基于新的关键词提取方法的快速文本分类系统*

罗 杰，陈 力，夏德麟，王 凯
（武汉大学 电子信息学院，湖北 武汉 430079）

摘 要：关键词的提取是进行计算机自动文本分类和其他文本数据挖掘应用的关键。系统从语言的词性角度考虑，对传统的最大匹配分词法进行了改进，提出一种基于动词、虚词和停用词三个较小词库的快速分词方法（FS），并利用 TFIDF 算法来筛选出关键词以完成将 Web 文档进行快速有效分类的目的。实验表明，该方法在不影响分类准确率的情况下，分类的速度明显提高。

关键词：计算机应用；中文信息处理；关键词提取；Web 文档分类
中图法分类号：TP391 **文献标识码**：A **文章编号**：1001-3695(2006)04-0032-03

Research on Fast Text Classifier Based on New Keywords Extraction Method

LUO Jie, CHEN Li, XIA De-lin, WANG Kai
(*School of Eletronic Information, Wuhan University, Wuhan Hubei* 430079, *China*)

Abstract: Keyword extraction is the sticking point for Automatic Classification and Text Data Mining Application. Taking traits of nature language into consideration, this paper provides a new way called Fast Segmentation (FS) which is based on verb, virtual words and stop words to improve traditional segmentation technique. Then, we filter result of FS by TFIDF[3] Algorithm so that we can classify Web text fast and efficiently. The experiment has indicated that without reducing the correct rate of classification, the speed of processing has improved distinctly.

Key words: Computer Application; Nature Language Processing; Keyword Extraction; Web Text Classification

随着 Web 技术的发展，WWW 已经发展成为拥有众多信息资源，发展日新月异，站点遍布全球的巨大信息服务网络。如何有效地将网络资源分门别类，提取有用信息再也不是单凭人工所能做到。利用计算机对网络资源的自动分类技术应运而生。这种对 Web 页面的分类往往首先将 Web 页面的结构化信息如 Head，Title，Body 等所含有的内容以文档的形式提取出来并将其转换成对文档的分类。对于一篇文档，它所拥有的关键词能够很好地反映文档的特征。将每一篇文档的关键词[1]挖掘出来，会给计算机快速高效检索，有效地组织资源等方面带来极大的帮助。关键词[2]的自动提取离不开切词技术，本文针对切词技术提出一些新的观点，即利用较低的数据库资源对文本中的关键词进行提取，接着使用 TFIDF 算法进行进一步的筛选找出分类用的关键词，进而在保证准确性的前提下提高分类的速度。该系统在一些个人用户端或者运算速度较低的个人掌上电脑上有很好的应用前景。

1 文档中文信息的切词

中文文本与英文文本不同，英文文章是以词为单位的，词与词间有固有的分隔符（空格）将其分开，所以对英文的文本分析处理时不存在分词上的需要；而中文是以字为单位，除了标点符号以外，句子中各个词语之间没有固有的分隔符（空格），因此对中文的文档进行词频统计和针对关键词的提取分类等处理之前必须进行切词，因此切词是进行中文信息处理的基础也是关键。

现有的分词算法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法及基于理解的分词方法和基于统计的分词方法。其中基于字符串匹配的分词方法是一种常用的分词方法，为许多中文搜索引擎和处理软件所采用，这种方法又叫作机械分词方法，它是按照一定的策略将待分析的汉字串与机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功，即将词识别出来。如常最大匹配法，亦称 MM 法。其基本思想是[3]：假设自动分词词典（或词库）中的最长词条是 i 字，则取被处理材料当前字符串序列中的前 i 字作为匹配字段，查找词典，若词典中存在这样的一个 i 字词，则匹配成功，匹配字段被作为一个词切分出来；如果在词典中找不到这样一个 i 字词，则匹配失败，匹配字段去掉最后一个字，剩下的字段重新进行匹配，这样循环进行下去，直到匹配成功，也就是完成一轮匹配，切分出一个词为止。

大部分这种分词的方法能够准确地分出中文的词，而该类方法的缺陷就在于它们过度依赖于使用“大容量”的机器词典，通过繁杂的统计分析和规则来达到准确分词的目的，这样势必会浪费大量的资源与时间。然而，对于某些特殊的应用（如文档的主题聚类），完全没有必要将所有词都准确地切分出来。下面我们就提出一种简便的中文文档关键词提取系统，以用于文档的主题分类。

1.1 新型机器词典的建立

名词和形容词的数量众多，在中文常用词所占的比例超过 70%，首先由于其数量众多，而且科学技术的发展使得新事物不断涌现，并对应于新词的不断出现，它们主要以描述新事物

收稿日期：2005-04-06；**修返日期**：2005-05-26
基金项目：国家自然科学基金资助项目（90204008）

内涵及外延的名词性新词和对应于新事物特征的形容词性新词出现[4]。同样无法一一收录。因此,我们可以借助相对变化较稳定的且"小容量"的动词库、虚词库以及停用词库并结合一定的语法规则来进行更准确切分。我们定义了用于分类的主题词,建立主题词库,实验中使用计算机、体育、经济、艺术几个类的主题词库。

该种分类方法能够帮我们找到用于分类用的关键词,同时达到对文本进行分类的目的,关键词提取的过程如图1所示。

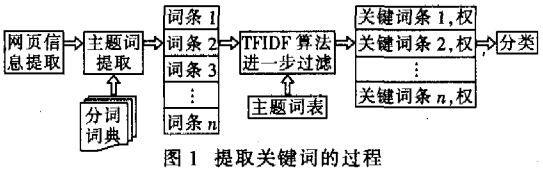

图1　提取关键词的过程

1.2　动词词典、虚词词典和停用词词典的建立依据

(1)动词词典。绝大多数句子通过主语(名词、代词为主)和谓语(动词为主)来反映该句话要表示的主要信息,即"某物表现了某种行为";同时,动词可很好地翻译句子的结构,如有些动词可接双宾语,像"我给你一份文件"中的"给"。对于动词中有一些专有动词也属于分类主题词,我们给出特殊的标记。例如,"编程"属于计算机类的主题词,标记的词将作为关键词参与文本主题分类。还有一些像"计算机"中的"计算",我们在词典里用特殊的索引标记,进行查找。

(2)虚词词典。在句子中反映句子结构,因而对规则的建立很有帮助,如"的",表明它的前面是修饰部分,后面则是中心语。虚词词典包括连词、介词、副词、助词。

(3)停用词词典。有利于消除歧义,同时这些词不具备太多信息。在切分中可以从句字中切除但不影响句子主干信息的提取,如数字等。停用词包括:数词、量词、代词、方位词、拟声词、叹词等;没有实际意义的动词,如"可能";一些太过于常用的名词,如"操作"等。对于一些在科技文章中经常出现的英文单词与字母也属于停用词,但是这些词在分类时也将作为关键词参与分类。

1.3　基于"小容量"词库的切分技术

首先,利用这三个"小容量"的词库分别对文档中的句子切分,并进行比对。规定规则符号为:①动词v;②虚词x;③停用词t;④未知词w(即利用某词库切分后不在该词库的部分)。

例如下面的句子,以MM切分为例:模式识别诞生于20世纪20年代,模式识别在60年代初迅速发展成一门学科。它所研究的理论和方法在很多学科和技术领域中得到了广泛的重视,推动了人工智能系统的发展,扩大了计算机应用的可能性。

(1)动词切分。模式识别/诞生/v于20t世纪20t年代,模式识别在x60年代初迅速/发展/v/成/v一门学科。它所/研究/v|的理论和方法在很多学科和技术领域中/得到/v了广泛的|/重视/v,/推动/v了人工智能系统的/发展/v,/扩大/v了计算机/应用/v|的/可能/t性(未标记的部分均属于集合W)。

(2)虚词切分。模式识别诞生/于/x20世纪20年代,模式识别/在/x60年代初迅速发展成一门学科。它所研究/的/x理论/和/x方法/在/x/很多/x学科/和/x技术领域中得到/了/x广泛/的/x重视,推动/了/x人工智能系统/的/x发展,扩大/了/x计算机应用/的/x可能性(未标记的部分均属于集合W)。

(3)停用词切分。模式识别诞生于/20世纪/t/20年代/t,模式识别在/60年代/t初迅速发展成/一门/t学科。它所研究的理论和方法在很多学科和技术领域中得到了广泛的重视,推动了人工智能系统的发展,扩大了计算机应用的可能性(未标记的部分均属于集合W)。

1.4　三种切分的比对及相关规则的建立

由于规则的建立需建立在语言现象和句法的基础上,因此规则的完善需大批语言学者的参与。现提出几条简单的规则,用于在实验中验证新方法的准确性。

设集合V={c|c经动词切分后被匹配为动词};X={c|c经虚词切分后被匹配为虚词};T={c|c经停用词切分后被匹配为停用词};W={c|c∈未知词}。z为字符串。句子Sentence=$z_1,z_2,z_3,\cdots,z_n$。

①If z_1z_2 ∈ one of V,X,T,while z_1(or z_2) ∈ one of V,X,T yet z_1(or z_2) $\cap z_1z_2=\emptyset$

Then making z_1(or z_2) belong to the set it once belong to; making z_2(or z_1) belong to set W.

②When every part has been tagged, if z∈vz, then add z into the set W.

经过规则①和比对之后,我们可得到V, X, T和W四个集合,其中W中的元素为:{模式识别,初迅速,学科,理论,方法,学科,技术领域中,人工智能系统,发展,计算机,性}现在把集合W作为初步的关键词自动提取词库。当然,通过筛选和添加更全面的停用词库以及完善规则[7]后,可以对W中的元素进一步提炼。

2　进一步过滤关键词:TFIDF算法及其改进

在前一节中,我们讲到利用分词技术以及已有的动词、虚词,以及停用词库把网页中的信息粗步提取出来,这些信息以词的形式被存储下来。我们可以计算所有被分出词在网页中的权重w,并由统计计算和人工调整得到,当$w\geq\lambda$时,将w对应的词x作为网页的特征项或是网页相关信息的主题词(因为经初步分割后主要以名词为主)。那么如何计算权重w呢?我们可以基于这样的事实:词x在网页P中出现的频率越大,同时在所有网页中出现的频率越小,则x对P的贡献程度越大,即TFIDF[6]:

$$\mathrm{TFIDF}(x_i,P)=\frac{TF(x_i,P)\cdot\log\left(\frac{n}{DF(x_i)}\right)}{\sqrt{\sum_{k=1}^{k=m}\left[TF(x_k,P)\log\left(\frac{n}{DF(x_k)}\right)\right]^2}}$$

其中,TFIDF(x,P)为词x在网页P中的权重;$TF(x,P)$为词x在网页P中出现的次数;n为网页总数;m为网页P中提取的词条x总数;$DF(x)$为词x在所有网页中出现的次数这样得到每一个词的权重,通过预先设定的权重λ,如果得到的TFIDF值大于预先设定的权重阈值,则说明该词为可能的关键词。

如果一系列的文档均属于同一类,则对于高频关键词无法利用该算法提取。例如属计算机类的一系列文档,由于"计算机"可能在所有文档中均出现。若该假设成立,则$\log(n/DF(x))=0$。因此,如果有一个已有主题词库的话,则在原词库的基础上利用三个词库并结合TFIDF算法可以很好地挖掘新的关键词。如图1所示,以计算机类文档为例,利用三个分词词典切分后,我们再通过TFIDF算法过滤掉一些对文本分

类没有影响的词条，如“方法”。而“计算机”由于存在于预先建立的主题词表内，不会被排除，并且赋予它们一个新的权值。规则描述如下：

IF TFIDF(x_i,P) > λ, THEN add x_i in keyword category $X = \{x_1, x_2, x_3, \cdots, x_i, \cdots, x_n\}$, ELSE IF x_i belongs to the keyword lexicon, THEN add x_i in keyword category $X = \{x_1, x_2, x_3, \cdots, x_i, \cdots, x_n\}$, using its original weight in the keyword category.

3 kNN(k-Nearest-Neighbor)分类算法

现在，我们就要利用每篇网页中筛选出来的关键词及其对应的权重来进行分类，这里介绍的是 kNN 分类算法[8,9]。

kNN 方法是一种基于文本特征向量空间模型表示的分类方法，它在文本分类上有较好的应用结果。它的实质是以特征属性权值作为特征空间的坐标系测度，先计算测试文本与训练文本之间的距离，然后依据测试文本与训练文本之间的远近来确定类别。结合 2 节的输出（网页的关键词及权重），具体算法步骤如下：

(1) 分类的文档（以计算机类为实验）为 m 个子类，C 为类别集合：$C = \{c_1, c_2, \cdots, c_m\}$。

(2) D 为文档集合 $X = \{x_1, x_2, \cdots, x_n\}$；将待分类的网页信息转换成文档的形式，用 X 表示未分类的文档集合：$Y = \{y_1, y_2, \cdots, y_n\}$ 文档 d_i 的 M 维特征向量为 $W_i = \{w_{i1}, w_{i2}, \cdots, w_{in}\}$，$w_{in}$ 为词 x_n 在 d_i 中的权重，若 d_i 不含词 x_n，则 w_{in} 为零。

(3) 则在训练文集中选出 K 个与新文本最相似的文本，计算公式为 $sim(d_i, d_j) = \dfrac{\sum_{k=1}^{M} W_{ik} \cdot W_{jk}}{\sqrt{(\sum_{k=1}^{M} W_{ik}^2) \cdot (\sum_{k=1}^{M} W_{jk}^2)}}$

(4) 若在(3)中 $W_{jk} = \{w_{jk1}, w_{jk2}, \cdots, w_{jkn}\}$ 为 c_j 内某篇文档的 M 维特征向量，则在新文本的 k 个近邻中，依次计算每类的权重，公式为 $p(y_i, c_j) = \sum_{d_m \in kNN} sim(y_i, d_m)\, y(d_m, c_j)$。其中 $y(d_m, c_j)$ 为类别属性函数，即如果 d_m 属于 c_j，则为 1，否则为 0。

(5) 若 $p(y_i, c_t) = \max p(y_i, c_j)$，则未分类文本 $y_i \in c_t$。

最终完成对文本的分类工作。

4 系统实现与结果分析

在本次实验中我们在中文常用词的基础上建立了拥有 5 134个词条的动词词库，拥有1 784个词条的虚词词库和小规模的停用词库和主题词表。实验中，我们应用由复旦大学提供的语料库（www. nlp. org. cn/docs/doclist. php? cat_id = 16& type = 15），从中提取计算机、体育、经济、艺术四个子类的各 800 篇文档训练和测试。

(1) 关键词提取效果测试。我们从计算机类中提取出 20 篇文档实验测试关键词的提取结果，将通过 TFIDF 方法过滤得到的关键词由人工筛选判断并与人工从文章中提取的关键词进行比较（图 2）。经比较后可知，在筛选出的关键词中对分类有用的关键词准确率为 83.7%，关键词筛选的遗漏率为 7.1%。

其中造成关键词提取不够完全准确的原因主要有三点：①仅仅使用动词和虚词库对某些词语造成过度切分，如“调查报告”会被拆分为“调查”和“报告”两词；②有一些词无法拆分出来造成句子过程或者不是真正的词。③有歧异存在（这在所有中文的切分中都存在也无法避免）。

(2) 分类结果。将切分和筛选后得到的关键词进行分类（表 1）。

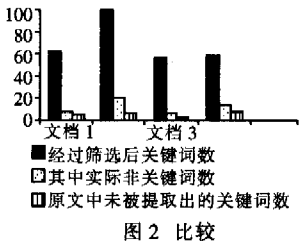

图 2 比较

表 1 对关键词分类

类别	计算机	体育	经济	艺术
样本数	800	800	800	800
原始样本数	500	500	500	500
测试验本数	300	300	300	300
准确率%	91.3	93	83.7	88.3

(3) 分析。对于系统的执行速度在实验中可发现即使未对程序进行优化的情况下也较普通的最大切分法有明显的提高。FS 方法虽然无法做到高准确度的提取关键词，但是它在不影响分类准确度的情况下能够胜任到中文文档的快速分类。

实验中的 kNN 分类算法在一定程度上较其他分类方法在分类过程中花费较多的时间，如果予以一定改进采用快速的分类算法[10]将进一步有效地提高系统的执行速度。

5 展望

我们的实验刚刚在起步阶段，系统设计还有许多不成熟，在以后的研究工作中我们将在词典中加入更多的语法规则，特别是虚词的规则以及动词的特殊情况，并且扩大停用词库，来进一步提高分词和关键词提取的准确率，系统在真正投入应用时应在准确率和速度上有所权衡。

参考文献：

[1] P Turney. Learning to Extract Keyphrases from Text[EB/OL]. National Research Council of Canada(1999), http://arxiv.org/ftp/cs/papers/0212/0212013.pdf.

[2] 沈小建，许景红. 清华同方主题词、分类号智能检索系统（医学专业）[J/OL]. 中国期刊网 CNKI 数字图书馆, 2005.

[3] http://linux.tcpip.com.cn/article/index.php? func = detail&par = 14&parentid = 159&start = 16&s = 0, 2005-03[EB/OL].

[4] Shiwen Yu, Xuefeng Zhu, Yunyun Zhang. The Specification of the Synthetic Knowledge-based of Contemporary Chinese[J]. Journal of Chinese Information Processing, 1996,10:1-22.

[5] 张义忠，赵明生. 基于内容的中文网页自动分类研究[J]. 信息与控制，2001,30(5).

[6] G Salton. Developments in Automatic Text Retrieval[J]. Science, 1991,253:974-979.

[7] 唐振民. 一种用于自动标引系统的主题词自动切分方法[J]. 南京理工大学学报，1995,19(5).

[8] 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究，2001,18(9):23-26.

[9] 边肇祺，等. 模式识别[M]. 北京：清华大学出版社, 1999.136-159.

[10] Giorgio Giacinto. Design of Effective Multiple Classifier Systems by Clustering of Classifiers[C]. International Conference on Pattern Recognition (ICPR'00).

作者简介：

罗杰(1982-)，男，福建福州人，本科生，研究方向为人工智能与数据挖掘；陈力(1982-)，男，湖北十堰人，本科生，研究方向为计算机网络与数据挖掘；夏德麟(1940-)，男，湖北武汉人，教授，研究方向为数据挖掘；王凯(1980-)，男，湖北武汉人，硕士生，研究方向为计算机网络等。