TFLD:一种中文文本关键词自动提取方法

管瑞霞[1,2],陆 蓓[1]
(1.杭州电子科技大学 计算机应用技术研究所,浙江 杭州 310018; 2.浙江育英职业技术学院,浙江 杭州 310013)

摘要:为了提高中文关键词提取的准确率和实用性,提出了一种改进了候选词权重计算的关键词提取算法 TFLD(term frequency, location & distance algorithm),利用候选词权重排序自学习,提高了提取关键词算法的效率。该方法采用词语词频统计、分布区域以及词语距离位序 3 种特征项,并使用最小均平方(LMS)法则训练算法模型的调整因子。实验结果表明,该方法提高了关键词提取的精度。

关键词:关键词提取;中文文本;中文信息处理
中图分类号:TP391 **文献标识码**:A **文章编号**:1001 - 4551(2010)09 - 0123 - 04

TFLD: a novel phrase-extraction method for Chinese text

GUAN Rui-xia[1,2], LU Bei[1]
(1. Institute of Computer Application Technology, Hangzhou Dianzi University, Hangzhou 310018, China;
2. Zhejiang Yuying College, Hangzhou 310013, China)

Abstract: Aiming at improving accuracy and practicality of key-phrase extraction for Chinese, a new algorithm was proposed, which named as TFLD(term frequency, location & distance algorithm), the calculation accuracy by obtaining a sorted candidate key word sequence was improved. Based on word frequency features including statistic of term frequency, term location and term distance, the least mean square (LMS) algorithm was trained to calculate the parameters for TFLD algorithm. The experimental results show that the proposed method improves the accuracy of key-phrase extraction in a considerable magnitude.

Key words: key-words extraction; Chinese text; Chinese information processing

0 引 言

互联网络积累了海量的文本信息,如何高效地检索文本信息成为亟需解决的技术问题。文本信息处理包括文本分类、文本聚类、文本挖掘和近似查询处理等内容,而文本关键词提取在上述方面有着广泛的应用,它不仅是进行这些工作不可缺少的基础和前提,也是互联网上信息建库的一项重要工作。英文文本的关键词自动标引的研究起步较早,已开发了一些相关系统。主要有 Turney 在 C4.5 决策树算法基础上实现的 GenEx 系统[1]。该系统使用遗传算法训练关键词提取器,然后提取器以文档为输入,经过处理后输出关键词[2];Frank 等人提出了基于朴素贝叶斯方法的提取算法,使用离散的短语特征值训练统计学习模型以获取输入参数,较好地实现了关键短语的自动提取[3]。

由于汉语没有显式的词边界,其关键词的自动抽取问题较英文文本的相同问题更为困难。为此需要依次对文本进行应用分词算法、词法分析、语法处理以及语义分析,使用最多的一种方法是基于 PAT Tree 结构获取新词[4]。另外李素建等人提出的最大熵模型利用各种成熟的语言学工具首先从文档中获取关键词候选项,提出了如何计算最优概率分布的方法,并建立了一个特征集合,再根据丰富的语言特征来判断候选项是否可以选做文档的关键词[5]。而基于语义的关键词提取算法在统计信息的基础上着重强调了语义对关键词判断的影响[6]。此外,基于词汇链的关键词提取算法[7]则在分词词频、文档反频、分词位置等基本统计手段的基础上,引入了词汇链的概念。

然而,现有中文文本关键词提取算法需要较大的空间代价,导致其实用性受限。实际应用系统工作通常基于高频词提取等手段,但文档中的关键字往往并

收稿日期:2009 - 11 - 25
作者简介:管瑞霞(1981 -),女,浙江瑞安人,主要从事中文信息处理方面的研究. E-mail: grx1981@163.com

不都是高频词。为了克服上述两个方面的缺陷,本研究提出一种结合分词词频统计、区域位置特征与分词距离位序特征的关键词提取算法 TFLD。

1 关键词提取的预处理

如何自动识别词的边界,从而将汉字序列切分为正确的词串的中文分词问题将是实现中文信息处理的基础问题[8-10]。同时利用分词技术作为关键词提取的前提时,则还需在分词的基础上进行相应的处理,如去除停用词以及进行词性过滤。分词预处理过程如图 1 所示。

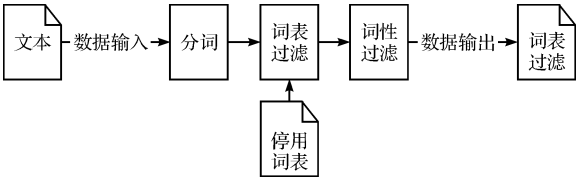

图 1 分词预处理过程

停用词是指那些不能反映主题的功能词,它们不但不能反映文献的主题,而且还会对关键词的抽取造成干扰,有必要将其滤除[11]。停用词确定为所有虚词以及标点符号,定义停用词表并根据这些表进行分词过滤是很好理解的。那么词性过滤的目的是什么呢?在汉语言中,能标识文本特性的往往是文本中的实词。而文本中的一些虚词,对于标识文本的类别特性并没有贡献。如果把这些对文本分类没有意义的虚词作为文本特征词,将会带来很大噪音,从而直接降低文本分类的效率和准确率。因此,在提取文本特征时,首先考虑剔除这些对文本分类没有用处的虚词。

2 关键词提取算法

2.1 特征项

经分词和词性过滤后得到文本的词语集合,选择其中的哪些词语作为文本的关键词需要考虑以下 3 个特征项。

(1) 词频:词频是对词的一种最简单的测度,也是最常用的参数之一。可以直接用它表示词在篇章或类别中的权重,这种处理方法假定一个词的重要程度与它出现的次数成比例。

(2) 区域位置:经研究发现[12],出现在标题中的词比出现在摘要中的词更能反映文献的主题,而出现在摘要中的词比出现在正文中的词更能反映文献主题,同时出现在首段中的词比其他段落中的词更能反映文献的主题。

(3) 分词距离次序:随着文本长度的增加,笔者发现利用词语第一次在文本中出现时距离文本开头的距离来衡量词语反映文本主题的价值也是很有意义的。因此,本研究将该因素引入关键词的抽取算法中,用来反映词语在文本中的权重。

2.2 候选词权重计算

综合考虑词语的词频,区域位置以及距离位序的因素,本研究提出了候选词权重计算函数如下:

$$weight_i = \alpha \times tf_i + \beta \times loc_i + \gamma \times dis_i \tag{1}$$

式中 $weight_i$—候选词 word_i 的权重;tf_i—其词频因子;loc_i—其区域位置因子;dis_i—其距离次序因子;α,β,γ—3 个因子的调节因子。

本研究针对词频因子采用公式 $tf_i = \ln f_i$(其中,f_i 为文本中该候选词的词频)。为了获取每个词语的位置信息,需要确定记录位置信息的方式以及各个位置在反映文章主题时的贡献程度。根据以往的研究结果以及实验的数据分析,本研究将文本位置区别为 3 种情况:当词语出现在标题时,其位置值设为 3.0;当词语出现在首个段落时,其位置值设为 2.0;当词语出现在其他文本段时,其位置权重值为 1.0。若一个词语在各个位置重复出现,则选取其最高的位置值。因而,在进行分词处理的同时,即可对每个获得的词语标记其位置值,然后通过下式获得候选词的位置权重:

$$loc_i = (w_i - 1)/(w_i + 1) \tag{2}$$

式中 w_i—候选词根据分词位置标记的位置值。

为了获得每个词语的距离信息,需要确定记录距离次序的方式。根据实验的数据分析,本研究通过一个线性函数:$val_i = a \times i + b$ 来标记各个分词距离次序值(其中 i 表示词语在文本中出现的次序,a,b 均为可调节的常数因子)。然后通过下式来计算候选词的距离次序权重:

$$dis_i = val_i / \ln val_i \tag{3}$$

式中 val_i—该分词第一次出现位置到文本开头的距离。

通过在子公式中引入对数函数,可以更好地刻画权重计算中特征项的非线性特点。

2.3 权重因子的训练

在确定了 3 个特征项影响词语权重的公式后,需要考虑如何确定调整因子 α,β,γ,使其更加合理地反映各个因素对权重的贡献程度。采用训练样例来自动学习调整各调整因子。为了保证提取算法的通用性,本研究从财经、信息、健康、体育、旅游、教育等多个领域分别选取 80 个训练文本。

本研究通过机器学习最小均方误差(LMS)训练法则[13]训练公式的调整因子。具体操作方法描述如下:

首先随机给定调整因子的值,然后依次计算出各个文本中各词语的权重,并将各个文本中的词语集合依次按其权重值由高到低排序。这里不妨设在第 k 次计算各文本词语权重并排序后第 i 个文本的词语集合为 $V(k,i)$,而该文本的训练词语排序恒记为 Vi。根据 $V(k,i)$ 和 Vi 中词语权重排序的差异性,设排序差值:

$$diff = \sum_{j=1}^{n} (sort_{(i,j)} - sort_{(k,i,j)}) \qquad (4)$$

式中　$sort_{(i,j)}$,$sort_{(k,i,j)}$—i 个文本中的第 j 个分词在训练排序集和第 k 次计算后的测试排序集中的排序次序。

然后,通过如下公式来调整各个调整因子(α,β,γ)的值:

$$w = w + \eta \times diff \times \sec \qquad (5)$$

式中　w—调整因子;η—一个很小的常数因子;sec—当前的测试因子(tf,loc,dis)的取值。

本研究提出了计算候选词权重的公式,根据权重大小对词语进行排序,并基于给定阈值提取合适的关键词集。为了获得合适的公式调整因子,本研究设计了对公式的训练过程,然后通过测试来验证该公式的合理性,其模型如图 2 所示。

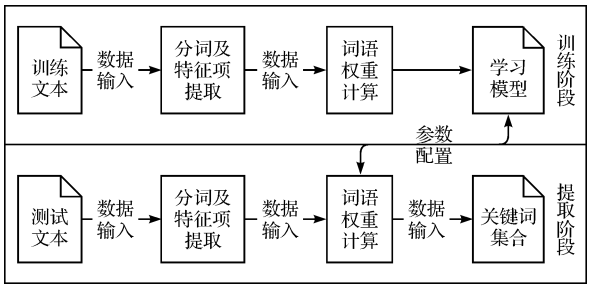

图 2　关键词提取算法模型

2.4　简单示例

本研究以一则报道京沪高铁 2010 年投入运营的简短新闻为测试样例,说明关键词提取算法的具体流程。该文本信息如下所示:

京沪高铁 2010 年投入运营

早报驻京记者　吴玉蓉

早报记者　杨洁

铁道部昨天宣布,国内第一条具有世界先进水平的高速铁路——京沪高速铁路预计今年开工建设,预计 5 年左右完成,并将于 2010 年投入运营。

铁道部新闻发言人王勇平说,京沪高速铁路正线全长约 1318 公里,与现在的京沪铁路走向大体并行,全线为新建双线,设计时速 350 公里,初期运营时速 300 公里,共设置北京、天津、济南、南京、上海等 21 个客运车站。每年单向输送 8 000 万人,京沪高速铁路是一条客运专线,在它建成后,既有京沪铁路将变为货运主线。届时,北京至上海高速列车全程运行时间只需 5 小时,比目前京沪间特快列车缩短 9 小时左右,年输送旅客单方向可达 8 000 余万人。而既有京沪线的单向年货运能力也将提高到 1.3 亿吨以上,从而满足京沪通道客货运输需求,从根本上解决京沪通道运输能力紧张的状况。此外,京沪高速铁路与时速 200 公里既有铁路兼容,时速不小于 200 公里列车可以在京沪高速铁路上运行。从上海去往哈尔滨、沈阳、包头、兰州、西安、成都、乌鲁木齐和从北京去往华东,均可大大缩短旅行时间。

通过本研究关键词提取算法计算所得候选词的权重如表 1 所示。

表 1　一则新闻文本的关键词自动提取:候选词权重结果表

候选词	铁路	高速	高铁	运营	时速
权重	7.916 49	7.047 45	5.291 48	5.020 63	4.508 57

若选择 4 个关键词,则根据权重递减排列并选用“铁路、高速、高铁、运营”作为此段新闻的关键词。同时,请 8 位中文专业研究生组成的人工阅读小组对本段新闻抽取 5 个关键词,其组成及顺序与表 1 相同。

3　实验分析

3.1　实验数据

本实验所使用的语料是搜狗提供的分类语料库,其中文档分为财经、信息、健康、体育、旅游、教育、招聘、文化和军事共 9 个类别。从各个类别中分别选取 80、20 个文档作为训练集和测试集。对于训练集和测试集,其关键词都是通过人工赋予的,根据每个文本的长度给予适当数目的关键词来标识该文本的内容。

3.2　评价方法

对关键词提取算法的评估办法是将算法提取的关键词与由 8 位新闻专业研究生组成的人工标注小组抽取的标准关键词作词法上的匹配。这里通过 3 个标准来衡量算法的效果,分别为查准率、查全率和 F 因子。其中查准率 p = 自动提取正确的关键词数目/自动提取的关键词数目,查全率 r = 自动提取正确的关键词数目/人工赋予的完全的关键词数目,而 F 因子综合考虑了查准率和查全率的因素[14]:$F_t = \frac{2rp}{r+p}$。

3.3　结果分析与比较

首先根据给定的训练集的大小不同,比较从测试集中分别提取 5 个和 10 个关键词时的性能。通过分析发现,当训练集大小比较少时,训练效果随文档数目增长而提高显著。而当训练集大小到达一定程度后,其后训练效果基本保持稳定,实验结果如图 3、图 4 所示。

同时本研究在选定相同的 30 篇训练文档的基础上,依次提取 5 个和 10 个关键词,与文献[11]所给的基于词频和分词位置影响分词权重的算法进行比较,

可以看出本研究的算法在一定程度上得到了优化，如图 5 所示。

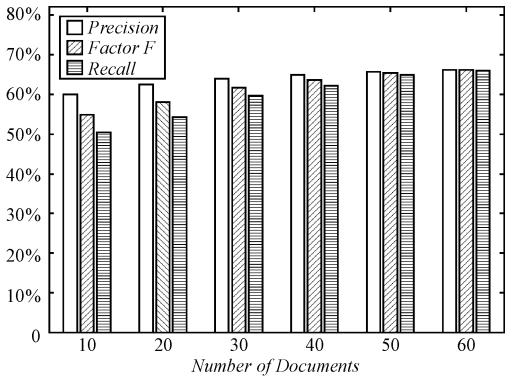

图 3 不同训练集大小下的评估因素取值(提取 5 个关键词)

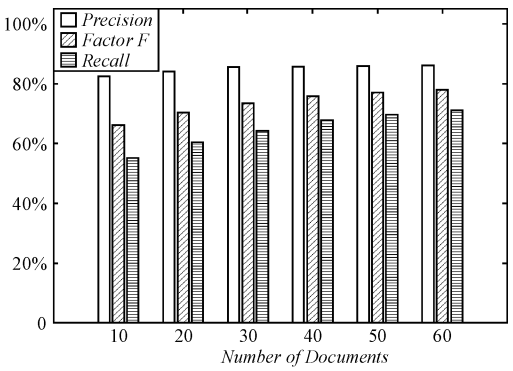

图 4 不同训练集大小下的评估因素取值(提取 10 个关键词)

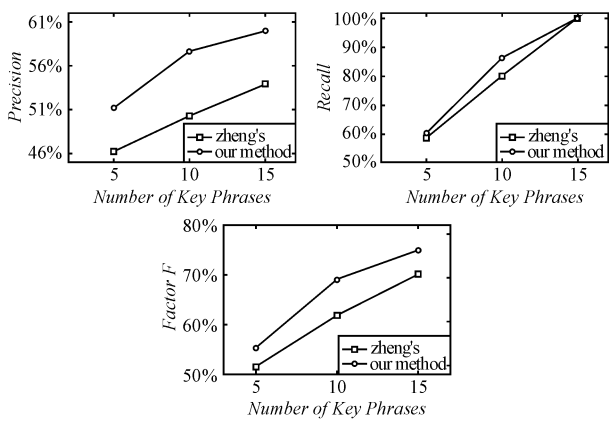

图 5 TFLD 算法与文献[11]的提取效果比较

由图 3 和图 4 可知，若提取相同数目的关键词，算法的查准率、查全率以及 F 因子随着训练集样本数目的增加，均有提高。说明通过更多训练，可改进算法的关键词提取能力。同时可注意到，随着训练样本集的进一步增加，查准率、查全率以及 F 因子的增长速度相应放慢，这说明了 LMS 训练法则中，算法的调整因子最终将趋于收敛。此外，图 4 中的查全率对比图 3，均有显著地改进，其原因在于相对于笔者所采用的语料库中的训练测试文本，5 ~ 10 个关键词是比较合适的。

通过分析图 5 发现，相对于文献[11]，本研究引入了分词距离作为特征项，同时采用 LMS 法则来训练提取算法公式，使得本研究的关键词提取算法在查准率、查全率和 F 因子的衡量上大概高出 3% ~ 5%。

4 结束语

本研究采用了词频、候选词区域位置以及距离次序作为衡量关键词权重的主要因素，并为每一个因素构建了非线性计算函数。同时，利用训练样例训练该公式的调整因子使得其更好地逼近训练取值，并与已有相关工作进行了比较实验，其结果表明该方法改进了文本关键词提取算法的查准率和查全率，并具有良好的扩展性，可应用于文本信息处理中的自动关键词提取。

参考文献(References)：

[1] TURNEY P D. Learning to Extract Key Phrases from Text [R]. NRC Technical Report ERB-1057, National Research Council, Canada, 1999:1 - 43.

[2] WHITLEY D. The GENITOR Algorithm and Selective Pressure[C]//Proceedings of the Third International Conference on Genetic Algorithms. California: Morgan Kaufmann, 1989:116 - 121.

[3] FRANK E, PAYNTER G W, WITTEN I H. Domain-Specific Key Phrase Extraction[C]//Proceedings of the 16th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: Morgan Kaufmann, 1999:668 - 673.

[4] CHIEN L F. PAT-tree-based Keyword Extraction for Chinese Information Retrieval[C]//Proceedings of 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Philadelphia: [s. n.], 1997:50 - 59.

[5] 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.

[6] 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.

[7] 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)：25 - 30.

[8] 金翔宇，孙正兴，张福炎. 一种中文文档的非受限无词典抽词方法[J]. 中文信息学报，2001，15(6)：33 - 39.

[9] NIE Jian-yun, GAO Jiang-feng, ZHANG Jian, et al. On the Use of Words and N-grams for Chinese Information Retrieval [C]//Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages. New York: ACM Press, 2000:141 - 148.

[10] 徐亚娟. 基于公安业务信息的文本挖掘技术研究与实现[D]. 杭州：浙江大学计算机学院，2008：22 - 40.

[11] 郑家恒，卢娇丽. 关键词抽取方法的研究[J]. 计算机工程，2005，18(9)：194 - 196.

[12] 何新贵，彭甫阳. 中文文本的关键词自动抽取和模糊分类[J]. 中文信息学报，1999，13(1)：9 - 15.

[13] MITCHELL T M. Machine Learning[M]. 曾华军，译. 1 版. 北京：机械工业出版社，2006：7 - 8.

[14] SEBASTIANI F. Machine learning in automated text categorization[J]. **ACM Computing Surveys**, 2002, 34(1):1 - 47.

[**编辑**：李 辉]