

深度词汇网络学习的文本聚类研究

易军凯 冯佳明 万 静*

(北京化工大学 信息科学与技术学院, 北京 100029)

摘 要:为改进已有中文文本聚类中数据非结构化导致的算法准确度不高及特征向量高维稀疏导致算法复杂度过高的现状,提出一种基于深度词汇网络学习的中文文本聚类算法,解决了优化数据非结构化带来的聚类结果准确性低及特征向量高维度带来的高复杂度问题。首先建立词汇网络用以抽取关键义原,以词语义原代替单词作为网络节点,不仅避免了语义消歧,同时考虑到词语间语义相似性与词汇相关性,使所提取的特征向量更能表现出文章的主旨,提高聚类效果;另一方面,训练深度学习网络对特征向量降维处理,在降维的同时保留尽可能多的信息,大大减低算法的执行时间。聚类质量检测方法(F -measure)的结果表明,本文算法比 k -means 算法在中文文本聚类中有更好的表现。

关键词:词汇网络;深度学习网络;中文文本聚类

中图分类号: TP391

引 言

随着网络的飞速发展,针对大规模数据的分析成为当前的一个研究热点。文本聚类分析是数据挖掘中的一项重要技术,可作为网络中海量的文本数据分析的基础。文本聚类分析能够挖掘出语料中潜在的规则,把文本分成不同的簇,协助实现对大规模数据的理解,已在搜索引擎、舆情分析等多个技术领域得到很好的应用。

文本特征提取是文本聚类分析的基础,其质量直接影响到聚类的效果^[1]。已有的文本特征提取方法包括基于评估函数方法、基于词语相关性方法和基于遗传算法方法等。在基于词语相关性的提取方法中,典型的方式为基于频繁集概念的特征提取^[2],即以多个词组成的词语片段代替单词作为文本的特征项,更能突显文本主旨,易于建立类标签,但该方法由于频繁候选项较多,在大数据量分析中效率往往不高。基于遗传算法的特征提取方法所提取的特征不仅反映文本本身,还反映了同类文本的共性^[3],其缺点在于算法复杂度较高。因此目前文本特征提取方法大多数采用基于评估函数的方法,如期望交叉熵、互信息、文本证据权等^[4],该方法

可以选取出对类间区分度较大的特征,但区分度的判断需依赖于已知类型,所以并不适合用于文本聚类。目前针对大规模数据特征提取方法的研究主要为基于统计计算评估函数的方法,统计方法具有算法简单、易于实现、过滤速度快、不依赖具体领域和语言等优点。文档频数是基于统计的方法中最典型的一种,该方法简单易行,受到许多研究者的关注;N-Gram 算法则是另外一种基于纯统计学文本分解技术的文本提取方法^[5],有效的避免了汉语分词障碍。但这类方法忽略了文本中词项的含义,同时也忽略了文本中的语法与文本结构,因此很难筛选出高质量的特征。

为了改善基于统计计算方法的缺点,本文提出一种基于深度词汇网络学习的文本聚类算法。通过词汇网络特征筛选与深度学习降维特性,一方面从语义相似度与词语相关性方面深层次的理解文本,无监督的筛选出具有类间高区分度的文本特征,加强文本聚类准确性,另一方面以分布式表示特征向量,降低特征向量维度,降低算法的复杂度。

1 基本概念的相关定义

1.1 语义网络

《同义词词林》包含了中文词汇中大部分中文词汇同义词、广义相关词等信息。该词典采用分层划分的形式,具备5层结构,随着层次的攀升,越高层次对词义的划分越细,当到达第5层时,同一分类

收稿日期: 2014-03-18

第一作者: 男,1972年生,教授

* 通讯联系人

E-mail: wanj@mail.buct.edu.cn

中都是词义相同或相近的词语。

定义 1 根据文本中的词汇在《同义词词林》中距离的远近,定义词语相似度 $Sim(i,j)$ 为^[6]

$$Sim(i,j) = \theta_l \cos \left(m \frac{\pi}{180} \right) \left(\frac{m-k+1}{m} \right) \quad (1)$$

其中 θ_l 为《同义词词林》第 l 层分支下的系数, m 为该层总分类数, k 为两分类的距离。

定义 2 将文本中各词的义原作为节点,当节点 i 与节点 j 之间相识度 $Sim(i,j) > Sim_{\min}$ 时 (Sim_{\min} 为最小相似度阈值),在两结点间建立一条边,将文本表示成图状数据结构,本文中称该图为语义网络。

1.2 词语相关性

本文中词语相关性分析主要基于频繁集概念^[7]。

定义 3 以文本中各词汇集合 $D = \{Word_1, Word_2, \dots, Word_n\}$ 作为数据集,在给定最小支持度 sup_{\min} 的情况下,对于项集 $X = \{Word_i, Word_j, \dots, Word_k\} \subseteq D$,若 X 的支持度 $sup(X) > sup_{\min}$,则 X 是数据集 D 上的频繁项集, $Word_i, Word_j, \dots, Word_k$ 具有词语相关性。

本文并不需要得到所有的频繁集,根据频繁集定义可知,如果 X 是频繁集,并且对于 $\forall Y (Y \subseteq D \wedge X \subset Y)$,均有 $sup(Y) < sup_{\min}$,则称 X 是数据集 D 上的最大频繁集,且 X 的非空真子集一定是频繁项集。

定义 4 得到某一最小支持度 sup_{\min} 下所有最大频繁集 $MFSI = \{MFS_1, MFS_2, \dots, MFS_m\}$ 后,在给定最小覆盖率 cov_{\min} 的情况下,对于项集 MFS_1, MFS_2 若 $Cov(MFS_1, MFS_2) > cov_{\min}$,则把 MFS_1, MFS_2 从 $MFSI$ 中去除,并把 $MFS_1 \cup MFS_2$ 加入 $MFSI$ 中,直到不存在任何项集满足合并条件,这样项集的集合称之为相关词集。

定义 5 若相关词集中各项集作为一个线性图,并加入到语义网络中,最终得到的图称之为词汇网络。

1.3 深度学习网络

深度学习网络 (DBM) 由 1 个输入层、 L 个隐藏层以及 1 个输出层组成^[8], $n, m_1, m_2, \dots, m_L, m$ 分别为它们的神经元数。其中每两层构成一个受限玻尔兹曼机 (RBM), 因此 DBM 中共有 L 个 RBM, 它们对应的权系数矩阵、可见层和隐层的偏置向量分别为 $W^{(l)}, a^{(l)}, b^{(l)} (l=1, 2, \dots, L)$, 其中 $W^{(l)} \in R_{m_l \times m_{l-1}}, a^{(l)} \in R_{m_{l-1}}, b^{(l)} \in R_{m_l}$ 。

在对深度学习网络的训练中,单独无监督的训练每个 RBM, 确保特征向量在映射到不同维空间时能够较多的保留特征信息。在 RBM 模型中,能量函数定义为

$$E(v, h | \theta) = - \sum_{i=1}^n a_i v_i - \sum_{j=1}^m b_j h_j - \sum_{i=1}^n \sum_{j=1}^m v_i W_{ij} h_j \quad (2)$$

其中 v 为输入向量, h 为输出层向量, θ 为模型参数。当参数确定后,基于能量函数,输入层与输出层的联合概率分布为

$$P(v, h | \theta) = \frac{e^{-E(v, h | \theta)}}{Z} \quad (3)$$

由于 RBM 的二分图结构,可知当给定输入层状态时,隐层中各隐单元激活状态是条件独立的,其激活概率为

$$P(h_j = 1 | v, \theta) = \vartheta \left(b_j + \sum_i v_i W_{ij} \right) \quad (4)$$

$$\vartheta(x) = \frac{1}{1 + \exp(-x)} \quad (5)$$

同理第 i 个输入层单元激活概率为

$$P(v_i = 1 | h, \theta) = \vartheta \left(a_i + \sum_j W_{ij} h_j \right) \quad (6)$$

对 RBM 网络的训练是通过 Gibbs 采样,可视层与隐层互为条件,通过参数梯度公式(7)、(8)、(9)不断更新参数,直至收敛。

$$\Delta W_{xj} = P(h_j = 1 | v^{(l)}) (v^{(l)})^T - P(h_j = 1 | \hat{v}^{(l)}) (\hat{v}^{(l)})^T \quad (7)$$

$$\Delta a = v^{(l)} - \hat{v}^{(l)} \quad (8)$$

$$\Delta b = P(h_j = 1 | v^{(l)}) - P(h_j = 1 | \hat{v}^{(l)}) \quad (9)$$

2 深度词汇网络学习算法

2.1 总体流程

深度词汇网络学习算法 (WDLB) 框架如图 1 所示。通过简化后的频繁集规则对文本中词汇进行分析,发现词语相关性。基于《同义词词林》构建出文本中词汇的语义网络,加入词语相关性信息,得到文本词汇网络。在词汇网络的基础上,利用 pagerank 算法^[9]生成文本特征向量,并以此作为深信度网络的输入进行降维。最终实现对文本的聚类。

2.2 词语相关性挖掘算法

在大规模的数据中,最大频繁集往往有数千上万甚至更多个,挖掘它们的候选集往往会带来很高

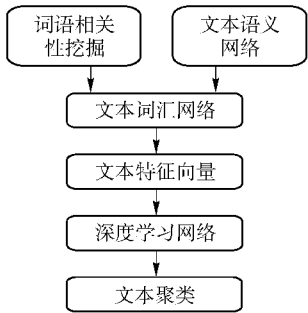


图 1 深度词汇网络学习算法框架

Fig. 1 WDLB algorithm framework

的复杂度和时间代价。考虑到算法并不需要所有的候选频繁集,而最大频繁集的非空子集一定是频繁集。本文利用《同义词词林》中义原的概念,以义原(同个义原代表多个语义相近的词)作为最小挖掘对象,并借鉴数学中随机抽样的方法,多次在不同的最小支持度(即 sup_{min})的条件下,随机抽取总数据集中 $n\%$ 的文章用以发现最大频繁集。在得到最大频繁集后,由于各频繁集存在很大的重复与冗余,这里基于文本词语相关性具有传递性假设,根据相似字符串匹配算法,对最大频繁集进行合并。假设现在有最大频繁集 $MFS_1 = \{a, b, c, d, e, f\}$, $MFS_2 = \{a, b, c, d, e, h\}$, $MFS_3 = \{e, f, g, h, i, j, k\}$,若最小覆盖率 $cov_{min} = 0.7$,合并后的最终结果为 $(MFS_1 \cup MFS_2) = \{a, b, c, d, e, f, h\}$, $MFS_3 = \{e, f, g, h, i, j, k\}$,最终得到一个相关词集 RS 。

2.3 词汇网络构建及特征筛选

文本信息的分析通常以词作为最小的分析单位,而文本中各词之间存在着语义相似度与词语相关性等信息。为了能够提取出可以代表文本自身、又能区别于其他类别的文本特征值,需要构建文本的词汇网络。词汇网络要考虑到词语间具有语义相似性与词语相关性的关系。算法采用义原作为节点,《同义词词林》中按照义原本身的含义,采用树形结构将其组织起来,形成一个义原树,这种义原树根据义原关系的远近,让各义原分布在不同层次上的各个节点中。首先,对文本中的所有词语进行义原标注,计算各义原之间语义相似度,当两义原之间相似度大于 Sim_{min} 时,为两义原节点之间建立联系,形成文本的语义网络图。然后再在语义网络的基础上加入词汇相关性,根据 2.2 节中得到的相关词集,为相关义原节点之间建立联系,各联系之间的指向根据文本中词频的大小确定(指向高词频词)。所构建的文本词汇网络中,词语在文本中的重要程度

则反映在网络中对应节点的链接度中,使用 pagerank 算法对各节点进行打分,选取网络中节点作为特征向量中的各项,并以其得分作为该项的值。通过词汇网络筛选出来的能够突出文本本身的语义特征,使得聚类簇可视化。

2.4 深度学习降维

深度学习是数据分布式表示的必然结果。分布式表示比一般的表示方式更加紧凑,不仅可以更好的体现出各概念间的相似性,而且在有限数据下能体现出更好的泛化性能。

本文将文本集中所有的特征向量作为一组未知的概率分布的输入,目标在于将其投影到低维分布中,且低维分布与输入样本尽可能地接近。

观察发现,由 2.3 节生成的特征向量维数较大,且各特征值之间存在组间相关性,影响后续聚类算法的效率与准确性。统计力学的结论表明,任何概率分布都可以转变成基于能量的模型。这里利用稀疏组受限波尔曼兹机^[10],将每篇文章的特征向量看作一个能量状态,则对文本的降维过程变为一个能量模型。通过对比散度(CD)学习法^[11],可以快速得到能量模型中的变量相关性,而得到的能量系统中能量最小时的解,即为所需要的目标解。以聚类参数 h 做为分组标准,把各隐层单元分为 h 的整数倍,例如假设 F 为 RBM 的隐层神经元集合,则分组个数为 $K(K = h \times n, n$ 为一整数),将分组的正则化系数加入到最终目标函数中,则最终目标函数为

$$\max_{W, b, c} \sum_{l=1}^L \lg P(\mathbf{v}^{(l)}) - \lambda \sum_{k=1}^K \sqrt{\sum_{m \in G_k} P(h_m = 1 | \mathbf{v}^{(l)})^2} \tag{10}$$

而在正则化系数作用在目标函数后,连接权增量式变为

$$\begin{aligned} \Delta W_{x_j} &= P(h_j = 1 | \mathbf{v}^{(l)}) P(h_j = 1 | \mathbf{v}^{(l)})^T - P(h_j = 1 | \hat{\mathbf{v}}^{(l)}) (\hat{\mathbf{v}}^{(l)})^T - \\ &\lambda \frac{P(h_j = 1 | \mathbf{v}^{(l)})^2 (1 - P(h_j = 1 | \mathbf{v}^{(l)})) \mathbf{v}^{(l)}}{\sqrt{\sum_{m \in G_k, m \neq j} P(h_m = 1 | \mathbf{v}^{(l)})^2 + P(h_j = 1 | \mathbf{v}^{(l)})^2}} \end{aligned} \tag{11}$$

由式(11)可看出,通过稀疏组深度学习网络对文本特征向量进行组内稀疏,不仅学习到了其自身的激活概率,还受组内其他神经元激活概率的影响,使得经过组稀疏深度学习网络降维后的特征向量仍然具有组稀疏的特征,更利于进行聚类分析。

3 实验结果与分析

实验进行条件为,Core(TM) i5-3317U CPU,4 G RAM,64 位 Window8 操作系统的 Matlab2012。实验语料取自 TanCorpV1.0、中国大百科及搜狗语料库。文本语料集如表 1 所示。

表 1 文本语料集
Table 1 Text corpus set

序号	类名	文章数量
1	体育	2805
2	军事	1990
3	卫生	1405
4	房产	900
5	教育	800
6	汽车	500
7	博物馆	300
总计		8400

3.1 文章特征词筛选

本节主要以 2.3 节中描述的算法对文本数据集中的文本数据进行词汇网络的构造,并提取关键词。算法中最小语义相似度 Sim_{min} 选取为 0.24, pagerank 算法中阻尼系数 α 选取为 0.85。由于展示所有结果需要大量篇幅,这里随机选取文本集中的 3 篇文章,其中两篇属于同一个类,另一篇来自不同的类,其中一篇原文如图 2,其词汇网络见图 3,经词汇网络提取关键词结果如表 2 所示。对比文章中关键词选取的准确性及其对之后聚类分析的贡献。

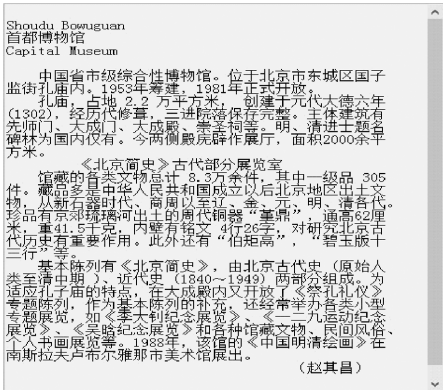


图 2 文章原文

Fig. 2 The original article

经过人工确认和比对,WDLB 算法所取得的文章关键义原可以很好的体现出文章的主题,同时可以作为类标签使得聚类结果可视化。表 3 是实验中

所用文章经过 WDLB 得到的特征向量与 Td-Idf 算法得到的特征向量之间欧式距离的对比。

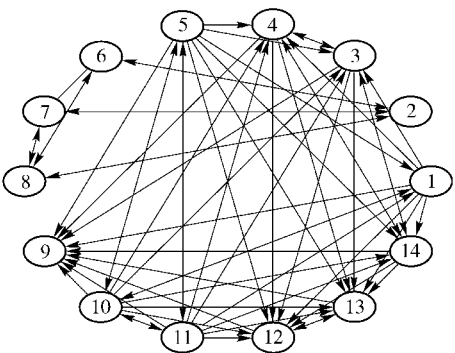


图 3 文章词汇网络

Fig. 3 The article vocabulary network

表 2 通过词汇网络提取关键词结果

Table 2 Text result of the feature selected based on vocabulary network

序号	义原	词林中义项
1	Hj47A01 =	展览 展出 展
2	Id08A01 =	陈列 陈放 罗列 摆 陈 位列 列支...
3	Ba01A17#	出土文物 文物
4	Jd04B02@	占地
5	Bn01A01 =	建筑 建筑物 构筑物
6	Dn01A54#	容积 面积 体积
7	Dd05A21#	面积 表面积 总面积
8	Hj40B01 =	储藏 贮藏 收藏 珍藏 保藏 藏...
9	Ba03A18#	工艺品 艺术品 展览品 陈列品...
10	Ba08A07#	文物 活化石 名物
11	Bn03A23 =	展览 展厅
12	Hg19A01 =	绘画 描画 描绘 画 绘 写 描 打 写生 点染 画画 作画
13	Dm07A08 =	博物馆 博物院
14	Bn01B38#	田径馆 训练馆 文史馆 军史馆...

表 3 欧式距离对比

Table 3 Euclidean distance comparison

算法类型	距离		
	军事 1-军事 2	军事 1-博物馆 1	军事 2-博物馆 1
WDLB	9.17	16.58	18.41
Td-Idf	13.02	15.56	13.69

由以上实验结果可以看出,经过词汇网络生成的文章特征向量能够减小同类文章之间的距离,同时增大不同类型文章之间的距离,可见这种方法能够提高同类型文章的凝聚度,提高文本的辨识度。

3.2 特征向量聚类特性

本节将 WDLB 算法与传统的 k -means 算法进行聚类对比实验,以证明 WDLB 算法的聚类有效性。由于 k -means 算法所选取的参数对结果有较大的影响,并有一定概率的随机性,这里采取多次取参进行实验,并取最大值进行对比。实验结果评定采用聚类质量检测方法 (F -measure),如式 (12) 所示。

$$F = 2PR / (P + R)$$

(12)

其中 P 为准确率, R 为召回率。

对两种算法分别以 $k = 7, 8, 9$ 时 3 个参数的实验得出的结果进行分析。由于 k -means 算法本身的簇中心随机性,结果具有一定的随机性,且有可能陷入局部最优。但相对于传统的 k -means 算法,词汇网络构造的文本特征向量不同类别中的文本区分度较大,无论采取基于密度选取初始中心的方法,还是最大距离法选取初始中心进行改进,WDLB 算法都比传统的 k -means 算法效果要好。

取 3 次实验中的最大值作为方法测试结果,如图 4 所示。

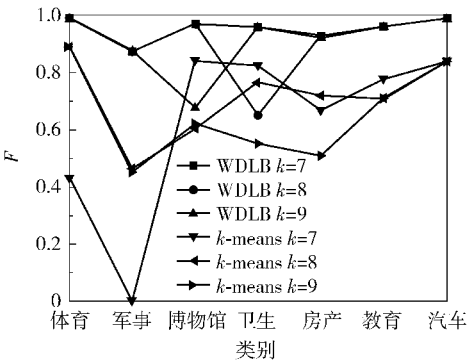


图 4 传统 k -means 算法与 WDLB 结果对比

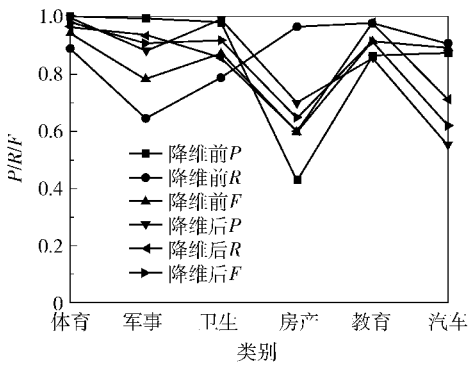
Fig. 4 Comparison of the results for k -means and WDLB

由以上结果可以很明显看出,基于词汇网络提取的文本特征向量比 Td-Idf 提取的文本特征向量有更好的聚类特性,且聚类稳定性较高,总体结果受参数影响较小。

3.3 深信度网络降维有效性

本节采取较大规模数据集对深信度网络进行降维测试,证明经过深信度网络降维后的文本特征向量仍然具有较好的聚类特性。深信度网络采用 3 层构建,第一层隐层单元数为 6000,第二层为 5000,第三层为 4000,收敛率设为 2.0,实验结果如图 5 所示。

由以上实验数据可以看出,在经过深信度网络



降维前特征向量维数 6342,聚类所用时间 38.66 s;降维后特征向量维数 4000,聚类所用时间 10.84 s。

图 5 使用深度学习降维前后结果对比

降维后的特征向量仍然具有较好的聚类效果,除此之外,经过降维后的数据大大降低了后续聚类算法的时间复杂度。因此本文算法可以在较大规模数据集上进行聚类分析。

4 结束语

在本文所提出的深度词汇网络学习算法中,通过词汇网络筛选出的特征词不仅能代表文章主题,为文本提供标签,使聚类可视化,同时以这些词作为特征项的特征向量,可增加各类文章的区分度;而采用深度学习网络对特征向量进行数据分布式表示,让数据更加紧凑,缩减了特征向量的存储空间,降低了聚类算法的时间复杂度,适用于文章主旨语义存在差异的文本聚类。不足之处在于此算法参数较多,算法结果容易受人为因素影响。

参考文献:

[1] Li Y J, Luo C N, Chung S M. Text clustering with feature selection by using statistical data[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(5): 641-652.

[2] Chen C L, Tseng F S C, Liang T. Mining fuzzy frequent itemsets for hierarchical document clustering[J]. Information Processing & Management, 2010, 46(2): 193-211.

[3] Song W, Li C H, Park S C. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures[J]. Expert Systems with Applications, 2009, 36(5): 9095-9104.

[4] 单丽莉,刘秉权,孙承杰. 文本分类中特征选择方法的比较与改进[J]. 哈尔滨工业大学学报, 2011(增刊)

- 1): 319–324.
- Shan L L, Liu B Q, Sun C J. Comparison and improvement of feature selection method for text categorization [J]. Journal of Harbin Institute of Technology, 2011 (Suppl 1): 319–324. (in Chinese)
- [5] Tomović A, Jančić P, Kešelj V. n-Gram-based classification and unsupervised hierarchical clustering of genome sequences[J]. Computer Methods and Programs in Biomedicine, 2006, 81(2): 137–153.
- [6] 田久乐, 赵蔚. 基于同义词林的词语相似度计算方法[J]. 吉林大学学报: 信息科学版, 2010, 28(6): 602–608.
- Tian J L, Zhao W. Words similarity algorithm based on Tongyici Cilin in Semantic WebAdaptive learning system [J]. Journal of Jilin University: Information Science, 2010, 28(6): 602–608. (in Chinese)
- [7] Han J W, Pei J, Yin Y W, et al. Mining frequent patterns without candidate generation: a frequent-pattern tree approach[J]. Data Mining and Knowledge Discovery, 2004, 8(1): 53–87.
- [8] 孙志军, 薛磊, 许阳明, 等. 深度学习研究综述[J]. 计算机应用研究, 2012, 29(8): 2806–2810.
- Sun Z J, Xue L, Xu Y M, et al. Overview of deep learning[J]. Application Research of Computer, 2012, 29(8): 2806–2810. (in Chinese)
- [9] Haveliwala T H. Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 784–796.
- [10] Luo H, Shen R M, Niu C Y, et al. Sparse group restricted Boltzmann machines[C]//Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, USA, 2011: 429–434.
- [11] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527–1554.

Text clustering based on deep vocabulary network learning

YI JunKai FENG JiaMing WAN Jing*

(College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: In the current methods of Chinese text clustering, most clustering algorithms are limited by the data scalability and algorithm efficiency. This paper presents a novel Chinese clustering algorithm based on deep vocabulary network learning. This algorithm has two key features. First, a vocabulary network is built to extract the key concepts, this not only resolves the “Word Sense Disambiguation” problem but also considers both syntactic and semantic information about words. Second, this algorithm use deep belief nets (DBN) to reduce the complexity of feature vectors, which maintains as much information as possible when distinguishing between features. This significantly reduces the processing times. The experimental results show that the new algorithm outperforms the well-known k -means clustering algorithm according to clustering quality measures.

Key words: vocabulary network; deep belief nets; Chinese text clustering