中文文本关键词提取算法①

Chinese Key Words Extraction Algorithm

张红鹰 (安徽财经大学 成人教育学院 安徽 蚌埠 233000)

摘 要： 本文主要研究关键词提取算法，在分析可能影响关键词提取的词语各种属性并将其量化的基础上，提出并实现了一种将分词与词性标注、文本预处理、线性加权算法、组合词生成与过滤、合并候选关键词等集成到一个完整框架中的模型算法。

关键词： 文本 关键词 提取

关键词提取文本自动处理的基础工作之一，本文设计出一种文本关键词提取算法，希望实现以下目标：(1)能够体现作者主要思想的重要词语而不仅仅是文档中的高频词语；(2)充分考虑分词系统对关键词提取的影响。

1 总体框架

文本关键词提取算法模型将分词与词性标注、文本预处理、线性加权算法、组合词生成与过滤、合并候选关键词等集成到一个完整的框架中，其中单词信息表和组合词信息表是关键的两个中间数据结构，生成的组合词不作为特例，而是以科学的方法给其赋予权值，与单词(由线性加权算法得到的词语)一起参与关键词的竞争，将两个链表合并得到最终的提取关键词的结果。

首先经过文本预处理、分词与词性标注系统的初步处理，然后进行线性加权算法。经过对中文文本词频，词性，词语所处位置等信息的分析，对加权因子进行量化，计算出每个词语的权值，然后按照权值大小排列实现候选关键词的提取，同时形成一级候选关键词作为提取最终关键词的第二层。

2 文本预处理

该算法处于关键词提取系统的第一层。文本关键词提取是在对其进行分词之后进行的处理。在分词的过程中，我们使用了统一的一些格式，比如每个词语词性标注后的字符都是词性+“ ”，每个标点之后的字符都是“/w”+“ ”这样就提供了一种断句的手段。

2.1 文本分段

不管是英文文本还是中文文本，段落结束的标志都是回车换行符，只要查找到回车换行符就可以视为分段的标志。

2.2 文本断句

研究断句的标志首先要研究标点符号的用法，从断句的角度来考虑，可以把标点符号分为两类：句末点号、右侧标点和其他标点。句末标点包括句号、问号、感叹号，断句时首先要找到句末点号，这是句子结束最主要的形式标记。

2.3 获得标题句

经过对大量文本的研究，得出判断标题句的如下规律[1]:

(1) 标题句与正文之间必然有回车换行符。

(2) 标题长度在 100 个字节以内。

(3) 文章首个回车换行符前面视为候选标题句，如果该句中没有标点；有标点但是没有句号且标点个数少于 4 个且句末没有指定的标点(分号和句号)。

3 线性加权

3.1 加权因子量化

(1) 词频：对于词频因子这里采用公式

① 基金项目:2007 年度教育部人文社科研究基金青年项目(07JC870006)
收稿时间:2008-12-18

$$freqi = \frac{fi}{1+fi}$$

其中，fi 表示词语 i 在一篇文章中的词频。该方法也叫非线性函数方法，它使词频因子随词频的增加而逐渐上升，当词语的词频逐渐增大时，函数逐渐向 1 收敛，即词语出现的次数越多，该词作为关键词的可能性越大。同时，可能性的增长又不是线性的，当词频特别高时，基本趋于稳定，比线性方法更加符合语言的实际。

(2) 词长：对于词长权重的处理函数如下：

$$lengthi = \frac{li}{Max(li)}$$

其中，li 表示词语 i 的词长，Max (li) 表示此文本所有词语的最大长度，对一篇固定的文本来说这个值是固定的。

(3) 词性：对于文章中的词 i 来说，从 i 的词性考虑，可得到以下权值计算公式：

$$posi = \begin{cases} 0.8 & \text{若 i 为名词动词或成语} \\ 0.6 & \text{若 } i \text{ 为形容词和副词} \\ 0 & \text{若 } i \text{ 为其他词性} \end{cases}$$

(4) 位置：为了获取每个词的位置信息，需要确定记录位置信息的方式以及各个位置的词在反映主题时的相对重要性。出现在标题中的词比出现在段首和段尾中的词更能反映文献的主题，而出现在段首中的词比出现在段尾中的词在反映文献主题方面更有价值，正文中的词比重最小。因此可以利用下列公式[2]：

$$addi = \frac{10\times(w1\times 5 + w2\times 3 + w3\times 2)}{L}$$

此处对词 w 在不同位置出现的次数赋予不同权值。

w1：词在标题中出现的次数　w2：词在段首出现的次数

w3：词在段尾出现的次数　L：文档中词的总数

(5) 互联网词典来自于对 SOGOU 搜索引擎所索引到的中文互联网语料的统计分析，统计所进行的时间是 2007 年 10 月，涉及到的互联网语料规模在 1 亿页面以上。它旨在给出基于互联网语料环境的高频词对应的词频、词性信息。

$$selecti = \frac{fi}{totalwordnum} \times fg$$

其中，fi 代表该词在互联网词频，fg 代表该词在当前文本中的词频，totalwordnum 代表互联网词典中词语的总个数(括重复的)，经统计约 301,865,939,512(30 多亿)，$\frac{fi}{totalwordnum}$ 代表选择因子，它表示词语出现在互联网词典中词语的频度与互联网词典总词数之比。代表词语的互联网词频属性权值，值越大该词是关键词的概率越大。

3.2 权重计算

经过以上因素的分析和量化，我们采用线性加权的方法，将以上因素归并到下面的权重计算公式中

Wi = A × freqi+B × lengthi+C × posi+D × addi+E×selecti

其中：Wi 为词 i 在文本中的权值；A、B、C、D、E 为比例系数，用来表明各因子在加权公式中的比重。确定这些比例系数可以利用大规模的语料库进行反向推理的方法，但由于语料库的选择以及此算法本身的研究性，其并不能代表所有领域的关键词提取情况，所以可以先采用模糊处理的方法。经过试验以及对语言学的研究[3]，词频在各种属性中是最重要的赋值为 1.5，其次是词性赋值为 1.1，位置和互联网词典赋值为 1.0，最后是词长由于它对关键词提取的影响受限于分词系统分出来词的长度和准确度，故对其赋值为 0.8。

最后，计算出每个词语的权值并降序排列即可得到候选关键词。但由于比例系数确定时采用了模糊处理的方法，因此，得到的关键词只能作为候选关键词。

4 组合词生成与过滤

由于现有分词算法难以全面考虑词在上下文中的关系，所以经过分词后文档中常常出现非完整的词串(如“计算机”拆分成“计算”和“机”)或者将特定文章中联系非常紧密的一个词拆分成两部分(如将“抗震救灾”拆分为“抗震”和“救灾”)，在提取关键词的时候必须对这种非完整词串和不考虑特定语境的词串进行取舍。

本算法设计了一种基于线性加权模糊处理结果的组合词生成过滤算法。首先提取前面线性加权处理结果的一定数目的候选关键词，仅对这些候选关键词提取“相邻词”构成组合词，同时在相邻词窗口的选择上借鉴了汉语言学中词语搭配的相关研究成果。

通过以上两个原则形成组合词，再用一定的规则

进行过滤，余下的组合词作为二级候选关键词，最后将两级候选关键词进行基于规则和算法的合并即可得到关键词。

4.1 一级候选关键词表

在线性加权模块中，已经得到按照词语总权重排序后的列表，这时候需要提取一定数目的词语构成候选关键词，经过试验发现排序在列表后面的词语已基本不能代表文章的意义，也就不可能成为关键词，一级候选关键词的生成原则如下：

(1) 过滤掉非特定词性的词语(名词、动词、形容词)或者在标题句、段首、段尾均未出现且词频为 1 的词语。

(2) 设文章的总词数为 Total 关键词的提取数目为 keywordnum 则 keywordnum，则应满足：如果 Total * 5% < 20 则 keywordnum = Total * 5%；如果 Total * 5% >= 20 则 keywordnum = 20 。

通过以上两个步骤提取得到的 keywordnum 个关键词作为一级候选关键词。

4.2 组合词生成

生成组合词其实是词语搭配问题，本算法在提取组合词的过程中，采用了结合一级候选关键词词性过滤和最佳窗口搭配的组合词算法。由于组合词的生成以相邻词为基础，因此必然在同一句话中才可能形成组合词。对于文档中的每一句话(设总词数为 N)进行遍历，对每句的算法流程如下所述：

(1) 设置当前词 now 为句子中的第一个词，pre 为空，代表 now 的前临词，after 为 now 后面的词，i=0。

(2) 若 now 在一级候选关键词列表中，其前临和后临词分别为 pre 和 after；若不在转(9)

(3) 若 pre 不在一级候选关键词列表中且 pre 非空，则 pre+now 形成一个组合词，组合词有一个分词(即它的成分中仅含一个在一级候选关键词列表中的词语，下同)， 组合词的权重为 now 的权重

(4) pre+now 组合是否符合规则(见下面文字说明)，符合则查找其是否在组合词结构体链表 houxuanzuhekeyword_GlobalList 中，若在则词频加 1，不在则加入到 houxuanzuhekeyword_Global List 中

(5) 若 after 在一级候选关键词列表中且非空，则 now+after 形成一个组合词，组合词有二个分词，组合词的权重为 now 和 after 的权重之和；若 after 不在一级候选关键词列表中且非空，则 now+after 形成一个组合词，组合词有一个分词，组合词的权重为 now 的权重。

(6) now+after 组合是否符合规则，符合则查找其是否在组合词结构体链表中，若在则词频加 1，不在则加入到组合词结构体链表中。

(7) 若 pre 在一级候选关键词列表中且 after 非空，则 pre+now+after 为一个组合词，由两个分词组成，权重为 pre,now 权重值和；若 pre，after 均非空且 after 在一级候选关键词列表中，则 pre+now+after 为一个组合词，由两个分词组成，权重为 now， after 权重值和；若 pre,after 均非空且不在列表中，则 pre+now+after 为一个组合词，由一个分词组成，权重为 now 的权重值。

(8) pre+now+after 是否符合规则，符合则查找其是否在组合词结构体链表中，若在，则词频加 1，不在，则加入到组合词结构体链表中转(10)。

(9) 若 pre,after 均在一级候选关键词列表中且非空，则 pre+now+after 为一个组合词，由两个分词组成，权重为 pre,after 权重值和；然后 pre+now+after 是否符合规则，符合则查找其是否在组合词结构体链表中，若在则词频加 1，不在则加入到组合词结构体链表中。

(10) i=i+1。

(11) pre,now,after 依次向后移动一个位置。若 i>N 结束，继续下一句；否则转(2)。

上面提到了组合词的过滤规则，这是本文根据实际试验结果和对语言学的研究建立的一个规则库，该库约定了词语之间的组合搭配规则，必须满足这些规则才可能组成有意义的候选组合关键词：

①对于两个词生成的组合词，若前面词的词性为形容词，则后面词只能是名词，即“形容词+名词”，其他的形式还有“动词+副词”，“动词+名词”，“名词+名词”，“名词+ 动词”五种形式。

②对于三个词生成的组合词，组合可以是下面的一种：“名词+名词+名词”，“名词+形容词+名词”，“名词+介词+名词”，“动词+名词+名词”，“名词+名词+动词”，“名词+动词+名词”，“形容词+名词+动词”，“名词+连词+名词”。

遍历文本的每一句话并应用上述流程，处理完毕

之后即得到一个候选组合关键词链表 houxuanzuhekeyword_GlobalList ，内部包含候选关键词的各种属性信息。

4.3 组合词过滤

组合词算法的第二步就是利用统计指标筛选达到一定搭配强度的候选词语作为搭配，本算法采用了词频统计的方法进行过滤。

在组合词生成阶段已经利用词性的语法规则过滤了很多不可能成为候选关键词的词语。下面需要进一步完善过滤规则：组合词词频为 1 的；词频选择率低于 0.3 的(词频选择率=组合词词频/分词的最小词频)；遍历组合词表，若词 A 是词 B 的子集则删除 A(如组合词中两个词，“安徽省长”和“安徽省长王金山”，则删除安徽省长)。

5 关键词生成

经过以上处理，得到两级候选关键词表(线性加权过滤得到的一级候选关键词表 A 和候选组合关键词表 B)，两个词表都有候选关键词和对应的词权重，但由于表 B 是有表 A 经过基于相邻词的算法生成的，因此两表有重复的内容，通过遍历两表筛选掉重复的词条，然后按照权值从大到小排列即可得到关键词，同时，可以由用户指定关键词的数目。

6 总结

本文主要针对关键词提取，提出并了一种将分词与词性标注、文本预处理、线性加权算法、组合词生成与过滤、合并候选关键词等集成到一个完整的框架中的模型算法，进而实现了一个关键词提取系统。经过试验，算法能够提取出文章中各个“单词”和组合词的信息，对关键词提取能够达到较高的召回率。

参考文献

1 张敏,耿焕同.一种利用 BC 方法的关键词自动提取算法研究.小型微型计算机系统, 2007,28(1):189－192.

2 方俊,郭雷,王晓东.基于语义的关键词提取算法.计算机科学, 2008,35(6):148－151.

3 王灿辉,张敏. 基于相邻词的中文关键词自动抽取. 广西师范大学学报, 2007,25(2):161－164.