第41卷 第6A期
2014年6月

计 算 机 科 学
Computer Science

Vol. 41 No. 6A
June 2014

基于短语的中文标签自动生成混合算法

刘 栋[1] 张彩环[2]

(洛阳师范学院信息技术学院 洛阳 471022)[1] (洛阳师范学院数学科学学院 洛阳 471022)[2]

摘 要 对中文文档标签生成的算法进行了研究,提出了一种中文文档标签生成的混合算法(Hybrid Tags Generation Algorithm)。鉴于短语在表达文档主题方面的优势,先进行短语模式匹配,然后利用短语的统计特性,综合考虑TF-IDF、词跨度和位置3个特征进行权重计算,从而抽取出权重较高的词语或短语作为标签。通过对实验数据的分析表明,该算法在查准率方面表现较好。通过人工比对可知,标签表达文档内容主题的效果相当或优于测试集标准答案的比率超过六成,取得了比较好的结果。

关键词 关键词抽取,标签生成,短语,中文标签,算法

中图法分类号 TP301.4 **文献标识码** A

Keyphrase-based Chinese Tags Generation Hybrid Algorithm

LIU Dong[1] ZHANG Cai-huan[2]

(Department of Information Technology, Luoyang Normal University, Luoyang 471022, China)[1]

(Department of Mathematics, Luoyang Normal University, Luoyang 471022, China)[2]

Abstract This work provided an algorithm HTGA(Hybrid Tags Generation Algorithm) to generate tags for Chinese documents, which extracts phrase chunks as candidate keywords, and considers other factors like TF, IDF, words span etc. Experiments show that this algorithm improves the accuracy of keyword extraction, and has a stable performance over various texts. Some samples were extracted and compared with the standard answers. There are more than 60% results that are as well as or better than the standard answers in reflection of document topics.

Keywords Keyword extraction, Tag generation, Keyphrase, Chinese tags, Algorithm

从语言的出现到文字的诞生,从简单的沟通、交流到信息的记录、传承,信息无处不在。近年来,随着信息技术的飞速发展和Internet的迅速普及,信息的表示形式呈现出多元化的趋势,除书籍、杂志等传统的传播媒介之外,人们更多接触的是网页、电子书,邮件,微博、短信、彩信等电子形式的信息资源。面对这些海量信息,如何快速有效地获取、存储、使用和管理这些资源,从海量信息中提取出有价值的信息,已经成为了迫切需要解决的问题。为了提高获取相关资源的效率,降低获取成本,搜索、阅读关键词(keywords又称为标签,tags)是一个有效的手段,它能帮助人们迅速决定是否有必要阅读全文,从而大大提高阅读的效率;对于信息的检索和管理来说,对这些关键词进行归纳整理,分门别类,就可以实现快速的查找和使用。但是,目前这些繁重的工作大部分都是由人工来完成的,人工处理效率低下,易受个人主观因素的影响,准确率不高。特别是随着现代社会信息量的爆发式增长,人工处理方式已经不能满足当今社会发展的实际需要,研究与开发自动、准确、高效的中文文档标签生成技术有着很强的应用需求。

标签生成起源于"自动标引"[1],文档标签自动生成(又称关键词自动提取)是指从文档中提取具有专指性且能反映文档主题的词语或短语,整个提取过程由计算机自动完成,很少或者几乎不用人工参与。这是信息处理中常用的一项关键技术。这项技术可以自动从文档中提取标签,帮助用户从大量、不规则的资源中快速、便捷地找到需要的信息,同时这也可以作为文本自动分类,信息检索、自动文摘等信息处理任务的前期处理工作。可以说关键词提取可以作为所有文本自动处理的基础与核心技术。

本文提出了一种中文文档标签生成的混合算法(Hybrid Tags Generation Algorithm)。基本思想是尽可能抽取,并且优先采用短语作为标签。先进行短语模式匹配,然后再利用短语的统计特性,综合考虑词跨度和位置因素进行权重计算,从而选择权重较高的词语或短语作为标签。下面本文第1节描述算法处理的流程以及算法的各个主要组成部分;第2节是算法的实验与分析;第3节简述了与本文相关的研究。

1 算法处理的流程及描述

1.1 算法流程

本文所提出的中文文档标签生成算法的中心思想是以提取能够表达文章中心思想的短语来作为文档标签的,需要对能否组成短语以及短语的长度进行判断,同时也要考虑这些

本文受2010年度河南省基础与前沿技术研究资助项目(102300410211),2009年河南省高等学校青年骨干教师资助项目(2009GGJS-105)资助。

刘 栋(1971—),男,博士,讲师,主要研究方向为信息检索、云计算、自然语言处理、数据挖掘等,E-mail:212616@qq.com;张彩环(1969—),女,博士,副教授,主要研究方向为计算数学、组合数学等。

短语的统计信息和重要程度，因此算法从中文自然语言处理的基本步骤——分词开始，分别进行基本短语模式的词性匹配、TF-IDF 计算、词跨度的计算和位置因素的计算，综合考虑各个特征，得出短语的最后权重，排序并取得权重最高的几个短语或词语作为文档标签。整个处理流程如图 1 所示。

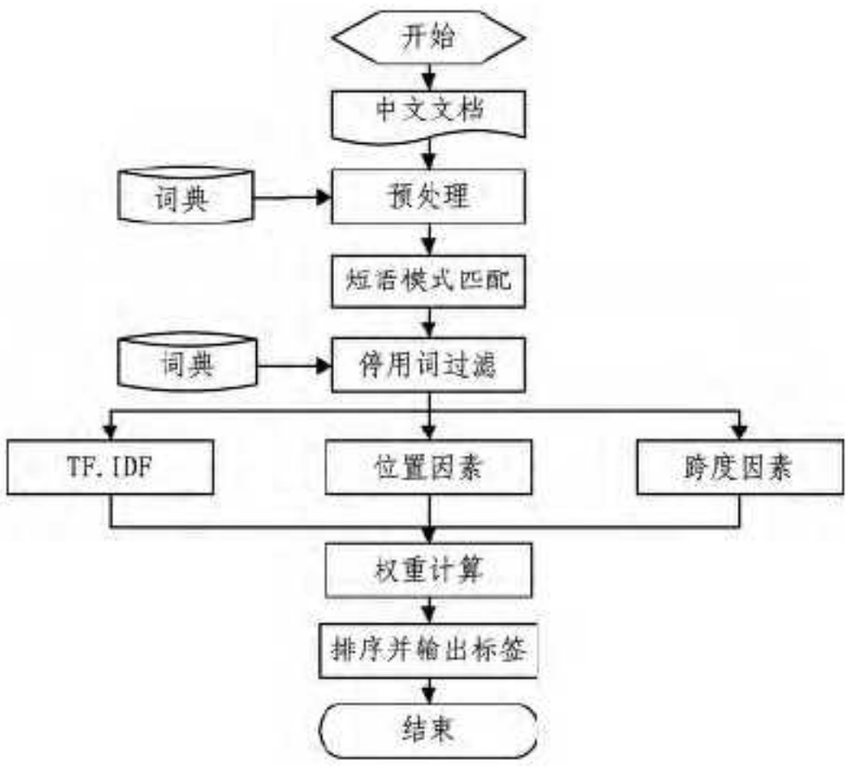

图 1　处理流程图

经过短语模式匹配之后再进行停用词过滤，这将过滤掉那些对文本主题不具备表达能力或者表达能力可以忽略的词，如功能词、标点符号等，需要注意的是这里的停用词过滤已经是第二次过滤了，因为在短语模式匹配过程中已经将很多不符合词性要求的词或短语过滤掉了，但仍然存在一些对区分文本没有意义的词，因此需要进行二次过滤。最后的权重计算，使用下面的模型来进行：

$$weight_i = \sum_{j=1}^{3} w_j Factor_{i,j}$$

由于本算法是先进行短语模式的匹配，接着综合考虑词频、词的跨度以及位置 3 个特征，在考虑语言习惯的前提下兼顾了统计与规则，因此我们把它称为混合标签生成算法，简称为 HTGA(Hybrid Tags Generation Algorithm)。

1.2　算法描述

1.2.1　文档预处理

预处理阶段的主要任务是对输入的文档类型、大小进行判断以及分词处理等，以便算法对其进行各种处理及统计：

1)对文档的类型进行判断，如果是可以转换为 TXT 的文档，则将其转换为文本文件，以 UTF-8 格式保存等待处理；否则取其文件名作为识别对象，以同样方式进行处理。

2)对文档的大小进行判断和处理。有些文档是长文档(大小在几 Mb 至几十 Mb 不等)，对于这样的文档无论是分词也好，还是进行各种统计都无法在可以接受的时间范围内对文档内容进行处理。因此，这里采用了对内容进行“切片”的方法，从文档中抽取适当的内容进行分析、处理，极大地提高了处理的效率，通过实验证明生成的结果也可以接受。

3)进行分词处理。用中文分词软件 SCWS[2] 进行分词，同时保留词性标注及分词的各种信息，留作以后处理使用。

1.2.2　短语模式匹配

关键短语是具有强文本表示功能的特征短语，所谓强文本表示功能是指在文本表示时，能够将文本的内容特征(例如领域类别、主题思想、中心意义等)鲜明地表示出来[5]。因此使用关键短语作为文档的标签是非常合适的。韩艳[6]在其学位论文中进行了统计分析，结论是大部分关键词都是以短语形式存在。关键短语含词数主要分布在 1-gram，2-gram、3-gram，超过 70% 以上的关键短语包含一个或两个词，这为我们的短语模式匹配提供了思路，在本研究中，我们将词数限制为 2。

SCWS 分词软件采用了北大词性标注版本。参考已有的关于短语的组成模式[12,13]，我们提炼出了自己的模式匹配规则，符合下述规则的选择为候选短语或关键词，否则就被过滤掉，不再考虑。模式规则见表 1。

表 1　短语匹配模式

词性匹配规则	说明	示例
nr/ns/nt/nz	人名，地名，机构团体，其他专有名词	“文化大革命”，“苏联”，“苏北”，“林肯”，“美国大学”
i	成语	“攀龙附凤”，“十全十美”
j	简称略语	“冤假错案”，“房地产”
l	习惯用语	“联产承包”，“摸着石头过河”，“差不多”，“蒙在鼓里”
en	英语	
a/an+n/ng	形容词+名词	“新规”“荒唐做法”“高级官员”“合法权益”“崇高理想”
a/an+v/vn	形容词+动词	“合法经营”“有序撤离”“懈怠延误”“恐怖袭击”“严峻考验”“轻微反弹”
n/ng+n/ng	名词+名词	“民工子女”
vn+vn	名动词+名动词	“违法犯罪”“工作独立”
v+v	动词+动词	“提起公诉”“实施监督”“排名下调”
v+n	动词+名词	“审理案件”“上升压力”“随迁子女”“常住人口”
vn+n	名动词+名词	“公办学校”“教育蓝皮书”
n+v	名词+动词	“第三方扩展”“竞争力排名”“自费求学”
n+vn	名词+名动词	“子女教育”“教育资源配置”“核心提示”

1.2.3　权重因子的度量

在抽取出短语之后本文考虑了 3 个特征作为权重因子，即：TF-IDF、词跨度、位置因子。在特征的选取和计算上我们参考文献[4]中的做法，但没有采用单纯的词频，而是采用了 TF-IDF 的计算方法。下面对 3 个特征的计算方法做详细介绍。

1)TF-IDF

本文中的文档频率 TF 采用了不同的规范化方法，其计算公式为：

$$tf_{i,j} = \frac{n_{i,j}}{1+n_{i,j}}$$

其中，$n_{i,j}$ 表示词 j 出现在文档 i 中的次数。

逆文档频率采用 SCWS 分词软件提供的 idf 值。它是由 SCWS 在分词算法的训练过程中计算出来的，我们对 idf 也做了类似词频的非线性处理，即：

$$idf_j = \frac{idf_j}{1+idf_j}$$

由此得出第一个特征的计算公式：

$$Factor_{1,j} = tf_{i,j} \times idf_j = \frac{n_{i,j}}{1+n_{i,j}} \times \frac{idf_j}{1+idf_j}$$

2)词跨度

词跨度是指词在文中首次出现和末次出现之间的距离。词跨度反映了词在文中出现的范围，相同条件下，词跨度越大，说明该词在文中被提及的范围越广，对文本主题就越重要，是体现词对所在文本重要程度的又一重要特征。本文也

沿用了文献[4]对于词跨度的计算方法，如下所示：

$$Factor_{2,j}=span_j=\frac{last_j-first_j+1}{sum}$$

词跨度的计算是用词 j 在文中最后一次出现的位置减去第一次出现的位置加 1 除以全文的词总数。

3)位置因子

词在文中所处的位置对于判断词的重要性也有着重要价值，位置不同对于文档主题的表达能力也不同，本文把词语所在位置分为头部(50 区)、中部(30 区)和尾部(10 区)。计算公式如下：

$$Factor_{3,j}=loc_j=\frac{v_j}{1+v_j}$$

其中，v_j 的值为词 j 的位置值。

2 实验与分析

2.1 实验环境

实验采用的机器配置为 Intel Core™ i5 CPU，2.53G 主频，4GB 内存的笔记本电脑，32 位的 Windows7 操作系统。开发环境使用了 PHP，方便起见，使用了 WampServer 2.2，其中包括 Apache 2.2.21，PHP 5.3.10，MySQL 5.5.20。

2.2 实验数据

实验所用的数据由 163.com 上的 13702 篇中文新闻作为文档集合[3]。这些新闻包括了科学、技术、政治、体育、文化、社会以及军事等多个主题。所有的新闻文档都被编辑者手工标记了关键词，这些关键词均来自于文档正文。

该文档集有 72900 个词项，关键词有 12405 个词项。每篇文档、标题 和摘要的平均长度为 971.7 个词、11.6 个词和 45.8 个词，平均每篇文档 2.4 个关键词。

2.3 结果分析

2.3.1 客观分析

实验对 13702 篇文档生成了 32797 个标签，平均每篇文档 2.39 个。最多时生成 7 个关键词，最少 1 个。图 2 显示了算法在生成不同关键词个数时的准确率、召回率以及 F1 值。可以看出，算法在生成 2～3 个关键词时效果最好，召回率最高达到 25%，准确率最高达到了 18%。

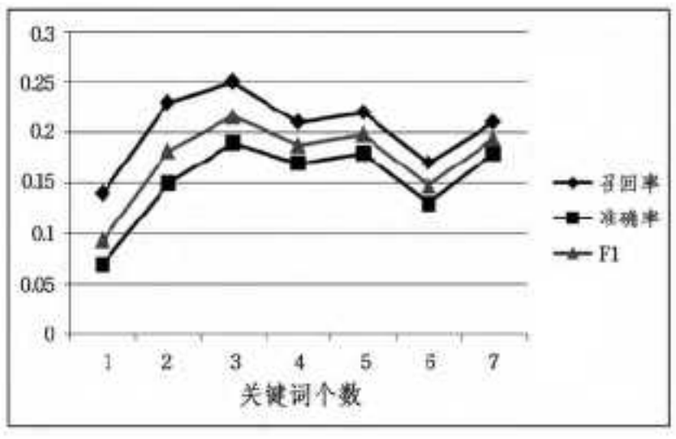

图 2 召回率、准确率及 F1

这种情况符合目前关键词抽取领域的现状，抽取 1 个关键词时召回率和准确率最低，这是因为对于 1 篇文档来说，能够表达主题思想的词或短语往往不止 1 个，在只抽取 1 个的情况下，要做到与标准答案完全相同，可能性的确要比多个时低。在抽取 7 个关键词时准确率和召回率又有所升高，本文分析的原因主要有两点：(1)在关键词比较多时，命中的几率就比较大；(2)有可能是概率性升高，之所以这么说是因为抽取 7 个关键词的文档只有 2 篇，6 个关键词的文档只有 1 篇，因此在较多关键词的情况下，还需要进一步测试算法的效果。

我们将 HTGA 算法抽取关键词的效果与另外两种无监督算法：TF-IDF[8]和 TextRank[7]进行了比较，图 3—图 5 分别显示了 3 种算法的表现，在召回率上 HTGA 算法表现不如其他两个，说明本文的算法在查"全"方面比较差，而在准确率上 HTGA 算法具有较为明显的优势，说明本文算法在查"准"方面表现较好。F1 值是召回率和准确率的调和平均值，总体来讲算法在抽取较少和较多两个比较极端的情况下表现不够稳定，而在常用的抽取 2～5 个关键词时取得了令人满意的效果。

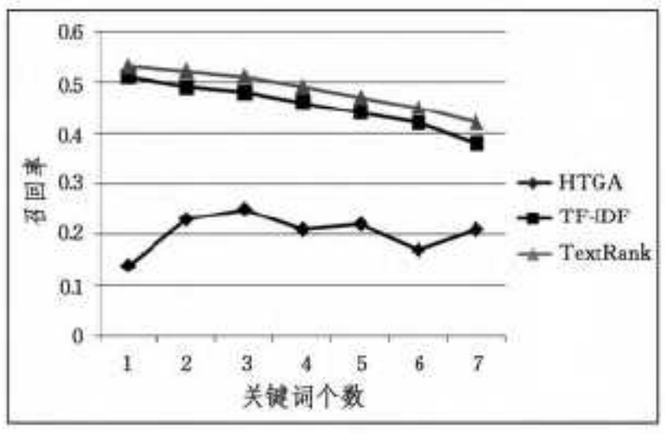

图 3 与其他算法的召回率比较

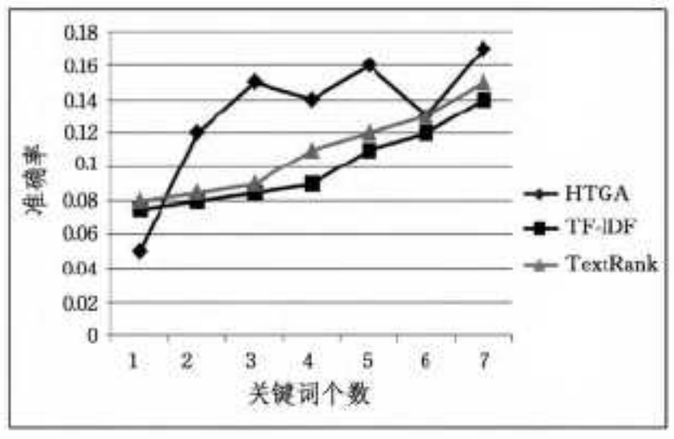

图 4 与其他算法准确率的比较

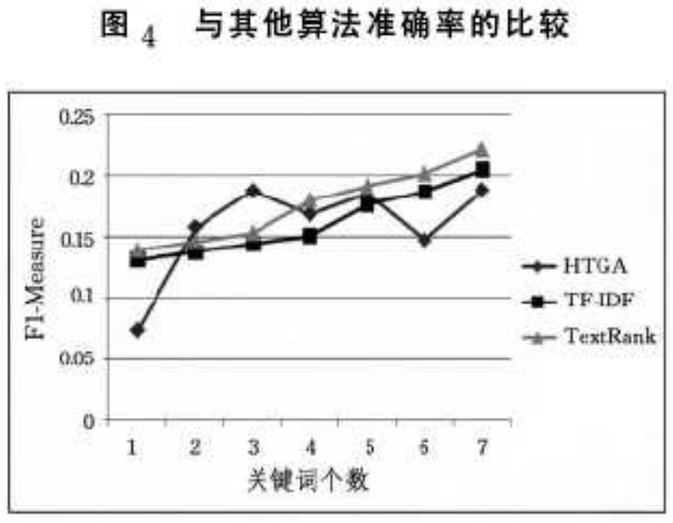

图 5 与其他算法 F1 值的比较

2.3.2 主观评价

由于 HTGA 算法是以短语作为抽取的首要目标，因此在生成的标签中比较难以做到与标准答案完全一致，这也是导致 HTGA 算法召回率比较低的原因之一。为了更深入地考察算法对于标签生成的效果，除了对生成效果采取客观计算进行评价之外，本文还对部分结果进行抽样，人工进行评价，把算法生成的标签与标准答案的关键词放在一起进行比较，人工主观地判断哪个结果更贴近文档内容，判断的结果分为三档：优、相当、差。从文档集中抽取 1000 篇文档，人工分析算法生成的标签，再与标准答案进行比较，我们得到了表 2 所列的结果。

表 2 主观评价结果

评价	篇数	占比
优	128	12.8%
相当	536	53.6%
差	336	33.6%

在评价的 1000 篇文档中抽取的关键词效果优于或者相

当于手工标注的关键词效果的总数是664篇，占总数的66.4%，下面通过实例对这3种情况加以说明。

对于第一种情况，也就是生成的标签优于人工标注，我们来看一个例子，表3中的第一个例子新闻标题为“荷兰海军将首次派遣潜艇赴亚丁湾打击海盗”，标准答案提供的关键词为海盗、潜艇、荷兰；算法标注的关键词为海盗、海军、护航行动。人工比较认为，算法标注的关键词中出现了“护航行动”的字眼，反映了新闻的主旨，因此优于“海盗、潜艇、荷兰”。另一例子产生了“罢免村主任”这样的短语作为关键词更能反映新闻的主题。

表3 优于标准答案举例

标题	算法抽取的关键词	标准答案
荷兰海军将首次派遣潜艇赴亚丁湾打击海盗	海盗，海军，护航行动	海盗，潜艇，荷兰
西安村民罢免村主任 选票未达全村总数一半失败	罢免村主任 人数	村主任 罢免

对于第2种情况，算法与标准答案的关键词不尽相同，但效果相当。如表4中的第一篇文档的标题为：“重庆打黑案中案：贪官借子女留学买房洗钱”，新闻给出的核心提示是：“日前，检察官对重庆方面经办贪官案件作出介绍，揭露了一系列“黑钱漂白”的手段，其中大肆买房和送子女留学是重要途径。”对比算法和标准答案的关键词可以看出，“洗钱犯罪 打黑 案件”和“漂白 贪官 黑钱”都反映了文档的内容实质，看到关键词就能够让人明白文档所要讲述的大致内容，因此我们认为两组关键词的效果相当。

表4 效果相当的情况举例

标题	算法抽取的关键词	标准答案
重庆打黑案中案：贪官借子女留学买房洗钱	洗钱犯罪 打黑 案件	漂白 贪官 黑钱
30名阿富汗塔利班人员向政府投诚	塔利班 人员 投诚	塔利班 投诚 赫拉特

对于第3种情况，我们也举两个例子，见表5，这两个例子的标题比较明显地表达了文档所描述的主要内容，第一例子的标题为“江苏海门副市长以40%利率强行向企业放贷”，标准答案的关键词是“副市长 强行 放贷”，可以说抓住了文档内容的主题；而算法抽取的关键词是“利息 检察机关 利率”，虽然也透露出文档内容的类型，但对于文档主题的表达却是不够清楚的；第二个例子标题为“汽车以旧换新政策实施期限将延长至年底”，主题也是十分清楚的，标准答案的关键词为“以旧换新 汽车”抓住了文档的主干，而算法抽取的关键词“汽车 政策”就比较含糊不清，看了之后只知道这篇文档是关于汽车政策的，但具体什么政策却不清晰，因此我们认为在这个例子中算法得到的结果不如标准答案的结果。

表5 算法不如标准答案的例子

标题	算法抽取的关键词	标准答案
江苏海门副市长以40%利率强行向企业放贷	利息 检察机关 利率	副市长 强行 放贷
汽车以旧换新政策实施期限将延长至年底	汽车 政策	以旧换新 汽车

3 相关研究

20世纪90年代以来，随着互联网普及出现了海量信息，如何有效地从这些海量信息中检索到有意义、有价值的内容成了人们迫切的要求，也诞生了Google、百度等这样的搜索引擎巨头。关键词可以让人们快速地了解一篇文档的主要内容，把握文档的主题，成为了人们从海量信息中检索内容的重要工具，也引起了广大研究者的重视。许多研究采用了基于语言分析的方法，通过研究词语与词语、词语与句子、词性之间的关系来抽取关键词，比如Hulth通过短语识别、Chunk识别等句法分析来获得关键短语[15]；文献[9-11]分别从语义、词汇链、词性3个方面的特征来抽取关键词；文献[14]提出一个用于自动标引的文献主题关键词提取方法，限于从已经标引的结构化语料库元数据的标题中提取关键词，这类方法需要不断增大的背景知识库才能保证提取的效果。

还有另外一类研究方法，将人工智能中的机器学习技术引入到关键词抽取中。Turney[16]提出了基于遗传算法的GenEx关键词自动抽取方法和基于机器学习和有监督的关键词自动抽取方法；Witten[17]提出了基于朴素贝叶斯的关键词自动抽取方法；这类方法通过训练数据进行模型训练，从而获得统计参数，然后对文档进行关键词抽取。还有应用如最大熵模型[18]、支持向量机(SVM)[19]、决策树等技术进行关键词提取，其主题思想是将关键词抽取问题转换为分类问题。文献[20]将关键短语抽取问题转化为序列标记问题，利用条件随机场(Conditional Random Field，CRF)进行关键词抽取。基于复杂网络理论的关键词抽取技术也得到了广泛的研究[25,26]。而采用简单统计方法则不需要训练数据，主要是利用文档中词语的统计信息来抽取文档的关键词，比如利用包括词频、TFIDF、词的同现信息Pat-Tree，或是上述某些统计方法的结合等[21-24]

结束语 本文对中文文档标签生成的算法进行了研究，提出了一种中文文档标签生成的混合算法(Hybrid Tags Generation Algorithm)。该算法采取了规则与统计相结合的方法，基本思想是提取短语作为候选的关键词，从实验评价来看，尽管生成的标签的召回率、准确率与其他算法相比优势并不明显，但是我们认识到这种简单的比较与标准答案是否相同的方法对于关键词提取来说并不合适。因此我们采取了主客观评价相结合的方法，从实验结果可以看出由于短语在表达文档主题方面的优势，提取的标签尤其是短语标签清晰地反映了文档的内容，效果优于或相当于人工抽取的比例超过六成。下一步的工作将继续围绕文档的标签生成技术来展开，进一步提高标签生成算法的性能和效率，并将其应用于搜索引擎、文本分类以及社交网络的分析等领域。

参考文献

[1] 章成志. 自动标引研究的回顾与展望[J]. 现代图书情报技术，2007(11)：33-39

[2] SCWS分词软件[OL]. http://www.xunsearch.com/scws/

[3] Liu Zhi-yuan，Chen Xin-xiong，Zheng Ya-bin，et al. Automatic keyphrase extraction by bridging vocabulary gap[C]// Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics. 2011：135-144

[4] 谢晋. 基于词跨度的中文文本关键词提取及在文本分类中的应用[D]. 杭州：浙江工业大学，2011

（下转第102页）

结束语　传统的语义网研究着眼于具有一定内部结构的静态对象的表示方法，我们的研究更注重词汇与语义的对应以及语义的动态生成。通过将意象图式提供的语义作为关系节点的语义，并加入属性空间表示各种基本属性，语义图具备了比传统语义网更强的语义表示能力。而将语义的表示和理解转化为特定语义操作序列的实现，则使语义图拥有更加灵活的推理能力和更强的语义扩展能力。

后续的研究包括如何处理一词多义现象，如何将隐喻加入到系统中以表示更丰富的抽象概念，以及如何恰当地表示两个对象之间的类属关系等。

参 考 文 献

[1] 危辉. 人工智能形式概念系统[M]. 北京：科学出版社，2011

[2] 袁毓林. 基于统计的语言处理模型的局限性[J]. 语言文字应用，2004，5(2)：99-107

[3] 袁毓林，陈振宇，张秀松，等. 从认知假设到计算分析和程序实现——一种认知语言学研究的计算范式与技术路线[J]. 当代语言学，2010，12(2)：97-114

[4] 陈振宇. 时间系统的认知模型与运算[M]. 上海：学林出版社，2007

[5] 陈振宇. 疑问系统的认知模型与运算[M]. 上海：学林出版社，2010

[6] Kedad Z，Métais E. Ontology-Based Data Cleaning [M]// Andersson B，Bergholtz M，Johannesson P. Natural Language Processing and Information Systems. Springer Berlin Heidelberg，2002：137-149

[7] Luger G F. 人工智能：复杂问题求解的结构和策略[M]. 史忠植，张银奎，赵志崑，等译. 北京：机械工业出版社，2006

[8] 朱海平. 基于概念图匹配的语义搜索[D]. 上海：上海交通大学，2006

[9] 王寅. 认知语言学[M]. 上海：上海外语教育出版社，2007

[10] Barwise J，Perry J. Semantic innocence and uncompromising situations[M]// Martinich A P，ed. The Philosophy of Language (2nd Edition). New York：Oxford University Press，1990：392-404

[11] Barwise J，Perry J. Situations and Attitudes[M]. Stanford，CA：CSLI Publications，1999

[12] 巴斯 D M. 进化心理学：心理的新科学(第二版)[M]. 熊哲宏，张勇，晏倩，译. 华东师范大学出版社，2007

[13] 朱跃. 语义论[M]. 北京：北京大学出版社，2006

[14] 张维鼎. 意义与认知范畴化[M]. 成都：四川大学出版社，2007

[15] Sowa J F. Conceptual structures：Information precessing in mind and machine[M]. Anddison-Wesley，1984

[16] Talmy L. Force dynamics in language and cognition [J]. Cognitive Science，1988，12(1)：49-100

[17] Sternberg R J. 认知心理学(第三版)[M]. 杨炳钧，陈燕，邹枝玲，译. 北京：中国轻工业出版社，2006

(上接第 90 页)

[5] 刘华. 基于关键短语的文本内容标引研究[D]. 北京：北京语言大学，2005

[6] 韩艳. 基于统计的中文文本关键短语自动抽取方法研究[D]. 苏州：苏州大学，2009

[7] Mihalcea R，Tarau P. TextRank：Bringing Order into Texts [C]// Proceedings of EMNLP. 2004：404-411

[8] 刘知远. 基于文档主题结构的关键词抽取方法研究[D]. 北京：清华大学，2011

[9] 方俊，郭雷，王晓东. 基于语义的关键词提取算法[J]. 计算机科学，2008，35(6)：148-151

[10] 索红光，刘玉树. 一种基于词汇链的关键词抽取方法[J]. 中文信息学报，2006，20(6)：25-30

[11] 胡燕，吴虎子，钟珞. 中文文本分类中基于词性的特征提取方法研究[J]. 武汉理工大学学报，2007，4

[12] 赵军，黄昌宁. 汉语基本名词短语结构分析模型[J]. 计算机学报，1999，22(2)：141-146

[13] 赵蕾蕾. 基于词和基本短语模式的特征提取方法[D]. 保定：河北大学，2009

[14] 王军. 词表的自动丰富——从元数据中提取关键词及其定位[J]. 中文信息学报，2005，19(6)：36-43

[15] Hulth A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge[C]// Proceedings of EMNLP. 2003：216-223

[16] Peter D. Turney，Learning Algorithms for Keyphrase Extraction [J]. Information Retrieval，2000，2(4)：303-336

[17] Frank E，Paynter G W，Witten I H，et al. Domain-specific Keyphrase Extraction[C]// Proceedings of IJCAI. 1999：668-673

[18] 李素建，王厚峰，俞士汶，等. 关键词自动标引的最大熵模型应用研究[J]. 计算机学报，2004，27(9)：1192-1197

[19] Zhang K，Xu H，Tang J，et al. Keyword Extraction Using Support Vector Machine[C]// Proc. of the Seventh International Conference on Web-Age Information Management (WAIM 2006). 2006：85-96

[20] Zhang Cheng-zhi，Wang Hui-lin，Liu Yao，et al. Automatic Keyword Extraction from Documents Using Conditional Random Fields[J]. Journal of Computational Information Systems，2008，4(3)：1169-1180

[21] 钱爱兵，江岚. 基于改进 TFIDF 的中文网页关键词抽取一以新闻网页为例[J]. 情报理论与实践，2008，6

[22] 郑家恒，卢娇丽. 关键词抽取方法的研究[J]. 计算机工程，2005，31(18)

[23] 都云程，周伟，韩艳铧，等. 基于字同现频率的关键词自动抽取[J]. 北京信息科技大学学报，2011，26(6)

[24] 肖根胜. 改进 TFIDF 和谱分割的关键词自动抽取方法研究[D]. 武汉：华中师范大学，2012

[25] 赵鹏，蔡庆生，王清毅，等. 一种基于复杂网络特征的中文文档关键词抽取算法[J]. 模式识别与人工智能，2007，20(6)

[26] 汪小帆，李翔，陈关荣. 复杂网络理论及其应用[M]. 北京：清华大学出版社，2006