介绍如何挖掘事件直接的关联关系并从数据库中检索事件。

4.1 事件检索模型

当检索用户发出一个基于关键词的检索时，事件检索引擎会在数据库里搜索相关的事件。

另一方面，存储在数据库中的事件可能彼此间相互关联，事件之间的关联关系可通过在事件数据库中挖掘事件任何属性之间的共现性而被发现。例如，如果一个命名实体的名字在2个事件中都出现，那么这2个事件就可以通过这个名字关联起来。数据库也会保存所有事件之间的关联关系。为了便于描述事件检索模型，首先给出如下定义和符号。

定义1 事件被表示为一个七元组 $E=\langle p,t,l,e,a,c,L\rangle$。其中 p,t,l,e,a 分别表示和事件有关的人("who"属性)、事件发生的时间("when"属性)、事件发生的地点("where"属性)、事件的结果("what"属性)、事件的动作("how"属性)。c 表示包含事件的原始上下文(即个人简历)。L 表示该事件与其它事件的所有关联关系。七元组的每个元素称为事件的属性。用符号 $e.attr$ 表示事件 e 的属性，其中 $attr$ 的取值范围是 p,t,l,e,a,c,L。

定义2 事件的关联类型(Link Type)是指：如果一对事件通过属性 $attr$ 发生关联关系，则这对事件的关联类型记为 LP_{attr}，其中 $attr$ 的取值范围是 p,t,l,c。

定义3 事件 e 和其它事件的关联关系的集合记为 $e.L$，一个关联关系项 $l\in e.L$ 被表示为元组 $l=\langle e,e',LT\rangle$，其中 e,e' 是发生关联关系的一对事件，LT 是关联类型。

给定一组已被抽取出的事件，事件之间的关联关系基于事件属性的共现性计算。例如，关联类型为 LT_p 的关联关系可以被形式化地定义为 $HashLink(l)$ 函数：

$$HashLink(l=\langle e,e',LT_p\rangle)=\begin{cases}\text{true}, & \text{if } e.p\cap e'.p\neq\emptyset\\ \text{false}, & \text{otherwise}\end{cases} \tag{7}$$

给定一对事件 e,e' 和关联类型 LT，如果函数 $HashLink(l=\langle e,e',LT_p\rangle)$ 返回 true，则关联关系项 $L=\langle e,e',LT\rangle$ 被加入到事件 e,e' 的关联关系集合 L 中。所有事件间的关联关系计算完毕后，就得到了事件关系图 $G=(V,E)$，其中 V 为事件的集合，E 为事件间关联关系的集合。

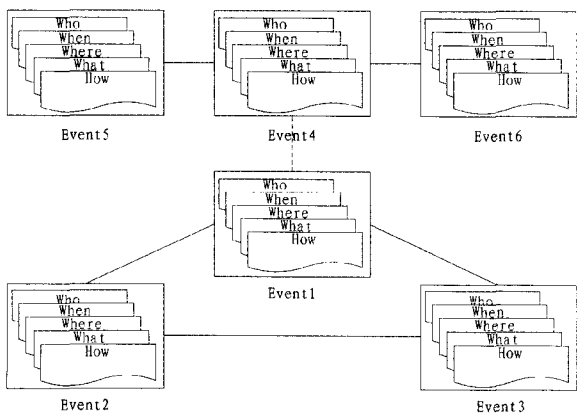

图3 事件关联关系图

图3显示了在事件关系图中如何通过关联关系将2组事件联系起来(图中的虚线所示)。图3显示了2组事件，每组内的事件彼此间有关联关系(图中实线所示)。例如，第一组的事件1、事件2、事件3之间可能通过事件相关人产生关联；第二组的事件4、事件5、事件6之间可能通过事件发生时间产生关联。假设事件1和事件4之间通过事件发生地点产生了关联，那么这2个事件组在图中就被连通。

4.2 事件重要性度量

基于计算得到的事件关联关系，可以定量计算事件的重要程度。事件重要性的计算基于如下假设：一个事件与其它事件的关联关系越多，其重要性越高；反之，其重要性越低。

因此，最直接和简单的方法就是将一个事件的关联关系数作为重要性度量。但是，事件之间关联关系的强度应该作为重要因素加以考虑。例如，2个事件发生的时间相隔越远，它们之间的关联强度越弱。我们把关联关系强度作为事件关系图中边的权重，并把事件关联关系项的强度记为 $W(l)$，将事件 e 的重要性度量记为 $l(e)$。$l(e)$ 的定义为：

$$l(e)=\sum_{l\in e.L}W(l) \tag{8}$$

即事件的重要性度量等于在事件关系图中从事件 e 出发的每条边的权重之和。

事件关联关系强度 $W(l)$ 则根据下列事件属性之间的相似度来计算：p(事件有关的人)、t(事件发生的时间)、l(事件发生的地点)、e(事件的结果)、c(包含事件的原始上下文)，分别记为 $S_p(e,e')$、$S_t(e,e')$、$S_l(e,e')$、$S_e(e,e')$ 和 $S_c(e,e')$。

事件有关人属性之间的相似度用 Jaccard 相似度度量。式(9)的含义是如果和2个事件都有关的人越多，事件有关人属性的相似度就越高。

$$S_p(e,e')=\frac{|e.p\cap e'.p|}{|e.p\cup e'.p|} \tag{9}$$

事件时间属性的相似度定义为：

$$S_t(e,e')=\frac{1}{1+e^{|e.t-e'.t|}} \tag{10}$$

为了计算事件地点属性之间的相似度，首先将事件的地点属性文本做分词处理，并将事件 e 的地点属性的分词结果记为 $T(e.l)$。事件地点属性之间的相似度定义为：

$$S_l(e,e')=\frac{|T(e.l)\cap T(e'.l)|}{|T(e.l)\cup T(e'.l)|} \tag{11}$$

类似地，事件结果属性的相似度定义为：

$$S_e(e,e')=\frac{|T(e.e)\cap T(e.e)|}{|T(e.e)\cup T(e.e)|} \tag{12}$$

事件上下文属性之间的相似度定义为：

$$S_c(e,e')=\begin{cases}1, & \text{if } e.c=e'.c\\ 0, & \text{otherwise}\end{cases} \tag{13}$$

上面定义的属性之间的相似度都在0～1之间，因此事件关联关系项 l 的强度 $W(l)$ 的定义为：

$$W(l)=\lambda_pS_p+\lambda_tS_t+\lambda_lS_l+\lambda_eS_e+\lambda_cS_c \tag{14}$$

式中，λ_p、λ_t、λ_l、λ_e 和 λ_c 为相应的权重。

4.3 事件排序

事件的重要性度量可以作为与用户查询无关的排序得分值。给定一个用户查询 q，将 q 的分词结果记为 $T(q)$，事件 e 的分词结果记为 $T(e)$。根据 q 和 e 的分词结果及词频信息，构建相应的单词向量 $V(q)$ 和 $V(e)$，则与查询相关的事件排

序得分值 $R(q,e)$定义为向量之间夹角的余弦值。因此，最终的排序得分 $S(q,e)$定义为：

$$S(q,e)=l(e)+R(q,e) \tag{15}$$

5 实验结果及分析

5.1 数据集准备

实验基于互联网上下载的 785 份英文个人简历(PDF 格式)进行。PDF 格式的简历内容利用开源软件 PDFBox(http://pdfbox.apache.org/)来解析。对 LCRF 和 HCRF 模型，分别使用 50%的个人简历作为模型的训练集和测试集。

另一方面，使用 785 份简历文档和 785 份非简历文档(选自路透社 RCV-1 英文文档集合)做个人简历文档分类测试。训练和测试使用的分类器是 SVM，同时使用 Naïve Bayes 分类器作为分类精度比较的基准。

5.2 主题爬虫和文本分类

我们实现了一个主题爬虫，以自动地从互联网上下载个人简历。与通用爬虫相比，主题爬虫必须能自动判断出一个文档是否为个人简历，我们使用基于 SVM 的分类器来自动判断文档的类型。主题爬虫自动找到个人简历，要经过下列步骤：

(1)对爬虫抓取的每个文档，首先进行分词处理，然后将文档表示成 tf/idf 加权的向量。为了减小单词词典的大小，采用了基于文档频率的方法进行特征提取，以过滤掉那些对于正确分类贡献不是很大的单词。

(2)利用 SVM 分类器将非个人简历文档过滤掉。

另外，定义了一些启发式规则以快速排除非简历文档，而无需调用 SVM 分类器。例如，如果一个文档没有包括下列任何单词之一，如“resume”、“curriculum”或“vitae”，就被认为是非简历文档而不用调用分类器。

5.3 性能评价标准

为了综合评估系统性能，我们采用不同的评价准则。对于 SVM 分类器，用分类精度作为评估准则。对于事件自动抽取的性能评估，对每个标记采用 $F1$ 评价标准，$F1$ 是对标记精度和召回率的综合评价。

如果定义 A 为被标记结果为 true positive 的单词数量，B 为被标记结果为 false negative 的单词数量，C 为被标记结果为 false positive 的单词数量，D 为被标记结果为 true negative 的单词数量，则 $F1$ 定义为精度(记为 *Precision*)和召回率(记为 *Recall*)的调和平均：

$$F1=\frac{1*Precision*Recall}{Precision+Recall} \tag{16}$$

其中，精度定义为：

$$Precision=\frac{A}{A+C} \tag{17}$$

召回率定义为：

$$Recall=\frac{A}{A+B} \tag{18}$$

5.4 文本分类实验结果

对于文本分类实验，从 785 份简历文档和 785 份非简历文档中按不同比例抽取文档，将其作为训练集进行分类实验，实验结果如表 4 所列。

表 4 分类实验结果

Training Set	Testing Set	Accuracy(SVM)	Accuracy(NB)
628(40%)	942(60%)	99.58%	99.04%
785(50%)	785(50%)	99.36%	98.98%
1256(80%)	314(20%)	99.36%	98.73%

从实验结果可以看出，SVM 分类器在精度上要好于 Naïve Bayes 分类器；而且我们的分类实验的精度都超过 99%。

5.5 事件抽取实验结果

事件抽取实验采用 LCRF 开源工具包 MALLET(http://mallet.cs.umass.edu/)和 HCRF 工具包 GRMM(http://mallet.cs.umass.edu/grmm/index.php)。实验效果的评估基于简历文档中的论文发表情况(publication)、工作经历(work experience)和教育背景(education) 3 类事件进行。训练集合大小为 162 个事件，测试集合大小为 541 个事件。为了评估标记结果，所有事件实例都由人工进行标注，以作为评估标准。同时采用 LCRF 和 HCRF 2 种模型进行实验。实验结果显示，HCRF 的 $F1$ 值要好于 LCRF，这是因为 HCRF 考虑到了简历中事件组织的层次结构。基于 LCRF 和 HCRF 的事件抽取结果如表 5、表 6 所列。

表 5 事件抽取结果(LCRF)

Event Attributes	Pre	Rec	F1
EDU_TIME	0.975	0.922	0.948
WORK_TIME	0.989	0.989	0.989
PUB_TIME	0.950	0.947	0.948
DEGREE	0.977	0.937	0.956
MAJOR	0.793	0.732	0.761
COMPANY	0.680	0.756	0.716
SCHOOL	0.686	0.828	0.750
JOBTITLE	0.962	0.726	0.828
AUTHOR	0.998	0.998	0.998
PAPERTITLE	0.903	0.952	0.927
JOURNAL	0.844	0.775	0.808
PROCEEDINGS	0.798	0.858	0.827
PAGES	0.997	0.999	0.998

表 6 事件抽取结果(HCRF)

Measure		Pre	Rec	F1
Block	Education	0.963	0.988	0.975
	Experience	1.000	0.977	0.988
	Publication	0.961	0.984	0.973
Event Attributes	EDU_TIME	0.884	0.946	0.914
	WORK_TIME	0.997	0.981	0.989
	PUB_TIME	0.953	0.932	0.942
	DEGREE	0.938	0.963	0.950
	MAJOR	0.851	0.874	0.863
	COMPANY	0.922	0.928	0.925
	SCHOOL	0.845	0.899	0.871
	JOBTITLE	0.846	0.839	0.842
	AUTHOR	0.996	0.996	0.996
	PAPERTITLE	0.911	0.980	0.944
	JOURNAL	0.911	0.822	0.864
	PROCEEDINGS	0.914	0.735	0.815
	PAGES	1.000	1.000	1.000

结束语 研究了如何从个人简历中提取事件以及如何检索事件。个人简历中包含的非结构化信息可以看作是按事件排序的事件序列。本文首先提出了基于五元组的事件表示模型，在此基础上，基于 HCRF 的事件抽取模型对简历中的事件块和事件属性进行标记，从而从非结构化的文本中自动抽取出事件。抽取的事件被存储在数据库中后，利用事件属性

的共现性发现事件之间的关联关系。同时提出了事件重要性度量模型以及事件排序算法。

下一步工作包括结合其他事件抽取模型提高事件抽取的精度；当自动抓取的个人简历和抽取出来的事件达到一定规模后，研究如何利用事件之间的链接关系实习自动的主题社区发现，如自动找到相同专业的个人社区等。

参 考 文 献

[1] Appan P, Sundaram H. Networked Multimedia Event Exploration[C]// Proceedings of the 12th Annual ACM International Conference on Multimedia. ACM Press, 2004: 40-47

[2] Arasu A, Garcia-Molina H. Extracting Structured Data from Web Pages[C]// Proceedings of 2003 ACM SIGMOD International Conference on Management of Data. ACM Press, 2003: 337-348

[3] Berger A L, Pietra D V J, Pietra D, et al. A Maximum Entropy Approach to Natural Language Processing[J]. Computational Linguistics, 1996, 22(1): 39-71

[4] Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998, 2, (2): 121-167

[5] Butter D, Liu L, Pu C. A Fully Automated Object Extraction System for the World Wide Web[C]// Proceedings of the 21th International Conference on Distributed Computing Systems (ICDCS 2001). IEEE Computer Society, 2001: 361-370

[6] Cai D, Yu S, Wen J R, et al. Blocked-based Web Search[C]// Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, 2004: 456-463

[7] Chieu H L, Lee Y K. Query based Event Extraction Along a Timeline[C]// Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, 2004: 425-432

[8] Fung G P C, Yu J X, Yu P S, et al. Parameter Free Bursty Events Detection in Text Streams[C]// Proceedings of the 31th International Conference on Very Large Data Bases (VLDB). VLDB Endowment, 2005: 181-192

[9] Fung G P C, Yu J X, Liu H, et al. Time-dependent Event Hierarchy Construction[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). ACM Press, 2007: 300-309

[10] Ghahramani Z, Jordan M I. Factorial Hidden Markov Models [J]. J. Machine Learning, 1997, 29(2/3): 245-273

[11] He X, Zemel R S, Carreira-Perpiñán M A. Multiscale Conditional Random Fields for Image Labeling[C]// Proceedings of 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2004, 2: 695-702

[12] Iria J, Ireson N, Ciravegna F. An Experimental Study on Boundary Classification Algorithms for Information Extraction Using SVM[C]// Proceedings of the EACL Workshop on Adaptive Text Extraction and Mining (ATEM). 2006: 17-24

[13] Jain R. Event Web[J]. IEEE Multimedia, 1999, 6(2): 1

[14] Jain R. Multimedia Electronic Chronicles[J]. IEEE Multimedia, 2003, 10(3): 102-103

[15] Kristjansson T, Culotta A, Viola P, et al. Interactive Information Extraction with Constrained Conditional Random Fields[C]// Proceedings of the 19th National Conference on Artificial Intelligence (AAAI). AAAI Press/The MIT Press, 2004: 412-418

[16] Kumar S, Hebert M. A Hierarchical Field Framework for Unified Context-based Classification[C]// Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV 2005). IEEE Computer Society, 2005: 1284-1291

[17] Lafferty J D, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Proceedings of the 18th International Conference on Machine Learning (ICML 2001). Morgan Kaufmann Publishers Inc, 2001: 282-289

[18] Li Y, Bontcheva K, Cunningham H. Using Uneven Margins SVM and Perception for Information Extraction[C]// Procee-dings of the 9th International Conference on Computational Na-tural Language Learning (CoNLL 2005). 2005: 72-79

[19] McCallum A, Freitag D, Pereira F C N. Maximum Entropy Markov Models for Information Extraction and Segmentation [C]// Proceedings of the 17th International Conference on Machine Learning (ICML 2000). Morgan Kaufmann Publishers Inc, 2000: 591-598

[20] Payne T R, Singh R, Sycara K P. Browsing Schedules-An Agent-based Approach to Navigating the Semantic Web[C]// Proceedings of the First International Semantic Web Conference on the Semantic Web (ISWC 2002). Springer 2002: 469-474

[21] Sarawagi S, William W C. Semi-Markov Conditional Random Fields for Information Extraction[C]// Proceedings of the 18th Annual Conference on Neural information Processing Systems (NIPS 2004). The MIT Press, 2004: 1185-1192

[22] Shaw R. Event Gazetteers for Navigating Humanities Resources [C]// Proceeding of the 2nd PhD Workshop on Information and Knowledge Management (PIKM 2008). ACM Press, 2008: 89-92

[23] Smith D A. Detecting and Browsing Events in Unstructured Text [C]// Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002). ACM Press, 2002: 73-80

[24] Sutton C, Mccallum A, Rohanimanesh K. Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data [J]. The Journal of Machine Learning Research, 2007(8): 693-723

[25] Tang J, Hong M, Li J, et al. Tree-structured Conditional Random Fields for Semantic Annotation[C]// Proceedings of the 5th International Semantic Web Conference (ISWC 2006). Springer 2006: 640-653

[26] Wu B, Singh R, Gupta P, et al. eVita: An Event-based Electronic Chronicle[C]// Proceedings of the 9th International Conference on Extending Database Technology (EDBT 2004). Springer 2004: 834-836

[27] Xin X, Li J, Tang J, et al. Academic Conference Homepage Understanding Using Constrained Hierarchical Conditional Random Fields[C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008). ACM Press, 2008: 1301-1310

[28] Zhu J, Nie Z, Wen J R, et al. 2D Conditional Random Fields for Web Information Extraction[C]// Proceedings of the 22nd International Conference on Machine Learning (ICML 2005). ACM Press, 2005: 1044-1051

(下转第 174 页)

TOPn准确率是指召回翻译的术语中前n个翻译项中存在正确翻译的术语比率。实验数据表明，TOP1准确率达到了97.4%，翻译对的质量令人满意。选取Fang Gaolin在文献[7]与文献[8]的方法及Zhang在文献[5]的方法进行对比，可以看出，本文提出的3种翻译验证模式确保了翻译的准确性，准确率方面，特别是TOP1准确率较Fang Gaolin及Zhang的方法有较大幅度的提高，并增强了系统的实用性和可靠性。

传统的基于搜索引擎的翻译获取系统，需要构建查询项进行搜索引擎查询，耗时较多。本文基于相关性反馈动态控制网页下载量，定义一个有效翻译递减率来刻画一个查询项返回结果中每页获取到的候选翻译数相对前一页的递减率，以动态终止查询项。实验中每个术语的平均耗时如表2所列。

表2 术语平均耗时

	位置约束
本文方法平均耗时	8.3s
Fang的方法平均耗时	9.2s
Zhang的方法平均耗时	15s

5.2 翻译获取

考察3种查询项模式各自返回的候选翻译的比率和候选翻译的正确率，如表3所列。

表3 查询项模式返回候选翻译比率及正确率

	位置约束	共现约束	相关词约束
比率	21%	66%	13%
正确率	41%	24%	17%

表3的结果表明，位置约束的查询项模式虽然获取的翻译数量较少，但是获取到的翻译的准确性最高。

通过3种翻译抽取模式抽取出的候选翻译的比率及候选翻译的正确率如表4所列。

表4 翻译抽取模式抽取的候选翻译比率及正确率

	基于模板	基于词典	基于位置
比率	7%	5%	88%
正确率	72%	68%	20%

表4的结果表明，基于模板模式和词典模式抽取出的翻译准确性较高，基于位置的抽取方法则可以最大限度地保证翻译的召回率。

5.3 候选翻译验证

表5 不同验证模式下的实验结果

	TOP1正确率	召回率
All	97.4%	90.0%
端类比对齐+构词法验证	93.8%	94.5%
双语对齐度+构词法验证	90.4%	92.2%
无验证	71.4%	99.2%
Fang的验证方法	82.0%	95.3%

表5列出了实验中候选翻译经过不同验证模式后，实验结果的正确率和召回率。表中的4组实验表明，在候选翻译没有经过验证时，TOP1正确率仅有71.4%，采用单一验证方法效果不佳。针对候选翻译的特点，将3种验证方法进行结合，可以在保证召回率的情况下，有效提高实验结果的准确性，TOP1准确率上升到97.4%。作为对比，采用Fang在文献[7]的验证方法对实验数据进行验证。Fang的方法是基于互信息的方法，去除候选数据中的前缀或后缀冗余信息。实验表明，验证后的TOP1正确率提高到82.0%。但由于去除冗余信息需要通过比较多个候选数据，因此其不适用于网页搜索结果较稀疏的术语，且它易受到无关干扰项的影响。

结束语 本文提出了一种通过网页获取专业术语翻译的方法。通过术语的部分翻译构建查询项，极大地提高了搜索引擎返回结果的相关性。利用多特征的翻译抽取方法，在保证召回率的基础上抽取出准确的候选翻译。最后提出了3种验证模型，排除候选翻译中的干扰项，保证了结果的准确性。实验结果表明，抽取出的翻译对准确性高，错误的干扰项极少，具有较好的实用价值。

参考文献

[1] Huang F, Vogel S, Waibel A. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization[C]// Proceedings of ACL 2003 Workshop on Multilingual and Mixed Language Named Entity Recognition. 2003: 9-16

[2] Huang F, Vogel S. Improved Named Entity Translation and Bilingual Named Entity Extraction[C]// Proceedings of ICMI. 2002: 253-258

[3] Zhang Y, Vines P. Detection and Translation of OOV Terms Prior to Query Time[C]// Proceedings of SIGIR. 2004: 524-525

[4] Cao Gui-hong, Gao Jian-feng, Nie Jian-yun. A System to Mine Large-scale Bilingual Dictionaries from Monolingual Web Pages [C]// Proceedings of MT Summit XI. 2007

[5] Zhang Y, Vines P. Using the Web for Automated Translation Extraction in Cross-language Information Retrieval [C]// Proceedings of SIGIR. 2004: 162-169

[6] Huang F, Zhang Y, Vogel S. Mining Key Phrase Translations from Web Corpora [C]// Proceedings of HL T2EMNLP. 2005: 483-490

[7] Fang Gao-lin, Yu Hao, Nishino F. Chinese-English Term Translation Mining Based on Semantic Prediction [C]// Proceedings of the COL ING/ACL 2006 Main Conference Poster Sessions. 2006: 199-206

[8] Fang Gao-lin, Yu Hao. Web Translation Mining Based on Suffix Arrays[J]. Journal of Chinese Language and Computing, 2007, 17 (1): 1-141

[9] 吕学强，吴宏林，姚天顺. 无双语词典的英汉词对齐[J]. 计算机学报，2004，27(8)：1036-1045

[10] 何彦璋. 从Web中获取中文术语的英文翻译的方法研究与实现[D]. 北京：北京航空航天大学，2008

[11] 符建辉，曹存根，王石. 基于区分词的汉语隐喻短语识别[J]. 计算机科学，2010，37(10)：193-196

(上接第160页)

[29] Zhu J, Nie Z, Wen J R, et al. Simultaneous Record Detection and Attribute Labeling in Web Data Extraction[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(SIGKDD 2006). ACM Press, 2006: 494-503

[30] Zhu J, Nie Z, Zhang B, et al. Dynamic Hierarchical Markov Random Fields and Their Application to Web Data Extraction[C]// Proceedings of the 24th International Conference on Machine Learning(ICML 2007). ACM Press, 2007: 1175-1182