基于联合权重的多文档关键词抽取技术

杨洁，季铎，蔡东风，白宇

沈阳航空工业学院 自然语言处理研究室 沈阳 110034

E-mail: yj10141985@yahoo.com.cn

摘 要：本文对内容相近的多个文档抽取关键词。考虑到TF*IDF方法仅适于计算词语在单个文档中的权重，本文提出一种计算词语权重的方法ATF*PDF（Average Term Frequency * Proportional Document Frequency），此方法能计算词语在多文档中的权重。首先对文档词语进行权重计算选取候选关键词，然后结合词语之间的语义相似度进行关键词抽取。实验表明，此方法能有效抽取多个文档的关键词，同基于关键词的聚类标记方法相比，其准确率，召回率和F值均有较大提高。

关键词：ATF*PDF，联合权重，多文档，语义相似度

Keywords Extraction in Multi-Document Based On United Weight Technology

YANG Jie, JI Duo, CAI Dongfeng, BAI Yu

Natural Language Processing Research Laboratory, Shenyang Institute of Aeronautical Engineering, Shenyang, 110034

E-mail: yj10141985@yahoo.com.cn

Abstract: This paper extracts keywords of multi-document with close content. The method of TF*IDF can only be used for calculating words' weight of single document, the paper proposes a method that is named as ATF*PDF (Average Term Frequency * Proportional Document Frequency) for calculating words' weight of multi-document. First calculate word's weight and select candidates, and then combine with semantic similarity between words so as to extract keywords. The experimental results shows that this method can efficiently extract keywords that cover the topic of multi-document, the precision, recall and F-score are all improved compared with keyword-based cluster-labeling algorithm.

Key Words: ATF*PDF, united weigh, multi-document, semantic similarity

1 引言

互联网的迅速发展为用户提供了越来越丰富的信息，然而用户在获得有效信息的同时，也被越来越多的冗余信息所困扰，因此用户迫切需要一个可以帮助快速浏览信息的方法。其中一个方法是文档聚类，然而虽然用户可以通过聚类发现文档集合中隐含的层次结构，但却不容易理解聚类后每个文档类的主要内容，因此需要在聚类的基础上加上可以抽取多文档主题的文本抽取技术[1]。

多文档主题的表现形式有多文档文摘[2, 3, 4]和多文档关键词/关键短语等，而在实际应用中，用户更喜欢使用形式简洁的关键词/关键短语来表现多文档主题。除了可以表现多文档主题，[1]多文档关键词/短语还可以用于主题本体库建设，用户兴趣建模[5]，数字图书馆管理等方面。在多

[1]作者简介：杨洁（1983-），女，硕士，主要研究方向是自然语言处理、知识管理。

文档关键词/短语抽取技术中，国内外的研究主要技术有支持向量机方法（SVM）[6]，基于文本索引图的Core Phrase方法[7]，基于关键词的聚类标记方法[8]等。其中支持向量机方法是基于监督学习的方法，由于需要对目标文档集合给定候选关键词，并根据候选关键词对训练语料中多个文档集进行正负标记，然后在实验语料中运用支持向量机进行学习抽取关键词，因此该方法抽取关键词无法脱离用户的手工工作，只是一种半自动的关键词抽取方法；基于文本索引图的Core Phrase方法首先需要建立复杂的文本索引图，并不断更新文本索引图和候选关键短语列表，由于文本索引图结构复杂，导致建立和更新文本索引图的时间复杂度和空间复杂度较大；基于关键词的聚类标记方法首先抽取文档集合的质心向量，其中质心向量由多个候选关键词组成，然后选择质心向量中频次较大的数个词语为关键词，方法简单，易实现，被认为是多文档关键词抽取的典型方法。

本文提出了一种使用ATF*PDF （Average Term Frequency * Proportional Document Frequency）方法计算词语权重，并结合词语的相似度计算来抽取文档集关键词的方法。实验结果表明，本文方法具有较好的关键词抽取效果，准确率达到53%，召回率达到88%，F值达到62%。

本文组织情况如下：第2节介绍传统计算词语权重方法并引入本文提出的ATF*PDF方法；第3节介绍基于联合权重的关键词抽取方法；第4节是实验结果分析以及评测；最后是总结和展望。

2 基于ATF*PDF的词语权重计算

TF*IDF是一种目前比较流行的词语权重计算方法，在TF*IDF中词语的权重计算如下[9]：

$$w_{ji} = \frac{n_j}{n_{all}} * \log_2 \frac{N}{n_j} \qquad (1)$$

w_{ji}表示词语j在文档i中的权重，n_j代表词语j在文档i中出现的次数，n_{all}代表文档i中有意义的词语总数，N表示所有的文档数，n_j表示词语j有出现过的文档数。从公式(1)中可以看出TF*IDF方法仅计算词语在单个文档中的权重，并且词语权重是和词语出现的文档数成反比的，而反映多文档主题的关键词应该是在文档集的大部分文档中出现，且在大部分文档中出现次数较多的词语，词语的重要性是和词语出现的文档数成正比的。所以TF*IDF方法不适合计算词语在多个文档中的权重。另一种计算词语权重的方法TF*PDF（Term Frequency * Proportional Document Frequency）最初在ETTS（Emerging Topic Tracking System）[10]中被用于追踪互联网上用户感兴趣的信息。公式如下：

$$w_j = \sum_{d=1}^{d=D} \left| F_{jd} \right| \exp(\frac{n_{jd}}{N_d}) \qquad (2)$$

其中

$$\left| F_j \right| = \frac{F_j}{\sqrt{\sum_{k=1}^{k=K} F_k}} \qquad (3)$$

w_j为词j的权重，F_{jd}为词j在信息锥(information cone) d中的频率，n_{jd}为信息锥d中包含词j的网页数，N_d为信息锥d中的所有网页数，K为信息锥d中所有词语个数，D为信息锥个数。

由于TF*PDF方法最初用于互联网中的主题追踪时，把所有搜索引擎返回的网页根据所属的信

息锥进行分类来计算词语权重，并不适合抽取无类别信息的文档集合的关键词，因此本文提出了一种计算词语权重的ATF*PDF方法。由于集合中每个文档的大小不同，文档越大，词语在文档中出现的次数可能越多，词语在文档中的词频就越大。而本文方法文档集中词语的词频是针对整个集合的词频，为了降低文档大小对词频的影响，本文通过对词语在每个文档中的词频进行规范化求和取平均值来计算词语在文档集中的词频信息，规范方法如下：

$$|tf_{ji}| = \frac{tf_{ji}}{\sqrt{\sum_{i=1}^{n} tf_{ji}^2}} \qquad (4)$$

同时由于词语出现的文档数不同对文档集主题的反映度也不同[10]，出现的文档数多的词语对主题的反映度大于出现文档数少的词语对主题的反映度，所以pdf_i给在较多文档中出现的词语以更大的权重，为词语文档频率的指数级，

$$pdf_i = e^{\frac{n_i}{N}} \qquad (5)$$

其中，N为文档集合中所有的文档数，n_i为集合中存在词i的文档数，n为第j个文档中词语的个数，每个词只计一次，这样词语PDF值的范围从1（e^0）到2.718（e^1）。最终在ATF*PDF方法中词语i的权重w_i为

$$w_i = \frac{\sum_{j=1}^{N} |tf_{ji}|}{N} pdf_i \qquad (6)$$

由公式（6）可以看出，TF*PDF公式主要由两部分组成，一部分是词语在整个文档集合中的平均词频$\sum_{j=1}^{N} |tf_{ji}| / N$，另一部分是词语的比例文档频率$pdf_i$，其中$pdf_i$为指数级，词语的pdf值为文档频率的指数级时，相对和文档频率成线形关系时有更好的关键词抽取效果[1]。

3 基于联合权重的关键词抽取方法

本文首先对文档集进行分词，过滤停用词语处理，然后采用ATF*PDF方法来计算词语的权重，并按权重对词语进行排序，选择权重较大的词语为候选关键词。考虑到反映文档主题的关键词多为实词，本文对候选关键词进行词性标注，保留其中的名词，动词等实词。经过对实词性候选关键词进行观察，本文发现实词性候选关键词中具有语义非常相近的词语，在内容相近的文档集中，这些语义相近的词语可被认为表达同一概念，而不影响人们对文档集主题的理解。例如：现有某类多文档的实词性候选关键词列表：

保险 *养老* 人员 单位 *制度* 企业 *基本* *事业* 帮助 *基金* *社会* *保障* 社保 工作 *职工* *改革* *退休* 个人 *养老金* *试点* 职业 *今年* *管理* *农民* *待遇* *劳动* *农村* *规定* 扶助 机关 *月前* *上海* *问题* *统筹* *完善*

其中语义相近的词语有：

① *单位，企业，机关*

② 人员，个人

③ 保险，社保

在候选关键词中除了存在语义相近的词语，同时也存在同义词，例如上面候选关键词列表中“工作”和“职业”。在按词语权重排序的候选关键词列表中，这些语义相近的词语和同义词的位置可能比较靠后，在选择关键词时，这些词中的重要词语就会被遗漏。通过计算词语两两之间的语义相似度把语义相似度大于给定阈值的词语分为同一组，并按词语权重大小对每组词语排序，改变第一个词语的权重为该组所有词语的权重之和，其他词语的权重不变，本文称这种计算词语权重的方法为“联合权重”：假设有一组词语w’ ={$w_1, w_2 \cdots\cdots w_{n-1}$}，这些词语之间的语义相似度大于给定阈值，现有词语w_n，如果w_n和w’中某个词的相似度大于给定的阈值s，则将词语w_n放入词组w’中。对所有词语分组后，找出每组中ATF*PDF值最大的词语w_i，重新计算w_i的权重$weight_i=\sum weight_j$，j=1——n，其他词语权重不变。

在计算词语的相似性方面，研究者提出了多种方法，如利用语义词典（同义词词林，WordNet等）进行计算，本文借鉴刘群基于《知网》[11]的词语相似度计算方法[12]，相似度大于给定阈值的多个词语可被认为表达同一意思，被分到同一组。本文实验的算法流程如下：

输入：同一主题的文档集合={$d_1, d_2, d_3 \cdots\cdots d_n$}；
输出：权重最大的几个词语为关键词；
步骤：
1) 对文档集d进行分词，去停用词预处理；
2) 使用ATF*PDF计算词语权重，选择权重较大的词语为候选关键词；
3) 对候选关键词进行词性标注，保留其中的实词w={$w_1, w_2, \cdots\cdots w_n$}；
4) 对候选关键词w={$w_1, w_2, \cdots\cdots w_n$}分组
①选择候选关键词中的第一个词w_1作为单独的一组w_1’；
②候选关键词中未分组的任意词语w_i($i \in [1,n]$)和每个词组中的词语计算相似度；
③如果w_i和某个词组中任意一个词语的相似度大于给定的阈值，则把该词加入该词组；
④否则w_i单独放入一个新词组；
⑤重复上述过程②～④，直到候选关键词中的所有词语分组完成；
5) 对每组词语按权重排序并对权重求和sum，改变每组词语中第一个词的权重为sum；
6) 候选关键词根据权重排序；

4 实验及结果分析

4.1 实验语料

本文建立了一个较大规模的语料库，语料来自新浪网站的大约4000个网页，根据网页的所属领域把所有网页分为15大类，每个大类根据网页的主题再进一步细分为多个小类，本文每个小类中至少包含3个网页，最多50个网页，手工去除网页中的链接，导航等信息，并提取每个小类的关键词。

表1 实验语料

领域	文档类别数	领域	文档类别数	领域	文档类别数	领域	文档类别数
法治	9	旅游	8	汽车	7	社会	13
经济	12	文艺	8	体育	10	历史	9
电子	13	计算机	14	娱乐	9	教育	8
房产	11	军事	13	环境	10	总计	134

4.2 评价方法

目前，国内外尚无统一的评价方法，基本上都限于由专家人工目测，按照专家认可的程度打分，或者与已经人工标注好的测试集进行对比。本文采用第二种方法，使用准确率，召回率[13]来对实验结果进行评测，考虑到准确率、召回率反映实验不同方面的性能，本文同时采用F值来评测实验结果：

$$P=\frac{|A\cap H|}{|A|} \qquad R=\frac{|A\cap H|}{|H|} \qquad (7)$$

$$F=2\times P\times R/(P+R) \qquad (8)$$

其中P为准确率，R为召回率，A表示本文方法抽取的关键词，H为手工抽取的关键词，|A∩H|表示A，H两个集合的交集，|A|，|H|表示对应集合中包含的词语个数。

4.3 实验结果与分析

关于关键词抽取数量问题，手工标引主题时，一般用2~5个主题词，本文抽取候选关键词中的前五个实词为关键词。同时为了降低词语相似度计算的复杂度，本文按1:8的比例抽取权重最大的前40个词语为候选关键词[14]。为了确定词语相似度阈值s，本文在每个大类中抽取4个小类共60个文档集作学习样本。取阈值s范围为0.1到0.8，分别对不同s值时抽取的60个文档集的关键词求准确率平均值，召回率平均值，F值平均值，确定当s取0.5时实验结果最好，实验数据如图1所示：

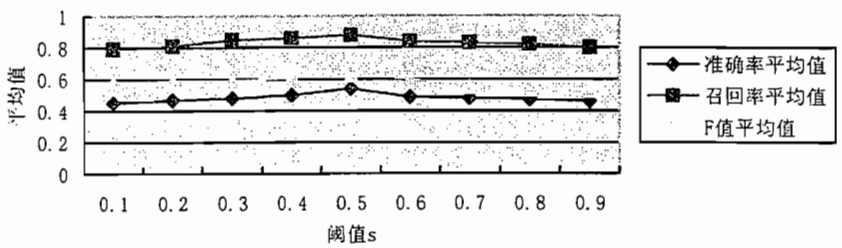

图1 60个文档集合在不同s时的准确率平均值，召回率平均值， F值平均值

本文共进行了三组实验。第一组（Baseline）为基于关键词的聚类标记方法，该方法首先计算文档集的质心向量，然后选择文档集质心中频次最大的前几个词语为关键词。第二组仅用ATF*PDF方法计算词语权重选择较大权重词语为关键词。第三组则是本文提出的使用ATF*PDF计算

词语权重，并结合词语相似度计算来选取关键词的方法。表2为三种关键词抽取方法在所有语料共134个文档集中的抽取结果。

表2 三种关键词抽取方法在所有语料共134个文档集中的抽取结果

关键词抽取方法	准确率平均值	召回率平均值	F 值平均值
基于关键词的聚类标记方法	0.47	0.58	0.54
ATF*PDF方法	0.49	0.75	0.60
基于联合权重的方法	0.53	0.88	0.62

由表2可以看出，用于计算词语权重的方法ATF*PDF的引入，有效提高了关键词抽取的性能，而词语相似度的加入使关键词抽取性能更进一步得到提高，基于联合权重的方法相对ATF*PDF方法准确率平均提高了4%，召回率平均提高13%，F值平均提高2%；相对基于关键词的聚类标记方法准确率平均提高了6%，召回率平均提高30%，F值平均提高8%。

为了观察基于联合权重的关键词抽取方法对不同领域文档集的关键词抽取效果，本文引入了针对每个领域大类的平均准确率AP，平均召回率AR，平均F值AF：

$$AP=\frac{1}{N}\sum_{i=1}^{N}P_i \qquad AR=\frac{1}{N}\sum_{i=1}^{N}R_i \qquad AF=\frac{1}{N}\sum_{i=1}^{N}F_i \tag{9}$$

N为不同领域文档集所包含的小类数。基于联合权重方法对不同领域文档集的关键词抽取结果如下：

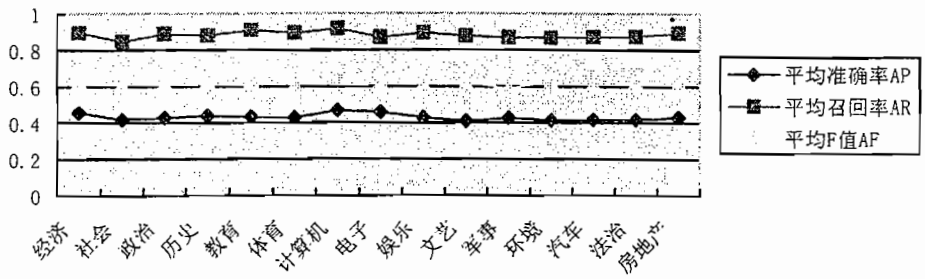

图2 基于联合权重的方法对每个领域文档集的抽取结果

从图2可以看出在不同领域使用上面三种评测方法时，各个曲线波动范围较小，本文所采用的基于联合权重方法抽取结果不受文档集所属领域的限制。

通过对实验结果和测试语料的全面分析，发现本文所采用的基于联合权重的关键词抽取方法主要存在的问题有下列几方面：

(1) 由于发现手工抽取文档集的关键词个数最多为5，所以本文对所有文档集合统一抽取5个关键词。然而由于实验语料分类粗细不均，不同文档集上手工提取的关键词个数不同，这样在很大程度上降低了抽取结果的准确率；

(2) 词语相似度计算的限制，本文采用基于《知网》的词语相似度计算，由于《知网》所收录的词汇有限，在对候选关键词列表进行调整时，由于部分词语位置的提前，那些不存在于《知网》中的重要词语由于不能同其他词语计算相似度，在候选关键词列表中的位置就被降低；

(3) 分词对关键词抽取结果的影响，本文采用哈工大分词工具分词，其中许多重要的词语由于分

词的错误没有被抽取到，例如某一文档集的关键词“权证”被分为“权-证”，显然分词后，这个词语会被作为两个单字过滤掉。同时通过对实验结果分析本文发现在134个文档集中有10个文档集由于分词的影响使得部分关键词漏选。

5 总结和展望

本文提出了一种基于联合权重的多文档关键词抽取方法，抽取结果不受文档集所属领域的限制，是比较有效的多文档关键词抽取方法。该方法使用ATF*PDF计算词语权重选取候选关键词，并且结合词语的相似度来计算词语的联合权重，提高重要词语在候选关键词中的排名。然而分词效果不好，语料分类不够精确，都影响了本文方法的抽取效果。因此本文下一步的工作是构建更加合理的实验语料，改进本文方法进一步提高关键词抽取的准确率。

参考文献

[1] Jilin Chen, Benyu Zhang, Dou Shen, Qiang Yang. Zheng Chen. Diverse Topic Phrase Extraction from Text Collection.Data Mining. 2006 .ICDM apos; 06.Sixth International Conference on Volume, Issue, Digital Object Identifier.

[2] 秦兵, 刘挺, 李生. 基于局部主题判定与抽取的多文档文摘技术. 自动化学报. 2004 年 06 期.

[3] 吴玲达, 雷震, 雷永林. 基于局部话题句群的事件相关多文档摘要研究. 计算机仿真, 2006 年 11 期.

[4] Khoo Khyou Bun, Mitsuru Ishizuka, Topic Extraction from News Archive Using TF*PDF Algorithm, The Third International Conference on Web Information Systems Engineering (WISE'02), 2002.

[5] 寇苏玲, 蔡庆生. 应用于用户兴趣建模的多文本关键词抽取研究. 计算机仿真，2007 年 02 期.

[6] Blaz Fortuna, Dunja Mladenic, Marko Grobelnik.Semi-Automatic Construction of Topic Ontology. ESWC 2005.

[7] Khaled M. Hammouda, Diego N. Matute, and Mohamed S. Kamel, CorePhrase: Keyphrase Extraction for Document Clustering, Machine Learning and Data Mining in Pattern Recognition (2005), pp. 265-274.

[8] Neto, J., Santos, A., Kaestner, C., Freitas, A. Document clustering and text summarization. In: Proc. 4th International Conference Practical Applications of Knowledge Discovery and Data Mining (PADD-2000), London, UK (2000) 41–55.

[9] Salton, G. (1991): Developments in Automatic Text Retrieval, Science, Vol 253, pages 974-979.

[10] K.B. Khoo and M. Ishizuka: “Emerging Topic Tracking System” In: Proc. Of Web Intelligent (WI 2001), LNAI 2198 (Springer), pp. 125-130, Maebashi, Japan. 2001.

[11] 董振东, 董强. 知网[EB/OL]. http://www.keenage.com, 1999-09-23/2004-03-06.

[12] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[J]. 计算语言学及中文信息处理, 2002, 7:59-76.

[13] 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.

[14] 索红光, 刘玉树, 曹淑英. 一种基于词汇链的关键词抽取方法. 中文信息学报, 2006 年 06 期.