第 38 卷 第 2 期
Vol.38 No.2

计 算 机 工 程
Computer Engineering

2012 年 1 月
January 2012

• 软件技术与数据库 •　　文章编号：1000—3428(2012)02—0057—03　　文献标识码：A　　中图分类号：TP393.08

信息过滤中基于统计与规则的关键词抽取研究

黄先珍[1]，杨玉珍[2,3]，刘培玉[2,3]

(1. 菏泽学院计算机与信息工程系，山东 菏泽 274015；2. 山东师范大学信息科学与工程学院，济南 250014；
3. 山东省分布式计算机软件新技术重点实验室，济南 250014)

摘　要：目前的研究大多把向量空间模型中特征项的选取与权重的计算分开，掩盖中文分词时产生的语义缺失，导致特征项区分度下降。为此，提出一种基于统计与规则的关键词抽取方法。利用句法规则提取出基本短语，以取代词袋模型中的词，考虑特征项位置、分布及语法角色等信息，综合加权计算特征项权重。实验结果表明，与现有方法相比，该方法能够更有效地进行文本信息过滤。

关键词：基本短语；合并规则；角色加权；分布加权；位置加权

Study of Keywords Extraction Based on Statistics and Rules in Information Filtering

HUANG Xian-zhen[1], YANG Yu-zhen[2,3], LIU Pei-yu[2,3]

(1. Department of Computer and Information Engineering, Heze University, Heze 274015, China;
2. School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China;
3. Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan 250014, China)

【Abstract】Currently, the items selection and calculation of weight are divided by most studies in Vector Space Model(VSM). Defects, such as the semantic vacancy of words after segmentation and low degree of differentiation based on the methods of frequency-based weight calculation, are caused. To overcome this shortcoming, a method of keywords extraction based on statistics and rules is proposed. The basic phrases are extracted by the rules of phrase syntax and instead of the words as terms in this method. Full account of feature frequency, position, distribution and grammatical role or other information, a joint feature weight function is constructed, to improve the differentiation of terms and weaken the semantic vacancy of words. Experimental results show that the keywords based on statistics and rules are more effective than others in the text information filtering.

【Key words】base phrase; merging rule; role weighted; distribution weighted; position weighted

DOI: 10.3969/j.issn.1000-3428.2012.02.018

1　概述

信息过滤是一项从动态信息流中自动获取相关信息的技术[1]。对于文本信息过滤主要用采用向量空间模型(Vector Space Model, VSM)对用户需求进行建模[2]，由词或 n-gram 项组成高维特征向量，其中，特征权重采用 TF-IDF、BM25 等方法进行估计。

文献[3]通过实验证明在英文文本集上基于词的标引是最适合文本分类一种标引方式，但对于中文此结论却不完全正确。此外，文献[4-5]也针对此问题进行了深入研究，并采取了相应的措施，但缺乏灵活性，受到应用领域的限制。不少学者对权重计算方法的研究也做了相当多的努力。文献[6]把特征项间的依赖关系引入到权重计算中；文献[7]利用特征重要程度对特征加权等。然而在权重计算的研究中，研究者重点强调频次等信息，却忽略了词的分布、词语的位置以及词语的角色等信息。

此外，目前大多把 BOW 中特征项的选取与权重的分开研究，从中相互假设彼此的正确性，这种研究不仅造成整个文本过滤系统精度下降，而且难以确认问题所在。

因此，本文在上述研究的基础上，利用句法规则抽取出基本短语，用此基本短语代替词作为 VSM 中的项，并综合特征项的频次、位置、分布及语法角色等信息，对主题词进行综合加权，构造一个联合权重评价函数，利用此函数评价特征项，最终抽取出有效词，达到提高文本信息过滤效率的目的。

2　基于统计的特征权重计算方法

2.1　特征角色权重

一个句子的主干部分是组成句子的关键。如例句 1 中“赛”、“龙舟”、“端午”、“习俗”这 4 个词，在句子中充当的语法角色不同，对成句的贡献度也各不相同。

例句 1

原句：赛龙舟是端午节的主要习俗。

切分结果：赛/v 龙舟/n 是/vshi 端午节/t 的/ude1 主要/b 习俗/n 。/wj

对例句 1 做依存文法分析，如图 1 所示。

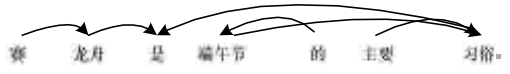

图 1　词语依存关系

由图 1 所示的依附关系可知，例句 1 的中心词为“龙舟”、“习俗”，还有一个特征区别词“端午节”，而其他的词均依附于它们构成一个完整的句子。这些词在句子中充当主

基金项目：国家自然科学基金资助项目(60873247)；山东省高新自主创新专项工程基金资助项目(2008ZZ28)

作者简介：黄先珍(1962－)，男，副教授，主研方向：智能计算，文本信息挖掘；杨玉珍，博士；刘培玉，教授、博士生导师

收稿日期：2011-07-20　**E-mail：**zscyyz@yahoo.com.cn

语、谓语或宾语，称它们为特征词语法角色，并利用语法角色对特征项加权。

对于一个句子而言，主干部分的构成一般由名词、动词及形容词构成，因此，对文本流采用倒排序的抽取中心词，即把文本流放到一个特殊栈中，制定修饰规则，并利用基本短语识别的相关规则[8]，提取出中心词。

中心词的权重体现在其他词汇对其依赖性上，因此，利用中心与其共现词汇的相对熵为中心词加权：

$$W_{role}=\sum_{i=1}^{n}p(t_i\mid t_{head})\mathrm{lb}\frac{p(t_i\mid t_{head})}{p(t_{head})} \tag{1}$$

其中，t_{head} 为中心词；t_i 为与 t_{head} 同现的有效词；$p(t_i\mid t_{head})$ 为以 t_{head} 为中心词的情形下的概率；$p(t_{head})$ 为 t_{head} 的概率。

由于对文档中的每个句子进行句法分析，提取中心词并不太现实，因此仅对标题、首尾段、首尾句进行句法分析，抽取中心词，进行中心词加权。

2.2 中心词关联加权

上述工作仅对文档中部分句子进行句法分析，对中心词加权；但是文档中仍然存在大量未进行句法分析的句子，因此，特征词的重要程度便倾向于位于首尾段与首尾句中的词项，致使特征发生偏移。因此，引入了中心词关联加权函数进行平滑。中心词关联加权是指对与中心词共同出现的有效词加权，称其为依附加权，如下所示：

$$W_{rela}=\sum_{j=1}^{n}(1/E)\times\frac{p(w_i,w_j)}{p(w_j)} \tag{2}$$

其中，$p(w_i,w_j)$ 为与中心词同现的概率；$p(w_j)$ 为词项的概率；$E=\sqrt{\frac{\sum_i^n(d_i-\bar{d})^2}{n-1}}$ 为两词同现的方差。

2.3 特征项分布加权

一个有效词在一个局部主题内分部愈均匀，愈能表达该主题，因此应该增加局部主题内分布较为均匀的项的权重。

由概率论知识，方差体现了随机变量取值的离散度，而样本方差是方差的无偏估计，因此选用如式(3)所示的统计量衡量特征在文档中的分布情况，称其为段内离散度：

$$PI_{ac}=\frac{\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(tf_i(t)-\overline{tf(t)})^2}}{tf(t)} \tag{3}$$

其中，$tf_i(t)$ 表示项 t 在第 i 句中出现的频度；m 为句子总数；$tf(t)$ 表示 t 在整篇文档出现的总频度；$\overline{tf(t)}=\frac{1}{m}\sum_{i=1}^{m}tf_i(t)$ 表示项 t 在各句中出现频度的平均值。

显然，$0\leqslant PI_{ac}$ 1。当 PI_{ac}=0 时，取得最小值，此时 t 在各个句子中均出现，最能代表文档主旨，即特征项的权重与其段内离散度呈反比，因此，对特征的分布信息权重作如下定义：

$$W_{dist}=1-PI_{ac} \tag{4}$$

2.4 特征位置加权

特征所处位置的不同，对文档的贡献度也不相同。例如特征出现在标题句等具有明显主旨的部分中，权重应该加强。定义特征的位置权重如下：

$$W_{pos}=\frac{p(t_i)}{p(t)} \tag{5}$$

其中，$p(t_i)$ 为特征出现在标题、首尾段、首尾句中的次数；$p(t)$ 为特征在整篇文档中的总次数。

同样，在文档中存在一些转折、总结性的句子，当特征词出现这些句子中出时，特征词的权重应该增加。利用同样的方法定义了特征词处于特殊句中的权重，如下所示：

$$W_{clue}=\frac{p(t_j)}{p(t)} \tag{6}$$

2.5 特征权重计算方法

通过上述加权计算，与传统 TF-IDF 方法计算所得的权重 W_{tf-idf} 相结合，形成最终的特征权重计算公式，定义如下：

$$W_i=(W_{role}+W_{rela}+W_{dist}+W_{pos}+W_{clue})\times W_{tf-idf} \tag{7}$$

3 基于规则的文本表示

对于一个句子的结构组成，其中起主要作用的为名词、动词和形容等，因此，信息过滤过程中也主要识别这些词类的短语。然而由于目前切词系统仍不够完善，因此容易产生切词歧义，如例句 2 所示。

例句 2

原句：文本过滤的过程实质上是一个文本分类的过程。

切分结果：信息/n 过滤/v 的/ude1 过程/n 实质/n 上/f 是/vshi 一个/mq 文本/n 分类/vi 的/ude1 过程/n 。/wj

例句 2 经过切分后，如“文本分类”、“文本过滤”之类的原本有特殊意义的 2 个词却被划分为 2 个不相干的词语，给分类造成一定困难。因此，有必要利用语法规则对词项进行合并，并利用合并后的短语代替 BOW 中的词，从而达到扩充语义的目的。

而短语按其长度又可分为两词短语、三词短语、多词短语，虽然短语长度越长，越能代表文本特征，但是过长的短语容易带数据稀疏的问题。因此，选取两字词或三字词的基本短语表示文本。所谓基本短语，是指具有独立语义单元的最小单位。如“现代化建设”是一个基本短语，而“长远发展的战略高度”不是一个基本短语。

图 2 描述了基本短语识别规则的制定过程。具体实现方法在文献[8]中进行了详细的阐述，由于篇幅有限，这里不再赘述。

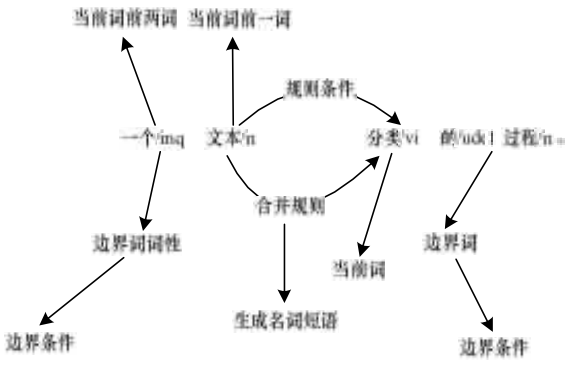

图 2 基本短语识别过程

4 实验结果与分析

4.1 评价指标

为了综合评价基于统计与规则的特征权重对过滤效果的影响，分别采用内部评价与外部评价 2 种方式综合评价实验结果。其中，内部评价采用有效词及高频词的覆盖率作为评价标准；外部评价是把抽取的关键词用于文本信息过滤，对比过滤效果，间接评价基于统计与规则的权重计算方法的有效性。

文本分类中通常使用查全率(*recall*)、查准率(*precision*)作为评估指标。对于单类别赋值，使用列联表计算，如表 1 所示。

表 1　二值分类列联表

文档	真正属该类文档数	真正不属该类文档数
判断为属该类的文档数	*a*	*b*
判断为不属该类的文档数	*c*	*d*

recall 和 *precision* 分别定义为：

$$recall = a/(a+c),\ precision = a/(a+b) \quad (8)$$

4.2　实验分析

本文实验选取复旦大学计算机与技术系国际数据库中心李荣陆提供的训练语料，20 个类别共 9 801 篇，由于军事等 14 个类别不满 700 篇，因此从中选取计算机、政治等 6 个类别作为训练语料。因为本项目主要用于信息过滤，所以又从网上自行搜集了色情、暴力 2 个类别，共组成 8 个类别共 7 541 篇文档。

4.2.1　关键词覆盖率评估

由于自然语言处理呈现面向真实语料的趋势，因此本文分别从人民网、凤凰网、新华网三大网站，随机下载了军事方面的新闻，选取 100 篇，共 7 684 个句子做关键词抽取，最后经过人工打分并与利用 TF-IDF 计算权重所抽取的关键词进行对比。

关键抽取中首选采用文献[9]提出的逻辑段落划分的方法对句子进行聚类，划分出各个局部主题。再利用本文制定的规则抽取出基本短语取代 BOW 中的词来表示文本，然后利用文献[10]改进的信息增益算法选取有效特征词，最后利用本文提出的权重计算公式计算特征项的权重，在不同压缩比下与 TF-IDF 权重计算方法进行对比，实验结果如表 2 所示。

表 2　2 种权重计算方法对比　　(%)

压缩比	有效词覆盖率		高频词覆盖率	
	本文方法	TF-IDF 方法	本文方法	TF-IDF 方法
10	27.23	22.62	46.47	34.67
20	49.10	36.45	68.45	58.45
30	62.20	51.20	81.23	70.34
40	68.40	60.11	84.51	78.45

从表 2 可见，随着压缩比的增长，有效词覆盖率和高频词的覆盖率增长的速度呈先慢后快的趋势增长，到压缩比为 40%时，有效词及高频词的覆盖程度趋于平缓。

从不同压缩比下的有效词覆盖率可以看出，利用联合权重得到有效词的覆盖率较高，并且在压缩为 30%左右时，有效词的覆盖率趋于稳定，而利用 TF-IDF 计算所得的有效词的覆盖率相对较低，在压缩比为 40%时逐渐趋于稳定。

高频词的覆盖率整体走向与有效词覆盖率走向趋势相同，综合对比，联合权重取得的关键词收敛较早，性能较好。

4.2.2　信息过滤实验结果分析

分别利用联合权重计算方法和 TF-IDF 方法计算特征权重，实验结果如图 3 所示。

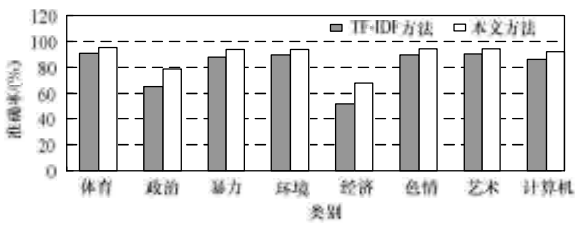

图 3　单类别准确率比较

从图 3 中可以看出，利用联合加权的方式所得到的关键词，其分类的准确率比 TF-IDF 方法在每个类别上均高出许多，这是因为 TF-IDF 方法仅靠频次衡量特征的重要性，所以在选择出特征项中不仅包含大量的有效信息，也包含了大量的重复信息，致使特征曲线较为平坦，导致最后分类精度的降低。

而本文提出的基于统计与规则的方法首先根据规则进行词项合并，再利用提取出基本短语代替 BOW 中的词，不仅增加了词项的语义描述性，而且在一定程度上消除了词项的冗余，减少了分类噪声，间接提高了特征项的区分度。其次联合权重不仅考虑了特征项的频次，而且综合考虑了特征项的分布、位置、语法角色等信息，并且利用同义词林消除同义项的同时，直接增加了特征项的区分度。因此，基于统计与规则的权重计算方法比传统的 TF-IDF 权重计算方法效果要好许多。

5　结束语

本文利用基本短语替代 BOW 中的词作为特征项，在一定程度上弥补了中文词法系统的不足。而且对特征项的位置、角色、分布等信息进行探讨，最终形成一个特征项联合权重计算方法，此方法不仅克服了单纯依靠频次抽取关键词中多冗余现象，而且增加了项的区分度，最终达到提高过滤效果的目的。

本文还存在一定的不足，如数据过度拟合、真实语料集较小等问题，这些问题对于信息过滤而言却是不忽视的。今后将深入研究特征项权重的计算方法以及噪声屏蔽等问题，以更好地提高过滤的效果。

参考文献

[1] 黄萱菁, 夏迎炬, 吴立德. 基于向量空间模型的文本过滤系统[J]. 软件学报, 2003, 14(3): 435-442.

[2] 洪　宇, 张　宇, 郑　伟, 等. 信息过滤中基于二元近似关系分布的噪声屏蔽算法[J]. 软件学报, 2008, 19(11): 2887-2898.

[3] Sebastiani F. Machine Learning in Automated Text Categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-47.

[4] 杨　震. 文本分类和聚类中若干问题的研究[D]. 北京: 北京邮电大学, 2007.

[5] 陈文亮, 朱靖波, 朱慕华, 等. 基于领域词典的文本特征表示[J]. 计算机研究与发展, 2005, 42(12): 2155-2160.

[6] Samer H, Rada M, Carmen B. Random-walk Term Weighting for Improved Text Classification[C]//Proc. of the 1st IEEE Int'l Conf. on Semantic Computing. Los Alamitos, USA: IEEE Computer Society, 2007: 242-249.

[7] 刘　赫, 刘大有, 裴志利, 等. 一种基于特征重要度的文本分类特征加权方法[J]. 计算机研究与发展, 2009, 46(10): 1693-1703.

[8] 杨玉珍, 刘培玉, 姜沛佩. 向量空间模型中结合句法的文本表示研究[J]. 计算机工程, 2011, 37(3): 58-60.

[9] 朱振方, 刘培玉, 王金龙. 一种基于语义特征的逻辑段落划分方法及应用[J]. 计算机科学, 2009, 36(12): 227-230.

[10] 杨玉珍, 刘培玉, 朱振方, 等. 应用特征分布信息的信息增益改进方法研究[J]. 山东大学学报: 理学版, 2009, 44(11): 48-51.

编辑　任吉慧