第25卷 第2期 广西师范大学学报:自然科学版 Vol. 25 No. 2
2007年6月 Journal of Guangxi Normal University:Natural Science Edition Jun. 2007

基于同义词词林的中文文本主题词提取

程 涛[1],施水才[1],王 霞[2],吕学强[1]

(1. 北京信息科技大学 中文信息处理研究中心,北京 100101;2. 抚顺市第十五中学,辽宁 抚顺 113006)

摘 要:中文文本主题词的提取可以浓缩一篇文章,可以提炼一个中文网页,还可以帮助实现网上广告与网页的精确匹配。提出了一种基于同义词词林的中文文本主题词提取方法,不仅考虑了传统的影响主题词语权重的因素,还考虑到了同义词、相关词以及下位词的出现对于词语权重的影响。实验表明,用该方法对中文文本进行主题词提取,准确率可达83.25%。

关键词:主题词提取;同义词词林;权值;同义词

中图分类号:TP391 **文献标识码**:A **文章编号**:1001-6600(2007)02-0145-04

在信息爆炸的时代,人们迫切希望能够迅速而准确地获得自己所需的信息。中文文本主题词提取为我们找到了一条出路。中文文本主题词提取可以帮助实现网页分类和文本分类,还可以帮助人们快速了解一篇文章的主要内容。另外,全球最大的基于网页内容的广告推介联盟——Google AdSence计划,在实现网页与广告关联的时候,也用到了主题词提取技术。

主题词提取方法主要有基于词典的提取方法、基于规则的提取方法以及基于统计的提取方法3种。词典法简单且容易快速实现,但是词典主要依靠人工构建和维护,实时性比较低。规则法又被称为文法剖析法,它对于英文的处理已经取得了非常好的效果,但是对于汉语的处理,由于基础资源的不足或难度太大,目前还很少见。统计法是目前使用最广泛,研究最深入的一种中文文本主题词提取方法。本文所提出的主题词提取方法就属于这一种。

1 同义词词林及其扩展版简介

《同义词词林》[1]原版是梅家驹先生等人人工构造的,并于1983年由上海辞书出版社第一次出版。原书收录词语近7万,全部按意义进行编排,所以它是一部类义词典。

同义词词林扩展版(以下简称"词林")是哈工大信息检索研究室在《同义词词林》的基础上,参照多部电子词典资源,按照人民日报语料库中词语的出现频度,只保留频度不低于3(小规模语料的统计结果)部分词语。最终,词林收录的词语共有77 343条。

词林按照树状层次结构把所有收录的词条组织到一起,把词汇分成大、中、小3类,大类有12个,中类有97个,小类有1 428个,小类下再划分词群。每个词群中的词语又进一步分成了若干个行,同一行的词语要么词义相同(有的词义十分接近),要么词义有很强的相关性。

基于同义词词林的中文文本主体词提取(Thematic Words Extraction Based on Tongyici Cilin,以下简称TWE)就是利用词林获得词语的同义词、相关词、上位词、下位词等相关信息的。

收稿日期:2006-12-18

基金项目:国家自然科学基金资助项目(60272084);北京市教育委员会科技发展计划重点项目(KZ200310772013);北京市教委项目(KM200510772008,KM200610772008)

作者简介:程涛(1982—),男,四川成都人,北京信息科技大学硕士研究生。

2 主题词提取词语权重计算方法

2.1 候选主题词的确定(表1)

表1 候选主题词的确定方法

Tab. 1 Determination of candidate thematic words

过滤规则	说明
词性	某些词性的词语不能成为主题词,如助词、介词、叹词、数词、副词等
停用词	即使满足词性规则,但如果它是停用词,仍不能成为主题词,如:“是”
词频	如果词频小于某个阈值,并且相关信息也小于某个阈值,它不能成为主题词

2.2 影响候选主题词权重的因素[2~4]

要从特定文献中准确、全面地抽出最能表达文献主题的词语,除了要有准确的分词结果作为基础,最重要的是对文献中的词在表达文献主旨方面的能力进行准确和全面的评价[2]。我们在设计主题词权重的计算方法时,作出了如下假设:

①词语在文献中出现的次数越多,它将越可能是主题词。

②词语在文献中出现的位置也将影响该词语的重要性。一般说来,一个词出现在标题、小标题、首末段、段首段末句和正文中所表达的重要性是不一样的。另外,一个词语出现在线索词语后面时,该词语的重要性也是不一样的。

③词语的长度也会影响到词语的重要性。对于名词而言,词语越长表示得越具体,表达主题的能力也越强。对于动词,一般2个字或3个字的动词更能表现主题。对于英文字串,单字母单词几乎没有什么意义,拥有2~10个字母的单词一般更能表现主题一些。

④一个词语所覆盖的段落数越多,它表现主题的能力就越强。

⑤英文字符串出现在中文文本中,也往往会和主题相关。

⑥如果一个词语的同义词在文中出现,那么这个词语将获得附加权重。

⑦如果一个词语的相关词在文中出现,那么这个词语将获得附加权重。

⑧如果一个词语的下位词在文中出现,那么这个词语将获得附加权重。

⑨一个词语即使在文献中没有出现,它仍有可能是主题词语。

2.3 权重的计算方法

基于以上假设来设计主题词的权重函数。目前,很多研究者都采用了TF-IDF公式来计算词条在文章中的权重[5]。但是这种方法只是考虑到了词语的出现次数,没有考虑其他因素对于词语权重的影响。本文在计算词语权重时,考虑了词频、位置、词长、覆盖段落以及相关信息,可以用权值函数表示如下:

$$\mathrm{Weight}=\mathrm{len}\times\mathrm{span}\times\sum_{i=1}^{k}f_i\times\mathrm{loc}_i+\sum_{j=1}^{h}\mathrm{fr}_j\times\mathrm{kind}_j。\tag{1}$$

其中Weight为该候选主题词的权重;len为词语权重基于词长的长度因子;span为词语权重基于覆盖段落的覆盖因子;$k=8$,位置种类,包括标题、小标题、首末段、首末句、首末段和首末句、正文、1类线索词语、2类线索词语等8种;f_i为词语在第i种位置上的出现次数;loc_i为位置因子;$h=3$,相关信息种类,指的是同义词、相关词和下位词;fr_j为词语的第i种信息的出现次数;kind_j为基于相关信息的相关因子。

长度因子、覆盖因子、位置因子和相关因子这4类参数的确定方法:首先查阅相关论文对这些问题的描述,获得经验参数;然后结合自己对汉语的了解,产生一组参数序列作为原始参数;通过不断地测试对这些参数进行调整,最后获得一组最优的参数序列。

2.4 权重排序

计算出所有候选词语的权重后,输出权重最大的前N个词语作为文献的主题词。在这个过程中,还需要做的就是要避免同义词同现。比如,在前N个词语中,不能同时出现“中国”和“中华人民共和国”,在处理时我们利用词林,只取权重较大的作为文献主题词。

3　系统的实现

3.1　主题词提取流程图

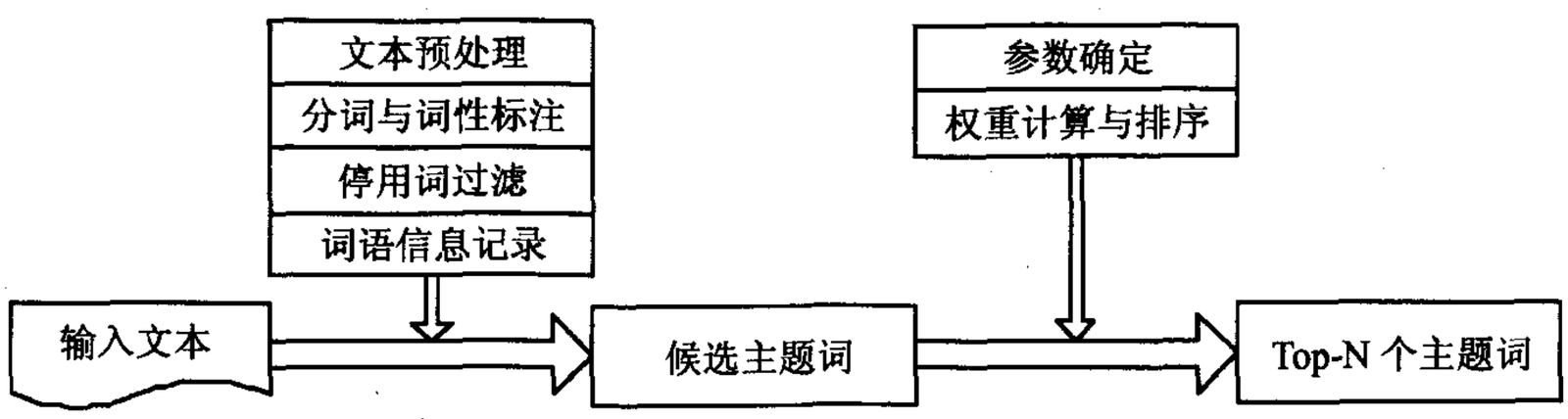

图 1　主题词提取流程图

Fig. 1　Flowchart of thematic words extraction

3.2　候选主题词语的存储结构[6,7]

在 TWE 中,要处理的数据量很大,要涉及很多的查找和插入操作,而且数据类型几乎是汉字。本系统采用 hash 表来存储这些信息,hash 表的大小是 256×16,用链地址法解决冲突。第一个汉字散列函数[6]是:

$$H(cc)=[(c_1\ll 5)\oplus c_1]\oplus c_2, \tag{2}$$

其中,c_1 和 c_2 为 cc 的高、低字节的机器内码值,⊕为异或运算符。

由于可能是单字词语,第二个汉字的参数借用了第一个汉字的 c_2,具体 hash 函数是:

$$H(c)=(c\oplus c_2)\&0x0F, \tag{3}$$

在存储候选主题词语时,先根据散列函数计算词语的散列地址,然后在链表中顺序查找链表。如果该词语已出现在该链表上,查找成功;否则将它置于链表的尾端。

4　实验结果

4.1　测试语料介绍

实验采用哈工大信息检索研究室单文档自动文摘语料库的 211 篇文章作为 TWE 的测试语料。该语料包含“奥运”相关文章 57 篇、记叙文 40 篇、说明文 40 篇、议论文 46 篇、应用文 18 篇、2003 年 863 评测语料 10 篇,语料共计 211 篇。

首先请几位在读研究生对这 211 篇文章进行人工主题词的提取,每篇文章提取 10 个主题词。由于主题词语提取的主观性相当强,并没有标准答案,所以在实验过程中,我们假定人工提取的 10 个主题词是标准答案。

4.2　实验过程及统计情况

用 TWE 对每篇文章提取 6 个主题词语,统计有多少个在人工提取的 10 个主题词语当中。本文采用准确率(查准率)作为衡量提取效果的依据,定义如下:

$$\text{准确率(查准率)}=\frac{\text{与标准答案中词语相同或同义的词语个数}}{\text{系统提取的词语总数}}。 \tag{4}$$

由于在权重排序时避免了同义词同现,所以如果提取出的主题词语与标准答案是同义词,本文也认为提取准确,表 2 是实验的统计情况。

表 2　TWE 的统计结果

Tab. 2　Statistic results of TWE

正确提取数	6 个	5 个	4 个	3 个	2 个	1 个	0 个
文章数	37	141	28	5	0	0	0
百分比/%	17.54	66.82	19.86	2.37	0	0	0
准确数	222	705	112	15	0	0	0
准确率	$\sum$提取准确的词语数/词语总数=1054/1266=83.25%						

4.3　实验结果分析与比较

通过统计结果(表 2)可以看出,用 TWE 提取主题词语的准确性比较

令人满意。我们查看了只提取出3个主题词的那5篇文章发现,那些文章是没有经过整理的,结构零乱,没有意义的空格和回车符较多,所以在分词与词性标注时就出现了不准确的情况。使用该方法,准确率高达83.25%。

5 结论与展望

本系统是在对现有的主题词提取方法进行总结之后,提出了一种基于同义词词林的提取算法,这实际上是一种基于语义的主题词提取算法。本文实验使用的211篇测试语料,内容全面,文体形式的覆盖面大,长度从540字节到13 135字节不等,具有很强的代表性。实验结果表明该方法对于中文文本主题词的研究有积极的推动作用。

但是,系统还有可以改进的地方。比如为了提高主题词提取的准确性,我们还可以将指代消解加到提取系统中。这样我们将离"机器理解文章"的目标更近一步。

另外,我们下一步准备将该方法应用于基于网页内容的广告推介的研究。

参 考 文 献:

[1] 梅家驹.同义词词林[M].2版.上海:上海辞书出版社,1996.

[2] 唐培丽,王树明,胡明.基于语义的汉语文献主题词提取算法研究[J].吉林大学学报:信息科学版,2005,23(5):535-540.

[3] 韩客松,王永成,沈洲,等.三个层面的中文文本主题自动提取研究[J].中文信息学报,2001,15(4):20-27.

[4] 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):194-196.

[5] 彭洪汇,林作铨.Internet上的搜索引擎和元搜索引擎[J].计算机科学,2002,29(9):1-12.

[6] 王忠效,范植华.汉字自适应散列分组查找算法[J].中文信息学报,2001,15(2):17-23.

[7] 王忠效,范植华.汉字异或动态散列分组查找算法[J].中文信息学报,1998,12(4):61-66.

Thematic Words Extracting from Chinese Text Based on Tongyici Cilin

CHENG Tao[1], SHI Shui-cai[1], WANG Xia[2], LÜ Xue-qiang[1]

(1. Chinese Information Processing Research Center, Beijing Information Science and Technology University, Beijing 100101, China; 2. Fushun No. 15 Middle School, Fushun 113006, China)

Abstract: Thematic words extraction from a Chinese text not only can concentrate an article, but also can extract main ideas from a Chinese Web and help to achieve precise matching between online advertisement and a webpage. The paper presents a method of thematic words extraction based on Tongyici Cilin. The method not only has taken traditional factors affecting the weight of a thematic word into account, but also has considered the factors such as the appearance of relevant words, synonymy and lower words. Experiments have confirmed that the accuracy rate of thematic word extraction from a Chinese text can reach 83.25% using this method.

Key words: thematic words extraction; tongyici cilin; weight; synonymy

(责任编辑 李小玲)