计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(7):245−250 [doi: 10.15888/j.cnki.csa.007508]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多特征融合 Single-Pass-SOM 组合模型的话题检测①

李丰男[1,2], 孟祥茹[2], 焦艳菲[3], 张琳琳[2], 刘　念[2]

[1](中国科学院大学 计算机控制与工程学院, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[3](沈阳高精数控智能技术股份有限公司, 沈阳 110168)
通讯作者: 李丰男, E-mail: 1148929576@qq.com

摘　要: 当今时代, 网络舆情传播速度快、影响力大, 而话题检测在网络舆情监管中有着不可替代的作用. 针对传统方法提取文本特征不完整和特征维度过高的问题, 本文提出了基于时间衰减因子的 LDA&&Word2Vec 文本表示模型, 将 LDA 模型的隐含主题特征和 Word2Vec 模型的语义特征进行加权融合, 并引入了时间衰减因子, 同时起到了降维和提高文本特征完整度的作用. 同时, 本文又提出了 Single-Pass-SOM 组合聚类模型, 该模型解决了 SOM 模型需要设定初始神经元的问题, 提高了话题聚类的精度. 实验结果表明, 本文提出的文本表示模型和文本聚类方法较传统方法拥有更好的话题检测效果.

关键词: 话题检测; 文本表示; SOM 聚类; Single-Pass 聚类; Single-Pass-SOM

引用格式: 李丰男,孟祥茹,焦艳菲,张琳琳,刘念.基于多特征融合 Single-Pass-SOM 组合模型的话题检测.计算机系统应用,2020,29(7):245–250. http://www.c-s-a.org.cn/1003-3254/7508.html

Topic Detection of Single-Pass-SOM Combination Model Based on Multi Feature

LI Feng-Nan[1,2], MENG Xiang-Ru[2], JIAO Yan-Fei[3], ZHANG Lin-Lin[2], LIU Nian[2]

[1](School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[3](Shenyang Golding NC Technology Co. Ltd., Shenyang 110168, China)

Abstract: Nowadays, internet public opinion has a rapid spread and great influence, and topic detection plays an irreplaceable role in the supervision of public opinion. Aiming at the problems of incomplete feature extraction and high feature dimension in traditional methods, this study proposes LDA&&Word2Vec text representation model based on time decay factor, which combines the hidden subject features by LDA model with the semantic features by Word2Vec model, and adds time decay factor, which can reduce the dimension and improve the integrity of text features. At the same time, this study proposes a Single-Pass-SOM clustering model, which solves the problem of setting initial neurons in SOM model, and improves the accuracy of topic clustering. Experimental results show that the text representation model and text clustering method proposed in this study have better topic detection effect than traditional methods.

Key words: topic detection; text representation; SOM clustering; Single-Pass clustering; Single-Pass-SOM

随着计算机技术和互联网的迅速发展, 越来越多的人习惯于通过互联网了解社会热点, 借助互联网发表个人的意见、看法和主张. 互联网已成为人们获取信息、发表意见、维护权益的重要场所. 因而, 如何监管舆情事件在互联网上的传播已成为一个具有现实意义的重大问题. 网络舆情具有传播速度快、影响力

① 收稿时间: 2019-12-18; 修改时间: 2020-01-14; 采用时间: 2020-01-22; csa 在线出版时间: 2020-07-03

大、参与性强的特点, 网民们的态度极易受到网络舆情传播方向的影响. 话题检测技术正是在这种情况下应运而生的. 它不仅能够帮助用户及时从海量数据中获取自己感兴趣的话题信息, 更能够帮助政府有关部门及时了解社会热点事件, 掌握社会舆论的方向, 这对于有效引导舆论、落实相关政策具有重大意义.

话题检测技术主要分为两大重要部分, 一是文本表示, 二是话题聚类. 文本表示是话题检测的基础. 传统的向量空间模型存在复杂度高、特征稀疏、噪声干扰严重等问题. 为了解决这些问题, 众多学者从不同方向进行了各种尝试. 路荣等[1]利用 LDA 话题模型有效解决了短文本的数据稀疏问题. 肖倩等[2]将 LDA 主题模型与卷积神经网络相结合, 摆脱了对语义信息的过度依赖. 李新盼[3]利用基于改进的 Word2Vec 和 tfidf 的文本表示模型, 有效解决了传统文本表示模型映射出的向量高维稀疏性和忽略语义相似度的问题. 但上述文本表示模型均只解决了某一方面的问题, 而未考虑尽可能包含全部文本信息. 在话题聚类方面, 陈艳红等[4]提出了一种基于信息熵和密度改进的 k-means 聚类算法, 降低了孤立点对算法性能的不利影响. 赵杨[5]将“话题簇代表”这一概念引入到 Single-Pass 聚类算法中, 降低了 Single-Pass 聚类算法的计算量. 传统的聚类算法在话题检测方面有着诸多应用, 但神经网络聚类在该方面的应用却较少.

针对以上问题, 本文提出了一种基于 Single-Pass 聚类和 SOM 神经网络聚类的话题检测方法. 该方法利用词向量获取文本的语义信息, 利用 LDA 话题模型获取文本的主题信息, 有效克服了文本聚类过程中特征维数高、数据稀疏的问题. 并考虑到时间推移对话题兴趣点的影响, 引入了时间衰减因子. 同时, 将 Single-Pass 聚类和 SOM 聚类相结合, 利用了 Single-Pass 聚类运算速度快且不需要提前设定聚类个数的优点, 先获得模糊聚类个数和权值矩阵, 并将其作为 SOM 聚类的初始神经元个数和连接权向量, 解决了 SOM 神经网络聚类需要提前确定初始神经元的问题, 进一步提高了话题检测的准确率和效率.

1 基于时间衰减因子的 LDA&&Word2Vec 文本表示模型

1.1 词向量模型

词向量是由 Hinton[6]提出的一种词语的特征表示, 它的基本思想是通过对大量未标注的文本数据进行无监督的语言模型训练, 将词语表示成一组低维实数向量, 以此来刻画词语的语义特征. Word2Vec 是由 Google 于 2013 年发布的词向量训练工具, 它能够从大规模未经标注的语料中高效地生成词的向量形式. 该模型可以通过减少训练过程中所需要的参数, 避免过拟合, 提升了训练效率. 因而本文为了获取文本的语义信息, 同时避免词向量的训练过程过于复杂, 采用了 Word2Vec 词向量方法进行文本的向量化表示.

1.2 LDA 主题模型

潜在狄利克雷分布 (Latent Dirichlet Allocation, LDA) 模型是由 Blei 等[7]于 2003 年提出的一种贝叶斯概率模型. 该模型具有优秀的话题建模能力, 能够有效实现文本的降维表示, 这些都促使其在话题检测领域得到了广泛应用.

LDA 模型是一个 3 层的文档生成模型, 主要结构包括文档、主题、词. 该模型基于这样的假设: 每个文档都是由多个隐含主题构成的, 而每个主题又是由多个相关词汇构成的, 其拓扑结构如图 1 所示.

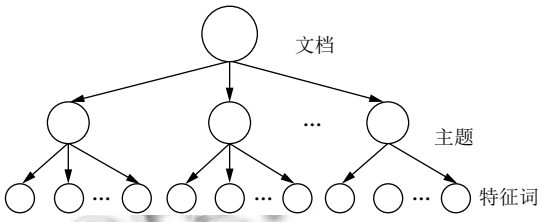

图 1 LDA 拓扑结构图

对于 LDA 主题模型, 在仅给定文本数据集的情况下, 可以采用 Gibbs 采样对模型未知参数进行估计, 进而得出文档-主题分布和主题-词分布.

1.3 文本相似度

由于基于词汇级别的语义特征向量只能对文本的浅层语义分布特征进行表示, 缺乏对主题信息的具体描述. 而基于 LDA 模型的主题特征向量恰好能对语义特征向量在特征表示上的不足进行补充. 因此, 本文采用多特征融合的方法结合了文本的主题特征和语义特征, 使得最终求得的文本相似度中综合考虑了文本的主题及语义信息, 具体过程如下所示:

(1) 采用 LDA 主题模型获取文本的主题特征, 根据主题特征采用 JS 距离来计算各文本主题分布的相似度. 利用 JS 距离公式计算文档 $p = p_1, p_2, \cdots, p_n$ 和文

档$q=q_1,q_2,\cdots,q_n$的主题相似度$sim_{LDA}(p,q)$如下:

$$D_{KL}(p,q)=\sum_{j=1}^{n}p_j\ln\frac{p_j}{q_j} \tag{1}$$

$$sim_{LDA}(p,q)=D_{JS}(p,q)=\frac{1}{2}\left[D_{KL}\left(p,\frac{p+q}{2}\right)+D_{KL}\left(q,\frac{p+q}{2}\right)\right] \tag{2}$$

其中, p和q为两个文本的主题概率向量, $D_{KL}(p,q)$为KL 距离. 由于其计算距离时不满足相似度对称性, 因此一般采用 JS 距离计算相似度.

(2) 采用 Word2Vec 词向量模型获取文本的语义特征, 根据语义特征采用余弦相似度来计算文本相似度. 利用余弦相似度公式计算文档$p=p_1,p_2,\cdots,p_n$和文档$q=q_1,q_2,\cdots,q_n$的语义相似度$sim_{W2V}(p,q)$如下:

$$sim_{W2V}(p,q)=\frac{\sum_{i=1}^{n}p_iq_i}{\sqrt{\sum_{i=1}^{n}p_i^2}\sqrt{\sum_{i=1}^{n}q_i^2}} \tag{3}$$

(3) 采用加权融合的方法结合文本的主题相似度和语义相似度, 文档$p=p_1,p_2,\cdots,p_n$和文档$q=q_1,q_2,\cdots,q_n$的文本相似度$sim(p,q)$的具体计算公式如下:

$$sim(p,q)=\alpha*sim_{LDA}(p,q)+\beta*sim_{W2V}(p,q) \tag{4}$$

其中, α和β分别表示$sim_{LDA}(p,q)$和$sim_{W2V}(p,q)$的权值, $\alpha+\beta=1$.

(4) 时间衰减因子同样是判断两个文本是否属于同一话题的重要因素. 两个文本的发布时间相隔越远, 这两个文本属于同一话题的可能性就越低, 那么, 就应该赋予较低的权重. 这是因为话题是具有一定生命周期的. 对于大众用户来说, 随着时间的推移其对该话题的兴趣点会慢慢淡化或者转移到新的话题上. 因此, 本文根据牛顿冷却定律设计了时间衰减因子, 用来表示大众对话题兴趣的下降. 本文设计时间衰减因子的计算公式如下:

$$T(t)=T(t_0)*e^{-k(t_0-t)} \tag{5}$$

其中, t_0, t分别表示两个文本的发布时间, k为衰减率, 表示大众对话题兴趣的下降速度.

将该时间衰减因子引入到本文的文本相似度计算中, 得到最终的相似度计算公式:

$$sim(p,q)=e^{-k(t_0-t)}*\alpha*sim_{LDA}(p,q)+e^{-k(t_0-t)}*\beta*sim_{W2V}(p,q) \tag{6}$$

2 文本聚类模型

2.1 Single-Pass 聚类算法

Single-Pass 算法是一种增量聚类算法, 它计算简单, 运行速度快, 且不需要预先指定聚类个数, 常应用于大规模文本聚类. 其基本思想是: 按照一定的顺序输入文本, 将第一个输入的文本作为第一个话题簇, 当后续文本继续输入时, 判断输入文本与已有话题簇的相似度, 选择输入文本与已有某个话题簇的最大相似度, 并判断是否满足相似度阈值要求, 满足则把输入文本归入到最大相似话题簇, 反之则说明输入文本与已有话题簇均为不同类别, 那么创建新的话题簇, 重复上述过程直到所有的文本处理结束.

虽然 Single-Pass 算法简单易懂, 并且在处理流数据时极具优势, 但它也存在一些缺点:

(1) 输入顺序对聚类结果的影响程度很大. 对于相同的文档集合, 不同的输入顺序很可能会导致不同的聚类结果.

(2) 聚类精度较低. Single-Pass 聚类算法仅仅遍历文本一次, 如果聚类结果出现偏差, 无法动态更新.

针对以上提出的不足, 本文对该算法做了以下改进:

(1) 在输入待聚类文本之前, 按照文本发布时间对其进行排序, 这符合话题的演变过程, 以此来减少不同输入顺序对聚类结果的影响.

(2) 因 Single-Pass 聚类算法聚类精度较低, 本文仅使用该算法进行粗聚类, 获取模糊聚类个数和中心点位置, 作为后续 SOM 神经网络算法的初始化参数.

2.2 SOM 神经网络聚类算法

自组织特征映射神经网络 (Self-Organizing feature Map, SOM) 是由 Kohonen[8]提出的一种无监督的竞争学习型前馈网络. 该模型能够通过神经网络学习获得数据的重要特征或内在规律, 从而将数据划分到不同的区域, 达到对数据聚类的效果. SOM 神经网络的网络结构仅由输入层和竞争层 (输出层) 构成. 输入层的每一个神经单元均与竞争层的每一神经单元相连接, 构成全互连的结构, 从而保证了输入层获取到的全部信息均能传输到竞争层. 其神经网络结构如图 2 所示:

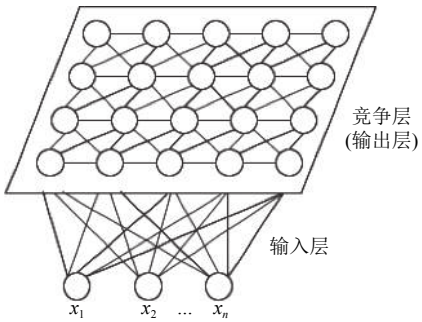

图 2 SOM 神经网络结构图

SOM 神经网络的训练过程主要分为竞争、合作和权值调整这 3 个阶段. 其算法流程图如图 3.

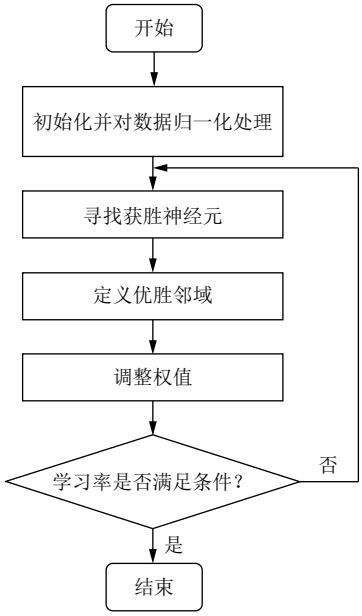

图 3 SOM 算法流程图

2.3 Single-Pass-SOM 组合聚类算法

SOM 神经网络聚类算法网络结构较为简单, 学习速度快, 具有较强的泛化能力, 适用于大规模数据的聚类. 但该模型也有一定的缺点, 在传统的 SOM 神经网络聚类算法中, 其权值的初始值是通过随机选择产生的, 这在一定程度上会影响该模型的聚类效果. 因此, 在初始化参数阶段, 本文提出采用 Single-Pass 聚类算法先进行粗聚类, 得到话题聚类的中心点, 将其作为 SOM 神经网络聚类算法权值的初始值, 使用 SOM 神经网络聚类算法进行细聚类, 得到最终的聚类结果. 具体流程如下:

(1) 按照文本的发布时间顺序输入待聚类的文本向量, 执行 Single-Pass 算法, 得到初始聚类数目 K 和初始权值矩阵 M.

(2) 将文本向量输入到 SOM 神经网络进行训练, 该神经网络采用 Single-Pass 算法确定的聚类数目 K 以及对应的权向量作为初始神经元个数和权向量.

(3) 获得 Single-Pass-SOM 组合聚类的结果, 并在此基础上进行相关分析.

Single-Pass-SOM 组合聚类算法结合了 SOM 网络和 Single-Pass 算法的优点, 同时弥补了各自的缺陷, 是一种较为理想的聚类方法.

3 实验分析

基于 Single-Pass-SOM 组合聚类算法, 本文构建了话题检测模型. 具体流程如图 4 所示.

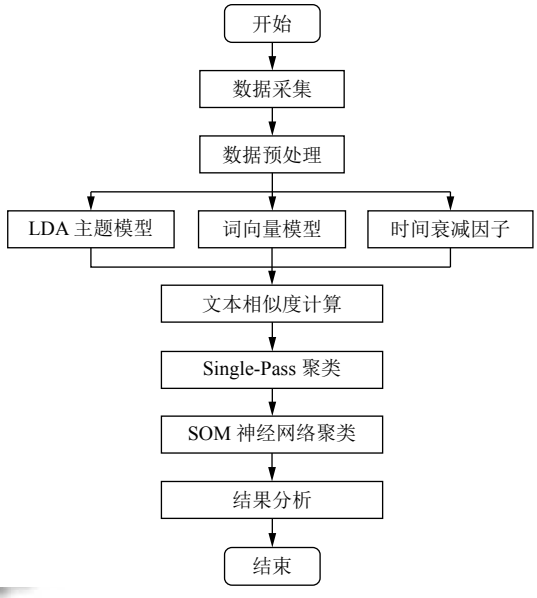

图 4 话题检测算法流程图

3.1 实验数据及其预处理

本文的实验数据为通过网络爬虫爬取的来自 20 个政府新闻门户网站以及新浪微博可供访问的从 2018 年 12 月到 2019 年 4 月的相关政策、政务新闻, 共约 10 万条文本数据.

通过网络爬虫获取到的原始数据含有大量的脏数据, 因此本文对获取到的实验数据进行了必要的预处理操作, 包括去除重复文本数据、分词、去除停用词、去除特殊符号等操作.

3.2 实验评价指标

在话题检测中常用的评价指标包括准确率 (P)、召回率 (R)、$F1$ 值. $F1$ 值是召回率和准确率的几何加

权均值, 可以更精确地衡量话题检测的精度, $F1$ 值越大, 话题检测效果越好. 计算公式如下:

$$P=\frac{TP}{TP+FP} \tag{7}$$

$$R=\frac{TP}{TP+FN} \tag{8}$$

$$F1=\frac{2*P*R}{P+R} \tag{9}$$

其中, TP 代表已检测到的正确的文档数, FP 代表已检测到的不正确的文档数, FN 代表未检测到的正确的文档数.

3.3 实验结果分析

3.3.1 不同文本表示模型对实验结果的影响

本实验分别选用 LDA 主题模型, Word2Vec 词向量模型, LDA&&Word2Vec 模型和本文提出的基于时间衰减因子的 LDA&&Word2Vec 文本表示模型进行性能对比. 在实验过程中, 首先采用这 4 种模型实现文本的向量表示, 再使用 Single-Pass-SOM 组合聚类模型进行文本话题检测, 并计算话题检测的准确率、召回率和 $F1$ 值, 最后, 对实验结果进行比较分析, 验证本文提出的文本表示模型的有效性. 实验结果如图所示, 图 5 为不同话题下 4 种文本表示模型话题检测结果的准确率、召回率和综合指标 $F1$ 值.

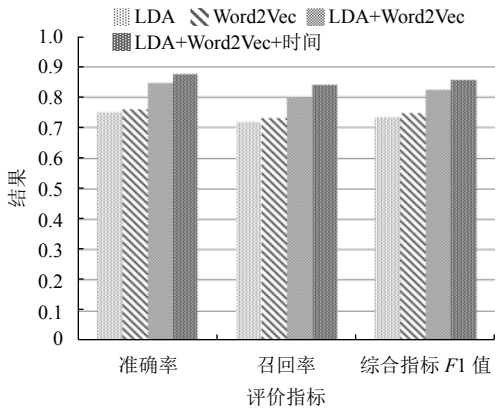

图 5 文本表示模型话题检测结果的 P、R 和 $F1$ 值

从图 5 可以看出本文提出的基于时间衰减因子的 LDA&&Word2Vec 文本表示模型无论是在准确率上, 还是在召回率上均优于其他 3 种文本表示模型. 从 $F1$ 值上看, 单独的 LDA 模型和 Word2Vec 词向量模型实验结果差异不大, 分别为 73.3% 和 74.6%. 而将这两种文本表示模型相结合的 LDA&&Word2Vec 模型, 由于综合了这两种模型的优点, 既通过 LDA 主题模型获取了文本的主题信息, 又通过 Word2Vec 词向量模型解决了文本数据稀疏和向量高维的问题, 其 $F1$ 值提高了 8.97%. 同时, 在 LDA&&Word2Vec 模型的基础上加入时间衰减因子, 考虑到了时间对话题检测效果的影响, 其 $F1$ 值又提高了 3.44%.

3.3.2 不同方法实验结果对比

本实验在对文本数据采用基于时间衰减因子的 LDA&&Word2Vec 模型进行文本向量表示的基础上, 对 Single-Pass 聚类模型、SOM 神经网络聚类模型和 Single-Pass-SOM 组合聚类模型 3 种聚类模型分别进行实验, 并对实验结果进行比较分析. 实验结果如图所示, 图 6 为不同话题下 3 种聚类模型话题检测结果的准确率、召回率和综合指标 $F1$ 值.

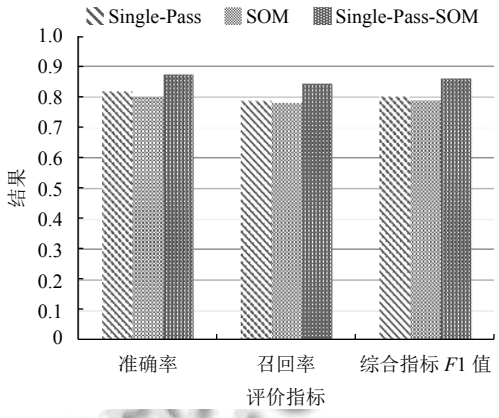

图 6 聚类模型话题检测结果的 P、R 和 $F1$ 值

从图 6 可以看出, 本文提出的 Single-Pass-SOM 组合聚类模型相比于单独的 Single-Pass 模型和 SOM 模型, 其在准确率、召回率和 $F1$ 值上均有更好的表现. 其在准确率上提高了 6%~7%, 在召回率上提高了 6%, 在综合指标 $F1$ 值上有 5%~7% 的提高. 原因在于 Single-Pass-SOM 组合聚类模型使用 Single-Pass 聚类模型解决了 SOM 神经网络模型初始化神经元设定的问题, 同时又用 SOM 神经网络模型提高了 Single-Pass 聚类模型的话题检测的精度.

4 总结

本文提出的 Single-Pass-SOM 组合聚类模型, 采用 LDA 主题模型和 Word2Vec 词向量模型从文本数据中获取更高阶的文本特征, 以此来解决特征维度

高、文本数据稀疏、主题不明显的问题, 并引入了时间衰减因子, 综合考虑了时间对话题兴趣衰减的影响. 同时, 很好地将 Single-Pass 聚类模型和 SOM 神经网络模型结合起来. 实验结果表明, 相对于单独的 Single-Pass 聚类和 SOM 神经网络聚类的话题检测方法, Single-Pass-SOM 组合聚类模型在准确率、召回率和 *F*1 值上均有明显提高.

参考文献

1 陈艳红, 向军, 刘嵩. 高校网络舆情分析的 K-Means 算法优化研究. 湖北民族学院学报 (自然科学版), 2018, 36(4): 442–447.

2 路荣, 项亮, 刘明荣, 等. 基于隐主题分析和文本聚类的微博客中新闻话题的发现. 模式识别与人工智能, 2012, 25(3): 382–387. [doi: 10.3969/j.issn.1003-6059.2012.03.004]

3 肖倩, 谢海涛, 刘平平. 一种融合 LDA 与 CNN 的社交媒体中热点舆情识别方法. 情报科学, 2019, 37(11): 27–33.

4 李新盼. 基于微博的网络舆情分析系统的设计与实现[硕士学位论文]. 成都: 电子科技大学, 2017.

5 赵杨. 面向热点话题的舆情演化分析方法研究[硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2018.

6 Hinton GE. Learning distributed representations of concepts. Proceedings of the 8th Annual Conference of the Cognitive Science Society. London, UK, 1986: 1–12.

7 Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3(3–4): 993–1022.

8 Kohonen T. The ‘neural’ phonetic typewriter. Computer, 1988, 21(3): 11–22. [doi: 10.1109/2.28]