基于事件的知识处理研究综述*)

周 文 刘宗田 孔庆苹
(上海大学计算机工程与科学学院 上海 200072)

摘 要 本文对近年来基于事件的知识处理研究进行了综述,从事件的定义开始,到事件的表示、提取方法和具体应用来说明该领域的研究进展。许多科学家认为人们是以事件为单位来体验和认识世界的,事件符合人们的正常认知规律,对事件的研究有广阔的前景,将成为基于概念的知识处理技术的必要补充和发展,为知识处理注入新的活力。

关键词 事件,知识表示,知识处理,本体,本体生成

A Survey of Event-based Knowledge Processing

ZHOU Wen LIU Zong-Tian KONG Qing-Ping
(School of Computer Engineering and Science, Shanghai University, Shanghai 200072)

Abstract It gives a survey of the literature about the research in recently on the Event based knowledge processing from the definition of event, the representation of event, the extraction of event to the application of event-based method. Many scientists think event accords with human normal cognitive rules and is the basic unit of human cognition. So it has wide foreground and will extend concept based knowledge processing techniques largely.

Keywords Event, Knowledge representation, Knowledge processing, Ontology, Ontology learning

很多认知科学家们认为,人们是以"事件"为单位来体验和认识世界的,事件符合人们的正常认知规律。然而,对事件的研究国内外都还刚刚起步。

1 "事件"的定义

"事件"(Event)起源于认知科学。近年来,"事件"的概念逐渐被计算语言学、人工智能、信息检索、信息抽取、自动文摘和自然语言处理等知识处理领域所采用。

"事件"暂时还没有统一的定义,许多学者都给出过"事件"的定义。WordNet中给出了很宽泛的"事件"定义:"在特定地点和时间发生的某件事"。一些语言学家给出了事件及其语义结构的定义,常包括目的、时间和外在条件。如:Chung[1]将事件定义为是由三个部分组成的术语:"谓词;事件框架,即谓词发生的时间段;事件界,即谓词发生的情况或者条件"。Pustejovsky[2]从语义理解的角度围绕动词及其属性给出了事件的定义。

在信息检索和信息提取领域,Allan和Yang等认为"事件"是"细化了的用于检索的主题"[3~5],Marsh[6]和Grishman[7]等人分别将事件信息与事件相关的特定的组织、人或者人为的实体关联起来,采用事件以提取预先指定的事件信息[6,7]。

在自动文摘领域,Filatova等[8]定义了"元事件",它是动词(或者动名词)及其连接起来的行为的主要组成部分(如参与者,地点,时间等)。从各领域对事件的定义来看,基于事件的研究常围绕动词及其属性,根据与事件类的关系对动词进行分类。基于事件的研究已经发展成为将"事件"看作一种知识表示的方式,它不同于基于概念的知识表示方法,重点在于对动词关注上,同时考虑动词与其关联成分的关系,并将事件扩展到更为抽象的事件类的层次上。

2 "事件"的表示模型

将"事件"作为一种知识表示方式,有学者提出了一些事件的表示模型。Nelson最早在1986年研究认知发展时,提出了事件表示模型[9]。他认为事件是包含了对象和关系的一个大的整体,它包括为取得某些结果,而有目的的行动的人对物体的作用,以及相互之间的作用。认为事件是动态的、复杂的、随着时间发生的。从认知科学的角度,将事件表示作为分类知识的基础。他结合基于脚本的知识表示方法[10],提出了一种基于脚本的通用事件表示方法(GER)。事件以"面"(slot)构成,这些面包括:行为人、行动和其他道具(props)构成。根据事件的不同,这些面可以有相应的值。同时,认为事件是具有结构的,也是有时间一因果顺序的,所以事件可以有层次化和时序化(因果化)两种组织方式,整体事件可以由若干个行为片段组成。事件知识表达了事件及其关系的信息。

Filatova等提出的"元事件"是由动词(或者动名词)和动词连接的行为的主要组成部分构成,这些行为的主要连接成分是指三类命名实体:参与者(人名和机构名)、地点和时间,事实上,元事件可以表示成三元组{命名实体,动词(或者动名词),命名实体}。

王寅[11]从语言学的角度,提出了事件域认知模型(ECM),并探讨了该模型的解释力。提出"事件"主要包括两大核心要素:行为和事体。一个事件域又包括若干要素。事件以动词为核心,只要有动作就要涉及到动作的发出者或接受者,事件域中的动作和个体之间就建立起了一定规律的搭配关系,因此BAB(事体+动作+事体)就构成一个事件域E的基本成分。基于此,语言就可形成各种基本句法构造,进而可以解释语句的构成、语汇的关系等等。

事件多元组模型提出,事件包括动词v和动词连接的高

*)本文受国家自然科学基金(60275022,60575035)资助。周 文 博士生,研究方向为人工智能、自动文摘等。刘宗田 教授,博导,从事人工智能、软件工程等方面的研究。

频名词或者命名实体(高频名词和命名实体用 n 表示),则一个事件可以表示为一个多元组:$\{n_1,\cdots,n_i,v,n_{i+1},\cdots,n_m\}$。相对于三元组,多元组更能体现一个事件的整体情况,避免丢失事件中的重要信息[12]。

综合以上各种事件表示模型,Nelson 的 GER 模型没有能结合更为有效的知识表示方法,知识提取的自动化程度不够高,从而限制了它的应用,使其停留在理论的层面上。王寅提出的 ECM 相当于将事件定义为一个三元组{事体,动作,事体},在三元组的基础上,有较强的分析与解释句法的能力,对语言中的句法进行分析和解释。EMC 不考虑时间、地点等事件发生的条件,从一方面看,这使得 ECM 模型更为抽象和概括,可不受具体的空间和时间的约束,但从计算机信息处理的角度看,却丢失了区别不同事件的重要组成成分。如,若两事件含有相同的{事体,动作,事体}三元组,但它们发生的时间和空间不同,那么,从 ECM 模型的角度,它们是两个相同事件,但对于信息抽取而言,它们是两个不同事件。显然,该模型需要进行扩展后才能应用于计算机信息处理领域中。从自然语言处理的角度 Filatova 提出的三元组较容易,通过分词和命名实体识别后,即可提取元事件,在应用中较容易实现,但三元组表示事件会丢失事件中包含的信息。事件多元组模型改善了这一状态,但相对三元组来说,提取时需要做句法分析,在效率上会受到影响。

3 "事件"的提取技术

语言学学者在考虑动词的同时考虑动词的结构信息,在此基础上发展了语言学上的事件理论。Pustejovsky 通过谓词分解和动词具体化的方式理解语句,提取事件。

Filatova 等将文本看作事件的集合,打破传统意义上将文本看作概念的集合的常规。通过自然语言处理和统计学的方法,从文本中提取元事件。但该方法在两上方面有待进一步完善,一方面,将事件固定地定义为三元组,在事件提取过程中,特别是对复杂的语句处理后,将丢失掉语句中大量的信息;另一方面,完全根据统计数据来确定事件,处理过程中没有考虑将语义的信息包含进去,事件提取的准确性和有效性都不高。

事件多元组提取方法[12]也是将文本看作事件的集合,通过自然语言处理技术,将文本断句分词标注后,提取出命名实体和高频名词,再采用句法分析的方法,找出核心动词作为该事件多元组的核心行为词,构成一个事件多元组,相对于三元组而言,更为灵活,也保留了更多的信息,但显然在处理过程中,需要多进行一道句法分析的工序。

此外,Hunter 给出了从结构数据中提取事件,他还给出了一种从时序数据中提取事件的技术[13]。Yang 在寻找一种新的问题回答技术[14]时,提出了将信息检索和机器学习技术应用于事件探测和追溯中,该技术可以从按年代顺序组织的新闻报道文件流中自动探测出新的事件,也可追溯用户感兴趣的发生在特定时期的事件。

国内的研究人员,如清华大学的吴平博、陈群秀等,湖南大学的梁晗等采用基于框架法的事件信息提取技术[15,16]。从事件语料中提炼出事件的框架知识,用该预先定义的框架对灾难性事件的不同侧面进行表示,定义侧面词,形成事件框架,利用该框架造抽取规则及侧面词的匹配方法,抽取文本的事件信息,在事件抽取时,绕过了深层句法分析的难题,降低了系统的实现难度。

中国科学院计算技术研究所的姜吉发等也给出了一种事件信息抽取模式的获取方法。主要通过定义三种模式:事件模式、事件触发模式和事件抽取模式。事件模式描述待抽取事件类型的各个角色及其相应语义约束;事件触发模式描述了某个待抽取事件类型的各个角色及其相应的可能的触发词,通过某个角色的触发词,找出该角色的候选描述语句;事件抽取模式用来指导从自由文本中进行实际的事件抽取。由这三个模式定义事件框架,用来通过关键词定位事件的候选描述语句。

这种采用事件分析的技术,是对采用关键词或者主题词匹配的方式的一种扩展,从技术角度而言,该方法的语义程度还可以进一步地提高和完善。

4 基于事件的应用技术

4.1 基于事件的本体生成技术

本体最初是一个哲学上的概念,十多年前被引入计算机领域中作为知识表示的方法并被广泛使用。Studer[19]在 Gruber [20]1993 年提出的定义和 Borst[21]1997 年提出的定义基础上,将本体定义为"共享概念模型的明确的形式化规范说明"。这意味着本体是某些应用领域的概念以及概念间关系的预先定义形式的表示方法[22]。

本体理论是当前信息领域和语言学领域研究的热点[21]。本体对于探索人的认知原理、发展自然语言理解技术和人机交互技术有重要的意义。

就其核心而言,本体是概念及概念与概念之间关系的一种表示方法。迄今,领域本体多采用手工或者半自动的方式获取。事实上,在现有本体的构建技术中,概念的获取相对较容易,但概念间关系由于大量蕴含于语句的动词中,因此难以用现有的技术自动获取,这使得概念间关系的获取成为本体建立和应用的瓶颈,导致本体构建耗时费力。

事件研究的核心是动词性概念,它围绕动词及其属性(相关的概念)展开,基于事件的理论技术的引入,将有助于突破本体技术的瓶颈,解决现有本体中概念与概念之间关系自动获取困难这一问题。

基于事件的本体半自动生成技术[12],以事件多元组模型和事件多元组提取技术为基础,解决本体中概念之间的关系获取难这一问题。具体的方法如图 1 所示。

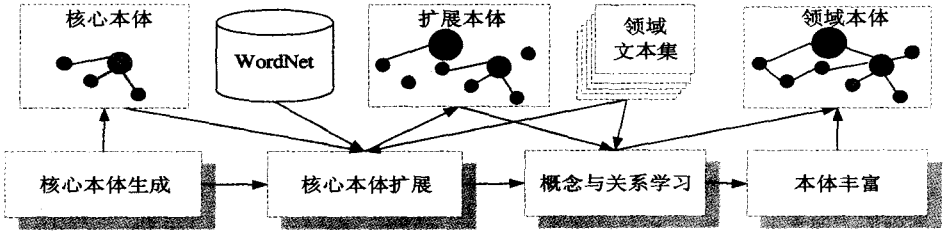

图 1 基于事件的本体生成过程

手工构建的核心本体通常仅包含领域的顶层概念；采用WordNet对生成的核心本体进行扩展，增加了核心本体中的概念，同时也将WordNet中概念之间的关系引入扩展本体中；将领域文本看作事件的集合，采用事件提取技术，提取领域文本中的事件多元组，为扩展本体添加了领域概念，更主要的是补充了扩展本体中概念和概念之间的关系，解决了本体构建中概念间关系获取难的问题。基于事件的领域本体生成系统架构如图2。

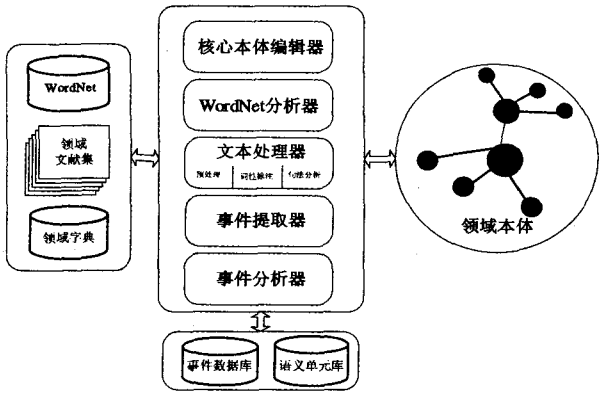

图2 基于事件的本体生成架构

核心本体编辑器用来辅助核心本体手工生成，WordNet分析器完成基于WordNet的概念和关系学习，通过文本处理器对领域文献集进行处理，生成的数据存入语义单元库，供事件提取器提取出事件多元组存入事件数据库中，再由事件分析器获取领域概念及概念之间的关系，丰富领域扩展本体进而生成领域本体。

Lin[23]在提出一种基于事件的本体检索技术时，构建的本体的顶层概念为事件的组成要件：如地点、事情、时间等，该事件的构成要件作为该本体中的主要分类。

4.2 基于事件的信息处理

基于事件的信息检索技术已有学者提出，Lin提出的事件本体的检索技术具备了一定程度上按事件要素对主题词进行扩展的能力，应用于检索技术可在一定程度上提高检索结果的质量。

吴平博等[15]探索了基于事件框架的文档智能检索，首先对事件文档进行聚类，然后从类向量中抽取框架的侧面词，生成完整的事件框架体系，比如，在空难事件的表示过程中，将空难的框架的侧面及侧面词进行收集，获得的侧面如：背景资料、原因调查、搜求过程等，相对应的侧面词分别为："概况、航空公司、介绍、型号等等"；"分析、责任、排除等等"；"救援、抢救、搜求等等"。利用框架对事件的不同侧面进行表达，并将这些侧面向量化，利用相关度评价函数得出文档与事件的相关程度，从而构成了一种分类体系，对事件的相关文档进行预测。

基于事件的自动文摘技术也获得了一些较突出的研究成果。哥伦比亚大学的Filatova[25]在她提出的事件三元组定义基础上，根据统计学的原理，对事件三元组的在文本中出现频率进行测定，提取出含有高频事件的语句，将它们组合成一篇摘要。该技术仅仅考虑单个事件出现的频率，没有进一步考虑事件间的联系，从而影响了应用的质量。

Daniel[26]通过对新闻主题子事件的识别来获取重要的信息以生成多文档的摘要。Li[27]在基于事件摘要技术的基础上，引入事件的内外部关联性的概念，不仅考虑事件本身成分的重要性，还通过事件间的关联性来考虑事件的重要性，进而提取出包含重要事件的语句形成摘要。

Lee[24]将本体技术与基于事件的技术结合起来用于网上中文新闻自动文摘，取得了实际效果。事件技术的引用仅仅扩展了概念在本体中的匹配范围，从原先的名词性概念，扩展到时间、地点等多方面的知识匹配，从而提高了文摘的质量。

结论和展望 综合以上对事件定义、事件理论模型、事件提取技术、事件技术应用的文献综述，可见已经有许多计算机领域的学者开始关注基于事件的技术。该领域存在的主要问题是：事件的定义尚不统一；事件的理论模型也很不统一、不够完善；事件的提取技术较传统，还是以自然语言处理技术结合统计学的技术为手段；事件技术与本体技术的结合已经有了萌芽，正有待于进行深入研究。

从前景上看，以研究动词为核心的事件技术的引入将有助于突破本体技术的瓶颈，解决现有本体中概念与概念关系自动获取困难这一问题，更有助于信息检索、自动文摘、信息抽取、问题回答等知识处理领域的理论创新与应用质量提高。

参 考 文 献

1 Chung S, Timberlake A. Tense, aspect, and mood. Language Typology and Syntactic Description, 1985(3): 202～258

2 Pustejovsky J. Events and the Semantics of Opposition. CSLI Publications, 2000. 445～482

3 Allan J, Papka R, Lavrenko V. On-line new event detection and tracking. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998. 37～45

4 Allan J, Carbonell J, Doddington G, Yamron J, Yang Y. Topic detection and tracking pilot study: Final report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998

5 Yang Y, Carbonell J G, Brown R D, Pierce T, Archibald B T, Liu X. Learning approaches for detecting and tracking news events. Intelligent Systems and Their Applications, IEEE [see also IEEE Intelligent Systems], 14: 32～43

6 Marsh E, Perzanowski D. MUC-7 evaluation of IE technology: Overview of results. In: Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998

7 Grishman R. Information Extraction: Techniques and Challenges. Information Extraction (International Summer School SCIE-97), 1997

8 Filatova E, Hatzivassiloglou V. Domain-independent detection, extraction, and labeling of atomic events. In: Proceedings of RANLP, Borovetz, Bulgaria, 2003. 145～152

9 Nelson K, Gruendel J. Event knowledge: structure and function in development. Erlbaum, 1986

10 Schank R C, Abelson R P. Scripts, plans, goals and understanding. Erlbaum, 1977

11 王寅. 事件域认知模型及其解释力. 现代外语, 2005, 18: 17～26

12 Zhou W, Liu Z, Zhao Y, Xu L, Qiang Y. A Semi-automatic Ontology Learning Based on WordNet and Event-based Natural Language Processing Technologies. In: Proceedings of ICIA'06 Conference (Accepted and Registered), 2006

13 Hunter J, McIntosh N. Knowledge-based event detection in complex time series data. Artificial Intelligence in Medicine, 1999. 271～280

（下转第184页）

4.3 实验过程

(1) 初选阶段:由于原始863IR网页数据集巨大,我们首先选择Query集的30个问题作为输入,通过带伪相关反馈的向量空间模型,从数据集获取针对每个Query的前50个返回文档,并以此作为小型测试用语料数据库。

(2) 检索阶段:采用向量空间模型进行上述三组实验。三组实验的不同之处在于参数选用,即是否采用主题词语加权和 $TF\times IDF$ 加权。最终取Query集中12个Query作为实验样本。每个Query根据与文档的夹角余弦作为相似度计算的工具,只返回前10个检索结果。

(3) 评估阶段:对每个Query返回结果,人工考察结果文档是否切合用户Query信息需求。采用TREC QA的平均倒数序和(Mean Reciprocal Rank, MRR)来评估不同方法的优劣,即对每个Query的前10个结果文档,取满足要求的文档在结果排序列表中序位的倒数和作为最终结果。该值越大,说明相关结果排在越靠前。

我们没有采用信息检索传统的查全率(recall)和查准率(precision)作为评判标准。这是因为我们手头没有标准答案集,相对16G原始语料,很难据此作出评估。同时,即使针对初选阶段构造的小型数据集,也难保证所有正确结果都已经在初选阶段被检出。

4.4 实验结果

采用以上三种加权方案对12个Qurey检索测试,得到三组不同结果,数据见表3。

表3 三种解决方案对12个Query检出结果MRR的比较

Query号	方案1	方案2	方案3	方案2较1高	方案3较1高	方案3较2高
1	1.5	1.91	1.99	27.33%	32.67%	4.19%
2	2.09	2.14	2.93	2.39%	40.19%	36.92%
3	1.12	1.05	1.92	−6.25%	71.43%	82.86%
4	0.5	0.61	0.75	22%	50%	22.95%
5	0.38	0.38	0.58	0%	52.63%	52.63%
6	0.1	1.23	1.74	1130%	1640%	41.46%
7	0.39	0.39	0.75	0%	92.31%	92.31%
8	1.2	1.2	1.13	0%	−5.83%	−5.83%
9	0.97	1.58	1.22	62.89%	25.77%	−22.78%
10	0.73	1.98	2.66	171.23%	264.38%	34.34%
11	2.25	2.35	2.29	4.44%	1.78%	−2.55%
12	1.08	1.41	2.55	30.56%	136.11%	80.85%
平均值	1.03	1.35	1.71	31.07%	66.02%	26.67%

从表3可以看出,直接使用所有非停用词(方案1)作为输入的检索效果不尽如人意,平均值为1.03;而只采用 $TF\times IDF$ 加权方式(方案2)平均值为1.35;同时利用词语在不同域和不同功能块加权及 $TF\times IDF$ 加权方式(方案3)的平均值为1.71。方案2比方案1提高31.07%,方案3比方案1提升66.02%;方案3比方案2提升26.67%。

从中发现,Query主题词语确实存在词频和主题区域的区分影响检出效果的情况。

结论 本文主要针对当前信息检索测试Query的SGML形式特征,区分不同主题描述域;对Narrative成段落的自然语言表述形式,采用正则表达式,划分Query背景块、核心请求块和排除条件块等不同主题功能块,使得Query信息内容词语能够根据描述域和功能块划分。对处于不同描述域和功能块中的词语实施不同的加权处理,结合 $TF\times IDF$ 传统词频统计方法,使得表述主题概念的Query词语得以强化。实验表明凸显Query主题词语和对信息内容词语的 $TF\times IDF$ 加权策略比单纯采用计数的 $TF\times IDF$ 加权效果更好。

参考文献

1 Baeza-Yates R, Ribeiro-Neto B, et al. Modern Information Retrieval. ACM Press, 1999

2 Dkaki T, Mothe J. Combining Positive and Negative Query Feedback in Passage Retrieval. In: Proceedings of Recherche d'Information Assistée par Ordinateur(RIAO'2004), 2004

3 Grootjen F A, van der Weide Th P. Conceptual Query Expansion:[Technical Report]. NIII-R0406. University of Nijmegen, The Netherlands, 2004

4 Navarro G. Query Languages. In: Baeza-Yates R, Ribeiro-Neto B. eds. Modern Information Retrieval. New York: ACM Press, 1999

5 Yin L. Topic Analysis and Answering Procedural Questions: [Technical Report ITRI-04-14]. University of Brighton, 2004

6 熊文新, 宋柔. 信息检索自然语言查询问句处理框架. 计算机科学, 2006(10)

7 熊文新, 宋柔. 信息检索用户查询语句的停用词过滤. 计算机工程, 2007(6)

8 张华平. 语言浅层分析与句子级新信息检测研究:[博士学位论文]. 北京: 中国科学院计算技术研究所, 2005

(上接第162页)

14 Yang H, Chua T-S, Wang S, Koh C-K. Structured use of external knowledge for event-based open domain question answering. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. ACM Press, Toronto, Canada, 2003

15 吴平博, 陈群秀, 马亮. 基于事件框架的事件相关文档的智能检索研究. 中文信息学报, 2003(7): 25～30

16 梁晗, 陈群秀, 吴平博. 基于事件框架的信息抽取系统. 中文信息学报, 2006(20): 40～46

17 Gruber T R. Toward principles for the design of ontologies used for knowledge sharing: A translation approach to portable ontologies. Knowledge Acquisition, 1993(5): 199～220

18 Borst W N. Construction of engineering ontologies for knowledge sharing and reuse:[PhD]. University of Twente, Enschede, 1997

19 Studer R, Benjamins VR, Fensel D. Knowledge engineer, principles and methods. Data and Knowledge Engineering, 1998, 25: 161～197

20 吴强, 刘宗田, 强宇. 基于本体的知识库推理研究. 计算机应用研究, 2005, 22: 50～52, 87

21 陆汝钤. 世纪之交的知识工程与知识科学. 清华大学出版社, 2001

22 周文, 刘宗田, 陈慧琼. FCA与本体结合研究的综述. 计算机科学, 2006, 33: 8～12

23 Lin H F, Liang J M. Event-based Ontology design for retrieving digital archives on human religious self-help consulting. Institute of Electrical and Electronics Engineers Computer Society, Piscataway, NJ 08855-1331, United States, Hong Kong, China, 2005. 522～527

24 Lee C S, Chen Y-J, Jian Z-W. Ontology-based fuzzy event extraction agent for Chinese e-news summarization. Expert Systems with Applications, 2003, 25: 431～447

25 Filatova E, Hatzivassiloglou V. Event-based Extractive summarization. In: Proceedings of ACL 2004 Workshop on Summarization, 2004. 104～111

26 Daniel N, Radev D, Allison T. Sub-event based Multi-document Summarization. In: Proceeding of the HLT-NAACL 2003 Workshop on Text Summarization, 2003. 9～16

27 Li W, Xu W, Wu M, Yuan C, Lu Q. Extractive Summarization using Inter-and Intra-Event Relevance. In: Proceedings of COLING-ACL, 2006