一种高效的用于话题检测的关键词元聚类方法

杨攀[1,2]，桂小林[1,2]，田丰[1,2]，王刚[1,3]
(1. 西安交通大学电子与信息工程学院，710049，西安；2. 陕西省计算机网络重点实验室，710049，西安；
3. 西安财经学院信息学院，710100，西安)

摘要：针对基于关键词元的话题内事件检测算法运行效率不高、不适合进行大规模文本话题检测的问题，提出了一种高效的关键词元聚类算法. 该算法在进行词元簇选择时，为簇间相似度分配权值，并借鉴正态分布函数评估词元簇的个数，提高词元簇的选择精度，从而减少所需的词元聚类次数. 实验结果表明，将改进的方法应用到舆情监控的话题检测中，能在不影响检测精度的前提下有效地提高算法的运行效率.

关键词：话题检测；关键词元；舆情监控
中图分类号：TP301 **文献标志码**：A **文章编号**：0253-987X(2012)10-0024-05

Efficient Keywords Clustering Method for Topic Detection

YANG Pan[1,2], GUI Xiaolin[1,2], TIAN Feng[1,2], WANG Gang[1,3]
(1. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
2. Shaanxi Province Key Laboratory of Computer Network, Xi'an 710049, China;
3. School of Information, Xi'an University of Finance and Economics, Xi'an 710100, China)

Abstract: An improved term-committee-based event identification algorithm is presented to meet the requirements of efficiency and accuracy in public opinion monitor system, where the original event identification algorithm can not be applied due to its lower efficiency. While the similarity between the clusters is calculated, the weight is taken into consideration simultaneously. Referencing the examples from normal curve, an evaluation algorithm is proposed to help choosing cluster with a proper term number, thus the improved algorithm only needs clustering once. The experiments indicate the operating efficiency for the required accuracy.

Keywords: topic detection; term-committee; public opinion monitor

网络舆情分析的任务是将Internet庞杂的新闻按照不同的话题聚类到一起，形成热点话题，为人们提供舆论导向，并将这一过程称为话题检测，也是文本聚类中的重要技术. 常用的聚类算法包括单遍聚类算法、k均值算法、层次聚类算法，以及基于密度的聚类算法，但在文本聚类中，文本向量的高维稀疏性以及近义词、多义词等问题的存在使得聚类效果大打折扣，同时舆情分析的实时性也对聚类算法的时间复杂度提出了更高的要求. 目前，国内外许多专家学者都在研究更有效的聚类和话题检测算法.

刘涛等人[1]提出了一种无监督的特征选择算法，通过在不同的k均值聚类结果上使用有监督的特征选择方法来选出最重要的特征词. 彭柳青等人[2]基于k均值均匀效应提出的聚类初始化算法能够降低噪声干扰，但它们都需要多次的聚类，因此不能满足话题检测的时效性要求. Hao等人[3]提出增加名称词的权值思想，Kumarand等人[4]在更广的范围对文档分类，从每个类别中挑选一些关键词，增加权重，进一步对相似度进行计算，但同样也存在效率的问题. 对文本聚类的相关研究还包括文献[5-9]

收稿日期：2012-04-12. 作者简介：杨攀(1987—)，男，博士生；桂小林(通信作者)，男，教授，博士生导师. 基金项目：国家自然科学基金资助项目(61172090)；国家科技重大专项课题(2012ZX03002001-004).
网络出版时间：2012-08-27 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20120827.1111.016.html

等.

张阔等人[10]提出了基于关键词元的话题内事件检测方法,对新闻事件中的关键词进行聚类,形成词元委员会,再利用词元委员会进行文档聚类. 该方法在新闻数量较少时,能达到较好的检测效果,但在舆情系统的实际应用中,由于新闻数量往往较大,该方法的计算复杂度较高,因此并不适用. 本文对其中的词元委员会查找算法进行了改进,以降低算法的时间复杂度,使其更好地应用于舆情检测系统中.

1　基于关键词元的话题检测方法

为表述方便,先对本文要用到的概念或表述进行定义:

(1)每篇新闻是一个文本文档,进行分词后,一篇新闻文档 d 可以表示为词元 w 的集合,即

$$d=\{w_1,w_2,\cdots,w_n\}$$

(2)词元的文档集合 $D_w=\{d\mid w\in d\}$,即 D_w 是所有包含词元 w 的文档集合;

(3)词元委员会 $T_C=\{w_1,w_2,\cdots,w_k\}$ 是一组叙述某一事件关键词的集合.

文献[10]的方法主要是解决话题内事件的检测问题,分为以下 4 个步骤.

步骤 1　新闻预处理. 对文档进行分词,统计词频,并计算每个 w 在 d 中的权重 $W_{d,w}$. 一篇新闻可以表示为一个权重向量

$$\boldsymbol{n}_{\mathrm{d}}=(W_{d,w1},W_{d,w2},\cdots,W_{d,wn})$$

步骤 2　查找词元委员会. 通常一个事件总会有一些关键词频繁地出现. 文献[10]的思想正是把这些词找出来,并代表该事件. 本文提出的改进方法可降低时间的复杂度.

步骤 3　寻找核心新闻簇. 对每个 T_C 建立一个对应的核心新闻簇 C_{TC},若一个 d 包含一半以上 T_C 中的词元,则将 d 加入到 C_{TC}.

步骤 4　新闻文档分配. 将每个新闻分配到最相似的核心新闻簇,若最相似新闻小于预定义阈值 θ_3,则新建新闻簇. 相似度可利用 cosine 距离公式计算,即 $\mathrm{sim}(d_1,d_2)=\cos(\boldsymbol{n}_{\mathrm{d1}},\boldsymbol{n}_{\mathrm{d2}})$.

2　改进的关键词元聚类算法

2.1　词元委员会评估标准

词元委员会的查找算法[10]:每找出一个词元委员会,至少要进行一次层次聚类,事件越多,需要进行的聚类次数就越多,于是仅考虑进行一次层次聚类. 对得到的词元簇进行筛选,而如何选择合适的词元簇也是算法的关键问题.

用 $E_{TC}(C)$ 表示词元簇 C 的评估值,则在文献[10]中

$$E_{TC}(C)=\lg\mid C\mid S_{\mathrm{a}}(C) \tag{1}$$

式中:$S_{\mathrm{a}}(C)$ 为簇内词元相似度的平均值. 选择 $E_{TC}(C)$ 值最高的簇作为词元委员会,但词元数量的选择并非越多越好,按照聚类的目标,所选择的簇应满足簇内相似度高、簇间相似度低且簇中词元数适量的条件. 本文对簇内、簇间相似度及簇内词元数的标准重新进行了量化评估,并作为选择词元委员会的依据. 词元簇 C 的簇内相似度继续用 $S_{\mathrm{a}}(C)$ 来评估. 对簇间相似度的评估,也要用到簇间的平均相似度,定义 U_{c} 表示聚类后得到的所有词元簇 C 的集合,G_{TC} 表示已找出的词元委员会集合,C 与 G_{TC} 的相似度评估所占权重为 w_1,与其他簇的相似度评估所占权重为 w_2,簇间相似度的评估值

$$E_{\mathrm{i}}(C)=1-(w_1S_{\mathrm{a}}(C,G_{TC})+w_2S_{\mathrm{a}}(C,(U_C\backslash G_{TC}))) \tag{2}$$

式中:w_1、w_2 的取值可利用层次分析法确定. 与 $S_{\mathrm{a}}(C,(U_C\backslash G_{TC}))$ 的重要性相比,$S_{\mathrm{a}}(C,G_{TC})$ 介于稍微重要与较强重要之间,因此可取 w_1、w_2 分别为 0.8 和 0.2.

对于 C 中词元数量的评估,词元数过多或过少都不利于新闻簇的形成,因此假设合理的区间为 $[x-y,x+y]$. 若词元数落在此区间内,则评估值 $E_{\mathrm{n}}(|C|)$ 较高,且越靠近中心越高,而当词元数落在区间外时,则评估值锐减. 这种中间高、两头低的形式与正态分布曲线十分类似,因此借鉴正态函数形式,定义 $E_{\mathrm{n}}(|C|)$ 的计算公式为

$$E_{\mathrm{n}}(\mid C\mid)=\lambda\mathrm{e}^{-(|C|-x)^2/2\sigma^2} \tag{3}$$

式中:λ、σ 为待定系数. 当 $|C|=x$ 时,$E_{\mathrm{n}}(|C|)=1$,为最大值,当 $\lambda=1$ 时,σ 的取值应使 $E_{\mathrm{n}}(|C|)$ 在区间边缘处锐减,即 $E'_{\mathrm{n}}(|C|)$ 在 $x-y$ 处的值最大(或在 $x+y$ 处的值最小),又可表示为

$$E''_{\mathrm{n}}(\mid C\mid)\mid_{|C|=x\pm y}=\frac{(\mid C\mid-x)^2-\sigma^2}{\sigma^4}\mathrm{e}^{-(|C|-x)^2/2\sigma^2}\mid_{|C|=x\pm y}=0$$

$$\sigma=y$$

综上所述,对于 C 的评估,将式(1)改为

$$E_{TC}(C)=S_{\mathrm{a}}(C)E_{\mathrm{i}}(C)E_{\mathrm{n}}(\mid C\mid) \tag{4}$$

2.2　改进的词元委员会查找算法

优化词元簇的评估方法,可以只进行一次聚类,即可提高运算效率.

改进的词元委员会查找算法的输入:词元列表 $W=\{w \| D_w|>1\}$,词元数区间$[x-y,x+y]$,阈值 θ_1、θ_2. 输出:词元委员会列表 L_{TC}. 改进算法的步骤如下.

步骤 1 将所有 W 内的词元加入到集合 R 中.

步骤 2 使用凝聚的层次聚类算法对 R 内的词元进行聚类,阈值设为 θ_1.

步骤 3 聚类后得到所有词元簇 C 的集合 U_C,对每个 $C_i \in U_C$,计算 $E_{TC}(C_i)$.

步骤 4 找出 $E_{TC}(C_j)$ 最大的词元簇 C_j 并加入到 L_{TC} 中,从 U_C 中删除 C_j 以及与 C_j 相似度大于 θ_2 的簇.

步骤 5 重新计算 U_C 中剩余词元簇的 $E_{TC}(C_i)$.

步骤 6 找出 $E_{TC}(C_j)$ 最大的词元簇 C_j,若 $E_{TC}(C_j)<\theta_2$,结束,返回 L_{TC};否则,将 C_j 加入到 L_{TC} 中,从 U_C 中删除 C_j 以及与 C_j 相似度大于 θ_2 的簇.

步骤 7 若集合 U_C 为空,结束,返回 L_{TC};否则,回到步骤 5.

2.3 算法的效率分析

原算法和改进算法的计算主要集中在凝聚式层次聚类上,层次聚类的时间复杂度一般为 $O(n^3)$,如果所有文档可以抽取出 n 个词,找出 m 个簇($m<n$),原算法最多可能需要 m 次层次聚类,因此原算法的时间复杂度为 $O(mn^3)$.

改进算法较原算法减少了层次聚类的次数,只进行一次层次聚类,因此算法中步骤 2 的时间复杂度为 $O(n^3)$. 算法步骤 3 中的簇内相似度和簇间相似度已在步骤 2 中计算出,因此这一步的时间复杂度为 $O(m)$. 算法步骤 4 寻找最大值的时间复杂度为$O(m)$,算法步骤 5~步骤 7 最多涉及 $m-1$ 次迭代,第 i 次迭代的计算复杂度为$O(m-i)$,因此所有迭代的时间复杂度为 $O(m^2)$. 综上分析,改进算法的时间复杂度为 $O(n^3)$,因此改进算法比原算法的效率更高. 随着文档数的增加,词元数 n 增大,m 也会相应增大,这时改进算法的效率优势将更加明显.

3 实 验

3.1 数据集及评价标准

本文采用 reuter21578 的数据集,从中挑选 20 个新闻类别,按照文献[11]的方法标注出话题和事件,关于标注标准与细节见文献[11]. 将每个类别中的新闻文档随机分割,形成 20 个训练集和 15 个测试集. 按照文献[11]的评价标准,准确率 p_c 指 2 个聚到一类的新闻确实属于同一类的概率,召回率 r_c 指 2 个同属于一类的新闻被聚到一类的概率. 计算 p_c 和 r_c,并合并为度量(主要评价标准)

$$F_1 = 2p_c r_c/(p_c + r_c) \tag{5}$$

3.2 实验结果及分析

为了验证改进方法的性能,首先需要确定改进方法中所有参数的取值,然后将改进方法与原方法进行比较.

3.2.1 *词元数区间的确定* 本文所期望的好的词元委员会是可使同一话题下的文档尽可能多地归到该词元委员会中,因此实验就是找出当同一话题下的文档归并到某一词元委员会的数量达到最大时,该词元委员会所含的词元数量,进而确定词元数区间$[x-y,x+y]$.

取相同话题下的文档为一个训练集,按本文方法分词并对词元进行凝聚式聚类,每一次词元的凝聚过程,都将该词元簇看做词元委员会,统计其核心新闻簇的文档数 n_f,直到聚类结束. 记录 n_f 的最大值及取得最大值时词元委员会的词元数量,可能同时会有多个临时词元簇取得最大值,应分别记录他们的词元数.

在所有训练集上找出取得最大 n_f 值时所对应的词元簇的词元数,图 1 是这些词元数的概率直方图,可以看出,词元数在 3~7 时,召回率较高,由此确定词元数区间为[3,7],则 $x=5$,$y=2$. 同时,从图 1 中可以发现,词元数为 1、2 时,由于数量过少,不适合作为词元委员会,因此可以将词元数小于 3 的临时簇直接剔除,以减少计算量.

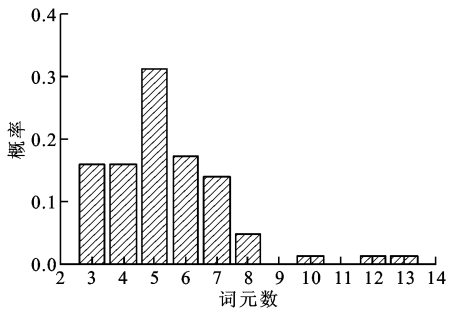

图 1 最大文档数所对应的词元数和概率

3.2.2 *阈值参数的确定* 改进的词元委员会查找算法需要阈值参数 θ_1、θ_2,分配新闻文档时需要阈值参数 θ_3,并由实验来确定这 3 个参数的取值. 分别取 θ_1 为 0.4~0.95,θ_2 为 0.2~0.7,θ_3 为 0.05~0.3,其步长均为 0.02,随机抽取 8 个训练集,来评估每次事件发现的结果.

图 2 是在 gas 训练集中 F_1 随 θ_1 的变化,由于

θ_2、θ_3 也在变化,因此相同的 θ_1 会有不同的 F_1 值,这里取 F_1 的平均值. 当 $0.7<\theta_1<0.76$ 时,F_1 的值最大且变化较平缓,当 $\theta_1>0.76$ 时,F_1 明显下降,由于 θ_1 的取值越大,凝聚聚类时的合并次数越少,效率也越高,因此将 gas 训练集中的 θ_1 取为 0.76.

采用同样的方法确定其他数据集的 θ_1,并将所有的 θ_1 取平均值. θ_2、θ_3 的确定方法也是类似的,由此确定 θ_1、θ_2 和 θ_3 的值分别为 0.74、0.39 和 0.16.

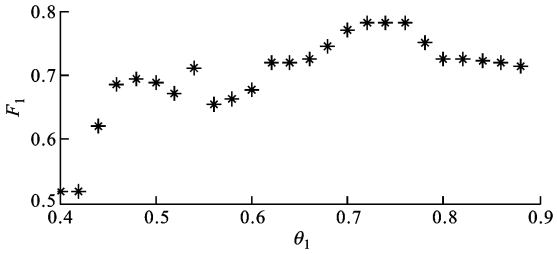

图 2　gas 训练集中 F_1 随 θ_1 的变化

3.2.3　*改进算法的性能测试*　为了测试改进算法在话题检测上的效果,在 3 个系统上进行了实验:①在 k 均值系统上采用 k 均值聚类方法;②在基于关键词元(TCB)系统上采用文献[3]所提出的基于关键词元的检测方法;③在改进的 TCB 系统上采用本文提出的改进的基于关键词元的检测方法.

对于 k 均值系统,当 k 值为实际话题数时 F_1 最高,虽然实际应用中指定的 k 值通常很难与话题数一致,但本文将其作为基线与其他系统进行比较,因此也将 k 值指定为当前测试集的话题数. 在训练集和测试集上分别对这 3 个系统的准确率、返回率、F_1 值和运行效率进行了比较,结果见表 1 和表 2.

由此可以看出,基于关键词的算法要比 k 均值算法中的 F_1 高很多,改进的算法与原算法相比,F_1 也略有提高,而且改进算法的运行效率更高. 从图 3 中可以看出,随着文本数量的增大,原基于关键词算法的运行时间几乎呈现几何级数的增长,而改进算法的增长较为缓慢. 显然,改进算法更加适用于处理文本数超过 1 000 的话题检测问题.

表 1　3 个系统在训练集上的评价

系统	p_c/%	r_c/%	F_1/%	t/ms
k 均值	48.84	42.90	45.68	160
TCB	64.52	75.56	69.61	709
改进 TCB	66.65	76.86	71.39	268

表 2　3 个系统在测试集上的评价

系统	p_c/%	r_c/%	F_1/%	t/ms
k 均值	59.96	50.39	54.76	152
TCB	63.24	73.16	67.84	413
改进 TCB	65.44	75.36	70.05	201

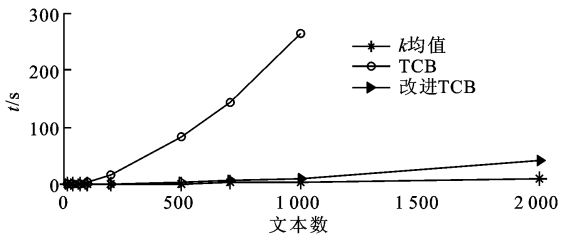

图 3　不同文本数下 3 个系统运行时间的比较

3.2.4　*改进算法的应用*　利用改进的 TCB 系统对文档集进行话题检测,形成话题集合,对每个话题中的文档再进行检测,形成粒度更细的事件,由此形成层次化的话题检测.

对 ship 测试集中的文档进行测试,检测出的话题中包括鹿特丹港口工人罢工和巴西海员罢工,可见文本方法可以将这 2 个类似的话题区分开. 对这 2 个话题进行更细粒度的检测,得到的层次化划分如图 4 所示. 可见,本文提出的方法可以对新闻文档集进行清晰有效的划分,更加方便对某方面的舆情进行重点监控.

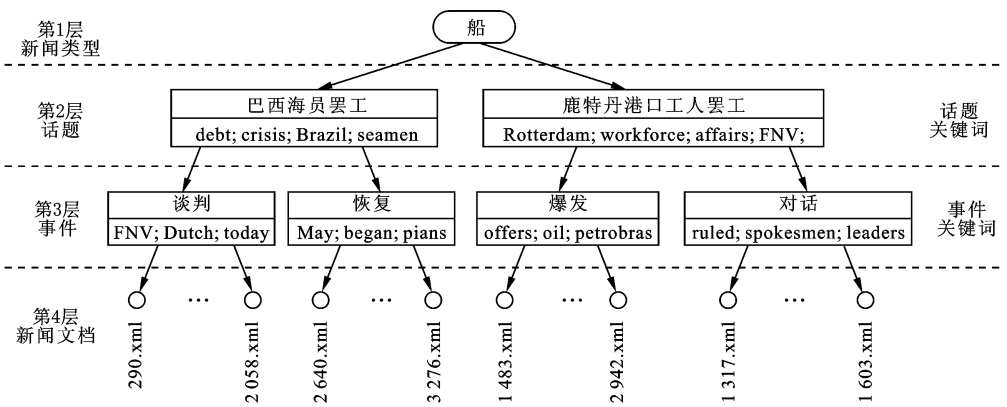

图 4　层次化的话题与事件检测

4 结 论

由于舆情监控需要高效率、高准确率和高召回率的话题检测方法，虽然关键词元的检测方法具有较高的准确率和召回率，但效率低，因此不适合进行较大数据量的话题检测. 本文对词元委员会查找算法进行了改进，重点优化了词元簇的评估方法，为词元委员会和其他簇的簇间相似度分配不同权重，并借鉴正态分布函数给出评估词元簇的计算式，最终在不降低 F_1 度量的条件下，大大地提高了算法的执行效率，为舆情监控提供了重要的技术支持.

参考文献：

[1] 刘涛，吴功宜，陈正. 一种高效的用于文本聚类的无监督特征选择算法 [J]. 计算机研究与发展，2005，42(3)：381-386.
LIU Tao, WU Gongyi, CHEN Zheng. An effective unsupervised feature selection method for text clustering [J]. Journal of Computer Research and Development, 2005, 42(3):381-386.

[2] 彭柳青，张军英，许进. 基于 k-Means 均匀效应的健壮聚类初始算法 [J]. 华中科技大学学报，2010，38(8)：73-76.
PENG Liuqing, ZHANG Junying, XU Jin. A robust algorithm for cluster initialization using uniform effect of k-means [J]. Journal of Huazhong University of Science and Technology, 2010,38(8):73-76.

[3] HAO Xiulan, HU Yunfa. Topic detection and tracking oriented to BBS [C]// Computer, Mechatronics, Control and Electronic Engineering (CMCE), 2010 International Conference on. Piscataway, NJ, USA: IEEE, 2010: 154-157.

[4] KUMARAN G, ALLAN J. Text classification and named entities for new event detection [C]// The 27th Annual International ACM SIGIR Conference. New York, USA: ACM, 2004: 297-304.

[5] 曾依灵，许洪波，吴高巍，等. 一种基于语料特性的聚类算法 [J]. 软件学报，2010，21(11)：2802-2813.
ZENG Yiling, XU Hongbo, WU Gaowei, et al. Clustering algorithm based on the distributions of intrinsic clusters [J]. Journal of Software, 2010, 21(11): 2802-2813.

[6] MOHD M, CRESTANI F, RUTHVEN I. Design of an interface for interactive topic detection and tracking [C]// Flexible Query Answering Systems 8th International Conference on. Berlin, German: Springer, 2009:227-238.

[7] HOUSHMAND M, NAGHIBZADEH M, ARABAN S. Reliability-based similarity aggregation in ontology matching [C]// Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on. Piscataway, NJ, USA: IEEE, 2010: 744-749.

[8] ALLAN J, CARBONELL J, DODDINGTON G, et al. Topic detection and tracking pilot study final report [C]// Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. San Francisco, USA: Morgan Kaufmann Publishers, 1998: 194-218.

[9] 张猛，王大玲，于戈. 一种基于自动阈值发现的文本聚类方法 [J]. 计算机研究与发展，2004，41(10)：1748-1753.
ZHANG Meng, WANG Daling, YU Ge. A text clustering method based on auto-selected threshold [J]. Journal of Computer Research and Development, 2004, 41(10): 1748-1753.

[10] 张阔，李涓子，吴刚，等. 基于关键词元的话题内事件检测 [J]. 计算机研究与发展，2009，46(2)：245-252.
ZHANG Kuo, LI Juanzi, WU Gang, et al. Term-committee-based event identification within topics [J]. Journal of Computer Research and Development, 2009,46(2):245-252.

[11] NALLAPATI R, FENG A, PENG F C, et al. Event threading within news topics [C]// Proc of the 13th ACM Conf on Information and Knowledge Management (CIKM). New York, USA: ACM, 2004: 446-453.

[本刊相关文献链接]

梁炎明，刘丁. 规则递归 T-S 模糊模型及其辨识方法. 2012，46(8)：54-58.

薛峰，周亚东，高峰，等. 一种突发性热点话题在线发现与跟踪方法. 2011，45(12)：64-69.

彭柳青，张军英. 一种鲁棒的子空间聚类算法. 2011，45(6)：13-19.

赵煜，蔡皖东，樊娜，等. 利用词汇分布相似度的中文词汇语义倾向性计算. 2009，43(6)：33-37.

张云，冯博琴. 利用标签的层次化搜索结果聚类方法. 2009，43(4)：18-21.

（编辑 管咏梅 武红江）