doi: 10.3969/j.issn.1003-4271.2014.01.28

# 基于网页特征的特征词提取技术

庞宁

(太原科技大学应用科学学院，山西 太原　030024)

**摘　要**：特征词提取是一项提炼整个 web 页面内容的实用技术，同时也为文本分类，信息抽取应用提供了技术支持. 在 web 页面内容上，利用段落间语义关系划分出网页内容的篇章结构，并以此为基础使用网页的元数据和特殊标签，设计了一个特征词的加权函数，综合考虑了词频、词长和位置因子，最后，实验对比了各类位置因子对系统的贡献度. 实验结果表明，改进方法的 $F_1$ 值比传统的 TFIDF 提取技术提高了 15.5%，其中，位置因子中的标题，关键词和摘要因素对系统的贡献最大.

**关键字**：特征词提取；网页；元数据；加权函数

**中图分类号**: TP391.1　　**文献标识码**: A　　**文章编号**: 1003-4271(2014)01-0137-05

由于计算机技术与网络的快速发展，各种信息以前所未有的速度每天在不断的生产更新，知识爆炸已经成为人类新的困惑. 如何从海量信息中提取出我们所需要的内容是自然语言处理技术中的新的焦点，因此能够有效反映文本内容的特征词提取技术受到了众多研究人员的重视，在各种相关领域中，例如，自动分类[1]、文本聚类[2]、文本过滤[3]等，都会看到该技术的应用.

## 1　相关研究

特征词可以认为是代表某类文本的类别领域词，是为了满足文献标引或检索工作的需要而从文章中萃取出来的、表示全文主题内容信息条目的单词或术语. 目前特征词提取技术大致可以分为三种：基于规则方法[4]，基于算法模型[5-6]和基于统计方法[7-9].

随着网络技术的发展，基于网页特征词提取开始受到关注. 为了尽量减少在提取过程中对文本内容结构的过分依赖，本文提出一种特征词抽取算法，除了考虑传统的词频、词长、位置等提取特征因子，还特别引入了网页元数据特征.

## 2　Web 文档的元数据

元数据在数据库领域和图书馆自动化系统中有着广泛应用. 随着对网络信息使用需求的不断增长，元数据逐渐开始用于描述 Web 文档. 目前，有些 Web 信息检索系统(如 Altavista 等 0 已经开始支持 HTML 中的 META 和 LINK 标记. 同时 Web 文档的作者也开始利用这些标记来指定若干简单的元数据(例如，东方网指定了 Description 和 keyword0. 而这些简单的元数据恰恰是特征词提取中所需要. 随着元数据使用的逐渐普及，越来越多的 Web 信息资源已经附有元数据，因此直接利用这些已有信息也是网络检索发展的趋势.

## 3　特征词提取算法

收稿日期：2013-11-22
作者简介：庞宁(1979-)，女，讲师，硕士，研究方向：自然语言处理.
基金项目：山西省自然科学基金(2012011011-4).

### 3.1 算法流程

本文研究的是基于网页元数据的一种提取算法，具体过程如图 1 所示. 首先将网页源文本利用 HTML 网页清洗技术去掉网页上的噪音，保留网页中的主题文本和超链接，利用网页上保留的重要标签信息对网页内容结构化，将其分为标题、关键词和摘要、正文、超链接，分别存储. 再利用分词软件将各部分文本分词，标注词性，仅保留文中名词和动词，这是因为特征词一般都是名词或动词，同时也避免高频虚词的干扰，第四步是将正文中的文本进行语义段落划分，即形成内容相近的若干子节，抽取各子节的子标题，进一步为提取各词的位置因子特征做准备，最后，计算各词的特征因子的值，利用权值函数，求出各词的权重，最后，按照权重值排序得到网页的特征词.

### 3.2 语义段落的生成机制

网页文本通常呈现半结构化的特点，为了更好地衡量每个候选特征词的位置因子，采用智能化的方法[10]对网页正文内容进行结构化，将内容相近的若干段落归为一个语义段落. 首先，通过计算每两个连续段落之间的语义距离来判断它们在内容上的相似程度. 假定文本任意两个连续段落 $pa_i$ 和 $pa_{i+1}$ 之间的语义相似度定义为:

$$sim(pa_i, pa_{i+1})= |pa_i \cap pa_{i+1}|/ |pa_i \cup pa_{i+1}| . \quad (1)$$

其中，$|pa_i \cap pa_{i+1}|$ 是$pa_i$ 和$pa_{i+1}$ 所具有的相同词的数目，$|pa_i \cup pa_{i+1}|$是$pa_i$ 和$pa_{i+1}$ 所有词的数目. 显然，段落相似度越大，说明二者在内容上的差异越小. 基于段落相似度，在相邻的段落上使用聚类算法进行合并. 具体而言，首先假设整篇文本是一个语义段落，从相似度最小的两个段落处断成两个新的语义段落，重复上述过程直至文本的语义段落的数目满足要求.

在各个语义段落中需要提取出一部分词代表该段的中心思想，类似于子标题的作用，做法是: 寻找在该语义段落中出现频率高的，而在其他语义段中的频率反而低的一些词借鉴TFIDF方法构造词频计算函数如下:

$$W_i=tf_i*lg((N_D+0.5)/n_i) . \quad (2)$$

其中，$tf_i$是 $t_i$ 在文本中的词频，$N_D$ 为文本中包含的所有段落数目，而$n_i$为文本中出现过词$t_i$ 的所有段落的数目. 这样，就得到$W_i$ 的一种可行的计算方法. 选取$W_i$值大的前10个代表该语义段落的子标题.

### 3.3 特征词权重的计算

#### 3.3.1 词长因子

词语的长度与词语的抽象度存在一定的联系，基本规律是词语的长度与意义具体化的关系呈反比，长度越短，意义越抽象、模糊，而通常需要更加具体的词语反映文本主题思想. 因此设计了如下的方法计算词长因子，

$$tl = len \Big/ \sqrt{(\max len)^2 + (\min len)^2} \quad . \quad (3)$$

其中 $len$ 是词 $t_i$ 的词长, $maxlen$ 为全文中最长的词长, $minlen$ 为全文中最短的词长.

#### 3.3.2 词频因子

通常研究人员更倾向于认为，在一篇文本中，高频词要比低频词更能反映主题，但事实上，词语的出现频率无法完全体现出该词对于文本分类的重要性，很多出现次数较少的专用名词反而更能反映文本的类别. 因此特别设计如下的词频因子计算方法，利用加权法克服了单纯考虑词语的出现次数的弊端.

$$tp = \sum t_w + 3\sum t_t + 2\sum t_l \quad . \quad (4)$$

其中 $t_w$ 是词 $t_i$ 在全文出现的次数, $t_t$ 是词 $t_i$ 在标题出现的次数, $t_l$ 是词 $t_i$ 在链接处出现的次数.

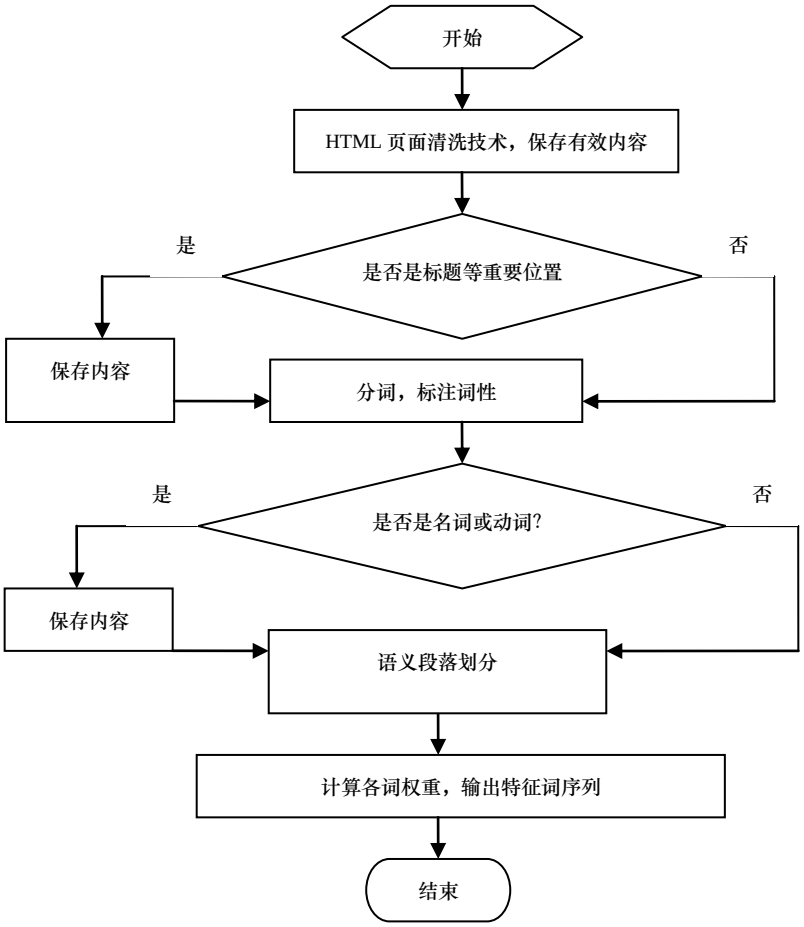


图 1　特征词提取算法流程图

Figure1　the flow chart of signature word extracting algorithm

### 3.3.3　位置因子

在文本中，不同位置上的词语所能蕴含的文本主题的作用是不同的．将网页文本按照体现主题内容的差别分为如下几种位置，见表 1.

表 1　位置因子具体描述表

Table1 specific description table of location factor

| 位置名称 | 文中的位置 | 体现主题信息量的能力 |
|---|---|---|
| 标题 | 位于网页源代码中的<title></title>之间 | 标题是要简明地将文本最有价值的内容概括提炼出，是最能反映全文的中心思想的语句. |
| 关键词、摘要 | 位于网页源代码中的元数据处, <meta name="description" content="">中""之间描述的是摘要，而<meta name="keywords" content="">是专门指向关键词的 | 摘要是文本观点的概括，是除全文以外，获得文本主要信息的重要途径，关键词是反映全文主题和最主要内容的有实质性意义的名词性术语．所以对于抽取特征词是非常重要的. |
| 超链接 | 位于<a></a>之间 | 超文本链接主要承担的是与网页主题相关近似的网页链接功能．多次出现在相关链接中的词更有可能是网页的特征词. |
| 语义段落子标题 | 位于网页正文处 | 是网页正文按照内容划分提炼出来代表各语义段落的主题思想的词语，类似于各部分子标题的作用. |
| 自然段首尾 | 位于网页源代码 之间的自然段落的段首和段尾 | 通常认为一个段落中的段首和段尾是要比段中的文字更具概括性，具有承上启下和总结本段的作用. |

为了体现出不同的位置上的词对于特征词提取结果的影响的差异，特别设计了式(40 所示的计算位置因子的函数,

$$tw=\begin{cases}\omega+\psi*f_w(t_i) & t_i\text{出现在表1所示的关键位置上}\\ 0 & t_i\text{出现在网页的其它位置上}\end{cases}. \quad (4)$$

$tw$表示词$t_i$的位置因子的计算函数，其中，$\omega$、$\psi$表示不同位置上的词语所含的信息量系数，经过大量实验，我们得到如表3-2的系数取值表，$f_w(t_i)$代表词$t_i$的信息量，具体计算公式见式(5)

$$f_w(t_i)=\frac{f_u(t_i)*\log_2(1+f_v(t_i)*l)}{\sqrt{\sum_{j=1}^{n}\sqrt{f_u(t_j)*\log_2(1+f_v(t_j)*l)}}}. \quad (5)$$

其中，$f_u(t_i)$表示词$t_i$在文本中的频数，$f_v(t_i)$表示词$t_i$的段落频数，$l$表示词长.

表 2 $\omega$、$\psi$系数取值表

Table2 factor value table of $\omega$、$\psi$

| $t_i$出现的位置 | $\omega$的取值 | $\psi$的取值 |
|---|---|---|
| 标题、关键词、摘要 | 1 | 0.5 |
| 语义段落的子标题 | 0 | 1.5 |
| 自然段落的段首和段尾 | 0 | 1.1 |
| 相关超链接 | 0 | 1.8 |

### 3.3.4 加权函数

综合上述三种特征因子，构造如下的特征词加权函数:

$$w(t_i)=2tw+tl+2tp. \quad (6)$$

其中，$w(t_i)$表示词 $t_i$ 在网页中作为特征词的权重值，而系数 2、1、2 分别用来表明位置因子(tw)，词长因子(tl)，词频因子(tp)在加权函数中的所占的比重.

## 4 实验

### 4.1 测试集和评价准则

为了避免评测时，由于测试人员的主观性带来的误差，我们选用网易网站提供的新闻网页，以该网站责任编辑自己提炼的核心提示作为评价标准，我们下载不同类别的新闻网页共 400 篇用于测试，分别计算召回率 *Recall*、准确率 *Precision*、$F_1$，以此评价实验系统的性能. 其定义如下:

*Precision*=x/y, *Recall*=x/z.

其中，x 表示系统正确识别的特征词的数目，y 表示系统所提取出的特征词总数，z 代表人工标注的全文的特征词总数.

$F_1$=(2×*Precision*×*Recall*) / (*Precision* +*Recall*).

### 4.2 实验结果与分析

分别对不同的特征因子的组合情况进行了评测对比，结果如表3所示.

表 3 特征因子组合情况表

Table3 feature factors combination table

| 系统性能 / 特征因子的选择情况 | *Recall* | *Precision* | $F_1$ |
|---|---|---|---|
| 仅加入词长、词频因子 | 18.6% | 19.3% | 18.9% |
| 加入词长、词频、位置因子 | 67.4% | 72.5% | 69.9% |
| 去掉位置因子中的超文本链接项 | 62.1% | 65.7% | 63.8% |
| 去掉位置因子中的自然段落的首尾项 | 65.4% | 71.3% | 68.2% |
| 去掉位置因子中的语义段落的子标题项 | 62.7% | 69.2% | 65.8% |
| 去掉位置因子中的标题、摘要、关键词项 | 28.1% | 37.6% | 32.2% |
| 传统的TFIDF提取方法 | 52.7% | 56.3% | 54.4% |

实验结果表明，对于网页特征词提取系统而言，仅仅依靠传统的词长、词频因子是无法满足提取需要的. 加

入位置因子可以使系统的$F_1$提高51%. 在位置因子中，各项特征对于系统的贡献度也不同，其中，去掉位置因子中的标题、摘要、关键词项会使系统的$F_1$降低37.7%，而去掉自然段落的首尾项仅会使系统降低1.7%. 与传统的TFIDF提取方法相比，添加位置因子的$F_1$提高了15.5%.

## 5 结论

本文是基于网页的标签的特征词提取，尤其是元数据和相关链接的标签，并采用自动生成语义段落的技术，将自动生成的网页内容结构结合传统的词频和词长因子，构建出一个综合多种因子的特征词计算公式，利用计算出各词的权重值求解出特征词. 该方法对文本格式无要求，实用性很广，不仅对格式规范的论文式文本有效，同样也适用于结构松散的网页文本. 但是,由于网页更新快的特点，有很多代表文本主题的关键词语没有被正确分词，从而进一步导致网页在提取特征词时的准确率降低.

**参考文献:**

[1] 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006, 17(9):1848-1859.
[2] 刘远超，王晓龙，徐志明，等.文档聚类综述[J]. 中文信息学报, 2006, 20(3):55-62.
[3] 尤文建，李绍滋，李堂秋.基于词汇链的文本过滤模型[J]. 计算机应用研究, 2003, 9:32-35.
[4] TURNEY PD. Learning to extract keyphrase from text[C]. National Research Council, Canada, 1999: 1057-1097.
[5] FRANK E, PAINTER GW. Domain-specific key phrase extraction[C]. Proceedings of the sixteenth international joint conference on artificial intelligence, Sweden, 1999: 668-673.
[6] 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.
[7] 徐建民，刘清江. 基于量化同义词关系的改进特征词提取方法[J]. 河北大学学报, 2010, 30(1):97-101.
[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): 25-30.
[9] 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.
[10] 王继成. 基于元数据的 Web 信息检索技术研究[D]. 南京：南京大学, 2000.

## Signature word extracting retrieval based on web feature

PANG Ning

(The School of Applied Sciences, Taiyuan University of Science and Technology, Taiyuan 030024, P.R.C.)

**Abstract**: Signature word extracting of the text is a useful technique which can abstract web page text, and it provides technical support for text classification, information extraction tasks. A web hierarchical structure is extracted through parsing the semantic relation between each adjacent paragraph in the web page contents. On the basis of the hierarchical structure, this paper uses the HTML metadata and special tags to design a weighting function, which is a combination of the factor of the frequency, length and location for a word. Meanwhile, an initial contrast analysis is carried out of various position factor about contributing degree to the system. Experimental results show that F1 value of improved method has increased by 15.5% than that of the traditional TFIDF extraction method. The contributing degree to the system of the title, abstract and keywords in the location factor are the largest.

**Key words:** signature word extracting; web; metadata; weighting function