心智与计算，Vol.5，No.2(2011)，48-54
文章编号：MC-2011-07
收稿日期：2011-08-15
出版日期：2011-09-30

一种改进的 KEA关键词抽取算法研究

陈　平，周昌乐，练睿婷
（厦门大学人工智能研究所，福建 厦门 361005）
chengfeng200641@163.com

摘要：本文在关键词抽取工具 KEA（Keyphrase Extraction Algorithm）的基础上，对候选关键词的选取方法及其特征属性抽取进行改进。考虑到 KEA 中使用的贝叶斯分类器对特征间的独立性假设引起的一些问题，本文采用了与 KEA 不同的机器学习方法—人工神经网络来训练模型。我们将改进后的模型应用于中文关键词抽取。实验结果表明，改进后的关键词抽取模型对于中文关键词的抽取效果要优于 KEA。

关键词：关键词抽取；KEA；机器学习
中图分类号：TP391　　文献标识码：A

An Improved Approach to Keyword Extraction Using KEA

CHEN Ping,ZHOU Chang-le,LIAN Rui-ting
(Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, China)
chengfeng200641@163.com

Abstract: The candidate keyword extraction method and the features for the classification have been improved on the Keyphrase Extraction Algorithm tool (KEA) in this paper. With respect to the independence assumption of different features, which are often inaccurate, on Bayes classifier, another machine learning algorithm –Artificial Neural Networks has been replaced in KEA. The improved approach has been applied on Chinese Keyphrase Extraction. The experimental results show that the improved approach works better than the original KEA for the Chinese keywords case.

Key words: keyword extraction;KEA;machine learning

1　引　言

随着因特网的迅速发展，我们碰到的电子文档越来越多，面对海量的网络资源，人们可能会迷失方向。如果文档提供了总结信息，我们就可以通过这些信息了解到文档的主要内容。一些文档会有作者列出的关键词，这些关键词就是非常有用的总结信息。它们是文档的浓缩，是对文档内容简洁精确的描述。它们还有很多更进一步的应用，比如文本分类，文本聚类，文本检索等等。

文档关键词一般都是作者或者专业的标注者手工标注的，但是并不是所有文档都会有已经标注好的

关键词，人工标注不仅费时费力，而且主观性强，抽取不当往往对下一步的应用造成消极影响，因此关键词的自动抽取具有一定的研究价值[1]。

2 相关研究

关键词抽取技术的研究已经相当广泛，具体可分为建立词关系树、简单统计、机器学习等几种[2]。

Chien L F 提出了基于 PAT 树的关键词抽取算法[3]，主要思想是采用 PAT 树结构，同时利用词之间的互信息来抽取中文关键词。从相关文章[4]的实验结果可以看出：该方法抽取关键词的效果较佳，但是构建 PAT 树的时间和空间成本太大，抽取效率相对较低。

简单统计方法主要是进行 N-Gram[5]、词频[6]、TFIDF[7]等统计信息获得关键词，这种方法简单易行，通用性强，但由于只是用几个简单的统计信息来判断是否为关键词的这一标准，使得抽取关键词准确率并不高。

机器学习方法主要是通过训练数据进行训练获得统计参数，进行关键词抽取，如 NB[8]、最大熵模型[9]、SVM[10]、GENEx[11]、KEA[1]等，这种方法不受句型限制，可以提取出未登录词，但会出现数据稀疏，过拟合学习的问题。

3 KEA 简介

KEA 是由 Eiber Frank 等人提出的用于实现关键词抽取的算法，该算法运用的是机器学习中的朴素贝叶斯分类器从已经标注出关键词的文档中学习出模型，然后应用训练好的模型从新文档中抽取出关键词[12]。

为了得到训练模型，首先需要一批已手工标注关键词的文档作为训练集，KEA 对每篇文档进行处理识别候选关键词，此过程主要用到标点去除、短语识别、停用词过滤、词干提取等技术，接着将所有文档的候选关键词作为候选关键词集合。

KEA 主要用到以下两个特征：

(1) TFIDF：TFIDF 是用来评估候选关键词对于语料库中的它所在文档的重要程度。词的重要性随着它在文档中出现的次数成正比增加，但同时随着它在语料库中出现的频率成反比下降。

(2) 词的首次出现位置：这个特征表示词在一篇文档中第一次出现的位置，因为如果一个词出现在文档的标题，摘要或者前面的介绍中，那么它是关键词的可能性就会比较大。具体计算是用这个词在文档中第一次出现时前面单词个数比上文档包含词的总个数。

对每一个候选关键词计算特征，并对计算得到的特征进行离散数值化处理形成特征向量。如果候选关键词在训练集中被标记为关键词，则此候选关键词将被标记为候选关键词集合中的正例，如果在训练集中被标记为非关键词，则此候选关键词将被标记为候选关键词集合中的反例。利用分类模型的思想，选取所有的候选关键词样本作为关键词模型训练样本集合。

用样本训练贝叶斯分类器得到关键词抽取模型，将此模型应用于新文档关键词抽取。当对新文档抽取关键词时，KEA 首先识别新文档的候选关键词，然后计算候选关键词特征，根据这些特征计算每个单

词的权值并排序，最后输出前 N 个单词作为关键词，其中 N 是你所设定的抽取关键词个数。KEA 算法的过程如图 1 所示：

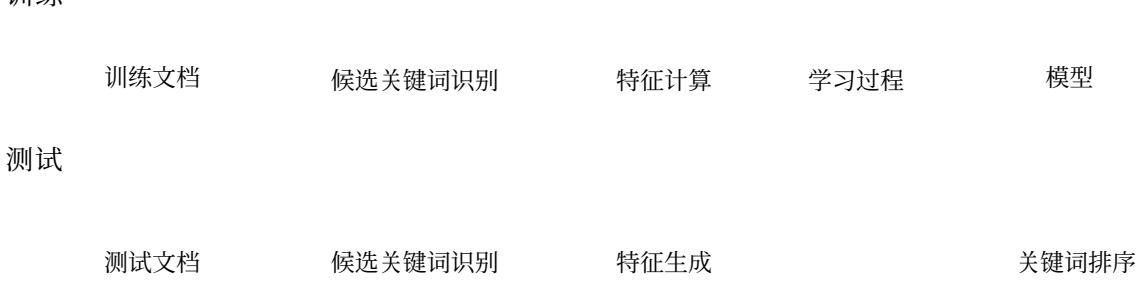

图 1　KEA 训练和抽取过程

Fig.1　Training and extraction process of KEA

KEA 主要是针对英文文档进行关键词抽取，它并没有跨语言的通用性。由于中英文的差异性，例如中文没有英文单词间的分隔符，KEA 对候选关键词识别方法并不适用于中文。KEA 只采用了采用词的频率，以及词在文档中首次出现的位置等全局的上下文信息作为机器学习算法的特征属性，这显然是不够的，缺少针对中文的特征组合。KEA 用的是贝叶斯分类器进行训练，贝叶斯分类器是一种假设独立的简单算法，在某些场合独立性假设可能非常合理，但是当特征间有着复杂的相关性时，比如下文我们将介绍的词的出现范围这个特征就与词的首次出现位置有着很大的关联性，这种分类器就会有明显的缺点，从而导致输出结果不够理想。

4　改进的 KEA 关键词抽取算法描述

本文在 KEA 基础上对候选关键词的识别方法，词的特征属性进行了改进，加入了中文分词标注等自然语言处理技术，用 BP 神经网络代替贝叶斯分类器来训练模型。下面将按照原 KEA 的流程详细介绍改进后的关键词抽取算法。

4.1 训练语料库的建立

由于会用到 TFIDF 这个具有语言依赖性的特征属性，所以我们需要用中文文档来进行训练和计算一个词语的 TFIDF 值。如果训练文档与包含这个词语的文档的语言是不同的，那么我们将无法计算这个词语的 TFIDF 值。我们的训练语料库是由理工、农业、医疗卫生、文史哲、政治军事与法律、教育与社会科学、电子技术与信息科学、经济与管理等领域各 40 篇中文论文组成，这样做的目的是保证改进后的算法具有领域独立性。这些论文都有由作者标注的关键词，但是这些关键词往往具有较强的主观性，所以我们在原作者标注的关键词基础上再手工标注添加了一些能反映论文主题的关键词，把这些关键词放到与论文同名以 key 为扩展名的文本文件中，这样我们的语料库就建好了。

4.2 候选关键词的识别

KEA 的候选关键词识别方法主要是根据英文语言特征而设计的，这显然不适用于中文，需要设计中文的候选关键词识别方法。由于中文并不像英文那样有天然的分隔符，因此首先需要对中文文档进行分词。我们用的是 fencibox[13]这个开源 JAVA 工具包对文档进行分词，这个工具包可以自定义词库，我们加入了一个包含 40 万词的词库，使得分词的准确率有很大提升。最后还要对分词后的文档进行过滤。具体的步骤如下：

(1) 对文档进行分词，采用的是 fencibox 中的全切分算法，分词后的文档会出现类型于英文的分隔符。

(2) 删除文档中的标点符号与数字。

(3) 利用中文停用词表过滤停用词。

(4) 由于关键词都是两字以上，删除文档分词后出现的单个字。

(5) 去掉只在文档中出现一次的词。

(6) 将文档中剩余的词作为候选关键词。

4.3 特征计算

基于机器学习的方法都需要开发一些属性或特征将样本抽象成特征向量，在这个阶段我们将对每个候选关键词计算特征。KEA 中只用到了 TFIDF 和词的首次出现位置两个特征，为了提高关键词抽取效果，我们在 KEA 基础上添加了适用于中文关键词的四个特征，因此每一个词总共要计算六个特征。下面是我们添加的特征：

(1) 词的最后出现位置：与首次出现位置类似，一个词在文档最后结尾中出现是关键词的可能性也会比较大，因为文档最后往往会出现一些总结性的词语，这些词语可以较好的概括文章的内容。这个特征是用词最后出现时前面词语总个数与文档总词语个数的比来计算的。

(2) 词出现范围特征：这个特征是词用来表示某个词在一篇文档中出现的跨度，一个词在文档中出现的范围可以反映出该词的使用频率，分布情况等。我们用词的最后出现位置值与词的首次出现位置值的差来计算该特征。

(3) 词的结尾部分是否名词：因为绝大部分关键词的结尾部分是名词（包括只含一个词的关键词），所以我们先用最小匹配分词法对候选关键词分词标注，如果最后一部分是名词，则此特征值设为 1，否则值为 0。

(4) 词的开头部分是否名词或形容词：我们发现关键词的开头部分大多数是名词或形容词，所以与上一步类似，我们也是先用最小匹配法对候选关键词分词标注，如果开头一个词是名词或形容词，此特征值为 1，否则值为 0。

4.4 训练：建立模型

我们用上一步生成的特征向量来训练模型，在向量集相同的情况下，用不同的机器学习方法训练产生的模型对中文关键词的抽取效果是不同的。KEA 使用的是贝叶斯分类器，此分类器虽然简单，但其对特征间的独立性假设使得我们的输出并不是很理想。我们研究发现用 BP（backpropagation）算法的多层感知器（multi-layer perceptron）可以很好的解决此问题，所以本文采用了此神经网络算法来训练模型，这个算法可以在数据挖掘工具包 weka 中找到。

4.5 抽取新文档的关键词

在这个阶段，我们首先按照 4.2 节的方法对新文档识别候选关键词，计算特征值（KEA 的 TFIDF、词的首次出现位置加上 4.3 节添加的 4 个特征），然后调用训练阶段生成的模型。这个模型首先计算每个候选单词是关键词的概率，再进行一些后续处理（按照概率对单词进行排序）最终抽取出关键词。

5 实验结果与分析

本文采用的是文献[14]的方法来对关键词抽取效果进行评价，国内的门户网站网易（www.163.com）提供的新闻网页一般含有责任编辑手动提炼的“核心提示”，这些核心提示就是非常有用的总结信息，若将核心提示作为标准即可对关键词抽取效果进行评价。我们随机从网易上按政治、财经、体育、军事、娱乐、社会等类别各搜集 20 篇，共 120 篇带有核心提示的新闻作为测试关键词抽取性能的数据集。为了保证实验的客观性，我们事先删除了测试集中每篇新闻文档的核心提示，只留下新闻正文。召回率 R 和准确率 P 常被用作评价抽取效果的指标，本实验采用的标题召回率 R 和核心提示准确率 P 分别定义如下：

在训练语料库和测试语料库完全相同的情况下，表 1 是 KEA 与本文算法对一则新闻抽取五个关键词时的结果比较。从表 1 可以看出，KEA 候选关键词识别方法明显不适合中文，它抽取出来的关键词只是一些句子片段，不符合我们对关键词的定义。本文在关键词识别阶段加入了中文分词与过滤技术，抽取结果是一些完整的词语或短语，与 KEA 相比有很大的改进。从表 1 可以观察到 KEA 抽取出来的句子片段都分布在新闻的前一部分，而忽略了后面的内容，这说明了 KEA 中第一次出现位置这个特征占了主导地位，TFIDF 这个特征没有发挥作用。相比而言，本文的特征组合全面的考虑到了一个词语或短语成为关键词所需具备的特征，抽取出来的关键词更加合理准确。

表 2 是 KEA 与本文改进算法抽取关键词的准确率与召回率统计。由表 2 可以看出，KEA 对中文文档抽取关键词的效果是很差的，它只是针对英文的语言特征而设计的，没有语言通用性。相比而言，本文改进的算法对中文关键词抽取效果提升十分明显，在抽取关键词个数相同的情况下，本文算法的召回率和准确率都远高于 KEA，说明改进后的算法对中文的适用性大大提高。但实验数据反映出准确率会随着关键词输出个数的增加而下降，这可能是我们测试新闻文档的篇幅短，关键词语个数本身就比较少的缘故。

表 1 KEA 与本文算法抽取结果比较

Tab.1 Results of KEA and extraction algorithm discussed in this paper

新闻标题	中石油集团未回应能源局副局长刘琦接掌传闻
新闻原文	针对国家能源局副局长刘琦将接任蒋洁敏中石油集团掌门人之位的传闻，昨天，中石油集团方面没有对传闻进行回答，只表示蒋洁敏仍为集团总经理。 值得玩味的是，中石化集团总经理苏树林调任福建省副书记前，针对他调任的传闻，中石化集团的回应与中石油类似。 蒋洁敏 1955 年出生，今年距央企负责人退休年龄仍有数年距离。他与苏树林类似，之前有地方政府任职经历，2000 年-2003 年曾任青海省副省长，后回归中石油，迄今担任中石油集团党组书记、总经理以及中石油股份董事长职务。消息人士称，蒋洁敏与苏树林一样，是懂企业运作、懂经济的人才。
新闻核心提示	针对国家能源局副局长刘琦将接任蒋洁敏中石油集团掌门人之位的传闻，昨天，中石油集团方面没有对传闻进行回答，只表示蒋洁敏仍为集团总经理。
KEA 抽取结果	针对国家能源局副局长刘琦将接任蒋洁敏中石油集团掌门人之位的传闻 昨天 中石油集团方面没有对传闻进行回答 只表示蒋洁敏仍为集团总经理 值得玩味的是
本文算法结果	中石油 总经理 树林 集团 副局长

表 2 抽取效果比较

Tab.2 The comparison of extraction effectiveness

抽取关键词个数	指标	3	5	7	10
KEA	R	0.0119	0.0143	0.0143	0.0193
	P	0.0333	0.0200	0.0155	0.0133
本文算法	R	0.1381	0.1746	0.2121	0.2517
	P	0.5444	0.4650	0.4369	0.3733

6 结束语

下一步我们将对分词算法与文档过滤技术进行改进，使候选关键词的识别能更加准确；我们将研究设计更适用于中文的特征属性组合，进一步提升中文关键词抽取的效果。

参考文献

[1] Steve Jones,Gordon W Paynter. Human evaluation of KEA,an automatic keyphrasing system[C]//Proceeding of the 1st ACM/IEEE-CS joint conference on Digital libraries.Roanoke,Virginia,United States,2001:148-156.

[2] 邓箴,包宏.改进的关键词抽取方法研究[J].计算机工程与设计,2009,30(20):4677-4680.

[3] Chien L F.PAT-tree-based keyword extraction for Chinese information retrieval[C]//Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information etrieval.Philadelphia,Pennsylvania,United States, 1997:50-58.

[4] Yang Wenfeng, Li Xing.Chinese keyword extraction based on max-duplicated Strings of the documents[C]//Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland, 2002:439-440.

[5] Cohen J D.Highlights: language-and domain-independent automatic indexing terms for abstracting[J]. Journal of the American Society for Information Science,1995,46(3):162-174.

[6] Luhn H P.A statistical approach to mechanized encoding and searching of literary information[J].IBM Journal of Research and Development,1957,1(4):309-317.

[7] Salton G,Yang C S,Yu C T.A theory of term importance in automatic text analysis[J].Journal of the American Society for Information Science,1975,26(1):33-44.

[8] Frank E,Paynter G W,Witten I H. Domain-specific keyphrase extraction[C]//Proceedings of the 16th International Joint Conference on Artificial intelligence. Madison,Wisconsin,United States,1998:517-523.

[9] 李素建，王厚峰，俞士汶.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.

[10]Zhang K,Xu H,Tang J.Keyword extraction using support vector machine[C]//Proceedings of the Seventh International Conference on Web-Age Information Management.Hong Kong,China,2006:85-96.

[11]Turney P D.Learning to extract keyphrases from text[R/OL].Ottawa:National Research Council of Canada,(1999-02-17)[2011-04-01]. http://arxiv.org/ftp/cs/papers/0212/0212013.pdf.

[12] Witten I H, Paynter G.W, Frank E,et al. Kea:practical automatic keyphrase extraction[C]//Proceedings of the fourth ACM conference on Digital libraries.Berkeley,California,United States,1999:254-255.

[13]何振宇.分词盒子项目主页[EB/OL].(2011-01-15)[2011-04-08].http://www.ithezi.com/fenci.

[14]李星华.中英文新闻网页关键词抽取技术研究[D].合肥:合肥工业大学,2009.

作者简介：

陈平，男，1987 年 10 月 1 日生，2009 年 6 月毕业于中南大学信息科学与工程学院获工学学士学位，现为厦门大学智能科学与技术系硕士研究生，主要研究方向为自然语言处理。