自动术语识别存在的问题及发展趋势综述*

祝清松[1,2]　冷伏海[1]

[1]中国科学院国家科学图书馆　北京 100190　[2]中国科学院研究生院　北京 100049

〔摘要〕自动术语识别对于以内容分析为主的情报研究具有重要作用。在目前研究的基础上,重点分析自动术语识别存在的问题,包括词性过滤难以兼顾召回率和准确率、单词术语和低频术语的识别未引起足够重视、术语识别领域相关性有待加强等。最后阐述自动术语识别的多特征融合、机器学习方法、高质量和高隶属度的领域术语识别、新术语识别、语义识别等发展趋势。

〔关键词〕自动术语识别　术语抽取　存在问题　发展趋势

〔分类号〕G35 TP391

Existing Problems and Developing Trends of Automatic Term Recognition

Zhu Qingsong[1,2]　Leng Fuhai[1]

[1]National Science Library, Chinese Academy of Sciences, Beijing 100190

[2]Graduate University, Chinese Academy of Sciences, Beijing 100049

〔**Abstract**〕Automatic term recognition (ATR) plays an important role in information research based on content analysis. According to the current studies on ATR, the paper raises some questions about ATR including precision and recall problems caused by pre-and-post filtering, single-word and low frequency term recognition, low field relevance. The paper also provides some trends about ATR including multi-features fusion, machine learning methods, high-quality and high-relevance field term recognition, new term recognition, semantic recognition and so on.

〔**Keywords**〕automatic term recognition　term extraction　existing problems　developing trends

1　引　言

自动术语识别(automatic term recognition, ATR)是指从文本中自动发现领域术语的过程[1]。它是一项具有重要作用的语言技术,在自然语言处理、机器翻译、知识抽取、文本挖掘、本体构建、信息检索等应用研究领域都有广泛应用。同时,自动术语识别属于内容分析方法的一种,从大量科技文献中抽取术语,对于把握学科领域的发展变化等情报研究工作具有重要的辅助作用。

自动术语识别常用的方法包括基于规则方法和基于统计方法[2-3]。基于规则方法是根据术语构成模式建立一套规则,选择匹配规则的词语作为领域术语。这种方法的最大缺陷是人工编写的规则不可能覆盖所有的语言学现象,领域依赖性很强。基于统计方法主要应用词频、TF-IDF、互信息、信息熵、log-likelihood、假设检验等统计特征,选择特征值符合阈值的词语作为领域术语。这种方法不受领域限制,但是对于单词术语和低频术语的识别并不理想,抽取的术语也存在较多噪声。基于两种方法的不同特点,两种方法结合使用成为主流方法,例如C-value和NC-value[4]。混合方法很好地利用了两种方法的优点,但是缺乏背景语料的支撑,在抽取领域术语时准确率较低[5]。

条件随机场[6]、规则学习算法[7]、朴素贝叶斯[8]、隐马尔科夫模型[9]等机器学习技术逐步应用到自动术语识别中来,通过对训练集文本特征的学习构造模型来进行术语识别。基于机器学习的方法是基于统计方法的一种,与传统的统计方法相比有其独特的优势,但是对训练语料的要求很高,语料训练也需要花费大量时间。

本文认为各种方法都有各自的特点,没有哪一种方法具有绝对的优势,应充分考虑不同的语言特征和

*本文系国家自然科学基金项目"科技创新演化分析理论与方法研究"(项目编号:70873123)研究成果之一。

收稿日期:2012-03-26　修回日期:2012-06-25　本文起止页码:104-109　本文责任编辑:杜杏叶

统计规律,将这些方法有机地融合起来才是提升自动术语识别准确率和召回率的有效方法。本文针对目前自动术语识别相关研究,梳理和分析目前存在的主要问题,并对其发展趋势进行总结和归纳,以把握自动术语识别的发展现状。

2 存在的问题

2.1 词性过滤难以兼顾召回率和准确率

词性是术语构成的一个重要特征,词性过滤是基于规则方法中一种重要的模式,已经广泛应用于自动术语识别。术语的构成遵循一定的语言学模式,通过对术语词性构成规则的分析,设定词性过滤规则,从而据此对候选术语进行词性过滤。

词性过滤包括开放式过滤模式和封闭式过滤模式[4]。开放式过滤模式对候选术语的词性要求宽松,比如 Frantzi 等使用了"Noun$^+$ Noun"或"(Adj|Noun)$^+$ Noun"或"((Adj|Noun)$^+$|((Adj|Noun)*(Noun-Prep)$^?$)(Adj|Noun)*)Noun"三种过滤规则,候选术语只要满足任何一种即可。由于候选术语可以包含形容词和介词,从而导致抽取非术语的几率增大。封闭式过滤模式要求候选术语必须符合规定的词性规则,比如 Dagan 等[10]使用 Noun$^+$ 过滤规则,候选术语必须是名词组成的术语。由于对候选术语的要求严格,从而导致术语丢失的几率增大。

词性过滤模式的选择对抽取结果的准确率和召回率有很大影响。开放式过滤模式提高了召回率,却降低了准确率。封闭式过滤模式提高了准确率,却降低了召回率。召回率和准确率很难兼顾,词性过滤规则过松或过严都会对抽取结果产生影响,这是词性过滤规则设定必须面临的问题,也是基于规则方法所面临的问题。语言表达是灵活多样的,人工所能编制的语言学规则是有限的,准确定义不产生噪声的规则存在一定困难。

另外,名词性术语识别限制了非名词性术语的抽取,尤其是中文术语的构成可能包含动词、数量词等丰富的语类,词性过滤规则应当给予充分考虑。比如,刘里等[11]针对领域现象术语一般都是动词性复合词的现象提出了一种基于分隔符和上下文术语的领域现象术语抽取方法。周浪[12]突破了名词短语的限制,针对不同类术语的特征,分别从不同角度获取术语语言规则,拓宽了语言规则,接纳了更多非名词性结构的专业术语。

2.2 单词术语的利用和低频术语的识别未引起足够重视

领域术语的长度越长,其专指度就越高,目前自动术语识别的研究主要集中在多词术语的抽取上,而对单词术语的识别未引起足够重视。虽然单词术语的数量无法与多词术语相比,但是不可否认单词术语是领域术语中不能忽视的重要部分。而且单词术语是构成多词术语的核心要素,其识别程度影响到多词术语的识别程度。

单词术语构造简单,术语本身可利用信息较少,无法像多词术语识别一样利用内部结构,只能借助外部辅助资源(各种领域或通用词典、语料库等)进行识别。比如,Gelbukh 等[13]认为多词术语识别应当分为两步,第一步识别特定领域的单词术语,第二步利用第一步的结果识别多词术语。其中,第一步利用极大似然相似度方法来对通用参考语料库处理得到单词术语;周浪[12]利用单词术语边界特征清晰的特征提出了一种介于模糊 C-均值聚类算法的中文单词术语识别方法,这种方法不需要外部资源的辅助,有一定程度的改善。

词频是重要的统计特征,一般认为词频越高,权重就越高,成为术语的可能性越大。但低频术语的识别没有引起足够的重视。高频统计是情报研究普遍应用的计量方法,高频反映了过去一段时间内学科的研究重点和研究热点,这是对过去的计量分析。而随着时间的推移,一方面学科的研究重点可能会发生变化,过去很少有人研究的方向可能成为新的研究热点,另一方面新的研究方向也会不断出现。低频词和突发词对把握学科发展变化具有重要意义,因此自动术语识别不能忽视对低频词的识别。目前非相关文献知识发现中文献集的识别大部分是基于高频短语的方法,Kostoff[14]对此提出了质疑,他认为这对于潜在知识发现是不利的,低频词语更有可能成为潜在发现。周浪等[15-16]提出了一种利用术语在语料中的词频分布变化程度来衡量其领域相关性的计算方法,将词频做归一化处理,将 TF-IDF 和样本方差相结合,在突出低频术语和基础术语的同时能够降低普通短语的影响。

2.3 术语识别的领域相关性有待加强

单元性(unithood)和术语性(termness)是术语的两个基本特征,单元性测度的是构词的可能性,术语性测度的是领域的相关性。互信息和极大似然等方法是典型的单元性测度方法,主要用于测度多词术语中单词

之间的关联强度。自动术语识别目前大部分的研究侧重于术语的单元性,即术语的语言结构完整性,而忽略了术语的术语性,即术语的领域特征相关性[17],从而导致识别出的术语领域相关性太差。术语的领域特征对于识别高质量的领域术语、构造领域词典、构造领域本体等具有重要意义,有利于提高领域术语识别的准确率和召回率。

术语的领域相关性是指术语和该领域具有很强的相关性,区别于通用术语。领域术语在该领域内出现的频率和在其他领域出现的频率有很大不同,一般将"在特定领域中出现频率较高或只出现在某个领域,在不相关领域中出现频率相对较低[1]"的术语定义为领域术语。目前的相关研究已经开始重视术语的领域相关性,从而进一步提高术语识别的准确率。如 TermExtractor[18]采用了领域相关性、领域一致性、词汇衔接性、结构相关度等各种方法对候选术语进行过滤,以提高候选术语的领域隶属度。傅继彬等[19]提出的基于语言特性的中文领域术语自动抽取算法,集成领域耦合性、领域相关性和领域一致性 3 种语言特性建立统计模型进行中文领域术语的自动抽取,对领域语料有更好的预测能力。

2.4 存在问题述评

- 词性过滤规则的选择影响自动术语识别的召回率和准确率。单纯利用词性过滤规则难以兼顾召回率和准确率,应当通过多种方法融合的途径来解决召回率和准确率的矛盾。例如,词性过滤规则的设定要保证召回率,避免术语的缺失;结合其他基于规则的方法和基于统计的方法进行候选术语的遴选,以保证术语的准确率,从而兼顾召回率和准确率。

- 单词术语和低频术语的识别因其所占比例较低而未引起足够重视。单词术语很多情况下是多词术语的重要组成部分,与其他词汇搭配成多词术语。比如单词术语"质谱仪"是多词术语"有机质谱仪"、"无机质谱仪"、"同位素质谱仪"等的核心要素或中心词。单词术语的准确识别,对于多词术语的准确识别奠定了基础,同样,多词术语的准确识别也为单词术语识别的边界确定起到了关键作用。低频术语有可能是领域的新术语,对于学科态势分析和领域知识演化等情报研究工作都很有价值。

- 术语区别于一般词语的重要特征就是其领域特征,强领域相关性术语蕴含丰富的领域知识,是自动术语识别的关键研究内容。很多研究已经对术语的领域性特征进行描述[17,19],主要包括术语组成词之间耦合性较高;与该领域有很强的相关性;在该领域分布较为均匀,而在其他领域分布不均匀等。传统领域术语识别一般要基于领域的术语词典,利用领域种子术语进行相似度、耦合度等的计算。领域术语词典的构建是提升领域术语识别水平的基础,其获取难度大、更新周期长等缺陷需要利用其他语言学和统计学方法进行弥补。针对领域术语的强相关性,可借鉴聚类算法来进行领域术语的描述,"领域词典结合聚类方法"为建立通用的术语领域性描述模型提供了一条途径。如谢芳[20]、李勇等[21]利用阿尔伯塔大学 Patrick Pantel 和 Dekang Lin 提出的 CBC 聚类方法,用递归方法寻找分布在相似空间的紧凑类 Committee,自动剔除领域相关性较弱的术语,效果比较明显。

3 发展趋势

3.1 多特征融合成为主流方法

术语有语言结构特征、分布特征、上下文特征、领域特征等各种特征,自动术语识别由只考虑单一特征逐渐转向考虑多个特征。成功的术语识别需要考虑到影响抽取结果的各个因素,尤其是术语的多元特征,多特征融合的自动术语识别模式将成为主流模式。但是并不是融合术语所有的特征就是最好的方法,应充分考虑特征之间的互作用,有效选取效果最佳的特征进行融合。

基于规则的方法主要利用的是术语的语言结构特征,从语言学的角度分析术语的构成规则,比如词性规则、组块规则以及停用词过滤规则等。基于统计的方法主要利用的是术语的统计分布特征,测度术语的单元性和术语性[22]。术语性测度更注重术语的领域特征,包括领域相关性、领域特殊性等。混合方法趋于多特征的融合,结合规则和统计方法,同时对术语的单元性和术语性进行测度,典型的是 C-value 方法,NC-value 方法充分考虑到了术语的上下文特征。

游宏梁等[1]提出一种基于加权投票的术语自动识别方法,该方法选择了 TF-IDF、C-value 和 TermExtractor 三个指标进行加权投票,通过实验证明了多指标加权投票方法的有效性。C-value 和 TermExtractor 各自都考虑到了术语的多个特征,尤其是后者充分考虑到术语的领域性,用到了 5 个过滤特征。Zhang 等[22]对 TF-IDf、Weirdness、C-value、Glossex 和 TermEx-

tractor 等 5 个算法进行比较和评价，对于多特征融合有很好的参考。多特征融合方法将成为自动术语识别的发展趋势，如何充分考虑术语的不同特征，如何有效利用各种排序算法，如何选取特征和算法权重使得组合结果最优，将成为多特征融合需要考虑的主要问题。

3.2 机器学习方法成为研究热点

自动术语识别的研究大多集中在语言学方法、统计分布方法及混合方法等，随着机器学习方法在其他领域的广泛引用，基于机器学习的自动术语识别逐渐成为新的研究方向和研究热点。机器学习中有很多算法如集成学习、粗糙集、决策树、朴素贝叶斯、神经网络、遗传算法、最大熵、K 近邻等已广泛应用于许多领域。跟自动术语识别有很大关联的包括本体构建、关键词提取、命名实体识别、自动标引、信息检索等方向。

机器学习方法包括有指导、无指导和弱指导，有指导方法是基于分类的思想，无指导方法是基于聚类的思想。有指导的机器学习方法能很好地利用术语的多特征，可取得较好的识别效果，最大的约束是训练语料的构建成本以及训练语料与测试语料的同构性要求方面[17]，而且训练集影响产生的规则，动态规则集的选择将成为关键。目前基于机器学习的自动术语识别还没有广泛开展，已有的研究也仅局限于有指导的机器学习方法，如条件随机场、隐马尔科夫模型等。机器学习中的其他方法将会逐渐融入到自动术语识别中来，并尝试无指导机器学习方法的应用。

Zhou 等[2]将多词术语识别的过程看成一个多状态齐次马尔可夫链，这种基于马尔科夫的方法计算速度快，而且可以获得较高的准确率和召回率，实验证明比词频、互信息、C-value 和 NC-value 抽取效果更好。并将考虑融入 WordNet 等知识和神经网络算法来提高识别效果。岑咏华[9]提出基于双层隐马尔科夫模型的中文泛术语识别方法，并认为结合互信息、对数似然和领域度等参数可在中文自动术语识别方面取得突破。Foo 等[7]利用规则学习方法训练规则，选择了词性、形态-语法描述、语法功能、语义信息等 10 个术语特征，并将这种方法由单语术语抽取转向双语术语抽取甚至多语术语抽取。

3.3 新术语识别、语义识别等需求强烈

自动术语识别是一项基础性的研究，对于自然语言处理的多个应用方向都有重要实践意义。但是目前自动术语识别的效果并不是特别理想，尤其是中文术语识别比英文术语识别更加复杂和困难。粗粒度的术语识别已经不能满足日益精确化的应用需求，术语识别的质量以及领域相关性必须进一步提升，高质量和高隶属度的领域术语识别将成为关键。高质量主要是指具有较高的准确率和召回率，高隶属度主要是指识别出的术语具有较强的领域特性，将领域术语同伪术语、一般词语等噪音区分开来的细粒度自动术语识别将成为研究的一个重点。同时，前面提到的低频术语的识别也将成为另外一个研究点。

科技文献的数量持续增加，网络信息指数增长，新术语随之不断涌现，对新术语识别的需求更加强烈。新术语对于把握学科发展动态等具有重要意义，新术语识别或扩充也是领域术语词典扩展的关键。荀恩东等[23]提出一种基于定义模式的新术语及定义的抽取方法，在支持向量机 SVM 中引入涉及术语和定义多种统计特征，尤其是句子隶属度的引入提高了 SVM 分类器的性能，为新术语抽取提供了一种较为有效的方法。新术语不同于新词，有更强的领域特征，网络新词监测等用到的有效方法可以尝试应用到新术语的识别中来。

目前自动术语识别的研究主要是识别出领域内的术语，抽取结果是术语列表。而术语之间是有很多关联的，不能孤立地看待术语，要重视术语之间存在的语义关系。就像主题词表中除了主题词还包括同义关系、属分关系、相关关系等词间关系，本体除了概念也包括属性以及概念之间的联系等。领域术语同样会存在术语的外延和内涵以及该术语相关的属性、特征等，如果自动术语识别能够将术语识别提升到语义层面，在识别出术语的同时能够将术语相关的语义知识抽取出来，对于相关的研究将有重要作用。姚贤明[24]利用知网 HowNet 实现领域术语内涵的自动获取，并针对未登录词和术语重复等问题提供了解决方案，另外还基于 SVM 的实例学习方法实现了领域术语的外延学习。

3.4 发展趋势述评

- 多特征融合策略是针对术语特征进行选择，同时充分利用单个方法的特点。例如 TF-IDF 方法考虑到单篇文档中的术语频次和术语在文档集中的分布特征，互信息方法考虑到术语构成的紧密程度等[1]，多特征整合策略可起到相互补充的作用，是当前优化自动术语识别性能的有力途径。特征选择方法的组合可以分为串行策略和并行策略，串行策略指依次使用选定

的方法,并行策略指同时使用选定的方法。两种策略都涉及到每种方法权重的问题,如果认为各种方法同样重要,则每个权重值为 1/n(n 种方法);如果认为方法的重要性不同,每种方法可设置不同权重,权重需要实验获得。

对于多特征融合策略效果的评价,可以有两种思路:①利用准确率和召回率综合评价术语识别的效果,可利用 F 值加权评价准确率和召回率。这种方法全面评价术语识别的效果,可与单一或其他组合方法进行比较评价。②利用 P@ N 方法对排序靠前的结果进行评价。这种方法考虑到某些方法会降低整个算法的性能,并且正确的结果在排序中更加靠前[1]。

- 基于机器学习术语识别的本质是机器通过对术语语料的学习生成模型,学习过程是对已有语料中术语的特征进行统计并产生规则,进而产生推理能力。各种机器学习方法的主要区别就在于学习策略的不同。术语具有词性、词频、词长、位置等多种规则和统计特征,符合机器学习的基本条件,支持向量机、隐马尔科夫模型、最大熵模型、条件随机场、支持向量机等均有研究表明其对术语识别的适用性,季培培等[17]系统描述了几种机器学习算法用于自动术语识别的思路和特点。本文认为基于机器学习术语识别的难点在于大规模细粒度领域术语语料库的构建和特征学习统计规则的选择。

- 新术语识别对于知识发现、知识演化、情报监测和态势分析等情报研究工作具有重要的实践意义。新术语识别的重点和难点在于术语新颖度的判别和新术语边界的探测。新颖度的度量是新术语识别的基础,需要进行严格的界定。新术语包含的子术语可能会是已有术语,如何探测新术语的边界对于准确识别新术语非常关键。

术语一般都不是孤立存在的,与上下文具有紧密的联系,将术语在语义上关联性较强的其他术语或词语和关系识别出来,对于内容情报研究具有重要意义。如某材料术语的出现可能会伴随其光学性能、力学性能、化学性能、热性能、电性能、磁性能等语义信息的出现,也可能会伴随其与其他现存材料的相同点和不同点,如果能将这些信息同术语一同识别出来,构成以该术语为核心的术语语义网,对情报研究工作将具有重要价值。语义识别分为两个层面:一是相关词语或术语的识别;二是术语与相关词语或术语的关系构建。重点在于相关性计算和关系构建,难点在于消歧和降噪。

4 结 论

自动术语识别作为自然语言处理中的一项基本技术,对以文本挖掘和内容分析为主的情报研究具有重要的支撑作用。随着情报研究逐渐由标题和摘要等基本元数据转向涵盖更多信息的文本内容,自动术语识别技术的进展直接影响到内容分析的深度,语义识别的水平也将成为关键。

目前自动术语识别的研究仍存在一定的问题,比如词性规则的过滤难以兼顾术语识别的召回率和准确率、单词术语和低频术语的识别未引起足够重视、术语识别的领域相关性有待加强等。针对这些问题已经有研究人员提出解决方案,对自动术语识别的改进提供了思路。

另外,自动术语识别的方法众多,针对术语的不同特征,多特征、多算法融合的思路已成为目前的主流方法,如何有效融合不同的特征识别算法将成为主要的发展趋势。机器学习方法在其他方向上的成功对于自动术语识别有很好的借鉴作用,基于机器学习的自动术语识别正在成为研究热点。随着情报研究应用需求的提升,对于高质量和高隶属度的领域术语识别、新术语识别、术语及其语义关系识别等将成为关键的研究点。

参考文献:

[1] 游宏梁,张巍,沈钧毅,等. 一种基于加权投票的术语自动识别方法[J]. 中文信息学报,2011,25(3):9-16.

[2] Zhou Zili, Wang Yanna, Gu Junzhong. Markov-based Automatic Term Extraction[C]// FBIE '08 Proceedings of the 2008 International Seminar on Future BioMedical Information Engineering. Washington D. C. : IEEE Computer Society, 2008: 86-89.

[3] 刘建华,张智雄,徐健,等. 自动术语识别——对科技文献进行文本挖掘的重要技术方法[J]. 现代图书情报技术,2008(8):12-17.

[4] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: The C-value/NC-value method[J]. International Journal of Digital Libraries, 2000, 3(2): 117-132.

[5] 翟笃风,刘柏嵩. 政务领域本体术语的自动抽取[J]. 现代图书情报技术,2010,26(4):59-65.

[6] 王海雄, 郭剑毅, 余正涛, 等. 基于 CRFs 的中文领域术语自动抽取研究[C]//第六届全国信息检索学术会议论文集. 北京: 中国中文信息学会, 2010: 505-512.

[7] Foo J, Merkel M. Using machine learning to perform automatic term recognition[C]// Proceedings of the LREC 2010 Workshop on Methods for Automatic Ccquisition of Language Resources and

Their Evaluation Methods. Linköping : Linköping University Electronic Press, 2010: 49 - 54.
[8] 刘成帅. 中文领域术语自动获取方法的研究[D]. 南京: 南京邮电大学,2011.
[9] 岑咏华,韩哲,季培培. 基于隐马尔科夫模型的中文术语识别研究[J]. 现代图书情报技术,2008,24(12): 54 - 58.
[10] Dagan I, Church K. Termight: Identifying and translating technical terminology[C]// ANLC '94 Proceedings of the Fourth Conference on Applied Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 1994: 34 - 40.
[11] 刘里,刘小明. 基于分隔符和上下文术语的领域现象术语抽取[J]. 华南理工大学学报(自然科学版),2011,39(7): 146 - 155.
[12] 周浪. 中文术语抽取若干问题研究[D]. 南京: 南京理工大学,2009.
[13] Gelbukh A, Sidorov G, Lavin - Villa E, et al. Automatic term extraction using log-likelihood based comparison with general reference corpus[C]// Lecture Notes in Computer Science. Berlin / Heidelberg : Springer - Verlag, 2010:248 - 255.
[14] Kostoff R N, Block J A, Solka J L, et al. Literature-related discovery[J]. Annual Review of Information Science and Technology, 2009,43(1): 241 - 285.
[15] 周浪,张亮,冯冲,等. 基于词频分布变化统计的术语抽取方法[J]. 计算机科学,2009,36(5): 177 - 180.
[16] 周浪,史树敏,冯冲,等. 基于多策略融合的中文术语抽取方法[J]. 情报学报,2010,29(3): 460 - 467.
[17] 季培培,鄢小燕,岑咏华. 面向领域中文文本信息处理的术语识别与抽取研究综述[J]. 图书情报工作,2010,54(16): 124 - 129.
[18] Sclano F, Velardi P. TermExtractor: A Web application to learn the shared terminology of emergent Web communities[C]// Enterprise Interoperability II New Challenges and Approaches. London: Springer, 2007: 287 - 290.
[19] 傅继彬,樊孝忠,毛金涛,等. 基于语言特性的中文领域术语抽取算法[J]. 北京理工大学学报,2010,30(3): 307 - 310.
[20] 谢芳. 特定领域术语的自动获取[D]. 武汉: 华中师范大学,2006.
[21] 李勇. 基于聚类方法对特定领域术语的自动筛选[J]. 计算机工程与科学,2008,30(2): 64 - 66.
[22] Zhang Ziqi, Iria J, Brewster C, et al. A comparative evaluation of term recognition algorithms[C]// Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08). Paris: European Language Resources Association, 2008.
[23] 荀恩东,李晟. 采用术语定义模式和多特征的新术语及定义识别方法[J]. 计算机研究与发展,2009,46(1): 62 - 68.
[24] 姚贤明. 领域概念自动抽取研究[D]. 昆明: 昆明理工大学,2010.

〔作者简介〕祝清松,男,1985 年生,博士研究生,发表论文 3 篇。
冷伏海,男,1963 年生,研究员,博士生导师,发表论文 60 余篇。

(上接第 53 页)

[3] Juha M, Helena A, Marko S. Simple semantics in topic detection and tracking[J]. Information Retrieval, 2004, 7(3/4): 347 - 368.
[4] Brants T, Chen F, Farahat A O. A system for new event detection [C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2003:330 - 337.
[5] Ji H, Grishman R. Refining event extraction through cross-document onference[C]//Moore J D, Teufel S, Allan J, et al. The 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Columbus: ACM Press, 2008: 254 - 262.
[6] Patwardhan S, Riloff E. Effective information extraction with semantic affinity patterns and relevant regions[C]//Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague: Association for Computational Linguistics Press, 2007: 717 - 727.
[7] Czech R, Patwardhan S, Riloff E. A unified model of phrasal and sentential evidence for information extraction[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP - 09). Stroudsburg: Association for Computational Linguistics Press, 2009:151 - 160.
[8] 王会珍, 朱靖波, 季铎, 等. 基于反馈学习自适应的中文话题追踪[J]. 中文信息学报, 2006, 20(3):92 - 98.
[9] 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]. 计算机技术与发展, 2006, 43(3): 489 - 495.
[10] 王煜, 白石, 王正欧. 用于 Web 文本分类的快速 KNN 算法[J]. 情报学报, 2007, 26(1):60 - 64.

〔作者简介〕姚长青,男,1974 年生,副研究员,博士,发表论文 10 余篇。
杜永萍,女,1977 年生,副教授,博士,硕士生导师,发表论文 20 余篇。