基于免疫的中文网络短文本聚类算法

贺 涛 [1,2] 曹先彬 [1,2] 谭 辉 [3]

摘 要 网络短文本聚类是网络内容安全的一种主要处理方法. 然而, 中文网络短文本固有的关键词词频低、存在大量变形词等特点, 使得难以直接使用现有面向长文本的聚类算法. 本文提出了一种面向中文网络短文本的基于免疫网络调节的聚类算法. 首先, 利用抽取的中文词语的 N-gram 片段的拼音序列来组成一个中文网络短文本的特征表示, 从而缓解关键词词频过低和存在变形词对聚类的影响; 然后, 将网络短文本集构建为一个动态网络, 利用免疫网络学习机制来自动发现网络短文本之间的内在关联, 获得合适的聚类结果. 测试实验表明, 相对于传统的聚类方法如 K-means, 本文的算法能够得到更好的中文网络短文本聚类效果.

关键词 网络内容安全, 中文网络短文本, 聚类, 免疫网络
中图分类号 TP399

An Immune Based Algorithm for Chinese Network Short Text Clustering

HE Tao[1,2] CAO Xian-Bin[1,2] TAN Hui[3]

Abstract Network short text clustering is a major technology in network content security. Since Chinese network short text is less of keywords and full of anomalous writings, the traditional text clustering method is not directly suitable for network short text clustering. This paper presents an immune network regulation based method to cluster Chinese network short texts. First, Chinese N-gram chunks are extracted and transformed to Chinese pinyin to form the feature representation to each Chinese network short text, so as to relieve these two characteristics′ bad influence on the clustering performance. Then, the network short text set is constructed as a dynamic network and an immune network learning mechanism is used to learn the similarity among short texts and therefore to gain a better clustering result. Experiments show our method can get better performance in Chinese network short text clustering, compared with traditional method such as K-means.

Key words Network content security, Chinese network short text, clustering, immune network

网络内容安全主要研究基于网络内容 (主要是文本) 的安全性问题, 它涉及到针对文本内容的获取、分类、聚类、话题发现与跟踪等关键技术. 其中, 聚类是实现网络内容安全的一个主要手段. 近年来, 随着即时通信、聊天室、BBS、博客等新的网络交流平台不断涌现, 针对这类应用背景下特定文本形式的聚类方法研究已经成为相关研究的一个重点[1−2].

在这类应用背景中, 待处理的文本在形式上已经发生了很大变化, 一般称之为网络短文本. 网络短文本是指那些出现在网络交流平台中、用少量词语表达的、可能会参杂不规范书写的简短的文本. 显然, 它与传统文本挖掘处理的文本 (我们称之为长文本) 在形式上明显不同. 网络短文本的固有特点对相应的聚类算法设计提出了更高的要求, 现有数据挖掘领域已取得较大成功的文本聚类算法还难以直接引用[3].

本文研究面向中文网络短文本的聚类算法. 中文网络短文本聚类面临的难点主要有: 1) 网络短文本中关键词词频过低. 这一方面导致无法直接使用现有文本处理中常用的基于词频的文本表示方法 (如向量空间模型) 来表示短文本; 同时, 由于难以提取短文本之间的语义、语用等信息, 常用的文本聚类算法如 K-means 等在短文本聚类中效果不佳[1]. 2) 中文网络短文本中存在大量的同义但不同形的变形词 (主要是由拼音输入法的使用导致的). 采用一般的特征提取方法时这些同义词会被看作是不同的特征, 这一方面导致中文网络短文本的表示不够准确, 另一方面会误导聚类结果.

目前, 专门针对网络短文本或中文网络短文本聚类的研究工作还较少. 其中, 在短文本关键词词频过低的处理方面, 提出过通过添加相关词汇来扩充

收稿日期 2008-05-08 收修改稿日期 2008-10-20
Received May 8, 2008; in revised form October 20, 2008
国家重点基础研究发展计划 (973 计划) (2004CB318109), 国家高技术研究发展计划 (863 计划) (2007AA11Z240), 教育部新世纪优秀人才支持计划 (NCET-07-0787) 资助
Supported by National Basic Research Program of China (973 Program) (2004CB318109), National High Technology Research and Development Program of China (863 Program) (2007AA11Z240), and Program for New Century Excellent Talents in University (NCET-07-0787)
1. 中国科学技术大学计算机科学与技术学院 合肥 230027 2. 安徽省计算与通讯软件重点实验室 合肥 230027 3. 哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001
1. School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027 2. Key Laboratory of Software in Computing and Communication, Anhui Province, Hefei 230027 3. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001
DOI: 10.3724/SP.J.1004.2009.00896

传统 TF-IDF 模型的方法[1] 和抽取短文本关键特征来降低关键词空间维度的方法[2]; 这些方法还不能充分挖掘出短文本间的内在关联, 同时也不能解决变形词对中文网络短文本聚类结果误导的难题. 而在变形词的处理方面, 虽然有一些针对短文本特性分析的工作, 但结合变形词有效处理的中文网络短文本聚类算法的研究还有待开展.

本文提出了一种基于免疫的中文网络短文本聚类算法. 我们首先设计了一种抽取分词后的中文 N-gram 片段的拼音序列来组成一个中文网络短文本的特征空间的表示方法. 利用片断的扩展可以扩大关键词空间, 解决关键词词频过低的问题; 采用拼音序列的特征表示可以部分解决变形词问题 (因为变形词主要是由拼音输入法的使用导致的). 然后, 将中文网络短文本集构建为一个动态网络, 利用免疫网络学习机制[4] 来自动发现短文本之间的相似性和内在关联, 从而降低变形词对聚类结果的误导, 获得合适的聚类结果. 与已有的短文本聚类算法如 K-means 相比, 实验表明本文的算法可以有效缓解关键词词频低、变形词多等对中文网络短文本聚类的影响, 得到更好的中文网络短文本聚类结果.

1 相关工作

网络短文本聚类首先需要分析网络短文本在表达形式上的固有特点, 然后设计相应的特征表示方法和针对性的聚类方法. 目前, 在中文网络短文本的表达形式分析方面取得了很多进展, 但还局限在语言学和自然语言处理范围之内. 在语言学领域, 已有一些针对中文网络语言基本特征的分析工作[5−6]. 通过对大量中文网络短文本的分析, 发现其中充斥着大量干扰信息, 包括缩写、同音字、拼音等, 拼音输入法的使用是造成此类现象的主要原因. 在自然语言处理领域, Xia 等对网络聊天语言的奇异性和动态性进行了分析和归纳, 并建立了一定规模的网络聊天语言语料库[7−8], 但其处理的是句子级别的网络聊天语言. 总之, 上述工作对网络短文本的形式特点进行了较好的归纳, 但目前还没有很好地用到中文网络短文本的聚类中. 总体而言, 目前专门研究网络短文本聚类算法的工作还较少. 代表性工作有: Wang 等针对即时通信消息的聚类提出了 WR-Kmeans 算法[1], 该方法先将即时通信消息分割成谈话片段, 然后向每个谈话片段中添加基于 HowNet 的相关词汇, 通过这种方法扩充了传统的 TF-IDF 模型, 可以在一定程度上缓解关键词词频过低带来的问题, 但是该方法并不能解决变形词对聚类结果误导的难题. He 等提出了一种基于中文块的中文短文本聚类方法[2], 先从原始语料中抽取中文块, 然后采用竞争者受惩罚竞争学习 (Rival penalized competitive learning, RPCL) 对中文短文本聚类; 该方法能抽取短文本的关键特征并有效降低短文本的维度, 达到改善中文短文本聚类效果的目的, 在抽取中文块时采用了基于 N-gram 的方法, 但直接针对原始语料进行; 与此不同, 本文将在分词后再抽取 N-gram 片段. 同时, 该方法也不能避免变形词对聚类结果的影响. 另外, 还有一些工作也涉及到短文本聚类, 但其处理的对象与本文针对的中文网络短文本在本质上是不同的. 例如, 黄永光等将那些用少量词语表达一定语义关系的书写不规范的文字称为变异短文本[3], 先把变异短文本转化为正规的短文本, 然后采用检索的思想, 将聚类问题转化为检索问题, 从而降低算法复杂度, 针对收集的手机短信的实验表明算法具有不错的效果. 另外, 王永恒等提出一种面向大规模数据库的基于频繁项的短文聚类方法[9]; 在处理大规模短文数据时, 文中指出该方法在效率和速度上都有优势, 但是该文处理的是文章摘要、电子邮件等短文, 有别于本文处理的中文网络短文本. 近年来, 人工免疫网络已被引入到文本聚类研究中, 并在长文本聚类中获得了一定的效果. Hang 等提出了一种基于免疫网络的用于 Web 文本聚类的算法[10], 提高了聚类过程中的动态适应性, 能自动发现新类. de Castro 等于 2000 年提出了一种用于数据聚类的人工免疫网络模型 aiNet[4]. Tang 等在 aiNet 模型的基础上, 先将原始数据训练成免疫网络的记忆矩阵, 然后再用 HAC 或 K-means 聚类算法对记忆矩阵进行聚类, 取得了不错的效果[11]. 现有基于人工免疫网络的文本聚类工作基本上还是针对长文本完成的.

2 基于免疫的中文网络短文本聚类

2.1 算法基本思路

本文提出的基于免疫的中文网络短文本聚类算法的基本思想是: 1) 首先构造出一种针对中文网络短文本的特征表示. 从前面的分析可知, 采用依赖于关键词词频的常用方法 (如向量空间模型) 是不合适的. 我们的表示方法是: 抽取中文网络短文本分词后的 N-gram 片段, 然后用各片段对应的拼音序列来组成该中文网络短文本的特征空间. 这一表示方法的优点是: 利用片断的扩展扩大了关键词空间和信息容量, 从而在一定程度上解决了关键词词频过低的问题; 同时, 鉴于变形词主要是由拼音输入法的使用导致的, 用拼音序列来表示也可以在很大程度上解决变形词问题. 2) 为中文网络短文本集构建一个动态网络, 一个中文网络短文本对应该网络的一个外部刺激, 网络将在不断刺激下动态调节, 最终生成的网络作为该短文本集的一种整体压缩表示, 尽可能准确地反映了该短文本集的结构特征 (即获

得短文本之间的相似性和内在关联); 该网络在用来支持聚类时, 可以获得满意的聚类结果. 本文算法的基本流程如图 1 所示.

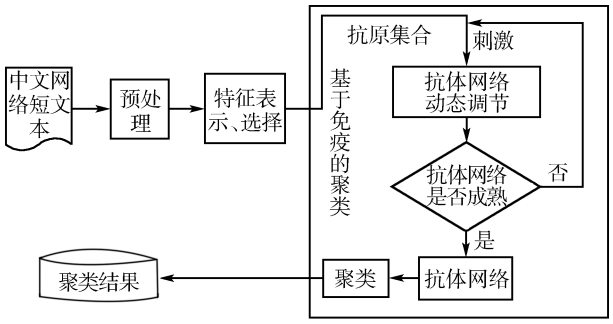


图 1 算法基本流程图

Fig. 1 Flowchart of the algorithm

2.2 中文网络短文本的特征表示

传统的文本表示方法不能正确表示中文网络短文本的内容. 因为在建立中文网络短文本的特征空间时, 首先, 分词程序会把一些网络新词分成片段, 而这些词语有可能就是中文网络短文本的重要特征; 其次, 对于一些变形词, 分词程序也会将其分割成片段. 这样, 如果分词后直接按照每个词汇建立一个统计特征, 可能导致中文网络短文本中重要信息的丢失. 例如: "他参加过澳运会" (实际表述为 "他参加过奥运会"), 采用 ICTCLAS 分词软件[12], 分词后的结果为 "他 参加 过 澳 运 会". 这样得到的 6 个特征词显然并不能获得句子表达的实际内容: "奥运会". 因此, 直接按每个词语作为一个特征来表述中文网络短文本是不合适的. 本文给出的中文网络短文本的特征表示方法为: 一个中文网络短文本, 首先对其分词, 再针对分词后的结果抽取出所有的 N-gram 片段; 然后将每一片断转换成相应的拼音序列作为一个特征串, 所有特征串的集合就组成了该中文网络短文本的特征表示. 这样的特征表示方法可以更好地描述出中文网络短文本的内在信息, 有利于它们之间的类别区分. 理论上 n 越大, 所携带的上下文信息越丰富; 然而过大的 n 将导致计算复杂度剧增. 为此, 本文采取折衷处理, 并考虑到中文词汇的长度, 我们采用 $n = 1, 2, 3$. 例如上面的例子中, 采用我们的方法就含有 15 个特征: "ta", "tacanjia", "tacanjiaguo", "canjia", "canjiaguo", "canjiaguoao", "guo", "guoao", "guoaoyun", "ao", "aoyun", "aoyunhui", "yun", "yunhui", "hui". 其中就含有我们关心的词汇 "aoyunhui". 另外, 考虑到即使对于一个很小的中文网络短文本集, 基于上述方法得到的特征数也可能很大, 因此有必要进行特征选择, 选出 L 个较好的特征来表示短文本. 本文采用文献 [13] 中提到的特征选择方法, 一个特征词的好坏可以用下式评价

$$q(t) = \sum_{j=1}^{m} {f_j}^2 - \frac{1}{m}\left(\sum_{j=1}^{m} f_j\right)^2 \tag{1}$$

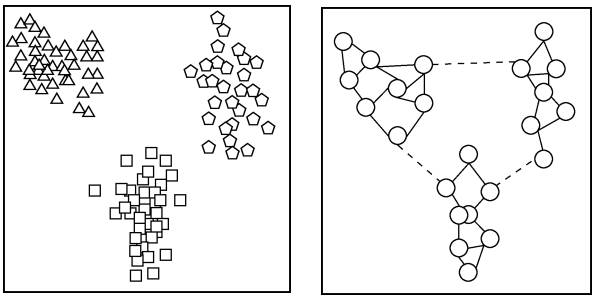
其中, m 是特征词 t 至少出现一次的短文本数, f_j 是特征词 t 在短文本 j 中的词频. 文献 [13] 指出: 针对文章摘要这样的特征词稀疏的短文本, 抽取出 15 % 的特征词来表示文本, 得到的聚类结果和选用全部特征来表示文本得到的聚类结果相当. 本文采用此方法选择出 q 值最大的 L 个特征串. 将每个中文网络短文本表示成一个二值 L 维特征向量.

2.3 基于免疫的聚类操作

算法在具体实现时有两个问题需要解决: 1) 采用什么样的网络学习机制来得到适当的动态网络? 2) 基于问题 1), 针对中文网络短文本的聚类, 如何进行针对性设计? 下面分别介绍我们采用的做法.

2.3.1 动态网络学习机制

鉴于免疫网络调节模型在常规的数据聚类中已经取得了一定的成功, 同时常规的数据聚类和本文的中文网络短文本聚类在机理上具有共通之处, 我们选用免疫网络调节模型中最经典的 aiNet 模型[4] 来设计本文的动态网络学习机制. 基本方法如下: 一个中文网络短文本对应于一个抗原, 我们的目的是构造一个记忆性抗体集 (以网络的形式组织) 以提取抗原集合的内在特征, 最终得到一个训练好的抗体网络, 它可以看作是抗原集合 (待聚类的中文网络短文本集) 的一个压缩影像; 然后就可以用这一抗体网络来指导原始中文网络短文本集的聚类. 压缩影像的含义包括: 1) 在抗体网络中, 抗体的空间分布反映出抗原的空间分布特点; 2) 抗体网络中包含的抗体数远少于原始中文网络短文本数, 因而在支持聚类时可以大大降低计算复杂度. 如图 2 所示.



(a) 一个包含三个类别的数据集 (a) A data set containing three categories

(b) 对应的抗体网络 (b) Corresponding antibody network

图 2 用于聚类的免疫网络

Fig. 2 Immune network for clustering

在上述方法中, 抗原和抗体都表示成 L 维的向量. 抗体和抗体之间的相似度用它们之间的距离来表示, 在图 2 中描述为节点之间的连接权; 而抗体和抗原之间的亲和力则描述为相似度的反比. 抗体通过竞争来获取生存的权利, 竞争通过亲和力来衡量. 与抗原亲和力高的抗体将进行克隆分裂, 亲和力低的抗体则被消除. 抗体和抗体之间的识别会产生网络的抑制. 通过一定的学习和调整机制最终留下一些记忆性抗体, 这些记忆抗体就表示原始中文网络短文本集的一种压缩表示. 抗体网络的学习过程简单描述如下:

1) 随机产生一些网络中的抗体作为最初的网络节点.

2) 依次把所有抗原呈给网络学习; 在每一次的学习中, 计算所选抗原与网络中抗体的亲和力.

3) 选择一定数量高亲和力的抗体, 根据亲和力进行克隆, 克隆的数目与亲和力成正比; 之后, 克隆产生的抗体会经历一个变异的过程: 亲和力越低, 变异率越高. 最后, 在这个克隆集合中, 选择亲和力高的抗体成为记忆抗体, 加入到原来的网络中.

4) 新的网络进行一些调整, 删除太相近的抗体. 另外, 还需要随机产生一些新的抗体加入到网络中.

2.3.2 针对性设计

上面粗略地描述了免疫网络的学习机制. 在用于中文网络短文本聚类时, 下面几点需要专门考虑: 1) 为保证抗体网络能充分学习给定的一个抗原 (中文网络短文本), 需要准确计算出抗原和抗体之间的亲和力. 由于中文网络短文本是二值型向量形式, aiNet 模型中采用的针对实数的计算公式不能使用, 因此需要给出专门的计算公式; 2) 在 aiNet 模型中, 只是采用简单的克隆和变异来保证抗体细胞的多样性, 但是网络的收敛速度很慢, 而且可能导致最终得到的抗体网络不能准确地反映抗原集合的空间分布特点. 因此, 为了保证最终得到的抗体网络能准确地反映原始中文网络短文本集合的结构特征, 得到合适的聚类结果, 需要对网络的学习过程进行改进. 本文的具体做法分别如下:

1) 亲和力计算

抗原和抗体之间的亲和力是指出现在抗原中的 "特征串" 能够被抗体识别的程度, 也就是出现在抗原中的词汇数, 能够被出现在抗体中的词汇数识别的比例. 在本文中, 抗原和抗体被表示成 L 维的二值向量, 它们之间的亲和力计算式为

$$D(\boldsymbol{Ab}, \boldsymbol{Ag}) = \frac{\sum_{i=1}^{L} \delta(Ab_i, Ag_i)}{|\boldsymbol{Ag}|} \tag{2}$$

其中

$$\delta(Ab_i, Ag_i) = \begin{cases} 1, & \text{若} \quad Ab_i = Ag_i = 1 \\ 0, & \text{否则} \end{cases}$$

而两个抗体之间的亲和力采用下式计算

$$S(\boldsymbol{Ab}^r, \boldsymbol{Ab}^s) = \frac{|\boldsymbol{Ab}^r| + |\boldsymbol{Ab}^s|}{\sum_{i=1}^{L} \delta(Ab_i^r, Ab_i^s)} \tag{3}$$

其中 δ 函数同上.

2) 网络学习过程

抗体网络的学习过程借鉴了文献 [14] 中提到的改进方法: 让每一个与抗原有相同特征的抗体均有机会进入抗体网络, 这样有利于最终生成的抗体网络尽可能准确地反映抗原集合的空间分布特点, 即获得中文网络短文本之间的相似性和内在关联.

2.4 算法描述

至此, 我们给出详细的算法描述. 为了描述方便, 首先引入下面的记号:

C: 包含 N_t 个抗体的矩阵;

M: 包含 N 个记忆抗体的矩阵, $(M \subseteq C)$;

N_c: 每个受激励抗体产生克隆的数目;

D: $(\boldsymbol{Ag} - \boldsymbol{Ab})$ 之间的亲和力;

S: $(\boldsymbol{Ab} - \boldsymbol{Ab})$ 之间的相似度;

n: n 个亲和度最高的抗体克隆变异;

ζ: 成熟抗体被选择的百分比;

σ_d, σ_s: 自然死亡阈值和抑制阈值.

算法的具体流程如下:

步骤 1. 令 $M = \phi$;

步骤 2. 通过特征选择将 N 个中文网络短文本转换成 N 个抗原, 每个抗原是一个 L- 维的 0-1 向量;

步骤 3. 随机产生一个抗体集合, 初始化抗体网络;

步骤 4. 对于每个抗原 i:

步骤 4.1. 计算抗原 i 同免疫网络中每个抗体的亲和度 d_{ij};

步骤 4.2. 选择 n 个亲和度最高的抗体, 构成临时抗体集合 tM;

步骤 4.3. 在 tM 中执行克隆操作. 克隆这 n 个细胞, 亲和度越大, N_c 越大;

步骤 4.4. 在 tM 中执行变异;

步骤 4.5. 计算 tM 中抗体与抗原 i 的亲和度 d_{kj};

步骤 4.6. 重新选择 ζ% 个高亲和度的抗体, 产生一个局部记忆抗体矩阵 M_p;

步骤 4.7. 删除 M_p 中亲和度 d_{kj} 低于 $1/\sigma_d$ 的抗体;

步骤 4.8. 把与抗原 i 有相同特征的抗体插入到 M_p 中;

步骤 4.9. 针对 M_p, 计算 $(\boldsymbol{Ab} - \boldsymbol{Ab})$ 之间的相似度 s_{ij};

步骤 4.10. 对 M_p 执行克隆抑制操作: 删除 $s_{ij} < \sigma_s$ 的抗体;

步骤 4.11. 将 M_p 中剩下的记忆抗体加入到 M 中. 计算 M 中所有记忆抗体的亲和力 s_{tk}, 在 M 上实行网络抑制;

删除过于相似的抗体 $s_{tk} < \sigma_s$;

步骤 5. 若不满足结束条件, 转步骤 4 执行; 进化到一定的代数, 或者当 M 达到一定的数量, 结束循环;

步骤 6. 使用 K-means 方法对 M 聚类;

步骤 7. 根据 M 中的抗体获得每个抗原的类别.

本算法在步骤 4.8 中把与抗原有相同特征的抗体添加进网络, 该步骤有利于最终生成的抗体网络尽可能准确地反映抗原集合的空间分布特点. 在步骤 4.11 中, 每一个抗原刺激网络后, 就把由它产生的临时抗体集合添加至网络中, 对网络进行抑制, 这相比于原始的 aiNet 把所有的抗原提呈后再把产生的抗体集合添加到网络进行抑制, 可以节省空间复杂度.

3 实验

3.1 数据准备

目前, 中文网络短文本领域并没有公开的数据集. 因此我们在网上下载了中文聊天数据, 进行人工筛选之后作为测试所用的中文网络短文本集. 每个短文本是一个简短的聊天片段, 包含若干句聊天语句, 但是内容都涉及到一个话题, 这些短文本中具备我们上面讲到的中文网络短文本的特性, 关键词词频低, 且充斥着同音别字、拼音等变形词. 实验中共用到了三个话题领域的数据, 表 1 给出了数据集的统计信息.

表 1 实验数据

Table 1 Experimental data

话题类别	NBA	金庸武侠	武器
短文本数目	102	113	82

实验中会对中文网络短文本进行预处理, 包含去噪和去停止词操作. 由于网络短文本中充斥着大量的干扰信息, 例如一些表情图示和特殊符号等, 这些干扰信息对聚类没有任何帮助, 所以首先对网络短文本进行去噪处理. 例如, 聊天信息中常出现的微笑图标, 实际的文本符号为 ":)", 类似这样的符号都需删除. 去噪处理后, 只保留有用的文本信息. 另外, 无论在汉语还是英语中, 都存在一些对文本内容识别意义不大的词, 称之为停止词. 最简单的如汉语中的 "的"、"啊" 等词, 它们没有具体的意义, 不能体现文本所表示的内容, 但几乎在所有文本中都出现, 如果在聚类中考虑这些词, 那么文本之间的相似性不能表现出内容的相似性, 而是一些无意义的相似性, 这不是我们所希望的. 为此我们建立了一个中文停止词词表, 通过这个词表去掉中文网络短文本中的停止词.

3.2 评价指标

采用信息检索领域的召回率、准确率和 F 值来衡量聚类算法的有效性. 原始的中文网络短文本按话题有三类, 记作 C_1, C_2, C_3. 然后将聚类之后的结果记作 D_1, D_2, D_3. 对每一个原始的话题类别 C_i, 找到一个与其有最大公共子集的聚类结果 D_j, 准确率和召回率分别为

$$P(C_i) = \frac{a}{b}, \quad R(C_i) = \frac{a}{c} \tag{4}$$

其中, a 是 D_j 中属于话题 C_i 的短文本数, b 是聚类得到的结果 D_j 中的中文网络短文本数, c 是话题类别 C_i 中的中文网络短文本数. 召回率和准确率是从两个不同侧面反映聚类的效果, F 值是一个把准确率和召回率结合起来的指标, 定义为

$$F = \frac{2 \times P \times R}{P + R} \tag{5}$$

3.3 实验结果与分析

为了充分检验本文算法的有效性, 我们设计了下面两组实验. 另外, 通过实验中的反复调整与测试, 在本文的算法中选用的参数为: $n = 4$, $\zeta\% = 0.4$, $\sigma_s = 5$, $\sigma_d = 0.5$.

实验 1. 与 K-means 算法的比较

K-means 算法是数据聚类中一个简单高效的算法, 也是目前在中文网络短文本聚类应用中最为常用的. 在本文算法中, 通过特征选择和抽取后, 每个短文本转换成一个 320 维的二值向量. 图 3 (见下页) 给出了两个聚类算法在召回率、准确率和 F 值三个指标上的性能比较; 其中, 图中数据是多组实验的平均值; K-means 曲线是指采用我们专门设计的中文网络短文本特征表示方法, 并采用 K-means 算法进行聚类的结果; 而 aiNet 曲线代表了采用本文提出的基于免疫的中文网络短文本聚类算法得到的结果.

从实验结果可以看出, 本文算法在准确率、召回率和 F 值方面, 都要优于 K-means 算法. 这是因为本文算法是利用免疫网络的学习机制发现短文本之间的相似性和内在关联, 最终生成的网络作为短文本集的一种压缩表示, 消除了冗余数据, 可以尽可能准确地反映原始短文本集的结构特征; 该网络用来支持聚类, 从而降低变形词对聚类结果的误导, 提高了所生成的类簇的质量, 获得了合适的聚类结果. 而 K-means 算法是一种基于划分的聚类算法, 直接对整个数据集进行聚类, 不能充分挖掘出短文本间的内在关联, 因此聚类结果不甚理想.

当然, 本文的基于免疫的聚类方法是一网络抑制、逐渐成熟的过程, 因此其训练需要花费较长的时

间. 但是, 当网络训练成熟后, 在用于短文本聚类时, 其聚类的时间花费和 K-means 算法是相同的. 可以说本文的方法是用事先的训练时间换取了较好的实际聚类效果.

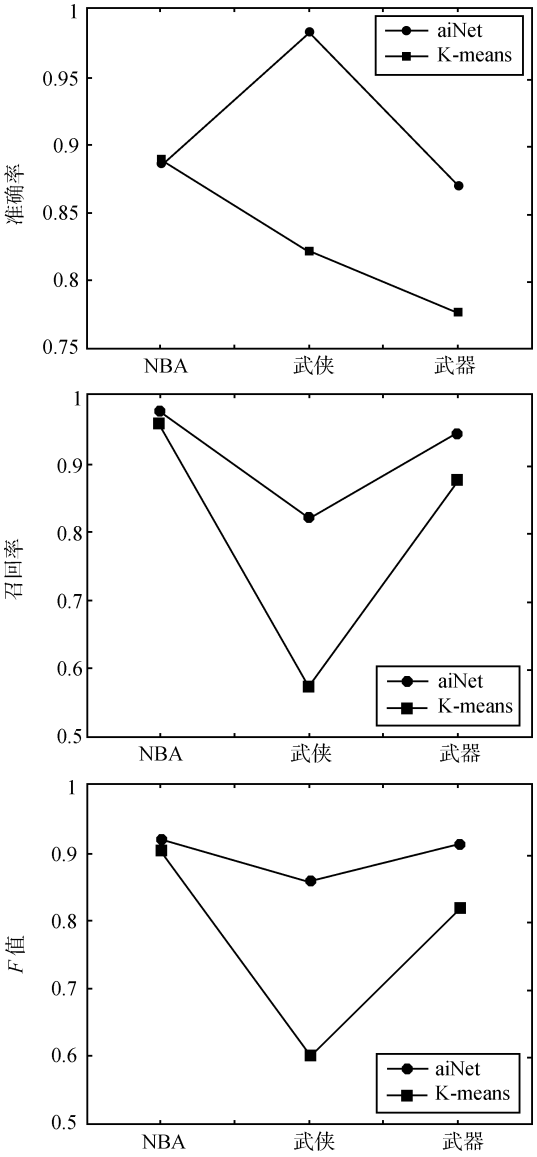


图 3　性能对比图

Fig. 3　Charts of the performance comparison

实验 2. 与直接选用原始词作为特征的比较

由于中文网络短文本内容较短, 而且存在干扰信息, 导致一些重要的词可能形式上不同, 但是表达的意思却是一样的, 因此直接按照每个词汇的形式建立一个统计度量, 每个特征的信息量会很少, 甚至会丢失网络短文本中的重要信息, 因此聚类效果应该不会理想. 为此, 我们改进了短文本的特征表示. 作为对比实验, 本文同时直接按照每个词汇的形式建立一个统计度量, 作为一个特征, 来做一组比较实验. 图 4 给出了比较结果, 图中准确率、召回率和 F 值是针对三个类别的数据的平均值.

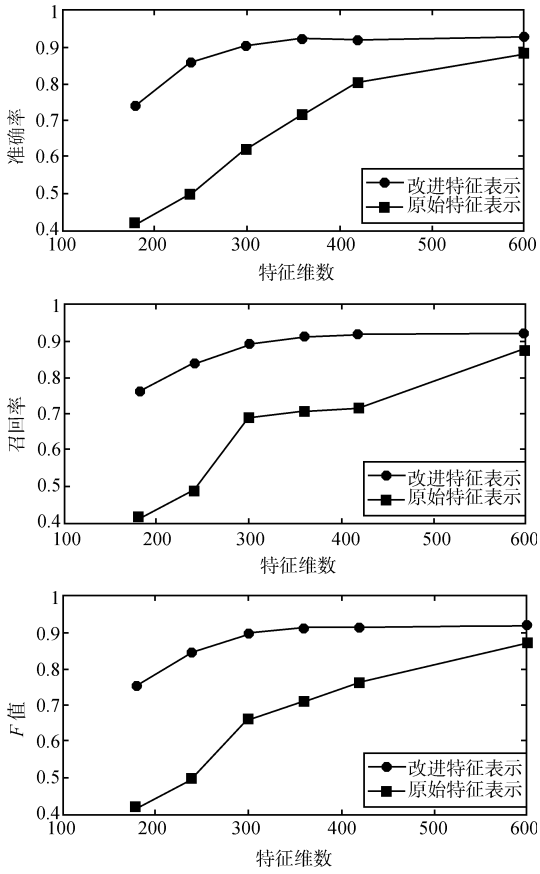


图 4　不同的特征表示时的性能对比图

Fig. 4　Charts of the performance comparison with different feature representations

可以看出: 采用我们专门设计的中文网络短文本特征表示方法, 在准确率、召回率和 F 值方面, 都要优于选用原始词汇作为特征的方法. 这是由于: 1) 测试中文网络短文本中存在一些变形词. 例如: 在话题类别 “NBA” 中, 有若干短文本中出现 “麦迪” (NBA 联盟中火箭队球员), 同时也有一些短文本中出现的是 “麦蒂”; 而这两个词实际上是同一个意思, 我们的特征表示方法会将它们处理成一个特征: “maidi”; 而在选用原始词汇作为特征时, 这就是两个不同的特征, 因此在被当成不同特征对待时就会影响聚类的结果. 2) 短文本集中有第 2.2 节例子中的现象, 分词会把一些关键特征分割开, 选用原始词作为特征时就会丢失这些关键特征.

从实验结果可以看出, 由于本文针对中文网络短文本的特征表示适应了中文网络短文本的特性, 因此具有较好的效果.

4　结束语

本文针对中文网络短文本关键词词频低、存在

大量变形词等特性, 首先对它的特征表示和抽取进行了专门设计, 采用了分词后抽取 N-gram 片段再转换成拼音序列作为一个特征的表示方法. 该方法可以使形式上不同但表达意思相同的短文本具有更多相同的特征, 扩充了每个短文本的特征数, 这样既解决了短文本中关键词词频低的问题, 又解决了部分变形词对聚类结果的误导. 然后, 通过免疫网络的学习机制自动地发现中文网络短文本之间的相似性, 获得短文本集的结构特征, 从而获得合适的聚类结果. 实验表明, 该算法相对于传统的聚类算法具有不错的效果, 是一种处理中文网络短文本聚类的有效算法.

当然, 中文网络短文本中存在大量的变形词, 本文提出的特征表示方法也只能解决部分变形词对聚类结果的误导. 因此, 在后续的工作中, 我们将研究如何避免其余形式的变形词对聚类结果的影响. 另外, 目前的实验中用到的中文网络短文本集相对较小, 将来需要在更大规模的数据环境中进行算法验证.

References

1 Wang L, Jia Y, Han W H. Instant message clustering based on extended vector space model. In: Proceedings of the 2nd International Symposium on Intelligence Computation and Applications. Wuhan, China: Springer, 2007. 435−443

2 He H, Chen B, Xu W R, Guo J. Short text feature extraction and clustering for web topic mining. In: Proceedings of the 3rd International Conference on Semantics, Knowledge and Grid. Washington D. C., USA: IEEE, 2007. 382−385

3 Huang Yong-Guang, Liu Ting, Che Wan-Xiang, Hu Xiao-Guang. A fast clustering algorithm for abnormal and short texts. *Journal of Chinese Information Processing*, 2007, **21**(2): 63−68
(黄永光, 刘挺, 车万翔, 胡晓光. 面向变异短文本的快速聚类算法. 中文信息学报, 2007, **21**(2): 63−68)

4 de Castro L N, Von Z F J. aiNet: an artificial immune network for data analysis. *Data Mining: A Heuristic Approach*. New York: Idea Group Publishing, 2001. 231−259

5 Ma Jing. Network language from linguistic perspective. *Journal of Northwestern Polytechnical University (Social Sciences)*, 2002, **22**(3): 52−56
(马静. 语言学视野中的网络语言. 西北工业大学学报 (社会科学版), 2002, **22**(3): 52−56)

6 Wu Chuan-Fei. A survey of China′s cyberlanguage study. *Journal of Social Science of Hunan Normal University*, 2003, **32**(6): 102−105
(吴传飞. 中国网络语言研究概观. 湖南师范大学社会科学学报, 2003, **32**(6): 102−105)

7 Xia Y Q, Wong K F. Anomaly detecting within dynamic Chinese chat text. In: Proceedings of New Text Workshop at the 11th Conference for European Chapter of the Association for Computational Linguistics. Trento, Italy: Acl Anthology Network, 2006. 48−55

8 Xia Y Q, Wong K F, Gao W. NIL is not nothing: recognition of Chinese network informal language expressions. In: Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing. Jeju Island, Republic of Korea: Acl Anthology Network, 2005. 95−102

9 Wang Yong-Heng, Jia Yan, Yang Shu-Qiang. Study on massive short documents clustering tchnology. *Computer Engineering*, 2007, **33**(14): 38−40
(王永恒, 贾焰, 杨树强. 海量短语信息文本聚类技术研究. 计算机工程, 2007, **33**(14): 38−40)

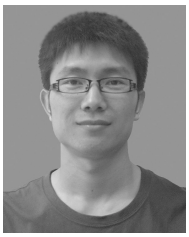
10 Hang X S, Dai H H. An immune network approach for web document clustering. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Beijing, China: IEEE, 2004. 278−284

11 Tang N, Vemuri V R. An artificial immune system approach to document clustering. In: Proceedings of the 20th ACM Symposium on Applied Computing. Santa Fe, USA: ACM, 2005. 918−922

12 Zhang H P, Yu H K, Xiong D Y, Liu Q. HHMM-based Chinese lexical analyzer ICTCLAS. In: Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing. Sapporo, Japan: Association for Computational Linguistics, 2003. 184−187

13 Dhillon I, Kogan J, Nicholas M. Feature selection and document clustering. *A Comprehensive Survey of Text Mining*. Berlin: Springer-Verlag, 2003. 73−100

14 Zhong Jiang, Wu Zhong-Fu, Wu Kai-Gui, Ou Ling. A novel dynamic clustering algorithm based on artificial immune network. *Acta Electronica Sinica*, 2004, **32**(8): 1268−1272
(钟将, 吴中福, 吴开贵, 欧灵. 基于人工免疫网络的动态聚类算法. 电子学报, 2004, **32**(8): 1268−1272)



贺 涛 中国科学技术大学计算机科学与技术学院硕士研究生. 主要研究方向为信息安全.
E-mail: mr.hetao@hotmail.com
(**HE Tao** Master student at the School of Computer Science and Technology, University of Science and Technology of China. His main research interest is information security.)

曹先彬 中国科学技术大学计算机科学与技术学院教授. 主要研究方向为计算智能、信息安全和智能交通系统. 本文通信作者. E-mail: xbcao@ustc.edu.cn
(**CAO Xian-Bin** Professor at the School of Computer Science and Technology, University of Science and Technology of China. His research interest covers computing intelligence, information security, and intelligent transportation system. Corresponding author of this paper.)

谭 辉 哈尔滨工业大学博士研究生, 高级工程师. 主要研究方向为智能信息处理、无线通信和信息安全.
E-mail: tanhui99@hit.edu.cn
(**TAN Hui** Senior engineer, Ph. D. candidate at the School of Computer Science and Technology, Harbin Institute of Technology. His research interest covers intelligent information processing, wireless communication, and information security.)