学术论文大纲中关键术语抽取方法研究*

何远标[1,2]　乐小虬[1]　张　帆[1,2]
[1](中国科学院国家科学图书馆　北京 100190)
[2](中国科学院大学　北京 100049)

摘要: 【**目的**】针对学术论文大纲内容精炼、层次性的特点, 研究从中抽取重要且具有实质意义术语的方法。【**方法**】结合语言学规则和术语词典从大纲各级标题中识别出候选术语集, 然后根据术语间的句法依存关系计算 tf-idf, 并利用大纲结构量化术语层级特征, 最后结合 tf-idf 与层级特征对候选术语进行排名, 选择出关键术语。【**结果**】实验证明, 该方法的候选术语识别 F 值达到 89.57%, 术语选择 F 值达到 36.89%。【**局限**】采用的术语抽取规则不完备, 且 tf-idf 计算过程中的权值设置仅使用经验值, 导致未能达到最优效果。【**结论**】该方法能有效抽取大纲中的关键术语, 适用于层级结构中的关键术语抽取。

关键词: 候选术语识别　候选术语选择　句法依存关系　层级特征
分类号: TP393

1　引　言

关键术语(Keyphrase)是指反映文章主旨的词或短语, 常在自动文摘、信息检索、文档聚类和自动问答等自然语言处理系统中表示文档[1-3]。关键术语抽取(Keyphrase Extraction)是从大量的文本中识别关键术语的过程, 抽取结果对自然语言处理应用有重要影响。传统的抽取方法主要针对全文(如文献[3-5]), 需要从大量的候选词中筛选出合适的词, 过程繁复, 准确率低。近年来, 部分研究从学术论文的逻辑结构(如标题、摘要、图表)中进行抽取, 抽取效率和准确率均有显著改进[6,7]。但由于论文各逻辑元素之间的粒度差异较大, 衡量不同元素间术语的重要度容易产生偏差。

学术论文大纲是由各级标题组成的一个层级结构, 用于展示各个要点或子主题[8], 其特点有: 保留论文的主要信息[9]; 各个标题的粒度处于同一数量级; 反映各子主题之间的层级关系。与全文和逻辑结构不同, 大纲包含的候选词数量少、质量高, 各级标题粒度差异不大, 因此大纲更适用于关键术语抽取。

本文在分析对比已有关键术语抽取方法的基础上, 针对学术论文大纲的特点, 利用标题间的层次关系, 结合语言学规则与统计分析方法进行关键术语抽取。

2　关键术语抽取研究现状

关键术语抽取需要根据相关特征进行判定, 具体过程分为两个步骤[2,3,10,11]:

(1) 候选术语识别, 利用统计学方法、语言学规则等方法初步将符合部分特征的词或词组抽取出来;

(2) 候选术语选择, 利用分类、排名等方法计算特征项, 筛选出最有代表性的术语作为关键术语。

2.1　术语特征

特征是指作为标志的显著特点[12], 是针对某个具体问题的测量指标, 其在不同的环境中表现的显著度有所差异。根据这个特性, 可将术语特征分为 5 类[13], 如表 1 所示。

这 5 类特征的测度范围依次扩大, 从不同的视角考量术语的专指性和概括性。实际应用中, 应针对文档集的特点及抽取方法的特性, 组合多个不同类别

收稿日期：2013-09-26
收修改稿日期：2013-12-12
*本文系国家科技支撑计划子课题“基于文献知识网络的领域学术关系研究与示范”(项目编号: 2011BAH10B06-04)的研究成果之一。

表 1　术语特征分类

类型	特征项	特性	适用环境
短语级特征	短语长度、词性、后缀	用于描述短语自身的属性，通常关键术语都会满足一定的条件，如词性多数为名词	术语外在形式比较规范的文档
标准特征	tf-idf、首次出现位置、段首、段尾、前 N 项、后 N 项	是有效的基准特征，从术语的分布统计的角度测量其专指性	术语重复出现率高的长文档
文档级特征	缩写、PMI[13]、句法特征	在文档范围内对术语进行测度，用于确定术语的主旨覆盖能力	上下文联系紧密的文档
语料库级特征	sf-isf、关键词	用于确定领域内关键术语	领域性强的文档集
外部知识库特征	维基词典、叙词表、本体	通过参考外部知识库进行术语识别	领域词汇变化不大的文档集

的特征对术语进行全面评估。

2.2　候选术语识别

候选术语识别是根据词或词组的外在形式判定，获得初步术语集的过程。传统的候选术语识别方法可以归纳为利用语言学规则、基于术语词典和统计分析三种[14,15]。

(1) 利用语言学规则。其核心是针对特定的语料环境，通过分析术语构成的特征，制定一系列共性规则及个性化规则来自动提取术语[14]。这种方法最大的缺陷是人工编写的规则不能覆盖所有的语言学现象[16]。

(2) 基于术语词典。该方法利用领域专业术语词典查找文本中的术语，词典一般由领域专家编撰，具有权威性。然而，词典的更新速度慢，不能识别学术论文中的衍生术语和新术语。

(3) 统计分析方法。核心思想是通过构建统计模型，综合计算词频、互信息、信息熵等特征值，抽取超过阈值的词串作为候选术语集。这种方法偏向于选择高频术语，容易忽略低频术语。

实际应用中，上述三种方法通常混合使用，互补优缺，改善识别效果，如语言学规则与词典结合，术语词典与统计分析结合。

2.3　候选术语选择

候选术语选择是根据术语的内在含义从候选集中筛选出最能代表文档主旨的术语的过程，选择方法分为监督和无监督两种[17-19]。

(1) 监督方法是将关键术语抽取看作分类问题[18,19]，通过分类器将文档中所有的候选术语分成关键术语与非关键术语两类。常用的分类法有朴素贝叶斯(如 Turney[20]、Nguyen 等[2,7]的研究)、条件随机场(如 Yu 等[21]的研究)、决策树(如 HaCohen-Kerner 等[4]的研究)等。监督方法的缺点在于需要大量的人工标注语料进行模型训练。

(2) 无监督方法则将关键术语抽取看作术语排名问题[18,19]，通过测度一系列的术语特征，按其综合得分进行排名，选择若干靠前的术语作为关键术语。其中，基于图的排名方法是最常用的方法，首先利用文档中术语之间的某种关系(如共现关系)构建一个图，然后利用随机漫步方法(Random Walk Techniques)计算每个术语的重要度，最后根据重要度排名选取关键术语[18]。Zhao 等[22]、Liu 等[18]在图排名基础上引入 PageRank 算法，取得良好效果。另外，还可直接组合多个术语特征值进行排名，如 Liao 等[19]的研究。无监督方法能灵活地组合各种术语特征来选择关键术语，通过简单的参数调整即可以适应不同特性的文档集。缺点在于其效果主要依赖于特征项的选取。

2.4　小　结

关键术语抽取包括术语特征选取、候选术语识别和候选术语选择三个子问题。在目前已有的术语特征项(见表 1)中，除了短语级特征之外，大多数针对长文本，且没有考虑到同一句子中不同成分术语的重要度，也没有考虑大纲层级给术语带来的影响。对于候选术语识别，目前的研究大多将多种方法混合使用以达到更好的识别效果；同时，大纲中的标题简短，语言学规则与词典结合的方法更适合。对于候选术语选择，由于学术论文具有明显的领域特征，使用监督方法需对各领域训练不同的分类器，灵活性低，语料标注难度大；而无监督方法主要依赖于特征项的选取。本文利用改进的 tf-idf 和层级特征项进行无监督术语选择。

3　学术论文大纲中关键术语抽取方法

3.1　方法思路

本文抽取关键术语的基本思路是：首先，针对术

语的词性特征、关键词特征和缩写词特征，使用语言学规则和术语词典结合的方法对大纲各级标题进行候选术语抽取，得到多个候选术语集。各术语集之间保留原大纲中的层级关系，本文称之为层级术语集。然后，利用句法依存关系计算各候选术语的 tf-idf (Term Frequency－Inverse Document Frequency)，并根据层级关系计算层级特征。最后，结合 tf-idf 和层级特征对候选术语进行排名，选择关键术语。

(1) 候选术语识别。术语基本上都是名词或名词性短语[3,7,11]，利用词性特征能够识别大部分候选术语。考虑到术语是一种结合紧密的固定或半固定的词或短语[14]，在仅有词性信息的条件下，难以判断是两个独立术语还是一个固定术语。为降低复杂度并保证准确率，将语料中的关键词和缩写词构建成术语词典，使用语言学规则和术语词典结合的方法进行候选术语识别。

(2) 候选术语选择。论文大纲的内容简短，传统特征如 tf-idf、出现位置、PMI 等表现不显著，本文使用基于句法依存关系的 tf-idf 和层级特征作为关键术语的主要考量特征。基于句法依存关系的 tf-idf 是结合句法特征及基于图排名算法来量化标题中的术语重要度。术语的支配或受支配关系越多，重要度越高。层级特征则用于量化不同大纲层级术语的主旨覆盖能力，层级越深，能力越弱。本文采用无监督方法进行候选术语选择，通过组合基于句法依存关系的 tf-idf 和层级特征对大纲所有的候选术语进行排名，选取 TopN 个作为关键术语。

3.2　基本流程

处理流程分两部分：候选术语识别，通过对论文大纲标题进行词性标注，并利用指定的词性规则识别出各标题中的候选术语，按大纲的层级关系组织起来得到层级候选术语集；候选术语选择，利用句法依存关系计算术语 tf-idf，并根据层级关系计算层级特征，最后结合 tf-idf 与层级特征进行术语排名，从层级候选术语集中筛选出关键术语。整体处理流程如图 1 所示。

以下将对候选术语识别和候选术语选择两个步骤进行介绍。

3.3　候选术语识别

使用语言学规则和术语词典结合的方法进行候选术语识别：首先将出现在词典中的词串标注为名词(NN)，在此基础上对其他部分进行词性标注，然后利用规则抽取符合指定词性特征的词串。在该过程中，术语词典与词性识别规则是决定抽取效果的核心要素。

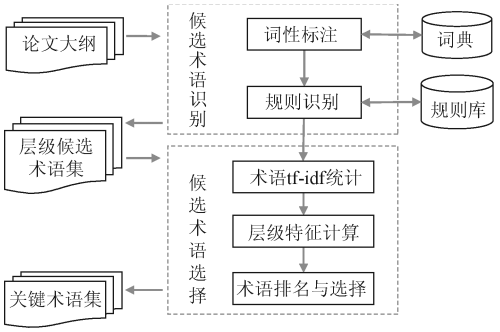

图 1　处理流程

(1) 术语词典

本文采用的词典数据来源于两方面：领域内论文的关键词，相对于外部知识库，关键词具有更新速度快、主题覆盖能力强等特点；论文大纲中的术语缩写及其原型，通过对大纲进行缩写检测得到，其内部结合紧密。

(2) 词性识别规则

名词性术语的词性组合主要有“JJ NN”、“NN”、“NN NN”、“VBN NN”几种[3]，本文设计如下正则表达式来匹配这些词性特征：

(JJ(\\w){0,1})*(NN(\\w){0,2})*NN(\\w)*　　(1)

3.4　候选术语选择

对大纲进行候选术语抽取后得到的层级术语集既有术语内容，又有层次关系信息，这两方面的特征决定术语的重要度，依此选择关键术语。

(1) 基于句法依存关系的 tf-idf 统计

在识别出候选术语的基础上，对大纲各级标题进行句法关系分析，将得到依存关系图。以文献[23]的标题“A Concept and Implementation of Higher-level XML Transformation Languages”为例，如图 2 所示，其中，有向弧的起始端表示支配者，箭头端表示从属者。

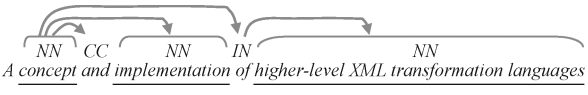

图 2　句法依存关系

根据句法依存关系，抽取形如<S,P,O>的三元关

系。S 为支配者, O 为从属者; P 表示三元关系类型, 用术语间的介词、连词、动词表示。不同类型的三元关系被赋予不同的权重因子(可设为经验值或通过监督学习得到), 用于计算支配者和从属者的重要度。抽取结果示例如下(从图 2 抽取):

S : concept ; P : and ; O : implementation;factor=1

S : concept ; P : of ; O : higher-level XML transformation languages;factor=1.25

S:implementation ; P : of ; O : higher-level XML transformation languages;factor=1.25

在三元组中, S 的词频默认为 1, O 的词频为 1×factor(factor 为权重因子)。术语 w 的词频是通过对<S,P,O>中词频累加并进行归一化得到。本文提出如下公式进行术语词频统计:

$$tf_w = \sum_{S\,is\,w} 1 + \sum_{O\,is\,w} 1 \times factor \quad (2)$$

$$tf_w = \frac{tf_w}{\sum_{w_i \in U} tf_{w_i}} \quad (3)$$

公式(3)中, U 为候选术语集。利用公式(2)和公式(3)对上述关系集进行词频统计的结果为:

tf(concept)=0.208

tf(Implementation)=0.208

tf(higher-level XML transformation languages)=0.385

在 idf 计算过程中, 将标题看作一个文档, 为防止低频新术语导致 idf 过高, 本文通过引入平滑因子来改善 idf 计算, 如公式(4)所示:

$$idf_w = \log\left(\frac{N}{df_w + 1} + e\right) \quad (4)$$

(2) 层级特征计算

层级特征是反映主题覆盖能力的测度指标。大纲描述的内容逐层深入和细化, 下级标题往往是上级标题的分面具体化描述或其子主题, 因此, 出现位于不同大纲级别的术语具有不同的主旨覆盖能力。一般来说, 层次越深其主题描述能力越弱, 重要度越低。本文提出公式(5)计算层级特征:

$$H\_Feature(h) = \prod_{i=1}^{h} \frac{1}{\ln(Count(i) + e)} \quad (5)$$

其中, h 是节点的层次深度, Count(i)是兄弟节点数量, e 为平滑因子。

(3) 候选术语排名与选择

结合基于句法依存关系的 tf-idf 与层级特征计算候选术语的综合特征得分排名, 选择得分 TopN 个术语作为关键术语。本文提出公式(6)计算综合得分:

$$Score(w) = tf_w \times idf_w \times H\_Feature(h), (w \in U(h)) \quad (6)$$

4 关键术语抽取实验

为验证上述文章大纲中关键术语抽取方法的有效性, 本文对候选术语识别和关键术语选择两个环节分别进行实验。实验数据是从 Elsevier 上以“Concept Hierarchy”为关键字, 时间为 2004 到 2013 年, 检索得到前 50 篇学术论文的大纲。通过人工标注, 得到大纲标题 443 条, 候选术语 921 个及大纲关键术语 150 个。另外, 抽取检索结果中的前 1 000 篇文章的大纲及关键词, 用于构建词典。

4.1 候选术语识别

采用 StanfordNLP 工具[24]对数据进行词性标注, 并通过正则表达式进行词性规则匹配, 匹配规则如 3.3 节所述。术语词典有两个: 关键词词典, 由检索结果的前 1 000 篇文章的关键词构成, 共 2 447 个词; 缩写词词典, 从这 1 000 篇文章的大纲中进行缩写检测得到, 共 412 个词(包括缩写及其原型)。缩写词词典、关键词词典和规则识别三种方法不同组合的实验评测结果如表 2 所示:

表 2 候选术语识别评测结果

方法	准确率	召回率	F 值
缩写词词典	85.11%	4.35%	8.28%
关键词词典	69.14%	12.19%	20.72%
关键词词典+缩写词词典	72.83%	13.71%	23.08%
规则识别	82.64%	88.03%	85.25%
规则识别+关键词词典+缩写词词典	88.11%	91.08%	89.57%

实验结果显示, 单独使用一种词典时, 缩写词词典的准确率最高; 而关键词词典因词数量较多, 召回率高于缩写词词典。综合使用两个词典可以提高对术语的识别率。另外, 使用规则识别的 F 值比使用词典的 F 值有较大提升, 说明名词规则能够覆盖大纲中大部分术语。综合两个词典和规则识别的方法获得最佳识别效果, F 值接近 90%。实验存在的不足有:

(1) 术语匹配规则不足以覆盖所有的术语构成特征, 如“shoulder and neck pain”词性为“NN CC NN NN”, 并不符合 3.3 节所述的名词规则;

(2) 部分关键术语本身存在多种切分, 如

“restricted Coulomb energy (RCE) neural network”可切分为｛“restricted Coulomb energy”, “RCE”, “neural network”｝或｛“restricted Coulomb energy (RCE) neural network”｝，本实验没有考虑不同切分之间的兼容处理。

4.2　候选术语选择

在识别关键术语的基础上，使用 StanfordNLP 工具[24]进行句法依存关系分析，根据术语间的支配关系提取<S,P,O>三元组，并计算 tf-idf 值。层级特征依据大纲的层级结构计算，计算方法见公式(5)。实验选取排名 Top1 和 Top3 的关键术语进行统计分析，评测结果如表 3 所示:

表 3　关键术语选择评测结果

方法	Top1			Top3		
	准确率	召回率	F 值	准确率	召回率	F 值
tf-idf	19.61%	19.23%	19.42%	23.53%	23.84%	23.68%
tf-idf+层级特征	37.25%	36.54%	36.89%	48.37%	49.01%	48.68%

传统关键术语抽取方法取得的 F 值为 30%左右(如 Berend 等[13]、Nguyen 等[7]、Medelyan 等[25]的研究)，本实验结合基于句法依存关系的 tf-idf 和层级特征进行排名选择，得到的识别率 F 值为 36.89%，若选取 Top3 的关键术语，F 值达到 48.68%。其中，采用 tf-idf 与层级特征结合的方法的识别效果(准确率、召回率、F 值)约为仅使用 tf-idf 的两倍。数据表明，基于句法依存关系的 tf-idf 和层级特征结合的关键术语选择方法能有效提取学术论文大纲中的关键术语，同时，层级特征反映术语的主题覆盖能力，对识别关键术语有重要作用。实验不足之处在于:

(1) 实验中<S,P,O>三元关系权重因子的取值为经验值，需改进调优;

(2) 未考虑论文大纲中各个层级自身的特殊性(如根节点和叶节点往往包含重要信息较多)，需进一步深入研究。

4.3　小　结

本实验使用语言学规则和术语词典结合的候选术语识别方法、基于句法依存关系的 tf-idf 和层级特征结合的关键术语选择方法从学术论文大纲中抽取关键术语，并取得良好效果。对于候选术语识别，使用词性特征和关键词、缩写词词典即可覆盖近 90%的术语，进一步改善可通过优化抽取规则以及扩大术语词典的方法来实现。对于候选术语选择，基于句法依存关系的 tf-idf 能有效地从短文档中筛选关键术语，结合论文大纲的层次结构特征的术语抽取结果的 Top1 F 值达 36.89%(Top3 为 48.68%)。优化三元关系权重因子以及层级特征计算方法均可提高关键术语选择的准确率。

5　结　语

学术论文大纲是内容的框架，包含论文的主要信息以及不同子主题之间的关系，充分利用这种关系能提高关键术语的抽取质量。本文分析了目前主流的关键术语抽取方法，针对学术论文大纲的特点，利用语言学规则和术语词典结合的候选术语识别方法，以及基于句法依存关系的 tf-idf 和层级特征结合的候选术语选择方法进行关键术语抽取。实验表明，这种方法取得了良好的抽取效果，适用于大纲或类似大纲的层级结构中的关键术语抽取。下一步工作将针对识别规则、词典及三元关系的权重等方面的不足进行改进，并深入研究论文大纲的层级特点，挖掘术语间的语义关系。

参考文献:

[1] Kim S N, Medelyan O, Kan M Y, et al. SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles[C]. In: Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval’10). Stroudsburg: Association for Computational Linguistics, 2010: 21-26.

[2] Nguyen T D, Kan M. Keyphrase Extraction in Scientific Publications [C]. In: Proceedings of the 10th International Conference on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers (ICADL’07). Berlin, Heidelberg: Springer-Verlag, 2007: 317-326.

[3] Kim S N, Kan M. Re-examining Automatic Keyphrase Extraction Approaches in Scientific Articles [C]. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE’09). Stroudsburg: Association for Computational Linguistics, 2009: 9-16.

[4] HaCohen-Kerner Y, Gross Z, Masa A. Automatic Extraction and Learning of Keyphrases from Scientific Articles [C]. In: Proceedings of the 6th International Conference on

Computational Linguistics and Intelligent Text Processing (CICLing'05). Berlin,Heidelberg: Springer-Verlag, 2005: 657-669.

[5] Planta E, Tonelli S. KX: A Flexible System for Keyphrase Extraction [C].In: Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10). Stroudsburg: Association for Computational Linguistics, 2010: 170-173.

[6] Alzahrani S, Palade V, Salim N, et al. Using Structural Information and Citation Evidence to Detect Significant Plagiarism Cases in Scientific Publications [J]. Journal of the American Society for Information Science and Technology, 2012, 63(2): 286-312.

[7] Nguyen T D, Luong M. WINGNUS: Keyphrase Extraction Utilizing Document Logical Structure [C]. In: Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10). Stroudsburg: Association for Computational Linguistics, 2010: 166-169.

[8] Wikipedia. Outline [EB/OL]. [2013-09-24]. http://en.wikipedia.org/wiki/Outline_(list)#cite_note-2.

[9] Alotaiby F, Foda S, Alkharashi I. New Approaches to Automatic Headline Generation for Arabic Documents [J]. Journal of Engineering and Computer Innovations, 2012, 3(1): 11-25.

[10] Nguyen C Q, Phan T T. An Ontology-based Approach for Key Phrase Extraction [C]. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Stroudsburg: Association for Computational Linguistics, 2009: 181-184.

[11] Lopez C, Prince V, Roche M. Automatic Titling of Electronic Documents with Noun Phrase Extraction [C]. In: Proceedings of 2010 International Conference of Soft Computing and Pattern Recognition (SoCPaR), Paris, France. IEEE, 2010: 168-171.

[12] 百度百科. 特征 [EB/OL]. [2013-09-24]. http://baike.baidu.com/view/1069886.htm.(Baidu Baike. Characteristic[EB/OL]. [2013-09-24]. http://baike.baidu.com/view/1069886.htm.)

[13] Berend G, Farkas R. SZTERGAK: Feature Engineering for Keyphrase Extraction [C]. In: Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10). Stroudsburg: Association for Computational Linguistics, 2010: 186-189.

[14] 计然. 计算机领域术语的自动获取和层次构建 [J]. 硅谷, 2011 (20): 29-30.(Ji Ran. Terminology Autoamtic Acquisiting and Hierarchy Building in the Field of Computer [J]. Silicon Valley, 2011(20): 29-30.)

[15] 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): 146-149, 155.(Liu Li, Liu Xiaoming. Extraction of Domain-Specific Phenomenal Terms Based on Separator and Contextual Terms [J]. Journal of South China University of Technology: Natural Science Edition, 2011,39(7):146-149, 155.)

[16] 祝清松, 冷伏海. 自动术语识别存在的问题及发展趋势综述 [J]. 图书情报工作, 2012, 56(18): 104-109.(Zhu Qingsong, Leng Fuhai. Existing Problems and Developing Trends of Automatic Term Recognition [J]. Library and Information Service, 2012, 56(18): 104-109.)

[17] Li D, Li S, Li W, et al. A Semi-supervised Key Phrase Extraction Approach: Learning from Title Phrases Through a Document Semantic Network [C]. In: Proceedings of the ACL 2010 Conference Short Papers. Stroudsburg: Association for Computational Linguistics, 2010: 296-300.

[18] Liu Z, Huang W, Zheng Y, et al. Automatic Keyphrase Extraction via Topic Decomposition [C]. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10). Stroudsburg: Association for Computational Linguistics, 2010: 366-376.

[19] Liao L, Huang H. Microblog Keyphrase Extraction Based on Similarity Features[C]. In: Proceedings of 2013 International Conference on Advanced Computer Science and Electronics Information (ICACSEI'13).2013.

[20] Tureney P D. Coherent Keyphrase Extraction via Web Mining [C]. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03). San Francisco: Morgan Kaufmann Publishers Inc., 2003: 434-439.

[21] Yu F, Xuan H, Zheng D. Key-Phrase Extraction Based on a Combination of CRF Model with Document Structure [C]. In: Proceedings of the 8th International Conference on Computational Intelligence and Security (CIS'12). Washington D C: IEEE Computer Society, 2012: 406-410.

[22] Zhao X, Jiang J, He J, et al. Topical Keyphrase Extraction from Twitter [C]. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT'11). 2011.

[23] Foetsch D, Pulvermueller E. A Concept and Implementation of Higher-level XML Transformation Languages [J]. Knowledge-Based Systems, 2009, 22(3): 186-194.

[24] The Stanford Natural Language Processing Group [EB/OL]. [2013-09-24]. http://nlp.stanford.edu.

[25] Medelyan O, Witten I H. Thesaurus Based Automatic Keyphrase Indexing [C]. In: Proceedings of the 6th

ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'06). New York: ACM, 2006: 296-297.

作者贡献声明：

何远标：负责调研，细化研究方向及技术方法路线，设计实验方案；负责实验，包括数据采集、清洗与结构化,编程及实验结果分析；论文撰写与最终版本修订；

乐小虬：提出研究方向和论文选题方向，就研究思路、实验方案及技术路线提供指导；

张帆：数据标注，部分编程及数据分析；参与论文修改。

(通讯作者：何远标 E-mail：bill_ho@foxmail.com)

Research on Keyphrase Extraction from Scholarly Article Outline

He Yuanbiao[1,2]　Le Xiaoqiu[1]　Zhang Fan[1,2]

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

[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] According to the succinct and hierarchical character of scholarly article outlines, this paper concentrates on finding a method to extract important and meaningful phrases from the outlines. [**Methods**] This paper first adopts a combined method of linguistic rules and terminology dictionaries to identify the candidate phrases. Then, it calculates tf-idf based on syntactic dependencies between phrases, and quantifies the hierarchical feature according to hierarchical structure of outline. At last, it combines the tf-idf and the hierarchical feature to rank candidate phrases, and selects the keyphrases. [**Results**] Experiments show that the F-score of the candidate phrases identification reaches 89.57%, and the F-score of candidate phrases selection reaches 36.89%. [**Limitations**] In this method, the inadequate phrase extraction rules and the empirical values involved in weight setting during tf-idf calculation lead to non-optimal effect. [**Conclusions**] This method can effectively extract the keyphrase from outlines, and is suitable for keyphrase extraction from hierarchical structure.

Keywords: Candidate phrases identification　Candidate phrases selection　Syntactic dependencies　Hierarchical feature