

基于事件的文本表示方法研究

廖涛^{1,2} 刘宗田¹ 王先传¹

(上海大学计算机工程与科学学院 上海 200072)¹ (安徽理工大学计算机科学与工程学院 淮南 232001)²

摘要 在传统文本表示模型的研究基础上,针对叙事类文本,考虑以事件作为基本语义单元,并结合图结构表示的特点,提出了一种基于事件的文本表示方法——事件网络。该方法利用事件和事件间的关系来表示文本,能够较大幅度地保留文本的结构信息及语义信息。实验结果表明,基于该方法的自动摘要取得了较好的效果。

关键词 文本表示,图结构,事件相似度,事件网络

中图法分类号 TP311 **文献标识码** A

Research on Event-based Method for Text Representation

LIAO Tao^{1,2} LIU Zong-tian¹ WANG Xian-chuan¹

(School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)¹

(School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China)²

Abstract By studying some traditional text representation models, this paper considered the event as a basic semantic unit for narrative texts, and presented a new event-based text representation method (Event-Network), which combines the characteristics of the graph structure. This method uses events and the relationship between events to represent the text, and can retain the structure information and semantic information of the text to a greater extent. The experimental results show that automatic summary based on the method has better performance.

Keywords Text representation, Graph structure, Event similarity, Event-network

1 引言

世界是物质的,物质的世界是动态的。认知科学家认为:人类以“事件”为单位进行记忆和理解现实世界。事件是随着时间变化的具体事实,涉及到多方面的事物概念,事件间具有本质的内在的联系^[1]。对于事件的研究,一直以来,深受学术界的高度重视。尤其近年来,对于事件的研究已经成为自然语言理解中的一个热点。由美国国防高级研究计划委员会主办的话题识别与跟踪评测会议的5个主要任务中就包括事件识别这一重要任务。另一个由美国国家标准技术研究所组织的自动内容抽取ACE(Automatic Content Extraction)评测会议也将事件抽取作为其主要任务。“事件”的概念正逐渐被计算语言学、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”是关联参与者、时间和地点等概念,且比“概念”粒度更大的语义单元,使用事件和事件间关系来表示文本更加符合知识表示的客观规律。如小说、人物传记、新闻报道等大量叙事类文本中都包含各类事件,对于这类文本可以考虑把事件作为基本语义单元,建立面向事件的文本知识表示,用以实现对句子和篇章的语义理解。本文在传统文本表示模型的研究基础上,结合对事件的定义和形式化表示,提出了一种基

于事件的文本表示方法——事件网络。

2 文本表示相关知识

文本表示是文本信息处理中的基础问题,是将实际文本内容变成机器内部表示结构的方法,即得到文本的形式化表示的方法,这种表示在一定程度上能够真实地反映出文档的内容(主题、领域和结构等)。

用来表示文档内容的特征项可以是各种类别,有词、短语,甚至是句子或句群等更高层次的单位。特征项也可以是相应词或短语的语义概念类。特征项的选择必须由处理速度、精度、存储空间等方面的具体要求来决定。选出的项越具有代表性,语言层次越高,项所包含的信息就越丰富,但分析的代价就越大,且受分析精度(如句法分析的准确率)的影响就越大。

目前的主要文本表示模型有:布尔模型、向量空间模型、概率模型和图结构表示模型等。

布尔模型(Boolean Model)是第一个被提出的模型,建立在布尔代数和经典的集合论的基础上^[2],构成特征项在文本中只有两种状态:出现或者不出现,相应地,每个特征项的权值为1或者0。布尔模型虽然简单、容易理解,但表示能力非常刚性,不能反映出特征项对文本语义的重要程度。

到稿日期:2012-02-29 返修日期:2012-06-17 本文受国家自然科学基金项目(60975033),安徽省高等学校优秀青年人才基金项目(2010SQRL050)资助。

廖涛(1977—),男,博士生,副教授,主要研究方向为Web数据挖掘、人工智能等,E-mail:tliao@aust.edu.cn;刘宗田(1946—),男,教授,博士生导师,主要研究方向为人工智能、软件工程等;王先传(1983—),男,博士生,主要研究方向为文本信息处理、问答技术。

向量空间模型(Vector Space Model, VSM)是目前文本表示的主要方法,该方法由 Salton 等首次提出。他们又于 1988 年提出了著名的 TF-IDF (Term Frequency-Inverse Document Frequency)公式法^[3],这种方法既考虑到了某个词在文档中的绝对词频,又考虑到了包含这个词文档的频率,是对单纯的词频权重方法的一种改进。向量空间模型的优点是将对文档内容的处理简化为向量空间中的向量运算,大大降低了问题的复杂性,提高了文本处理速度。其缺点是假设所有的特征项之间是相互独立的,无法发掘特征项之间的关联性,从而造成文本语义信息的丢失。

随着向量空间模型研究的发展以及 TREC 评测的推动,向量空间模型还存在很多变种,出现了一些新的表示模型,如概念表示模型、N-gram 表示模型、多词表达模型、句子表示模型和统计主题模型^[4]等。

概率模型(Probabilistic Model)由 Stephen Robertson 和 Spark Jones 等人提出^[5],是信息检索领域比较成熟的模型,经过多年的发展已在很多应用中取得不错的效果,尤其 OKAPI 模型在多届的 TREC 评测中都取得了很好的成绩。该模型基于概率排队理论:当文档按相关概率递减排列时可以获得最大的检索性能。该模型综合考虑了词频、文档频率和文档长度等因素,把文档和用户兴趣(查询)按照一定的概率关系融合,形成了著名的 OKAPI 公式。概率模型能更为准确地描述关键词和文档之间的相关关系,但这种模型对所处理的文本集依赖过强,而且处理问题过于简单。由于需要事先确定关键词和文档之间的相关概率,因此限制了其广泛使用,但它为检索系统的理论研究提供了依据。

上述模型虽在一定程度上实现了对文本的形式化表示,但均未考虑特征项在文本中出现的顺序,缺乏对特征词语上下文环境的理解,从而忽略了文本结构信息和词语之间的语义联系。因此,从自然语言的角度出发,这些模型并不能很好地表达文本的语义信息,是不完善的。

针对这些缺陷,近年来许多学者提出了基于图结构模型的文本表示。

在国外,基于图结构模型的文本表示方法近年来已成为研究热点。Schenker 等^[6]首次将图结构引入到文本表示中,他们将文本中的英文特征项作为节点,以节点间的邻接共现关系为边进行构图,并用 3 种位置名称定义边的类别。Svetlana 针对词条间的相关性,在文献^[7]中提出基于辅助词典 VerbNet 和 WordNet 的文档概念图表示模型。Bhoopesh 和 Pushpak 在文献^[8]中提出了根据 UNL 图来构造代表文档的特征向量,并采用 SOM 技术对文本进行聚类。Inderjeet 和 Eric 在文献^[9]中也提出了用于多文档摘要提取的文档图模型表示方法。Jin 等^[10]提出了基于图的表示方法,其能够包含词序、词频、词的共现和词的上下文,该模型用于反恐语料集的信息检索和文本挖掘,对比其他的一些检索模型有很好表现和效果。

国内,周昭涛等^[11]尝试在文本表示中引入序关系,用图结构来表示文本,提出了一种新的文本表示模型——图表示模型,并对该模型的表示效果进行了验证。实验结果表明,该表示模型仍达不到 VSM 模型所取得的表示效果。吴江宁和刘巧凤^[12]提出了一种考虑词间语义和语序信息的基于图结

构的中文文本表示方法,其思想是将文本的特征项表示成图结构中的节点,特征项间的关系表示成节点间的有向边。据此,可将文本表示成由节点和有向边组成的图结构,并且将边赋予权重信息。该模型较大程度地保留了原文本的结构信息及语义信息。

可以看出,基于图结构的文本表示从某种程度来说是一种语义网络,节点表示对象,节点间的边表示对象间的关系,图结构能够考虑文本中特征项之间的序关系,在一定程度上避免了文本语义信息的损失。本文将事件作为基本语义单元,并结合图结构表示模型特点,提出了一种基于事件的文本表示方法——事件网络。

3 事件与事件相似度

不同的应用领域,对事件有不同的定义。在信息检索和提取领域,事件被认为是细化了的用于检索的主题^[13,14]。在自动文摘领域,Filatova 等^[15]采用调查问卷的方式,给出了事件的三元组定义 (n_i, t, n_j) , n_i 和 n_j 分别为与事件项 t 这一动词/动名词前后相邻的命名实体,这些命名实体包括人名、机构名、地点和时间共 4 类。周文等^[16]提出了事件多元组的表示形式,事件多元组由一个事件项和多个事件论元构成,事件项为事件中的动作,限定它为语句中的谓语动词。尽管不同领域对事件的定义不尽相同,但是它们都离不开行为(一般都由动词或动名词来描述)、参与事件的对象、时间和地点等基本要素。综上所述,给出了以下相关定义。

定义 1 事件指在某个特定的时间和环境下发生的、由若干角色参与、表现出若干动作特征的一件事情。形式上,事件可表示为 e , 定义为一个六元组^[17]:

$$e = (A, O, T, V, P, L)$$

式中,事件六元组中的元素称为事件要素,分别表示动作、对象、时间、环境、断言和语言表现。

事件由各事件要素构成,而从文本中可以直接获得的事件要素主要包括:动作 A 、对象 O (主体和客体)、时间 T 、环境 V ,可利用这 4 个事件要素来计算事件的相似度。

定义 2(事件相似度) 两个事件 e_m 和 e_n 的相似度记为 $\text{Sim}(e_m, e_n)$, 定义如下:

$$\text{Sim}(e_m, e_n) = \sum_{i=1}^4 w_i s(x_{im}, x_{in}), x_i \in \{A, O, T, V\}$$

式中, x_{im}, x_{in} 分别表示事件 e_m 和 e_n 的第 i 个事件要素, $s(x_{im}, x_{in})$ 表示事件要素 x_{im}, x_{in} 的相似度,取值如下:①两要素是同义词,如事件 e_m 的环境要素是上海,事件 e_n 的环境要素是沪,则 $s(x_{im}, x_{in}) = 1$;②两要素存在包含关系或上下位关系,如事件 e_m 的时间要素是 7 月 23 日晚,事件 e_n 的时间要素是 7 月 23 日 20 时 34 分,则认为 $s(x_{im}, x_{in}) = 0.5$;③其他情况,两要素的相似度 $s(x_{im}, x_{in}) = 0$ 。 w_i 表示事件各要素在计算事件相似度时的权重, $\sum w_i = 1$ 。由此得到,事件间的相似度是一个 0 到 1 之间的实数,0 (或低于某一定义的阈值)表示两个事件完全不相似,1 表示两个事件完全相似,即同一个事件。

在本文中,考虑各事件要素对事件的描述能力,在这里设定 $w_1 = 0.5, w_2 = 0.3, w_{3,4} = 0.1$ 。通过实验观察,当事件相似度 $\text{Sim}(e_m, e_n) \geq 0.7$ 时,可以认为 e_m 和 e_n 是相似的。

重计算出句子的权重;②将原文中的所有句子按权重高低降序排列,权重最高的若干句子被确定为摘要句;③将所有摘要句按照它们在原文中的出现顺序输出。

在评价标准上,本文使用准确率 P 、召回率 R 和 F 值来评价生成摘要的质量。假设人工摘要出来的句子集为 x ,自动摘要出来的句子集为 y ,则 $P=(x \cap y)/y, R=(x \cap y)/x, F=2rp/(r+p)$ 。

在实验中,本文从 CEC 语料库的 5 个事件类中各随机抽取 10 篇,共 50 篇文本语料,分别使用微软 Word 自动摘要方法和本文方法来生成摘要,摘要压缩比为 0.3。实验结果如表 3 所列。

表 3 本文方法与 MS Word 方法的实验对比

类型	MS Word			本文方法		
	P	R	F	P	R	F
地震	0.48	0.42	0.45	0.62	0.56	0.59
交通事故	0.52	0.47	0.49	0.66	0.55	0.60
恐怖袭击	0.54	0.46	0.50	0.68	0.60	0.64
食物中毒	0.45	0.43	0.44	0.67	0.57	0.62
火灾	0.55	0.51	0.53	0.63	0.55	0.59
平均	0.51	0.46	0.48	0.65	0.57	0.61

通过实验对比 MS Word 自动摘要方法可以看出,本文方法的摘要质量明显要更好一些。

此外,将本文实验结果与国内外其他一些自动摘要的研究成果进行对比,其结果如表 4 所列。

表 4 本文实验结果与其他研究结果比较

作者	实验数据	实验结果		
		P	R	F
王志琪 ^[18]	2003 国家 863 自动文摘系统评价测试集,共 10 篇文档	0.60	0.59	0.59
葛斌 ^[19]	27 篇新闻类和 20 篇文学类文章,20 篇科技论文,共 67 篇	0.57	0.54	0.56
Jaruskulchai ^[20]	从 Ziff-Davis 语料库中选取了 10 篇文章	0.60	0.44	0.50
本文方法	从 CEC 语料库中的 5 个事件类中各随机抽取 10 篇,共 50 篇文本语料	0.65	0.57	0.61

从表 4 可以看出,对比国内外的一些其他自动摘要方法,本文提出的基于事件网络的摘要抽取方法虽然有着一定的局限性(只适用于包含大量各类事件的叙事类文本),但也取得了较好的效果。

结束语 本文针对包含大量各类事件的叙事类文本,考虑把事件作为基本语义单元,建立面向事件的文本知识表示,提出了一种基于事件的文本表示方法——事件网络,并将其用于文本自动摘要中检验。实验结果表明,事件网络使用事件和事件间关系来表示文本更加符合知识表示的客观规律,能够较好地表达文本的语义信息。当然,本文提出的基于事件的文本表示方法也存在很多不足,事件网络构建过于简单,未考虑事件间存在的其他多种语义关系,如因果、分类关系等,这都需要我们进一步深入研究,不断对该方法进行完善。

参 考 文 献

[1] Chen X. Why did John Herschel fail to understand polarization? The differences between object and event concepts[J]. Studies in

History and Philosophy of Science,2003,34:491-513

[2] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval [M]. Addison-Wesley- Longman,Reading,MA,1999

[3] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval [J]. Information Processing and Management, 1988,24(5):513-523

[4] Wei Xing,Croft W. LDA-based document models for ad-hoc retrieval[C]//Proceedings of the 29th SIGIR Conference. Seattle, Washington, USA,2006:178-185

[5] Robertson S E, Walker S, Jones S, et al. Okapi TREC-4[C]// Proceeding of The 4th Text Retrieval Conference. 1996:500-236

[6] Schenker A, Last M, Bunke H, et al. Graph Representations for Web Documents Clustering[C]//LNCS on Pattern Recognition and Image Analysis. Springer Berlin,2003:935-942

[7] Hensman S. Construction of conceptual graph representation of texts[C]//Proceeding of the Student Research Workshop at HL T-NAACL. 2004:49-54

[8] Bhoopesh,Pushpak. Text clustering using Semantics[C]//Proceeding of the 11th International World Wide Web Conference. 2002:78-84

[9] Mani I, Bloedorn E. Multi-document Summarization by Graph Search and Matching[C]//Proceeding of the 15th National Conference on Artificial Intelligence. 1997:622-628

[10] Jin W, Srihari R K. Graph-based Text Representation and Knowledge Discovery[C]//Proceedings of the 2007 ACM Symposium on Applied Computing. Seoul, Korea, New York, NY, USA:ACM,2007:807-811

[11] 周昭涛,卜东波,程学旗. 文本的图表示初探[J]. 中文信息学报, 2005,19(2):36-43

[12] 吴江宁,刘巧凤. 基于图结构的中文文本表示方法研究[J]. 情报学报,2010,29(4):618-624

[13] 于满泉,骆卫华,许洪波,等. 话题识别与跟踪中的层次化话题识别技术研究[J]. 计算机研究与发展,2006,43(3):489-495

[14] 贾自艳,何清,张海俊,等. 一种基于动态进化模型的事件探测和追踪算法[J]. 计算机研究与发展,2004,41(7):1273-1280

[15] Filatova E, Hatzivassiloglou V. Domain independent detection, extraction, and labeling of atomic events [C]//Proceedings of RANLP. Borovetz, Bulgaria,2003:145-152

[16] Zhou W, Liu Z T. Event-based Knowledge Acquisition for Ontology Learning[C]//Proceedings of The 6th IEEE International Conference on Cognitive Information. Lake Tahoe, California, USA,2007:498-501

[17] 刘宗田,黄美丽,周文,等. 面向事件的本体研究[J]. 计算机科学,2009,36(11):189-192,199

[18] 王志琪,王永成,刘传汉. 基于互增强关系的自动文摘句子加权方法[J]. 上海交通大学学报,2007,41(8):1297-1300

[19] 葛斌,李芳芳,李阜,等. 基于无向图构建策略的主题句抽取[J]. 计算机科学,2011,38(5):181-185

[20] Jaruskulchai C, Kruengkrai C. Generic text summarization using local and global properties of sentences[C]//Proceedings of the IEEE/WIC International Conference on Web Intelligence. Piscataway, USA:IEEE Press,2003:201-206