文章编号: 1003-0077(2009)01-0063-08

基于分离模型的中文关键词提取算法研究

罗准辰,王 挺

(国防科学技术大学 计算机学院,湖南 长沙 410073)

摘 要: 关键词提取在自动文摘、信息检索、文本分类、文本聚类等方面具有十分重要的作用。通常所说的关键词实际上有相当一部分是关键的短语和未登录词,而这部分关键词的抽取是十分困难的问题。该文提出将关键词提取分为两个问题进行处理: 关键单词提取和关键词串提取,设计了一种基于分离模型的中文关键词提取算法。该算法并针对关键单词提取和关键词串提取这两个问题设计了不同的特征以提高抽取的准确性。实验表明,相对于传统的关键词提取算法,基于分离模型的中文关键词提取算法效果更好。

关键词: 计算机应用;中文信息处理;关键词提取;关键词串;分离模型;互信息;词串边界参数表

中图分类号: TP391.1 **文献标识码**: A

Research on the Chinese Keyword Extraction Algorithm Based on Separate Models

LUO Zhun-chen, WANG Ting

(School of Computer, National University of Defense Technology, Changsha, Hunan 410073, China)

Abstract: Keyword extraction plays an important role in information retrieval, automatic summarizing, text clustering, and text classification, etc. A significant portion of keywords usually extracted are actually key phrases or the words not recorded yet, which makes the keyword extraction more difficult. This paper argues that the keyword extraction can be treated as two problems: extracting key words and extracting key phrases. A keyword extraction algorithm based on separate models was proposed, with different features developed for the two mentioned problems so as to improve the accuracy of keywords extracted from the Chinese documents. The experiment results show that the proposed algorithm has a better performance compared with the traditional keyword extraction algorithms.

Key words: computer application; Chinese information processing; keyword extraction; keyphrases; separate model; mutual information; word-sequence boundary

1 引言

随着信息时代的发展,信息的表达方式日益多样化,其中文本信息是一种不可替代的方式。随着网络上文本信息的爆炸式增长,手工获取所需的文本信息的难度日益增大,如何提高信息访问的效率成了一个越来越重要的课题。为了对海量文本信息进行有效地组织和处理,研究人员在自动文摘、信息检索、文本分类、文本聚类等方面进行了大量研究,而这些研究都涉及到一个关键的基础性问题,即如何从文本中提取关键词。

关键词高度概括了文本的主要内容,易于使不同的读者判断出文本是否是自己需要的内容。不仅如此,由于关键词十分精练,故可以利用关键词以很小的计算代价进行文本相关性度量,从而高效地进行信息检索、文本聚类和分类等处理。在这方面应用最广泛的还是文本检索。用户在搜索引擎中输入关键词,系统将出现此关键词的所有文本返回给用户。国外对于关键词的研究起步较早,已经建立了

投稿日期: 2008-04-18 **定稿日期**: 2008-09-26

基金项目: 国家自然科学基金资助项目(60403050);新世纪优秀人才支持计划资助项目(NCET-06-0926)

作者简介: 罗准辰(1984—),男,硕士生,研究方向为自然语言处理;王挺(1970—),男,教授,博士生导师,研究方向为自然语言处理。

一些实用或实验系统。Turney 等[1]设计了 GenEx 系统,它将遗传算法和 C4.5 决策树机器学习方法用于关键短语的提取;Witten 等[2]开发了系统 KEA,它采用朴素贝叶斯技术对短语离散的特征值进行训练,获取模型的权值,以完成下一步从文档中提取关键短语的任务。在实际研究和应用中,通常所说的关键词实际上有相当一部分是短语。短语比词更具有概括能力,包含的信息更加丰富,研究关键词短语的提取具有更加重要的意义[3]。Turney 和 Witten 的研究都把文本中连续出现的几个词序列看成候选关键词短语,但并未充分考虑这些词序列是否符合人们习惯认可的短语形式。一种比较常见的研究方法是通过统计 N-gram 词性匹配模式的方法来提取关键词短语;另外一个相关的研究领域是 Chunk 的自动识别,但 Anette helth 指出通过自动识别的方法难以获得符合人们习惯的关键词短语,为此她人工总结了 56 个词性匹配模式,用于英文关键词短语的自动提取[4]。从国内看,由于汉语语言本身的特点,没有显式的词边界,为关键词自动标引任务又增加了一定的难度。目前主要的工作包括:基于 PAT Tree 结构获取新词,并采用互信息等统计方法对文档的关键词进行标引,但获取候选词选用的 PAT Tree,它的建立用计算机实现时需要大量的空间消耗[5];李素建等[6]提出的利用最大熵模型进行关键词自动标引的方法,由于特征选择和特征参数估计时不够准确,造成关键词自动标引应用时不够理想;王军[7]提出了一种用于自动标引的文献主题关键词抽取方法,它限于从已标引的结构化语料库中元数据的标题中抽取关键词;索红光等[8]提出了利用《知网》知识库构建词汇链的方法,但这种方法只适用于收录在《知网》中的关键词。

虽然国内外研究关键词提取的方法很多,但存在的难点依然是“关键”的度量与“词”的选择上。其中对于一些“关键”的度量方法无法应用于短语是研究者普遍遇到的问题。通常所说的关键词实际上有相当一部分是关键的短语和未登录词,而这部分关键词的抽取是十分困难的问题。本文提出将关键词提取分为两个问题进行处理:关键单词提取和关键词串提取,设计了一种基于分离模型的中文关键词提取算法。该算法并针对关键单词提取和关键词串提取这两个问题设计了不同的特征以提高抽取的准确性。

本文第 2 部分介绍了关键词串的定义以及如何通过互信息与词串参数表识别词串;第 3 部分结合关键词串的定义详细介绍了基于分离模型的关键词提取算法以及特征选取的问题;第 4 部分说明了对分离模型进行评估的实验方法;第 5 部分给出了实验结果,并进行了比较和分析;最后对全文进行了总结。

2 关键词串的定义

严格意义上的关键词仅含一个词,而关键词短语至少含两个词,但人们通常习惯把关键词与关键词短语统称为关键词(有时统称为关键短语)。为了对不同意义的关键词加以区别,我们在本文以“关键单词”表示仅含一个词的关键词,仍然以“关键词”表示通常意义上的关键词,即包括关键单词与关键词短语。汉语文本中词无天然的分割符,而关键词提取技术大都先依赖词典分词,结果造成一些未登录词被切分成多个词典中的词。本文把这些未登录词以及短语统称为词串。汉语中的关键词则可分为关键单词与关键词串。未登录词与短语有相同的特点,它们在分词时都被切分成由几个词典中的词组成的词序列。与其他词序列相比,词串在相邻词之间结合更加紧凑。但未登录词与短语又是不同的,短语有一定的语法结构,而未登录词本质上还是一个词。基于未登录词与短语的相同点与不同点,本文采用互信息与构造词串边界参数表的方法识别词串。

2.1 互信息

互信息 MI(Mutual Information)是统计模型中衡量两个随机变量 X 和 Y 之间关联程度的常用参数,它反映了两变量之间结合的紧密程度,互信息越大说明 X 和 Y 之间存在比较紧密的二元搭配关系,互信息越小说明 X 和 Y 之间基本没有结合关系。

直观上可以根据互信息对任意长度的词序列紧密程度进行度量,具体如下:

$$MI(w_1 w_2 w_3 \cdots w_{n-1} w_n) = \mathrm{Min}(MI(w_1 w_2), MI(w_2 w_3) \cdots MI(w_{n-1} w_n)) \quad (1)$$

$$MI(w_{i-1} w_i) = \log \frac{P(w_{i-1} w_i)}{p(w_{i-1}) \times p(w_i)} \quad (2)$$

$$p(w_{i-1} w_i) = \frac{n(w_{i-1} w_i)}{n(w)} \quad (3)$$

$$p(w_{i-1}) = \frac{n(w_{i-1})}{n(w)} \quad (4)$$

$$p(w_i) = \frac{n(w_i)}{n(w)} \quad (5)$$

其中 $MI(w_1 w_2 w_3 \cdots w_{n-1} w_n)$ 表示词序列 $w_1 w_2 w_3 \cdots w_{n-1} w_n$ 的结合的紧密程度, w_i 表示词, $n(w_i)$ 表示 w_i 在文本中出现次数, $n(w)$ 表示文本中的词数。

词串是一种结合紧密的词序列。如果词序列结合得越紧密,则该词序列越有可能是词串。词序列类似于一种链式结构,链的强度由链中最薄弱的环节确定。因此词序列结合的紧密程度由所有相邻两个词之间互信息的最小值决定。

2.2　词串边界参数表

词串是由一些连续出现的词典词组成,而词串的串头词与串尾词都有一些共同的特点。比如,经常以副词、助词形式存在的词典中的词很少作为词串的串头词与串尾词,而有部分词典词却经常作为词串的串头与串尾。因此,我们构造了词串边界参数表,近似评估了所有词典词作为词串串头和串尾的可能性。如果某个词在串头参数表中权值越大,则该词作为词串串头词的可能性越大,权值越小,则该词作为词串串头词的可能性越小。串尾参数表同样如此。

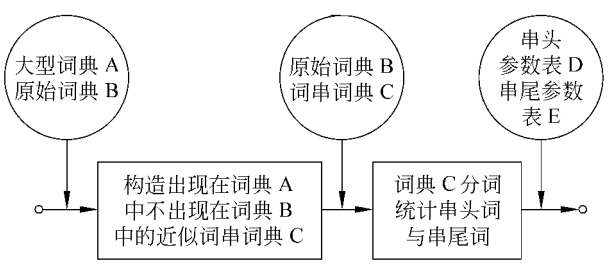


图 1　词串边界参数表构造过程

如图 1 所示,我们把一个拥有 548 387 个词条的词典作为大型词典 A;标准分词器 S 中的词典作为原始词典 B,此词典拥有 108 750 个词条,标准分词器 S 采用最长逆向匹配算法进行分词。词典 A 不仅包含所有词典 B 中的词条,还包含许多人们日常生活经常用到的词串。我们从词典 A 中过滤掉所有出现在词典 B 中的词条,得到近似的词串词典 C。接着利用分词器 S 对词典 C 中所有词条进行分词,统计词典 B 中所有词条作为词典 C 中串头词与串尾词的数目,依次作为词典词的权值,从而生成串头参数表 D 和串尾参数表 E。

3　基于分离模型的关键词提取算法

我们把关键词提取看成一个分类问题,即文本中每个候选关键词是属于关键词还是属于非关键词。利用机器学习的方法,通过输入一批已标注是否为关键词的训练样本,训练一个关键词分类模型,通过此模型对新的候选关键词进行是否为关键词的判断。

3.1　生成候选关键单词与候选关键词串

汉语中的关键词提取必须首先分词。但不是所有的词都适合作为候选关键单词,其中数字、标点符号都应该过滤。而对于候选的关键词串来说,同样并不是每个词串都适合作为候选关键词串。我们选取词数大于 1 小于 5 的词串作为候选关键词串,删除其中存在标点、开头词或结尾词是数字的词串。英文中的关键词提取技术在选择候选关键词时,把开头词或结尾词是停用词的候选关键词过滤[2]。我们以同样的方法对中文中候选关键词的选择问题进行了实验,实验结果表明此方法在过滤掉 45 %左右的非关键词的情况下,关键词的丢失率不到 1.5 %。因此在中文中我们采用此方法选择候选关键单词与候选关键词串。

3.2　分离模型

传统的关键词抽取研究中,关键单词样本与关键词串样本是不加区别的。通过同时对所有标注好的关键单词样本与关键词串样本进行训练形成一个整体模型。然后以此模型来判断其他未标注的候选关键单词与候选关键词串。然而正如我们在第 2 部分介绍的那样,词串类似一种链式结构,其本身具有一定的结构特点,不应简单地把词与词串等同,而应该把它们分开考虑。正是因为传统的研究中把词与词串一同训练,使得许多“关键”特征无法在词与词串上通用,或者忽略了词与词串各自所特有的有效特征。因此我们针对词和词串的不同特性设计相应的特征,并把关键单词样本集合与关键词串样本集合分别进行学习和训练,以获得关键单词模型与关键词串模型。在应用这两个模型抽取文本关键单词和关键词串时,将根据两个不同的模型分别对候选关键单词与候选关键词串进行判断。此分离模型不但可以根据词与词串的不同特点添加不同的“关键”特征,而且在相同的条件下比整体模型效果更好(本文第 5 部分的实验结果证实了这一点)。

3.3　特征选取

由于分离模型是对词与词串分别建立模型,所以在“关键”特征的选取上,两个模型可以选取不同

的特征。在 Witten 等[2]开发的 KEA 系统中,候选关键词的 TF ×IDF 值与首次出现的位置 POS 是判断候选关键词是否为关键词最有效的特征,我们同样选取这两个特征并都应用于词和词串两个模型的建立。但特征 TF ×IDF 有两个缺点:(1)对于需要提取关键词的短文本来说,它们的候选关键词的 TF ×IDF 值相对比长文本小,这是因为同一个候选关键词在短文本中的词频比长文本小;(2)由于 IDF (反转文档频率)是数据集中出现该候选关键词的文档数目的倒数,可能一些无意义的候选关键词由于相对集中出现在少量文档中而使得 IDF 值过大,影响了文本中候选关键词的提取。针对 TF ×IDF 的不足,我们另外选取了两个特征:

® NWT(Number Words of Text):文本中所含的词数,通过该特征可以解决小文本中候选关键词 TF ×IDF 值相对较小的问题;

® TF × IF (Term Frequency × Inverse Frequency):候选关键词在一篇文档中出现的频率与它在整个数据文档集中词频倒数的积,通过该特征克服了 TF ×IDF 的第二个缺点。

如表 1 所示,对于候选关键单词与候选关键词串来说,以上四个特征都能作为判断它们是否为关键单词或关键词串的属性特征,并以此构造分类模型。另外长度 LEN、互信息 MI、串头参数 HB、串尾参数 TB 可以单独作为候选关键词串的属性特征,具体如下:

® LEN(Length):词串所含词数;

® MI(Mutual Information):互信息(2.1 节),通过该特征可以度量候选关键词串中词序列结合的紧密程度,互信息值越大,词序列结合得越紧密且越可能成为词串,成为词串的可能性越大,则该候选关键词串成为关键词串的可能性越大,反之亦然;

® HB 与 TB:串头与串尾参数(第 2 部分),候选关键词串中串头词或串尾词在词串边界参数表中权值越大,则候选关键词串作为词串可能性越大,因此该候选关键词串作为关键词串的可能性也越大;反之亦然。

表 1 特征基本信息

特征编号	特征名称	特征意义	适用模型
(1)	TF ×IDF	词频与反转文档频率的积	关键单词模型、关键词串模型
(2)	POS	首次出现位置	关键单词模型、关键词串模型
(3)	NWT	文本所含词数	关键单词模型、关键词串模型
(4)	TF ×IF	词频与反转频率的积	关键单词模型、关键词串模型
(5)	LEN	词串所含词数	关键词串模型
(6)	MI	互信息	关键词串模型
(7)	HB	串头参数	关键词串模型
(8)	TB	串尾参数	关键词串模型

4 实验方法

前面介绍了关键词串的定义以及分离模型的本质,即把关键词提取分成关键单词提取与关键词串提取两个问题。如何更好地利用分离模型完成关键词提取任务,我们做了一些探索,提出了两种以分离模型为基础的实验方法:分类实验、评分实验。下面对两种方法的具体实现过程分别进行介绍。

4.1 分类实验

选取一批已手工标注关键词的文档作为训练集。同时对每一个文档生成候选关键单词与候选关键词串,并以此作为每一个文档的关键单词候选项集合与关键词串候选项集合。每一个候选项按照表 1 计算特征,形成特征向量。如果候选关键单词或候选关键词串属于手工标注的关键词,则为正例,否则为反例。选取所有的候选关键单词样本作为关键单词模型训练样本集合,选取所有的候选关键词串样本作为关键词串模型训练样本集合。选取所有的候选关键单词样本与候选关键词串样本作为整体模型训练样本集合。当然候选关键单词与候选关键词串生成的特征向量长度是不同的。因为整体模型中的候选关键单词无法抽取对应的 LEN、MI、HB 和 TB 特征,而整体模型又必须能判定候选关键单词,所以候选关键单词在抽取这几个特征时选定默认值。我们假定每个候选关键单词样本的 LEN 为 1、MI 为 0、HB 与 TB 为 0。接着我们利用 LIBSVM[9]对三个训练样本集合进行训练,获得关键单词模型、关键词串模型、整体模型。

对于新文档,首先自动获得候选关键单词集合与候选关键词串集合。然后对于每一个候选关键单词分别假设其为关键单词,并根据该候选关键单词的特征获得特征向量,最后利用关键单词模型对候选关键单词进行是否为关键单词的判断。候选关键

词串同样也如此。而整体模型可以同时判断候选关键单词与候选关键词串。

4.2　评分实验

在 LIBSVM 的二分类问题中,新样本的分类是通过模型中的分类器评分判定的。基于 LIBSVM 的实现原理,我们修改了 LIBSVM 的部分代码,使得 LIBSVM 训练出的分类模型可以对新样本成为正例的可能性评分。

与分类实验中构造训练模型方法一样,我们同样选取一批已手工标注关键词的文档作为训练集构造了关键单词评分器、关键词串评分器、整体评分器。对于新文档中的候选关键单词,计算该候选关键单词的特征并形成特征向量,利用关键单词评分器对其评分,分值越高,该候选关键单词越可能是关键单词;分值越低,则越可能是非关键单词。利用关键词串评分器对候选关键词串评分类似,而整体评分器可以同时对候选关键单词与候选关键词串评分。

5　实验结果及分析

我们从 Web 网站中抓取了博客网页作为关键词提取测试的语料。因为每篇博客中都有 tag 标签,可以看成作者手工标注的关键词。我们选取了其中拥有 5 个 tag 标签的中文博客,总共有 2 096 篇。每篇博客的平均词数为 1 270。由于很多 tag 标签并没有出现在它自己的博客中,因此所有语料总共只拥有 9 339 个 tag 标签。我们选取其中 1 572 篇博客作为训练集,剩下的 524 篇博客作为测试集。

我们利用 LIBSVM 对训练集中的候选关键单词与候选关键词串按照表 1 选取的特征进行训练,但由于每篇文本的非关键词数目远远多于关键词数目,使得训练样本的正例与反例极不平衡。为此我们采用 Chong Huang[10]的方法,随机地在反例样本集合中选取样本,使得训练集中正例与反例的数目基本为 1 ∶1,具体数目见表 2。

表 2　分类实验训练集中正例与反例的具体数目

关键单词模型		关键词串模型		整体模型	
正例数目	反例数目	正例数目	反例数目	正例数目	反例数目
5 478	5 516	1 154	1 128	6 632	6 644

按照分类实验的方法训练出关键单词模型、关键词串模型、整体模型。然后分别对测试集进行测试,结果如表 3、表 4。

表 3　分类实验候选关键单词测试结果

特征选取(编号表示)	关键单词模型				整体模型			
	正例准确率/ %	反例准确率/ %	整体准确率/ %	整体 F1 值/ %	正例准确率/ %	反例准确率/ %	整体准确率/ %	整体 F1 值/ %
(1) (2)	80.091 0	90.085 5	90.005 9	84.759 5	81.114 9	88.708 3	88.646 8	84.713 8
(1) (2) (3)	79.010 2	92.987 7	92.876 4	85.384 0	77.246 9	93.622 6	93.492 2	84.593 2
(1) (2) (3) (4)	81.513 1	92.231 3	92.145 8	86.503 9	82.935 2	91.444 8	91.377 0	86.951 7

表 4　分类实验候选关键词串测试结果

特征选取(编号表示)	关键词串模型				整体模型			
	正例准确率/ %	反例准确率/ %	整体准确率/ %	整体 F1 值/ %	正例准确率/ %	反例准确率/ %	整体准确率/ %	整体 F1 值/ %
(1) (2)	77.684 2	87.127 6	87.119 5	82.131 8	75.578 7	85.973 9	85.965 0	80.439 4
(1) (2) (3)	75.368 4	95.281 5	95.263 5	84.156 1	75.157 9	91.200 0	91.186 2	82.399 8
(1) (2) (3) (4)	81.052 6	93.697 2	93.686 4	86.912 8	77.473 7	88.861 3	88.851 5	82.773 4
(1) (2) (3) (4) (5)	80.201 5	95.292 9	95.280 0	87.092 9	76.421 1	93.599 2	93.584 5	84.129 1
(1) (2) (3) (4) (5) (6)	80.201 5	95.421 5	95.408 5	87.146 6	79.578 9	92.207 2	92.196 5	85.434 7
(1) (2) (3) (4) (5) (6) (7) (8)	79.157 9	95.933 6	95.919 2	86.736 2	79.368 4	92.181 9	92.176 0	85.294 1

测试集中包含 1 758 个关键单词、219 072 个非关键单词、475 个关键词串、555 772 个非关键词串。

从表 3 与表 4 中我们可以看出,分离模型与整体模型在候选关键词串测试中,模型选取相同特征的条件下,分离模型比整体模型效果更好,整体 F1 值提高了 1 %～3 %。而这种提高在候选关键单词中并不非常显著。实验数据总体上说明,在选取同样特征的情况下,基于分离模型的关键词提取比整体模型好。从表 3 我们还可以看出,添加特征(3) NWT 时,关键单词模型的整体 F1 值为 85.384 %,比不添加特征(3) NWT 时关键单词模型的整体 F1 值 84.759 5 %高。同时添加特征(3) NWT 和特征(4) TF ×IF 时,关键单词模型的整体 F1 值为 86.503 9 %,整体模型的整体 F1 值为 86.957 1 %,这比相同条件下不添加特征(3) NWT 和特征(4) TF ×IF 时,整体 F1 值高,这说明特征 NWT 与 TF ×IF 对于关键单词提取是有帮助的。同样从表 4 中我们可以得出这两个特征对于关键词串提取也是有意义的结论。对于添加特征(5) LEN,它提高了提取关键词串的整体 F1 值。对于在关键词串提取中添加互信息与词串边界参数表特征,从表 4 可以看出,互信息提高了两个模型测试的整体 F1 值,这说明互信息对于关键词串提取是有作用的。虽然词串边界参数表特征使得关键词串提取整体 F1 值下降,但整体准确率为 95.919 2 %,它比不添加此特征的整体准确率高,这对于关键词串提取中如何过滤掉大量的非关键词串是有帮助的。

我们按评分实验的方法以同样的训练集合构造了关键单词评分器、关键词串评分器、整体评分器。利用关键单词评分器与整体评分器对测试集中每篇博客的所有候选关键单词评分,然后抽取每篇博客中前 N 个评分分数最高的候选关键单词,比较前 N 个候选关键单词中平均关键单词数目,得到图 2,同样利用关键词串评分器与整体评分器对候选关键词串进行评分测试,得到图 3、4。

从图 2、3、4 我们可以看出在输出相同数目的候选关键单词或关键词串以及选取相同的特征时,利用关键词串评分器输出的平均关键词串数目明显高于利用整体评分器输出的平均关键词串数目,而利用关键单词评分器输出的平均关键单词数目略高于利用整体评分器输出的平均关键单词数目。这说明基于分类模型思想构造的关键单词评分器和关键词串评分器对候选关键单词和候选关键词串评分排序时,关键单词和关键词串的排名普遍比整体评分器评分排名高。从图 2、3 我们可以看出对于添加 NWT 与 FEP 特征,输出的平均关键单词或关键词串数目在输出相同的候选关键单词或关键词串数目时增加明显。这进一步说明 NWT 与 FEP 特征对于提高关键单词和关键词串的排名是有意义的。在图 4 中我们可以看到对于词串评分器,互信息与词串边界参数表特征使得在输出相同数目的候选关键词串时输出的平均关键词串数目增加,而对于整体评分器却减少了。平均关键词串数目增加说明互信

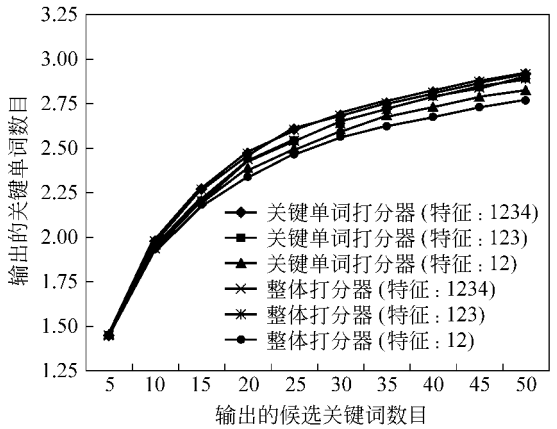


图 2 关键单词评分器测试结果

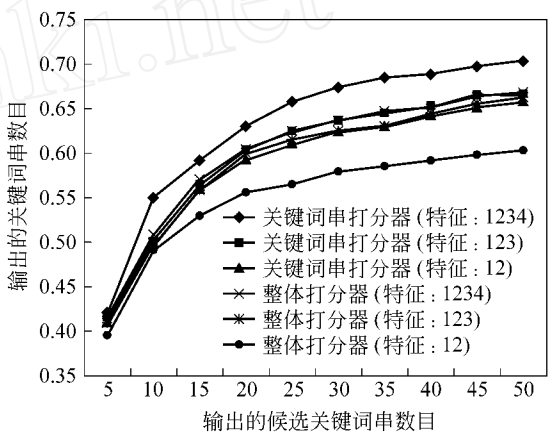


图 3 关键词串评分器测试结果(1)

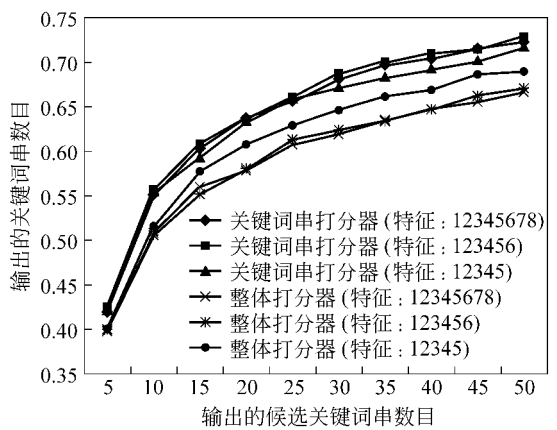


图 4 关键词串评分器测试结果(2)

息与词串边界参数表特征对于关键词串提取是有作用的。平均关键词串数目减少的原因可能是,在整体评分器构造时,关键单词样本的互信息与词串边界参数表特征都给了默认值,影响了整体评分器对于关键词串的评分效果,但这也从另一侧面说明关键单词提取与关键词串提取本应该看成不同的问题。

在很多关键词提取的实际应用中,需要的是能够表述文档内容的关键概念,并不关心提取的关键词是关键单词还是关键词串。因此为了找到最合适关键词,需要将关键单词和关键词串进行统一排序,以选出最佳结果。为此,我们把关键单词评分器与关键词串评分器合成一体,形成了一个综合评分器。综合评分器利用关键单词评分器与关键词串评分器分别对候选关键单词与候选关键词串评分,然后将候选关键单词与候选关键词串按评分的分值由高到低进行整体排名。我们仍然对相同的测试集选取不同的特征,将该评分排序方法和整体模型的排序结果进行比较,分别输出前 5 个与前 15 个分数最高的候选关键词,结果如图 5、6。

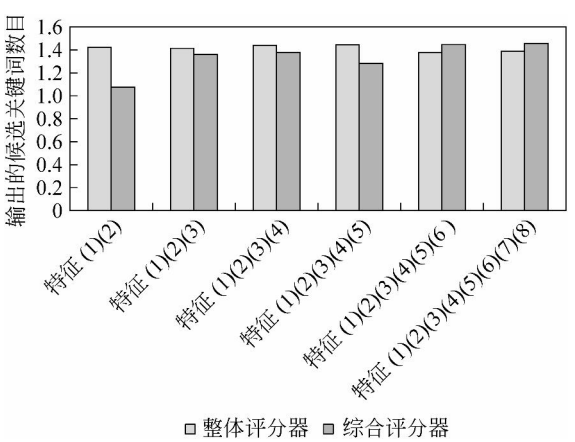


图 5　输出前 5 个候选关键词比较

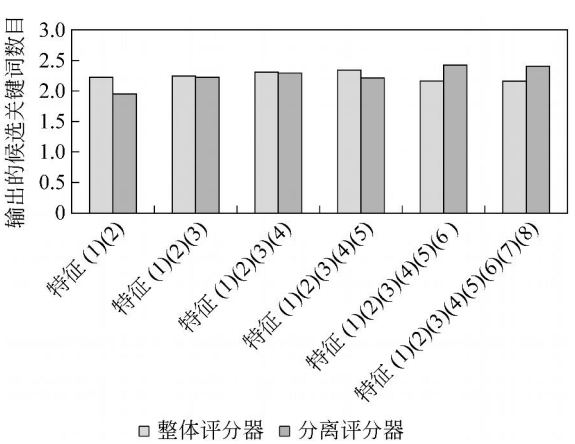


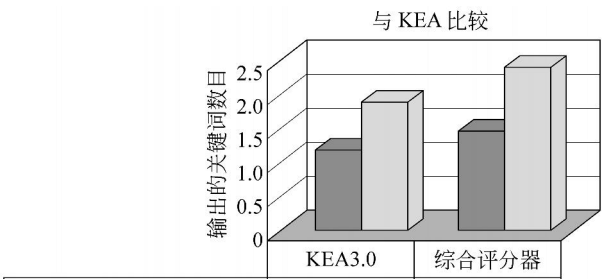
图 6　输出前 15 个候选关键词比较

从图 5、6 我们可以看出所有的特征的挖掘对于关键词提取是有效的。虽然在添加 LEN 特征时输出的平均关键词数目减少,但在前面的分类实验中已证明 LEN 特征对于候选关键词是否为关键词的判断是有意义的,只是对于候选关键词的评分排序问题效果略有下降。从图 5、6 我们还可以看出在选取 1～5 号全部或部分特征(见表 1)时综合评分器对于关键词的评分效果不如整体评分器。这说明虽然综合评分器单独对关键单词或关键词串评分时效果好,但这种评分器如果只选取 1～5 号全部或部分特征对关键词的整体评分效果不如整体评分器好。不过,如果我们添加互信息与词串边界参数表特征时,综合评分器对于关键词的评分效果比整体评分器好,而且在选取 1～8 号全部特征时,综合评分器输出前 5 个候选关键词时,平均正确数目为 1.45 个;在输出前 15 个候选关键词时,平均正确数目为2.41 个,达到了最佳效果。综合图 5 和图 6,分离模型的最好结果为选取 1～8 号全部特征 ,整体模型的最好结果为选取 1～8 号全部特征,这说明基于分离模型的关键词提取算法比传统的关键词提取算法好。

选取 1～8 号全部特征构造的综合评分器达到了最好的效果,我们以此与 KEA 进行提取效果的总体比较。KEA 是一个基于文本的关键词提取算法,它是由新西兰怀卡托大学开发[2]。KEA 通过计算每个候选关键词的 TF ×IDF 与 POS 特征,利用机器学习算法来判断每个候选关键词成为关键词的可能性。它首选需要通过一批以标注好关键词的文档集合作为训练集进行训练,然后通过 Bayes 算法得到训练模型,以此对新的候选文档对每个候选关键词成为关键词可能性进行评分。通过输入需要的候选关键词数目 n,KEA 自动提取新文档中评分最高的前 n 个候选关键词。

我们利用 KEA3.0[11] 进行比较实验。将其提取候选关键词的长度设置为 1～4,候选关键词最小词频设置为 1。输出测试集中每篇文档分数最高的前 5 个与前 15 个候选关键词,并将其与综合评分器进行比较,我们仍然利用 Blog 语料进行比较实验。结果如图 7。

我们从图 7 可以看出,无论在输出候选关键词数目为 5 还是 15 时,综合评分器对于关键词的提取效果都明显好于 KEA3.0。这说明基于分离模型的关键词提取算法优于 KEA 算法。



	KEA3.0	综合评分器
输出前 5 个候选关键词	1.18	1.45
输出前 15 个候选关键词	1.89	2.41

图 7 输出前 5 个及前 15 个候选关键词比较

6 结束语

本文从关键词的组成结构出发,将传统的关键词提取分为关键单词提取与关键词串提取。首先给出了关键词串的定义,说明了将关键词分为关键单词与关键词串的原因。然后提出了以互信息与构造词串边界参数表的方法识别词串的方法,并详细说明了基于分离模型的关键词提取算法。最后通过实验验证了分离模型的有效性,并对特征的选取进行了深入分析。

虽然基于分离模型的关键词提取对于单独的关键单词提取和关键词串提取效果明显,但对于关键词整体的提取效果却在选取部分特征时不如传统的关键词提取算法。未来进一步的研究将主要围绕如何将分离模型集合成整体、提高关键词提取的整体效果展开,并对分离模型怎样选取特征作深入的研究。

参考文献:

[1] Turney P . D . Learning to extract keyphrases from text[R]. National Research Council, Canada, NRC Technical Report ERB-1057,1999.

[2] Witten I. H., Paynter G. W., Frank E., Gutwin C., Nevill—Manning C. G. KEA: Practical automatic keyphrase extraction[C]// Proceedings of the 4th ACM conference on Digital libraries, Berkeley, California, US,1999: 254-256.

[3] 刘远超, 王晓龙, 徐志明, 刘秉权. 基于粗集理论的中文关键词短语构成规则挖掘[J]. 电子学报, 2007, 35(2): 371-374.

[4] Anette Helth. Combining machine learning and natural language processing for automatic keyword extraction [D]. Stockholm: Department of computer and systems sciences, Stockholm University,2004: 35-38.

[5] Yang Wen-Feng. 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.

[6] 李素建, 王厚峰, 俞士汶, 辛乘胜. 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.

[7] 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.

[8] 索红光, 刘玉树, 曹淑英. 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): 27-32.

[9] Chang C. LIBSVM: a library for support vector machines[EB/OL]. 2006. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[10] Huang C., Tian Y., Zhou Z., Ling C., Huang T. Keyphrase extraction using semantic networks structure analysis[C]// Sixth IEEE International Conference on Data Mining, Hong Kong, China,2006: 275-284.

[11] Frank E. KEA: Keyphrase Extraction Algorithm [EB/OL]. 1999. Software available at http://www.nzdl.org/Kea/download.html.