基于域加权聚类算法的网络舆情热点话题探测*

□ 陆伟 刘屹 孟睿 陈英杰 / 武汉大学信息资源研究中心 武汉 430072

摘要：面对自由无序的网络舆情信息，对舆情组织方式的研究体现出重要研究意义。文章提出一种网络舆情信息的组织方法，采用域加权的方式，通过一种single-pass增量算法聚类实现面向主题的舆情信息组织，即对新闻主题或新闻事件有较强表达能力的域进行加权处理以突出该主题或事件，再以无监督自动化的方式对无序的网络舆情信息进行聚类，进而发现热点话题，达到话题探测的目的。实验结果显示，聚类类簇均基于主题或事件，可以代表一个话题，F-measure评价值在85%以上，也进一步表明了本研究方法的有效性。

关键词：网络舆情，域加权，热点话题发现，聚类算法
DOI：10.3772/j.issn.1673—2286.2011.08.011

1 引言

中国互联网络信息中心（CNNIC）2010年7月发布《第26次中国互联网络发展状况统计报告》[1]，《报告》显示，截至2010年6月底，我国网民规模达4.2亿人，互联网普及率持续上升增至31.8%。网络正在成为人们获取与发布信息的主要渠道。

随着互联网的发展，网络逐渐成为舆情信息传播的主要载体，通过天津社科院的刘毅[2,3]、王来华[4,5]和毕宏音[6]、华东师范大学的许鑫[7,8]等人对网络舆情理论及应用前景的研究，笔者从舆情与舆论概念上的区别[5]出发，从信息分析的角度理解王来华教授关于舆情到舆论的转化过程[5]，可以认为舆论是对舆情进行信息组织整理后发现和得到的。

由于网络舆情的特殊性质[2]，以监督的手段进行组织是不符合实际情况的，所以笔者将在本文介绍一种无监督聚类的方法实现对舆情的组织。

2 相关工作

关于网络舆情话题发现，国内外已经取得了一定的研究成果。其中最具代表性的是话题检测与跟踪（Topic Detection and Tracking，简写为TDT）。TDT是一项面向新闻媒体信息流进行未知话题识别和已知话题跟踪的信息处理技术。James Allan[9]和Schultz[10]等人较早对话题探测与追踪问题进行研究，他们采用向量空间模型（简称为VSM）描述报道的特征空间，根据特征在文本中的概率分布估计权重，利用余弦夹角衡量报道之间的相似性。此外，Leek[11]和Yamron[12]将参与检测的两篇报道分别看作一个话题和一篇报道，采用语言模型（简称为LM）描述报道产生于话题的概率，并通过调换两篇报道的角色分别从两个方向估计它们的产生概率，最终的相关性则依据这两种概率分布。VSM和LM存在的主要缺陷在于特征空间的数据稀疏性，通常解决这一问题的方法是数据平滑技术和特征扩展技术。

Kumaran[13]、James Allan[14]、Yiming Yang[15]和Lam[16]等学者使用自然语言处理（NLP）技术辅助统计策略解决话题探测问题。其中最常用的NLP技术是命名实体识别。比如Kumaran[13]以Yiming Yang的分类方法[15]为统计框架，将报道描述成三种向量空间，分别为全集特征向量、仅包含NE的特征向量和排除NE的特征向

* 基金项目：本项目为教育部人文社会科学规划项目“专家专长智能识别与检索系统实现研究”（项目编号：09yjA870021）和教育部人文社科重点研究基地重大项目“面向细粒度的网络信息检索模型及框架构建研究”成果之一。

量。最终Kumaran对比了三种向量空间模型对新事件检测的影响，并验证NE极大地促进了事件之间的区分。

国内也有很多学者对热点话题探测与追踪问题进行研究，万小军[17]等提出了在线新闻主题探测方法，引入了创新阀值的概念，当文档与类簇向量间的相似度大于相似度阀值，但小于创新阀值时，将该文档归入该类而不用更改类簇的中心向量。

中山大学黄晓斌与赵超[18]提出了一种通过文本挖掘处理网络舆情信息，并作分析预测的方法。他们通过代理软件抓取新闻数据然后人工筛选，再利用数据挖掘工具Text Analyst进行分析，该方法的缺点是对人工筛选的要求比较严格。华东师范大学的王伟与徐鑫[19]提出一种在舆情分析中利用二次聚类的方法发现舆论热点，即通过先随机抽取网页样本集合进行聚类，选取关注的单个网页簇利用词的信息增益进行特征词抽取后，对全部网页使用抽取后的网页特征向量进行二次聚类，得到相关度较为纯粹的网络舆情网页集，从而用该网页代表一个舆论热点。南京大学的王昊与苏新宁[20]在研究中论证了条件随机场（Conditional Random Fields，CRFs）优于隐马尔可夫模型（Hidden Markov Model，HMM）和最大熵模型（Max Entropy Model，MEM），并在此基础上提出一种基于CRFs的角色标注模型。利用该模型，对新闻或论坛讨论帖的标题进行角色标注，通过对人名出现次数的统计结合人名的背景进行舆情关注点的发现。

在聚类算法上，也有很多研究者提出了改进方案。文献[21]中提出了二阶聚类的层次聚类算法，先找到每日热点簇，再利用增量聚类算法发现热点事件。文献[22,23]中结合基于密度的聚类算法和K-means算法的优点，改进了K-means算法中初始聚类中心选择的随机性问题。此外，基于主题的聚类方法也是热点新闻探测的一个研究方向，LDA、LSI等主题模型方法在新闻聚类上的应用也是当前新闻主题发现的重要研究领域。

本文中，笔者引入了TDT的思想，采用域加权的方式将主题相关域突显出来，通过聚类的方式组织舆情信息，进而达到发现网络舆情中热点话题的目的。

3 研究方法

3.1 基本聚类算法

考虑到舆情信息的动态性和时间序列特征，本文采用Single-pass增量聚类算法[24]，即预设一个聚类阀值（Clustering Threshold），按文档生成时间顺序处理输入的每篇文档，初始以第一篇文档为种子创建第一个类簇，对于每一篇输入的新文档，与以前生成的所有类簇进行相似比较，如果该文档与之前的某个类簇的相似度值大于聚类阈值时，那么该文档将属于该类簇；否则，将以该文档为种子创建一个新的主题类簇。

在计算相似度值时，本文主要采用文本信息处理中常用的余弦相似度的公式：

$$\mathrm{Sim}(D_1,D_2)=\cos\theta=\frac{\sum_{k=1}^{n}W_{1k}\times W_{2k}}{\sqrt{\left(\sum_{k=1}^{n}W_{1k}^2\right)\left(\sum_{k=1}^{n}W_{2k}^2\right)}} \quad (1)$$

其中W_{1k}和W_{2k}表示两文本向量。

3.2 确定加权域

文档中不同的域对于聚类的作用效果会有所差异，根据经验笔者选择了标题域、全文域和实体域三部分作为加权域。标题作为文本内容的直接体现，通常都包含主题的关键词或者直接表明主题，并且标题的属性决定了标题的高度概括性与简短性，可以作为加权考虑的域之一。舆情信息通常会包含5W1H（Who人物、When时间、Why目的、Where地点、What对象、How方法）中的若干项，而其中人物、地点和时间等信息对于确定主题亦具有重要作用，文献[13]的研究也说明这一点，因此实体信息也是本研究考虑加权的内容之一。

3.3 域加权方案

通过加权的方式，计算相似程度时就具有了一定的容错能力，即使某一部分权值很低甚至为0，但是通过其他部分的权值依然可以判断是否相似，是否属于同一主题或描述同一事件。并且通过对上述3种域的加权，突出了主题相关的词，使得聚类基于主题或主要描述事件。

在具体加权方式上，陆伟[25]曾提及3种计算文档权重的方法，由于第一种方法过于简单，计算过于朴素，存在不少问题；而第三种域词频加权法，笔者认

为该方法应该在监督环境下进行，对于本文需要无监督的聚类不太适用。所以本文采取第二种方案用各个域权重得分之和作为文档权重。

计算标题相似度、全文相似度以及实体相似度，再对3个相似度以一种权值组合的数值代表两文档的实际相似度，加权公式如下：

$$Sim = \alpha * Sim(title) + \beta * Sim(content) + \lambda * Sim(entity) \quad (2)$$

其中*Sim(title)*表示标题相似度，*Sim(content)*表示全文相似度，*Sim(entity)*表示实体相似度。α、β、λ分别表示其所占的权重，且 α＋β ＋λ =1.0。

*Sim(title)*即标题的相似度，由于标题信息量有限，所以对标题向量化时单独处理。在去除停用词的基础上建立标题的动态词表，由此词表建立向量，在只考虑特征词共现程度的基础上用公式(1)计算得到*Sim(title)*。

全文相似度*Sim(content)*由全文本域组成的向量经过公式（1）计算得出。

*Sim(entity)*即实体相似度，实体信息由ICTCLAS分词器[26]中实体标注功能得出，选取其中的NR（人名）、NS（地名）、NT（机构名）作为实体特征。实体相似度计算基于词共现方法，当两文本具有相同的实体特征时，实体相似度权值value提升1.0，最后用两篇文档中出现的实体最大数对求得的相似度进行归一化处理，即：

$$similar = \frac{value}{maxnum} \quad (3)$$

value为最终累和得到的权值，maxnum为两文档中共含有的实体数目。

在判断两文本相似时，本文认定1.0为完全相同，即全文相似度、标题相似度、实体相似度都为1.0。

计算相似度的方法有多种，通常在文本向量计算中使用的是余弦相似度公式，在几何中常用的则是欧几里得距离。使用余弦相似度计算时不会放大数据对象重要部分的作用，而欧几里得距离则在一定程度上放大了较大元素误差在距离测度中的作用[27]。但是通常的欧氏距离不考虑文本词数的影响，导致高频词对短文本和长文本的相似度贡献很高，这样对查准率会很低。而采用归一化的欧氏距离，就消减了文本长度的影响，但是通过公式不难推导出：

$$NormalizedEuclidean = \sqrt{2-2Cosine} \quad (4)$$

所以归一化欧式距离与余弦距离是等价的，因此本文即采用余弦相似度公式进行计算。

3.4 热点话题发现

将混乱的舆情信息用基于域加权的聚类算法组织之后，具有相同或相近特征的文章聚为一类，形成了若干类簇。由于在聚类过程中突出了新闻的主题特征，所以每一类都应当是由某一新闻主题或事件的相关新闻组成的簇。这样就可以从舆情研究的角度加以利用，对某一主题或事件的新闻数目在时间度量上的演化情况以及在信息来源（站点、博客、BBS等）上的演化情况进行分析（对在时间轴上的演化再次投影在某个来源站点上细化研究）达到热点话题探测的目的。本文认定最后聚类结果中的每个类簇为一个话题，并根据类簇的大小确定热点话题。

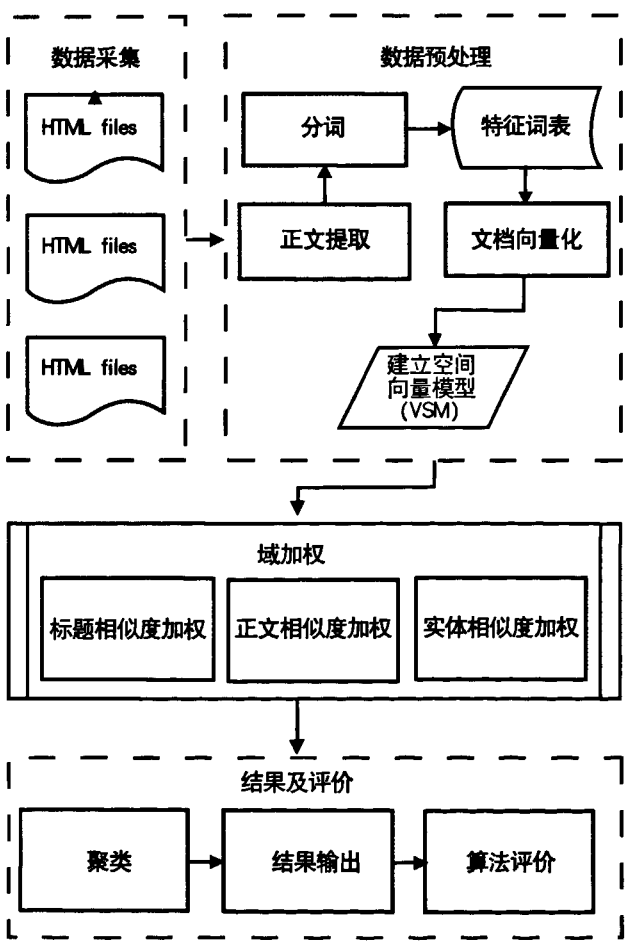

图1 系统流程图

3.5 其他细节处理

（1）抽取特征词

因为维度过高会导致“维灾难”[28]，所以要抽取特征词进行降维，缩小主题相关词的探查范围，在一定程度上避免因为向量过度稀疏或者冗余信息导致结果不好的问题。

考虑到实时抽取特征向量对聚类效率的影响，笔者通过对一个大语料库处理得到特征词列表，该特征词表存有特征词以及IDF，该IDF的计算公式[29]如下，即：

$$\mathrm{IDF}(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \qquad (4)$$

其中，N为数据集中的全部文档数，$n(q_i)$为数据集中包含特征词q_i的文档数目。

假设在生成该特征词表时使用的语料库足够大，则该特征词表可以包含所有汉语中具有语用或语义信息的词语，所以当有新数据集需要聚类时，重新生成一个特征词表与原特征词表相比特征词数量的改变极其微小，特征词的IDF值的变动极小，对结果的影响可以忽略不计。这样只需计算每篇文档中包含的特征词的TF，再从特征词表中查得IDF值，以TF*IDF为权值生成文档向量。为了突出实体（人名、机构名等）在聚类中的影响，计算TF-IDF权值时根据ICTCLAS标注出的词性，给不同的词赋予不同的权值，例如：nr（人名）权值为8、nt（机构）权值为6。在实验二中将给出实验结果，以确定静态特征词表对聚类结果的影响。

（2）建立VSM模型

根据特征词表生成词-文本向量：

$$Vector(i)=[V(1),V(2),\cdots V(m)] \qquad (5)$$

其中*V(m)*为第*m*个特征词的权值，*V(m)*的计算公式如下：

$$V(i)=tf(w_i)\times idf(w_i) \qquad (6)$$

*V(i)*为第i维的权值，w_i为第*i*维表示的特征词，*tf(w)*为词w_i在一篇文档中的词频，*idf(w)*为词w_i在一个文档集中的逆文档频。

（3）聚类结果输出

聚类完成后，为了给每个聚类进行恰当的描述，从每个聚类中抽取出一组可以表征该聚类的关键词，对于该关键词的抽取，可以在聚类中共现度最高的词中选择TF-IDF权值最高的若干词。这样用这组关键词就可以大致描述该聚类所描述的事件。

（4）聚类评价

完成上述整体过程后，为了对聚类的效果进行衡量，需要一个评价体系。本文采用外部评价法，首先抽取部分数据，手工标注出所属事件类别，根据评价结果判断聚类优劣。评价方法选用F度量值（F-measure）[30]，该方法是通过聚类前的分类标记与聚类后的聚类标记建的查全率与查准率来评价聚类效果的。即首先给抽样中的文本，按其所讲述不同事件标记不同的类别标号，在聚类之后，赋予新闻相应的聚类标号，统计聚类中某类号的数目，根据信息检索中查准率（precision）与查全率（recall）的思想来进行评价。一个聚类j及与此相关的分类i的precision与recall定义为：

$$P=\mathrm{precision}(i,j)=\frac{N_{ij}}{N_i} \qquad (7)$$

$$R=\mathrm{recall}(i,j)=\frac{N_{ij}}{N_j} \qquad (8)$$

其中N_{ij}是在聚类j中分类i的数目；N_j是聚类j中所有对象的数目；N_i是分类i中所有对象的数目。分类i的F-measure定义为:

$$F(i)=\frac{2P\times R}{(P+R)} \qquad (9)$$

对分类i而言，哪个聚类的F-measure值高，就认为该聚类代表分类i的映射。换句话说，F-measure可看成分类i的评判分值。对聚类结果来说，其中F-measure可由每个分类i的F-measure加权平均得到：

$$F=\frac{\sum[|i|\times F(i)]}{\sum|i|} \qquad (10)$$

其中|i|为分类i中的文本数。

4 实验描述与结果评价

本文测试实验所用数据是以“武汉大学”为搜索关键词在元搜索引擎中采集得到的结果。

4.1 实验一

通过对比实验，仅标题相似度、仅全文相似度、

仅实体相似度、标题+全文相似度和标题+全文+实体相似度的聚类，通过聚类评价的F-measure值以及平均查全率及平均查准率来评估聚类。

首先在原始数据中挑选了8个新闻事件，其中有：1 蚁族报告，2 武大割喉案，3 台州官员子女高考加分事件，4 传统仪式过端午节，5 华科大根叔演讲，6 预科诈骗调查，7 武汉蓝藻爆发，8汪晖学术剽窃事件，9 厅官妻子被打事件。根据事件抽取了新闻326篇，再向其中掺入无关新闻54篇，编号为0类，共组成共380篇文档的数据集，进行测试，结果如表1。

通过上述实验可以得出如下结论：

1）本文的聚类算法总体上是有效的

考虑类簇文本数的加权F值可以达到0.867，并且算数平均的查全率达到99%，查准率也有89%。

2）域加权策略取得了较好的效果

通过与仅标题、仅全文、仅实体的聚类方式进行比较，综合加权的F值分别提升了206.4%、10.8%和292.3%。当然，仅用标题和实体的效果较差，在一定程度上不具有可比性，但其比仅全文法仍有10.8%的提升，效果明显。

3）实体对文本相似性有影响，但效果不如预期

通过后两组对比实验，发现考虑实体影响仅有2%的增长，小于预期。原因可能有二，其一是全文+实体的效果已经很好，提升空间有限；其二可能在于采用实体加权时会有一些干扰实体出现，影响了效果，但具体如何仍然需要进一步实验分析。

表1 加权方案效果对比实验

类别	文档数	标题			全文			实体			标题+全文			标题+全文+实体		
		查准率	查全率	F值	查准率	查全率	F值	查准率	查全率	F值	查准率	查全率	F值	查准率	查全率	F值
0	54	0.259	0.06	0.098	0.074	1	0.138	0.944	0.135	0.237	0.074	1	0.138	0.074	1	0.138
1	64	0.953	0.263	0.412	0.953	1	0.976	1	0.169	0.29	0.984	1	0.992	0.984	1	0.992
2	70	0.771	0.232	0.357	0.971	0.66	0.786	1	0.186	0.313	1	0.986	0.992	1	0.986	0.993
3	37	0.649	0.103	0.178	1	1	1	1	0.098	0.179	1	1	1	1	1	1
4	46	0.674	0.134	0.223	0.956	0.978	0.967	1	0.122	0.217	0.957	0.978	0.966	0.957	0.978	0.967
5	30	0.567	0.073	0.129	0.967	0.967	0.967	1	0.079	0.147	0.967	0.967	0.967	0.967	0.967	0.966
6	18	0.667	0.5	0.571	1	1	1	1	0.047	0.091	1	1	1	1	1	1
7	6	0.667	0.017	0.034	1	1	1	1	0.015	0.031	0.833	0.135	0.233	1	1	1
8	23	0.696	0.069	0.125	0.957	1	0.978	1	0.061	0.115	0.957	1	0.978	0.957	1	0.978
9	32	0.5	0.571	0.533	1	0.311	0.474	1	0.084	0.156	1	0.864	0.927	1	1	1
平均		0.640	0.202	0.266	0.888	0.892	0.829	0.994	0.100	0.178	0.877	0.893	0.819	0.894	0.993	0.903
加权F值				0.283			0.782			0.221			0.849			0.867

实验结果显示，仅采用标题聚类的效果很不理想，原因可能在于：

1）标题不准确，页面提取以及文本信息预处理过程中识别不准确，标题完全错误，与主题无关。

2）标题信息量不足，标题本身概括程度太高，并且概括方式以能使读者理解为目的而形式（用词）多样，主题直接相关信息尤其是关键词较少。

3）分词器产生的影响，由于标题中关键词未识别或错误识别，导致标题的文本向量极其稀疏。

实验结果显示，标题+全文的效果比仅采用全文的效果提升明显，说明对标题进行加权是有效果的。但是在人工检验聚类结果时依然发现了一些错误，笔者认为原因在于：虽然全文是原始信息，但是不同的文本包含的信息量不同，可能包含有大量冗余信息。比如：仅包含主题关键信息的短文本与对同主题详细扩展描述的长文本计算相似度时，笔者希望相似度能大于阈值，划分为同一类，但是由于短文本生成的向量过于稀疏，关键词对主题的表达不能显现出来，从而相似度低于阈值。

此外，由于是Single-pass增量聚类算法，所以当某文本包含2个主题时，会导致这两个主题的类边界模糊，甚至混合。仅标题和仅实体的情况下就含有这种情况。

4.2 实验二

通常涉及VSM模型都需要为目标数据集生成专属的特征词表，在增量式聚类中加入新文本就需要重新生成词表。本实验特征词表的策略是静态特征词表，即由大量数据集生成一个静态的特征词表，固定IDF值，对任意数据集聚类时都利用该静态特征词表来生成文本向量。下列实验证明不同特征词表策略对结果的影响。

另标注240篇不同事件类别的文档，其中有：1 学者证实曹操墓，2 方舟子炮轰刘维宁，3 白沙洲，4 曙光学子，5 曝光体检枪手，6 全球气温上涨1.1度，7 数字鸿沟，8 易中天与李泽厚，9 北极科考。共9个事件类别，共200篇，另加入无关新闻40篇。

其中实验一380篇文档的数据集是从5082篇文档的特征库中抽取出来标注的，而实验二240篇文档的数据集是从1416篇文档的特征库中抽取出来标注的，其中6497维的特征库为两特征库的并集。

该实验表明在有足够大量文本后，得到的含有IDF值的特征词表，对增量聚类的效果影响很小，所以使用静态的特征词表对于聚类效果没有很大影响。但当新数据增加足够多后是否会有更大影响，尚不能确定。因此，笔者建议，如有可能，特征词表的IDF值等需在新文档两增加到一定量时进行更新。

5 结束语

本文采用域加权的方式，通过一种single-pass增量算法聚类实现面向主题的舆情信息聚类，实验结果表明该方法是有效的。但是，本方法仍然存在些缺陷：首先从4.1中可以看出虽然考虑到实体的影响因素，但是对于聚类评价的F值的提升并不大，所以对实体的处理方式后续仍然需要改进，如何发现文档中的关键性实体是本研究要进一步考虑的问题。此外，本方法未考虑效率问题，其在超大规模数据集上的效率是否在可接受范围内，还需要后续的实验检验。

表2 对固定IDF表的可行性验证

测试数据集大小	训练特征库大小	特征词数目	F值
240	1416	10804	0.8503
240	5082	9741	0.8434
240	6497	44824	0.8503
380	1416	10804	0.8613
380	5082	9741	0.8613
380	6497	44824	0.8667

参考文献

[1] 中国互联网络信息中心. 第26次中国互联网络发展状况统计报告[R/OL]. [2011-04-10]. .http://research.cnnic.cn/html/1279171593d2348.html.
[2] 刘毅. 略论网络舆情的概念、特点、表达与传播[J]. 理论界,2007(1):11-12.
[3] 刘毅. 内容分析法在网络舆情信息分析中的应用[J]. 天津大学学报:社会科学版,2006(4):307-310.
[4] 王来华. 舆情研究概念：理论、方法和实现热点[M]. 天津:天津社会科学院出版社,2006.
[5] 王来华. 论网络舆情与舆论的转换及其影响[J]. 天津社会科学,2008(4):66-69.
[6] 毕宏音. 网民的网络舆情主题特征研究[J]. 广西社会科学,2008(7):166-169.
[7] 许鑫,章成志,李雯静. 国内网络舆情研究的回顾与展望[J]. 情报理论与实践,2009(3):115-120.
[8] 徐鑫,张志成. 互联网舆情分析及应用研究[J]. 情报科学,2008(8):1194-1200.
[9] ALLAN J, LAVRENKO V, SWAN R. Explorations within topic tracking and detection [C]// Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic Massachusetts,2002,197-224.
[10] SCHULTZ J M, LIBERMAN M Y. Towards an universal dictionary for multi-language IR applications [C]// Topic Detection and Tracking: Event2based Information Organization. Kluwer Academic: Massachusetts,2002,225-241.
[11] YAMRON J, GILLICK L, VAN MULBREGT P, et al. Statistical models of topical content [C]// Topic Detection and Tracking: Event2based Information Organization. Kluwer Academic: Massachusetts,2002,115-134.
[12] LEEK T, SCHWARTZ R M, SISTA S. Probabilistic approaches to topic detection and t racking [C]// Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic: Massachusetts,2002,67-83.
[13] KUMARAN G, ALLAN J. Text classification and named entities for new event detection [C]// Proceedings of the SIGIR Conference on Research and Development in Information Retrieval. Sheffield, South Yorkshire: ACM,2004,297-304.
[14] ALLAN J, JIN H, RAJMAN M, et al. Topic-based novelty detection [C]// Proceedings of the Johns Hopkins Summer Workshop. CLSP,Baltimore,1999.
[15] YANG Y, CARBONELL J, JIN C. Topic-conditioned novelty detection [C]// HAND D, et al. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press,2002,688-693.
[16] LAM W, MENG H, WONG K, et al. Using contextual analysis for news event detection [J]. International Journal on Intelligent Systems, 2001,16(4):525-546.
[17] 万小军,杨建武. 在线新闻主题检测系统的设计与应用[J]. 华南理工大学学报,2004,32(11):42-46.
[18] 黄晓斌,赵超. 文本挖掘在网络舆情信息分析中的应用[J]. 情报科学,2009(1):94-99.
[19] 王伟,徐鑫. 基于聚类的网络舆情热点发现及分析[J]. 现代图书情报技术,2009(3):74-79.
[20] 王灵,苏新宁. 基于CRFs的角色标注人名识别模型在网络舆情分析中的应用[J]. 情报学报,2009(1):88-96.
[21] 刘翌莹,何婷婷. 热点事件发现及事件内容特征自动抽取研究[D]. 武汉:华中师范大学,2009.
[22] 李若鹏,李翔,林祥,等. 基于DK算法的互联网热点主动发现研究与实现[J]. 计算机技术与发展,2008,18(9):1-4.
[23] 黄宇栋,李翔,林祥. 互联网媒体信息热点主动发现技术研究与应用[J]. 计算机技术与发展,2009,19(5):1-4.
[24] PAPKA R. On-line new event detection, clustering, and tracing [D]. MA: University of Massachusetts Amherst,1999.
[25] 陆伟,夏立新. 基于okapi的XML信息检索实现研究[J]. 中国图书馆学报,2006(4):60-64.
[26] 张华平. ICTCLAS分词器[OL]. [2011-04-20]. http://www.nlpir.org/?action-category-catid-23.
[27] 张宇,刘雨东,计钊. 向量相似度测度方法[J]. 声学技术,2009,28(4):532-536.
[28] 数据挖掘导论[M/OL]. [2011-04-20]. http://book.csdn.net/bookfiles/327.
[29] ROBERTSON S E, WALKER S. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval [C]// Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 1994,345-354.
[30] 杨燕,靳蕃,KAMEL M. 聚类有效性评价综述[J]. 计算机应用研究,2008,25(6):1630-1632.

作者简介

陆伟（1974-），武汉大学信息管理系教授，博士生导师，研究方向为信息检索、Web智能挖掘、数字图书馆、知识管理。E-mail: reedwhu@gmail.com
刘屹（1990-），武汉大学信息管理与信息系统专业本科生。E-mail: whu.louis@gmail.com
孟睿（1990-），武汉大学信息管理与信息系统专业本科生。E-mail: memray0@gmail.com
陈英杰（1988-），武汉大学信息管理系情报学硕士研究生。E-mail: herochen04@gmail.com

Hot Topics Detection of Web Public Opinion Based on Field-weighted Clustering Algorithm

Lu Wei, Liu Yi, Meng Rui, Chen Yingjie / The Center for the Studies of Information Resources of Wuhan University, Wuhan, 430072

Abstract: The research of information organization shows great significance when dealing with large amount of unordered web public opinion information. In this paper, we introduce a new organization method for web public opinion. We highlight the subject by weighting text fields which are more effective to express the theme, and then deal with unsupervised clustering. By analyzing public opinion information, we realize the purpose of topic detection. In this method, the F-measure is more than 85%, which shows the effectiveness of letting clusters represent themes.

Keywords: Web public opinion, Field-weighted, Hot topics detection, Clustering algorithm