DOI: 10.3969/j.issn.1007-2861.2013.07.003

基于关联语义链网络的文本聚类方法

何 祥, 骆祥峰
(上海大学 计算机工程与科学学院, 上海 200444)

摘要: 基于关联语义链网络提出了一种自适应分裂的文本聚类方法. 该方法通过从关联语义链网络中检测出各个社团结构作为文本集中的类别, 以避免对聚类数目的预先确定. 同时, 针对高维稀疏的词向量导致的文本之间或文本与类之间相似性低的问题, 将关联语义链网络中词与词之间的关联关系映射到文本与类之间的关联关系中去, 以增强文本与类之间关系的强度. 通过与其他主要聚类方法进行实验对比, 发现该聚类方法不仅能够对文本集合进行准确的聚类, 而且能够较准确地确定聚类中心数目和识别出文本集中的话题信息.

关键词: 文本聚类; 关联语义链网络; 社区检测
中图分类号: TP 391 **文献标志码:** A **文章编号:** 1007-2861(2014)02-0190-09

Document Clustering Method Based on Association Link Network

HE Xiang, LUO Xiang-feng
(School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract: This paper proposes a document clustering method with adaptive divisions based on association link network. Instead of explicitly offering the number of cluster centers in the traditional document clustering algorithms, categories were acquired automatically by detecting the community structure in association link network. Simultaneously, with the consideration of the high-dimension and sparse word vectors that result in low similarities between the documents, the relationships were mapped between words in association link network to the relationships between the documents. Through the experimental comparisons with other clustering methods, it was found that the proposed clustering method not only obtains a high aggregation accuracy, but also are good at adaptively discovering the number of cluster centers and distinguishing categories of topics.

Key words: document clustering; association link network; community detection

文本聚类是聚类分析领域的一个重要研究分支, 是聚类方法在文本发掘中的重要应用. 传统的聚类方法在处理文本集时存在一些普遍的问题, 如聚类中心难以确定, 高维稀疏的文本向量导致它们之间的相似性过低, 以及难以利用文本中蕴含的语义关系等, 因而导致聚类效果较差.

针对以上问题, 本研究基于关联语义链网络 (一般指词汇关联语义链网络) 词汇节点之间语义关联的特性, 提出了一种自适应分裂的文本聚类方法. 本方法利用网络社区检测算法, 从

收稿日期: 2012-12-01
基金项目: 国家自然科学基金资助项目(61071110)
通信作者: 骆祥峰(1970—), 男, 研究员, 博士, 研究方向为海量网络信息处理、认知信息学与人工智能等.
E-mail: luoxf@shu.edu.cn

关联语义链网络中检测社团结构进行类的分裂, 以自动发现文本集的类信息; 在度量文本与类的相似性时, 考虑到词向量的高维稀疏性导致的相似性低的缺点, 将关联语义链网络中词与词之间的关联关系映射到文本与类的关联关系上, 以增强文本与类之间的相关性. 通过以上几个关键步骤, 本文本聚类方法不仅能够得到准确的聚类结果, 还能够准确识别文本中的类别信息.

1　相关工作

1.1　文本聚类

目前, 文本聚类主要采用传统的聚类方法或改进的传统聚类方法. 常用的方法有基于划分的聚类方法[1-4]、基于层次的聚类方法[5-7]和基于自组织映射 (self-organization map, SOM) 的聚类方法[8-10]. 通常这些方法不能或难以准确地确定聚类中心, 即不能自动发现文本集的类别信息. 其次, 传统的文本聚类方法是基于文本的向量空间模型表示, 一个文本集可能会有几十万个词来表示, 而每篇文本只有很少的词会被用到, 即高维的词特征空间存在着内在的稀疏性, 使得针对低维数据的传统聚类方法在高维空间中的聚类效果大大下降. 另外, 传统的聚类方法只是将文本中的词当成普通特征, 没有考虑文本本身特有的结构特征 (如词与词之的关联特征), 从而导致聚类准确率较低.

1.2　关联语义链网络介绍

关联语义链网络是一种基于关联语义的网络资源组织模型[11]. 词汇关联语义链网络是由文本关键词之间的语义关联链接而构成的网络. Luo 等[12]提出了一种不借助外部领域知识, 自动生成词汇关联语义链网络的方法, 即使用关联规则的方法从文本中发现文本关键词之间的关联关系. Xu 等[13]通过对词汇语义链网络的特征分析发现, 语义相关的节点具有明显的语义聚集特征.

构建关联语义链网络最重要的是正确评估节点之间的关系. 文献 [12] 构建的关联语义链网络是一种有向网络. 由于本聚类方法在从关联语义链网络中检测社区时, 只考虑了对无向网络的操作, 因此提出一种简化的公式度量两个词汇之间的无向关系. 具体地, 关于两个词 a, b 关联强度的计算公式为

$$P(a,b)=\frac{C(a,b)}{\sqrt{\mathrm{DF}(a)\mathrm{DF}(b)}}, \tag{1}$$

式中, $C(a,b)$ 表示词 a 与 b 的共现概率, 用以统计它们出现在同一篇文本中的频率; $\mathrm{DF}(a)$ 和 $\mathrm{DF}(b)$ 分别表示词 a 和 b 出现在文本中的概率, 用以统计它们在文本集中的篇频.

本研究在对文本集的聚类过程中, 不断重构类的词汇关联语义链网络, 并从中检测社区自动发现文本集的类别信息, 同时用特征词网中节点之间的关系增强文本与类的相关性, 用以解决文本与类相似性过低的问题.

1.3　词汇关联语义链网络的社团结构

网络的社团结构是大多数网络共同的特性. 所谓的社团结构, 也就是说网络通常是由若干个社区构成, 而这个社区是由多个具有相似或关联的节点组成的团所构成. 每个社区内部节点之间的连接呈很紧密的状态, 而每个社区之间的连接与社区内部节点之间的连接相对地就显得很稀疏. 在词汇关联语义链网络中, 语义相关的节点聚集在一起, 即表达同一个话题或领域的关键词连接的可能性比较大或关系强度高, 往往会聚集成一个社区; 而不同领域或话题的词汇连接的可能性低或关系强度低, 且分散在不同社区. 因此, 可以应用网络社区发现算法从词汇关联语义链网络中挖掘出不同社团的子网络(见图 1).

目前, 网络的社区检测算法非常多, 但大多数算法因过高的时间复杂度导致在实际环境中

难以被应用. 标签传播算法[14]是一种接近线性时间复杂度的网络社区检测算法, 其基本思想如下: 某个节点所属的社区标签决定于其邻居节点的标签, 即与邻居节点的关联度和所有邻居节点中出现较多的标签有关. 该算法初始赋予每个节点唯一的标签, 以自身网络结构作为导引, 不断迭代更新节点标签, 直到所有节点的标签不再变化为止. 鉴于该算法良好的时间性能以及能够自适应停止的特点, 本研究选用该算法检测词汇关联语义链网络中的社团.

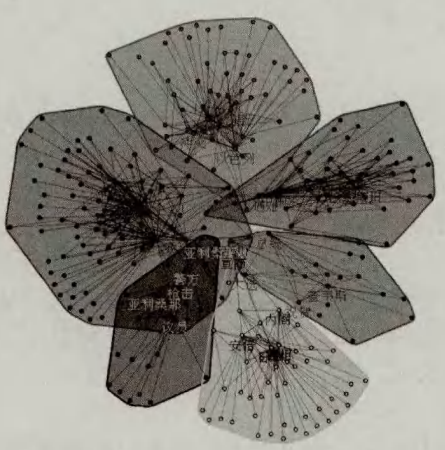

图 1 关联语义链网络的社区检测示意图

Fig. 1 Diagram of community detection from association link network

2 基于关联语义链网络的文本聚类方法

2.1 相关术语

(1) 特征词网. 一个词汇关联语义链网络, 属于某个特定的类, 其中的节点是类中关键词, 边表示两个词在所属类中的关联强度.

(2) 类. 在本研究中, 一个类不仅包含一个相似或相关文本集, 还包括一个关联语义链网络作为它的特征词网.

(3) 类的分裂. 将某个类分成多个子类的过程. 本研究采用网络社区检测的方法, 从类的特征网络中提取出多个社团的子网络. 这些子网络可作为新的子类的特征网络.

(4) 初始类. 最初将所有的文本看作一个类, 并使用文本标题集对该类构建关联语义链网络, 以词作为它的特征网络.

(5) 终止类. 再也不能分裂成多个子类的类, 即从特征词网中检测出的社区数目少于两个的类; 相反, 还能继续分裂成多个子类的类称为非终止类.

(6) 第 i 层类簇. 只有一个初始类的类簇称为第 0 层类簇; 第 $i-1$ 层类簇中的所有非终止类分别分裂得到的所有子类, 加上第 $i-1$ 层的所有终止类形成的类簇称为第 i 层类簇.

2.2 基于关联语义链网络的文本聚类过程

本聚类方法是一种分裂的层次聚类算法. 首先构建初始类, 对初始类分裂产生新的子类, 将文本映射到子类中并重构子类的特征网络, 新生成的子类继续分裂出更多的子类, 直到所有子类成为终止类或达到了指定的最大分裂层次后停止. 由于各个类可能分裂出相似的子类, 因此最后将相似类合并, 最终得到的所有类便构成了对的聚类划分. 详细步骤参见如下的算法流程.

(1) 初始化. i 表示的当前聚类的层次, cluster_i 表示第 i 层的类簇; 构建初始类, 并置

cluster_0={初始类}, 置 $i=0$, 跳到下一步执行.

(2) 创建 cluster_{i+1}. 初始化 cluster_{i+1}={ cluster_i 中的终止类}.

(3) 分裂非终止类. 对 cluster_i 中的每个非终止类 C 做如下分裂处理: ① 发现子类, 从特征词网中检测社区, 若能检测出一个以上的社团, 以此生成多个子类, 继续执行 ②; 否则标识类 C 为终止类, 跳出分裂过程; ② 映射文本到子类, 计算 C 中的文本与各个子类的相关度, 将文本映射到最相关的子类中去; ③ 删除文本数少于一定数目的子类, 并将这些类中的文本标志为未分类状态; ④ 重构子类的特征词网, 使用子类中映射到的文本集重构它的特征词网; ⑤ 将新生成的子类添加到 cluster_{i+1} 中.

(4) 映射未分类文本. 计算未分类文本与 cluster_{i+1} 中新子类的相关度, 将文本映射到最相关的子类中去.

(5) 终止条件. $i+=1$, 若 i 小于设定的最大分裂层次, 或没有新的子类产生, 跳到下一步; 否则跳回步骤 (2) 执行.

(6) 合并相似类. 计算 cluster_i 中的两个类之间的相似度, 若两个类之间的相似度小于一定阈值, 合并两个类.

(7) 输出 clusers_i 作为最终的类簇.

2.3　初始类的特征词网构建

初始类分裂的目的是发现文本中主要的骨干类, 它是文本聚类过程中最为关键的一步. 通常, 每个子类的特征网络都是从上层父类的特征网络中分裂而来. 然而, 初始类的特征词网比较特殊, 由于它没有父类, 需要从所有文本中构建出来, 因此应用关联语义链网络的构建方法从文本集中构建出它的特征词网.

由于网络的社区发现过程是一个比较耗时的过程, 如果从文本正文构建关键语义链网络, 则过多的正文关键词将会导致整个算法的性能低下. 在通常情况下, 文本标题概括了文本的主要内容, 标题中的词汇也必定是体现文章主要内容的词汇; 同时, 标题的词汇量少, 整个词向量的维度低. 因此, 相比从正文中构建关键词网络, 直接从标题中构建关键词网络不仅能够减少运算时间, 而且能有效地避免正文噪音词汇的干扰, 从而使后续的网络社区发现效果更好.

2.4　文本与类的相关度计算

文本的映射是将文本分类到最相关的类中去, 计算文本与类的相关度是该步骤的核心. 如果将文本和类都分别看成是词的特征向量, 则通常采用两个词向量的夹角余弦来度量两个词向量的相关性. 但是该方法在词特征比较稀疏的情况下, 相似度的过低导致了文本不能分类. 为了增强相关度, 本研究结合了类的特征词网增强文本与类的关系权重.

由于类的特征网络揭示了词与词之间的关联关系, 因此, 可以利用词与词之间的关联关系映射到文本与类之间的关联性. 具体地, 对于文本 d 和类 C, 其关联性可表示为

$$w_{\text{ass}}(d,C)=\sum_{i\in d}\sum_{j\in C}w(i,d)w(j,C)P_C(i,j), \tag{2}$$

式中, $w(i,d)$ 表示词 i 在 d 中的权重; $P_C(i,j)$ 表示词 i 与 j 在类 C 中的关联强度; $w(j,C)$ 表示词 j 在类 C 中的权重, 用特征词网中节点的强度 (即带权节点的度) 归一化表示. 最后, 将文本 d 与类 C 的相关度表示为相似性和关联性, 即

$$w(d,C)=w_{\text{sim}}(d,C)+w_{\text{ass}}(d,C), \tag{3}$$

式中, $w_{\text{sim}}(d,C)$ 表示文本之间的 d 与 C 的相似性, 用文本与类的词向量的夹角余弦表示.

2.5 特征词网的重构

重构类的特征词网是指加强对本类有重要意义的词对的关联性或减弱不重要的词对的关联性, 从而使得特征网络能更加精确地表达本类的局部信息. 本研究综合了词与词之间已有的关联度、在本类中的关联概率以及在全局文本中的关联概率, 提出了一种词与词之间关系的计算公式.

对于所有文本, 基于上述方法从中挖掘词的关联关系, 并构建全局的关键词关联语义链网络. 在该词网络中, 两个节点 a 和 b 表示为 $R_G(a,b)$. 对于每个类 C, 使用式 (2) 构建的本地词汇的关联语义链网络, 其中节点 a 和 b 的关联度表示为 $P_C(a,b)$, a, b 在类 C 中的关联度的更新公式为

$$P_C(a,b) = (R_C(a,b) + P'_C(a,b)) \frac{R_C(a,b)}{R_G(a,b)}, \tag{4}$$

式中, 词 a, b 的关联度可分为两部分的乘积: ① 从父类中继承而来的关联度 $P'_C(a,b)$ 与在本类文本集中计算得出的关联度 $R_C(a,b)$ 之和; ② $R_C(a,b)$ 与在所有文本的关联强度 $R_G(a,b)$ 的比值, 即如果在本类文本中关联概率越大而在全局关联概率低, 则说明该对词在本类的关联强度大.

3 实 验

3.1 实验说明

本研究从腾讯新闻网站的国际专题频道采集了 2012 年发生的 11 个热门话题作为主要的实验数据集, 各个话题的内容及文本数如表 1 所示. 对这些网页提取正文、分词, 保留动词和名词, 使用 TF-IDF (term frequency-inverse document frequency) 公式提取文本关键词, 最终将每篇网页表示成一个标题关键词的集合和一个正文 (带权重的) 关键词的集合.

表 1 实验数据集

Table 1 Experimental dataset

编号	新闻话题类	文本数
1	飓风“桑迪”	220
2	美中情局长丑闻	51
3	穆巴拉克陷入昏迷	40
4	巴基斯坦客机坠毁	40
5	尼日利亚客机坠毁	59
6	委内瑞拉总统选举	39
7	多国爆发反美示威	132
8	伊朗地震	47
9	以色列空袭加沙	198
10	亚利桑那州枪击案	80
11	2012年日本大选	316

3.2 文本聚类的对比验证

目前, 评价聚类准确性的方法一般为比较聚类算法得到的类簇与标注类簇的相似性或距离的方法, 即如果聚类算法得到的类簇与实际类簇的相似度越高或距离越小, 则聚类准确性就越高. 评估类簇的相似性一般采用以下 3 种指标.

(1) 类簇之间的互信息[15-16]是一种基于信息论的评价指标. 两个类簇的互信息量越大, 说明越相似. 本实验中使用了修正后的归一化互信息指标 (index of normalized mutual information, ANMI).

(2) 元素对计数的 F-measure 方法[17]是将 F-measure 应用到元素对的集合中去. 该指标将精度定义为正确聚类的元素对与聚类结果中所有元素对的比值; 召回率为正确聚类的元素对与标注类簇中所有元素对的比值.

(3) V-measure[18]是近年来提出的一种基于条件熵的聚类评价方法. 它通过定义 2 种聚类类簇的一致性 (homogeneity) 和完整性 (compeleteness) 来计算它们之间的相似性.

为了与其他聚类算法进行对比, 本研究使用了 3 种基本聚类算法: K-means 聚类算法[1]、基于凝聚的层次聚类算法[5]以及 SOM 聚类算法[8]对本实验数据集进行聚类. 通过调整其他聚类算法的参数 (如计算聚类个数、向量的距离等), 分别得到了它们各自最佳的聚类效果 (见表 2). 从表中可知, 层次聚类和 SOM 聚类算法的聚类效果比较差, 这是因为文本的词向量过于稀疏. K-means 算法的聚类准确度在指定聚类数为 8 时达到了最佳聚类效果. 本算法基于标签传播的网络社区检测算法因其固有的随机性导致每次聚类结果都有所偏差, 但从平均值和对应的标准差可以看出, 各种指标值波动很小, 聚类比较稳定. 本聚类方法在聚类准确率上接近于 K-means 聚类的最佳效果, 同时能自适应地发现文本集中大致的类别数.

表 2　各聚类在最佳情况下的评价指标值

Table 2　Evaluation index values in optimal condition

算法	类数	ANMI	F-measure	V-measure
K-means	8	0.832	0.887	0.871
层次	42	0.576	0.447	0.69
SOM	39	0.271	0.271	0.36
本算法	12.5±1.9	0.821±0.038	0.872±0.051	0.848±0.024

为了证明不断分裂能提高聚类准确率, 本实验随机挑选了一组实验结果观察本聚类方法在各层分裂后的聚类评价指标值(见图 2). 虽然第一层分裂后聚类效果比较差, 但随着逐层地分裂, 在类数目越来越接近实际类数目时, 各项评价指标显示聚类越来越准确. 可见, 合并相似类确实能使聚类的准确率有所提高.

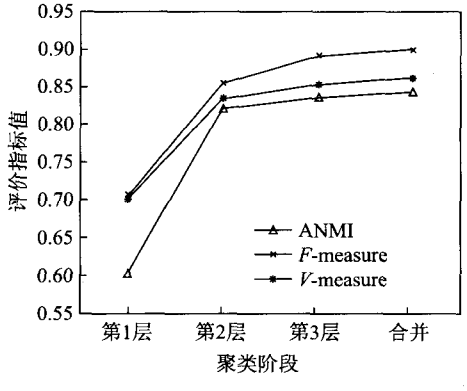

图 2　本聚类方法在不同阶段下聚类准确率的变化趋势

Fig. 2　Changing trend of clustering accuracy on different phases of our method

3.3　类的识别效果

本研究从各个类的特征词网中提取出重要的关键词, 并查看各个类的关键词能否真实地反映出实际类.

表 3 显示了某组实验结果从初始特征词网中发现的 6 个社区以及各个社区的前 5 个重要的关键词. 与表 1 比较可见, 这 6 个社区分别代表了实验数据集中的 6 个主要话题, 说明了基于

社区的检测方法确实能够较好地自动地发现文本集中的类别信息.

表 3 初始类的关联语义链网络中发现的社区

Table 3 Communities discovered from the initial cluster

编号	重要关键词				
1	美国	枪击	议员	利比亚	奥巴马
2	飓风	桑迪	纽约	取消	海岸
3	日本	安倍	选举	众院	内阁
4	穆巴拉克	昏迷	埃及	转院	临床
5	尼日利	亚客机	坠毁	巴基斯坦	中国
6	加沙	以色列	哈马斯	空袭	轰炸

对最终形成的类簇, 从每个类的特征词网中提取出前 10 个重要关键词, 如图 3 所示, 其中左边为聚集类的文本数和关键词, 右边为新闻文本的话题类. 通过观察关键词手动地对它们建立对应关系, 可以看出, 聚集成的 9 个类分别对应了 9 个实际话题, 其中话题"巴基斯坦客机坠毁"没有表现出来, 通过进一步查看正文后发现, 由于它与话题"尼日利亚客机坠毁"极为相似, 文本易产生混淆; 话题"委内瑞拉总统选举"没有被提取出来, 这是因为它包含的文本数 (39 篇) 过少, 它的绝大多数文本被混在较大的话题"2012 日本大选"中. 总之, 本聚类方法能够准确地识别文本集中的类.

图 3 本聚类方法聚集而成的类与实际话题类的对比

Fig. 3 Comparison clustering results by our method with actual clusters

为了进一步检测本方法对类的聚集能力, 本研究通过向原实验数据集中添入无关的噪音文本来验证它的抗干扰能力. 具体步骤如下: ① 从腾讯新闻各个版面上采集网页文本, 随机抽取指定数目的文本插入到实验数据集中去; ② 对这些文本聚类; ③ 聚类完成后, 剔除噪音文本, 查看原文本集是否还具有高的聚类准确率. 从上述实验可知, 由于 3 种聚类评价指标表现趋势一致, 因此本研究简单选用了一种 ANMI 指标. 表 4 显示了依次加入各种不同比例的噪音数据后, 本聚类方法和 K-means 聚类方法 (K 值设定为 7~12) 在多次实验后产生的聚类结果的 ANMI 均值. 从表中可以看出, 随着噪音文本的增加, K-means 方法的聚类准确率有较大的降低; 而噪音文本对文本提出的聚类方法在聚类准确率上并没有造成多大影响. 因此, 本聚类方法具有较好的抗干扰能力.

表 4 加入不同比例噪音数据后, 本聚类方法和K-means方法产生的聚类结果的ANMI均值

Table 4 Average ANMI on our method and K-means when mixed with different proportions of noises

	本聚类方法	K-means
无噪音	0.843	0.838
5%噪音	0.846	0.767
10%噪音	0.831	0.745
15%噪音	0.835	0.733
20%噪音	0.821	0.714

3.4 时间性能分析

最后, 对本聚类方法进行时间性能分析. 这是针对不同规模的文本集分别应用基于关联语义链网络的文本聚类方法进行聚类. 图 4 所示为在不同文本数目下聚类的平均运行时间. 从拟合的曲线函数可以看出, 本聚类方法具有较好的时间复杂性.

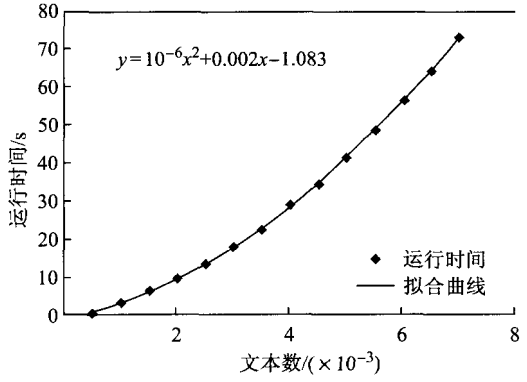

图 4 在不同文本数目下聚类的运行时间

Fig. 4 Running time of clustering different numbers of documents

4 结　论

关联语义链网络是一种将资源进行语义关联的网络资源组织模型. 本研究将词汇关联语义链网络应用到文本聚类中, 并以此作为类的特征词网, 提出了一种自适应分裂的文本聚类方法. 本聚类方法的主要特点有如下几点.

(1) 不需要人工指定类别数, 只需通过从关联语义链网络中检测社团结构以自动发现文本集中的类, 从而避免了传统聚类算法对聚类中心数目的确定.

(2) 利用关联语义链网络中词与词之间的关联关系增强了文本与类的相关度, 从而在一定程度上解决了高维稀疏词向量之间相似度低的问题.

(3) 通过与其他聚类算法的实验对比, 可知本聚类方法在保持高聚类准确率的同时, 还在话题的识别效果上具有较大的优势.

(4) 通过实验方法验证了本聚类方法的运行时间具有可伸缩性, 且与文本数成近线性比例关系.

参考文献:

[1] MacQueen J. Some methods for classification and analysis of multivariate observations [C]// Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967:

14.

[2] 张世博. 基于优化初始中心点的 K-means 文本聚类算法 [J]. 计算机与数字工程, 2011, 39(10): 30-31.

[3] 张霞, 王素贞, 尹怡欣, 等. 基于模糊粒度 K-means 文本聚类算法研究 [J]. 计算机科学, 2010, 37(2): 209-211.

[4] 汪中, 刘贵全, 陈恩红. 一种优化初始中心点的 K-means 算法 [J]. 模式识别与人工智能, 2009, 22(2): 299-304.

[5] DEFAYS D. An efficient algorithm for a complete link method [J]. The Computer Journal, 1977, 20(4): 364-366.

[6] FUNG B C, WANG K, ESTER M. Hierarchical document clustering using frequent itemsets [C]// Proceedings of the SIAM International Conference on Data Mining. 2003: 59-70.

[7] 常鹏, 冯楠, 马辉. 一种基于词共现的文档聚类算法 [J]. 计算机工程, 2012, 38(2): 213-214.

[8] BAKUS J, HUSSIN M, KAMEL M. A SOM-based document clustering using phrases [C]// Proceedings of the 9th International Conference. 2002: 2212-2216.

[9] ROMERO F P, PERALTA A, SOTO A, et al. Fuzzy optimized self-organizing maps and their application to document clustering [J]. Soft Computing-A Fusion of Foundations, Methodologies and Applications, 2010, 14(8): 857-867.

[10] 张立文, 徐家宁, 李进, 等. 基于免疫网络和 SOM 的文本聚类算法研究 [J]. 计算机应用与软件, 2010, 27(5): 118-120.

[11] LUO X, XU Z, YU J, et al. Building association link network for semantic link on web resources [J]. Automation Science and Engineering, 2011, 8(3): 482-494.

[12] LUO X, YAN K, CHEN X. Automatic discovery of semantic relations based on association rule [J]. Journal of Software, 2008, 3(8): 11-18.

[13] XU Z, LUO X, LU W. Association link network: an incremental semantic data model on organizing web resources [C]// Proceeding ICPAD'09 Proceedings of the 2009 15th International Conference on Parallel and Distributed Systems. 2009: 793-798.

[14] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks [J]. Physical Review E, 2007, 76: 036106.

[15] DANON L, DÍAZ-GUILERA A, DUCH J, et al. Comparing community structure identification [J]. Journal of Statistical Mechanics: Theory and Experiment, 2005: 09008.

[16] VINH N X, EPPS J, BAILEY J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance [J]. The Journal of Machine Learning Research, 2010, 11(10): 2837-2854.

[17] STEINBACH M, KARYPIS G, KUMAR V. A comparison of document clustering techniques [C]// KDD Workshop on Text Mining. 2000: 525-526.

[18] ROSENBERG A, HIRSCHBERG J. V-measure: a conditional entropy-based external cluster evaluation measure [C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007: 410-420.

本文彩色版可登陆本刊网站查询: http://www.journal.shu.edu.cn