文章编号:1673-9469(2022)01-0092-07 DOI: 10.3969/j.issn.1673-9469.2022.01.014

融合 LDA 和 GloVe 模型的病症文本聚类算法

吴 迪,赵玉凤

(河北工程大学 信息与电气工程学院,河北 邯郸 056038)

摘要:针对隐含狄利克雷分布(LDA)模型特征提取时忽略语义信息的问题,提出一种融合 LDA 和全局文本表示(GloVe)模型的病症文本聚类算法 LG&K-Medoide。首先,利用 LDA 对病症文本数据建模,采用 JS(Jensen-Shannon)距离计算文本相似度;其次,利用 GloVe 对病症文本数据建模获取词向量,根据病症词性贡献度,对词向量权重进行标注,采用余弦距离计算基于 GloVe 建模加权的文本相似度;最后,将两种相似度进行结合,改进距离公式,实现 K-Medoide 聚类。实验结果表明,LG&K-Medoide 算法较基于 LDA,LDA+TF-IDF,LDA+Word2Vec 模型的聚类算法具有较高的精度。

关键词:病症文本;LDA;GloVe;相似度结合;聚类

中图分类号:TG391 **文献标识码**:A

Disease Text Clustering Algorithm Based on LDA and GloVe Model

WU Di, ZHAO Yufeng

(School of Information and Electrical Engineering, Hebei University of Engineering, Handan, Hebei 056038, China)

Abstract: Aiming at solving the problem of ignoring semantic information in LDA model feature extraction, a disease text clustering algorithm LG&K-Medoide based on LDA and GloVe model was proposed. First, LDA was used to model the disease text data, and the JS distance was used to calculate the text similarity; second, GloVe was used to model the disease text data to obtain the word vector, the weight of the word vector was labeled according to the contribution to part of speech from disease text, and the cosine distance was used to calculate weighted text similarity based on GloVe modeling; finally, the two similarities are combined to improve the distance formula to realize K-Medoide clustering. The experimental results show that the LG&K-Medoide algorithm has higher accuracy than the clustering algorithm based on LDA, LDA+TF-IDF and LDA+Word2Vec models.

Key words: disease text; LDA; GloVe; similarity combined finite; clustering

随着医疗信息平台日益普及,医疗数据日益丰富,整个社会对医疗信息的需求巨大[1]。本文采用主题模型技术,对病症文本数据进行深入分析,有助于患者根据自身病症了解所患疾病,辅助医生进行临床决策[2],提高医院医疗服务质量,为总结各类病症发展趋势以及自主诊断发展等有着巨大的价值。主题模型是一种识别文本集潜在主题信息的概率生成模型[3]。与潜在语义分析(Latent Semantic Analysis, LSA)、概率隐性语义分析(Probabilistic Latent Semantic Analysis, PLSA)相比较,隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)主题模型化解了随着文本数量直线增加而产生的的过度拟合问题,实现了概率化[4]。闫俊伢等[5]利用 LDA 模型表示文本,将其输入到 K-means 中进行聚类分析。Kim 等[6]提出了利用 LDA 模型提取关键词,并计算词频-逆文档频度(Term Frequency-Inverse Document Frequency, TF-IDF)值,将其应用到 K-means 聚类算法获取主题相似的文本内容。王少鹏等[7]提出了一种基于 LDA 模型的文本聚类算法,该算法将 TF-IDF 和 LDA 建模后进行相似度融合,从而实现文本聚类。由于 TF-IDF 所构建的矩阵具有较高的稀疏性[8],

收稿日期:2021-06-21

基金项目:河北省自然科学基金资助项目(F2020402003,F2019402428)

作者简介:吴迪(1984-),女,河北邯郸人,博士,副教授,从事数据挖掘、文本聚类、自然语言处理方面的研究。

2013 年 Google 发布了 Word2Vec 词向量训练工具,其不仅具有降维效果还能够表示词的语义信息[9-10]。Chen 等[11]提出了将文本用 LDA 进行表示后,利用 Word2Vec 计算主题之间的语义关联,以提高关键词准确性。Kim 等[12]提出了一种基于 Word2Vec 的建模方法,该方法基于 Word2Vec 和 K-means 聚类能够提高捕捉和表示语料库文本的能力。郑恒毅等[13]利用 LDA 主题模型提取特征词隐含主题,Word2Vec 获取特征词向量,将两者融合以实现文本聚类。Word2Vec 提出不久,Pennington J 等在 2014 年提出了全局文本表示(Global Vectors for Word Representation, GloVe)模型,其充分考虑了语料中的统计信息,使其能够携带更多的语义信息。王欣研等[14]提出了将 LDA 和 GloVe 模型进行主题语义关联,通过 LDA 识别主题并基于 GloVe 相似性获取主题语义关联。李少华等[15]提出了 GV-LDA 模型,该模型在 LDA 建模前,利用 GloVe 提取词向量,将相似性较高的词替换以降低稀疏性。

综上所述,传统聚类方法特征提取时,在局部上下文窗口训练模型,忽略了文本集中的部分统计信息。因此,本文提出一种融合 LDA 和 GloVe 模型的病症文本聚类算法。LDA 建模后利用 JS 距离计算相似度,以提取基于主题表示的文本相似度;GloVe 建模后利用余弦距离计算相似度,以提取基于词向量表示的文本相似度;将两者结合后进行 K-Medoide 聚类,以提高病症文本聚类精度。

1　问题定义

词性是以语法特征作为主要依据同时兼顾词语意义的划分结果。对于病症文本数据而言,名词与其它词的重要性不同,名词一般为概括性词语,具有代表性,因此,词性可作为病症文本特征词提取时重要的衡量指标之一。为了提高医疗名词这类蕴含主要特征的单词对病症文本相似度影响,在 GloVe 词向量建模时应把词的词性因素考虑进去,提出词性贡献权重。利用 GloVe 建模后得到病症词向量,并根据相应词性,对词向量权重进行标注,进而计算病症文本向量。

定义 1(词性贡献权重)假设 $\bar{\omega}_{\text{GloVe}}$ 表示 GloVe 建模后获得的病症词向量,μ_i 表示第 i 种词性的贡献权重,则基于词性贡献权重的病症词向量 $\bar{\omega}_{\text{GloVe\_ps}}$ 公式如下:

$$\bar{\omega}_{\text{GloVe\_ps}} = \mu_i \cdot \bar{\omega}_{\text{GloVe}} \tag{1}$$

式中,i 的取值范围为 $i = 1,2$,μ_1 表示名词贡献度,μ_2 表示其它词贡献度。设定 $\mu_1 = 1$,$\mu_2 = 0.5$。

对病症文本数据进行聚类时,利用相似度计算出病症文本与各个聚簇中心之间的距离,进而判断该病症文本所属簇,因此距离计算对聚类效果尤为重要,故本文提出相似度结合距离。根据 LDA 和融合词性的 GloVe 分别对病症文本建模后,利用 JS 和余弦距离计算得到文本相似度,并将其进行结合。

定义 2(相似度结合)假设 $F = \{f_1, f_2, \cdots, f_n\}$ 表示病症文本数据集 $\{f_1, f_2, \cdots, f_n\}$,$f_i$ 表示第 i 个病症文本,c_K 表示聚簇中心,λ 表示线性结合系数,则文本 f_i 与 c_K 的相似度 $\text{Dt}(f_i, c_K)$ 公式如下:

$$\text{Dt}(f_i, c_K) = \lambda \cdot \text{Dt}_{\text{LDA}}(f_i, c_K) + (1 - \lambda) \cdot \text{Dt}_{\text{GloVe}}(f_i, c_K) \tag{2}$$

其中,λ 取值范围为 $0 < \lambda < 1$,i 的取值范围为 $i = 1,2,\cdots,n$。

2　融合 LDA 和 GloVe 模型的病症文本聚类算法

针对 LDA 主题建模时忽略语义信息以及病症文本词性贡献度不同的问题,本文提出了一种融合 LDA 和 GloVe 模型的病症文本聚类算法。首先,收集病症文本数据,构建医疗专业词汇分词词典,并对病症文本数据进行预处理;其次,利用 LDA 主题模型对预处理后的病症文本集建模,得到基于主题的文本表示,采用 JS 距离计算文本相似度;再次,利用 GloVe 全局向量模型对预处理后的病症文本集建模,根据病症词性贡献度赋予词向量权重,得到基于词向量的文本表示,采用余弦距离计算文本相似度;最后,将两种相似度结合应用到 K-Medoide 聚类中,进而得到病症文本数据的聚类结果。具体框架如图 1 所示。

2.1　病症文本数据预处理

病症文本数据集预处理主要包括筛选、中文分词、词性标注、去停用词四个部分。针对病症文本数据包含诸多医学专有词汇的问题,构建医疗专业词汇分词词典对病症数据进行分词,以提高病症分析结果的准确度。具体流程如图 2 所示。

2.2　LDA&GloVe 建模

对病症文本数据建模时分为基于 LDA 主题建模和基于 GloVe 词向量建模的文本相似性度量两

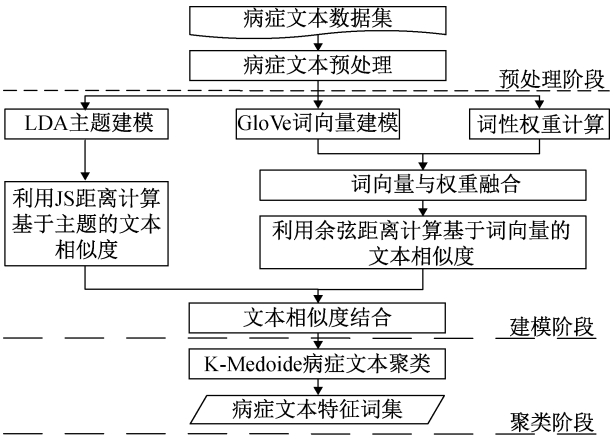

图1 融合LDA和GloVe模型的病症文本聚类算法框架图

Fig. 1 Framework diagram of disease text clustering algorithm based on LDA and GloVe model

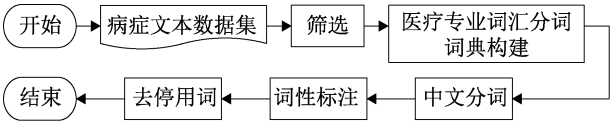

图2 病症文本数据集预处理流程图

Fig. 2 Flowchart for preprocessing of disease text data set

种部分。

(1) 利用困惑度选择最优主题数目,对病症文本数据集进行LDA主题建模,使用JS距离计算基于主题表示的文本相似度。

(2) 对病症文本数据集进行GloVe词向量建模,根据病症词性贡献度不同,对词向量权重进行标注,使用余弦距离计算基于词向量表示的文本相似度。

2.2.1 LDA文本相似性度量

本文利用Gibbs采样算法求解主题分布和词分布。LDA为无监督模型,建模前需先确定α、β、K三个超参数,α、β选取默认值,主题数的选取直接影响LDA模型对病症文本数据的释义情况,本文利用困惑度确定K的最优值,困惑度最小时表示建模结果最理想。病症文本F的困惑度公式perplexity(F)如下:

$$perplexity(\mathrm{F}) = \exp\left\{-\frac{\sum_{\mathrm{f}=1}^{M}\log P(\mathrm{W_f})}{\sum_{\mathrm{f}=1}^{M}N_{\mathrm{f}}}\right\} \quad (3)$$

式中,M表示病症文本语料库大小,N_{f}表示文本f中的词数量,$\mathrm{W_f}$表示文本f中的词,$p(\mathrm{W_f})$表示词$\mathrm{W_f}$产生的概率。

LDA主题建模完成后,对于每一篇病症文本,在其文本-主题分布$p(\mathrm{t}|\mathrm{f})$的最大概率主题下选取主题-词分布$p(\mathrm{w}|\mathrm{t})$概率中前6个词作为该文本的特征词,这样即可以最大化保留文本语义又能够降低算法复杂度。基于LDA主题模型的文本向量可用文本-主题概率分布表示,文本$\mathrm{f}_{i_\mathrm{LDA}}$计算公式如下:

$$\mathrm{f}_{i\_\mathrm{LDA}} = \{p(\mathrm{t}_1 \mid \mathrm{f}_i), p(\mathrm{t}_2 \mid \mathrm{f}_i), \cdots, p(\mathrm{t}_K \mid \mathrm{f}_i)\} \quad (4)$$

两篇病症文本间可通过计算这两个文本-主题分布之间的相似性来实现。本文使用JS距离计算基于LDA主题建模的相似度,文本$\mathrm{f}_{i_\mathrm{LDA}}$和$\mathrm{f}_{j_\mathrm{LDA}}$的相似度$\mathrm{Dt_{js}}(\mathrm{f}_{i_\mathrm{LDA}}, \mathrm{f}_{j_\mathrm{LDA}})$计算公式如下:

$$\mathrm{Dt_{js}}(\mathrm{f}_{i\_\mathrm{LDA}}, \mathrm{f}_{j\_\mathrm{LDA}}) = \frac{1}{2}\left[\mathrm{Dt_{KL}}\left(\mathrm{f}_i, \frac{\mathrm{f}_i + \mathrm{f}_j}{2}\right) + \mathrm{Dt_{KL}}\left(\mathrm{f}_j, \frac{\mathrm{f}_i + \mathrm{f}_j}{2}\right)\right] \quad (5)$$

基于LDA建模和JS距离(LDA-JS)的相似性如算法1所示。

算法1:基于LDA-JS的相似性算法

输入:病症文本数据集$F = \{\mathrm{f}_1, \mathrm{f}_2, \cdots, \mathrm{f}_M\}$,超参数$\alpha$和$\beta$

输出:基于LDA的文本相似度$\mathrm{Dt_{LDA}}(\mathrm{f}_i, \mathrm{f}_j)$

步骤1:根据公式(4)确定主题数K;

步骤2:for $\mathrm{f}_m \in F$ do

步骤3:从α分布取样生成f_m的主题分布θ_m

步骤4:从θ_m中取样生成第n个词的主题$\mathrm{t}_{m,n}$

步骤5:从β分布取样生成主题对应的词分布$\eta_{\mathrm{t}_{m,n}}$

步骤6:end for

步骤7:统计主题和特征词的共现频率,得到主题分布和主题-词分布

步骤8:根据公式(5)为每条病症文本选取特征词并对其进行向量化表示

步骤9:根据公式(6)计算基于LDA的文本相似度并输出

2.2.2 GloVe文本相似性度量

GloVe病症词向量的训练过程如下:首先,扫描整个病症文本数据集,根据上下文窗口的大小,统计目标词与上下文词在整个病症文本数据集中共同出现的次数,构造词共现矩阵。然后,以矩阵作为输入,使用最小二乘法作为损失函数,通过训练使损失函数值最小时得到最终的词向量。

GloVe模型作者Pennington进行对比实验得出词向量维度和上下文窗口大小的经验值分别为300和8。

病症文本数据集中,由于病症词的词性贡献

度不同,在 GloVe 词向量建模完成后,根据其词性贡献度对词向量权重进行标注,基于词性贡献度的病症文本向量 $\bar{f}_{\text{GloVe\_ps}}$ 公式如下:

$$\bar{f}_{\text{GloVe\_ps}} = \sum_{i=1}^{n} \bar{\omega}_{i\_\text{GloVe\_ps}} \tag{6}$$

利用余弦距离计算基于 GloVe 词向量建模的文本相似度,文本 $f_{i\_\text{GloVe\_ps}}$ 和 $f_{j\_\text{GloVe\_ps}}$ 的相似度 $\text{Dt}_{\text{GloVe\_ps}}(f_{i\_\text{GloVe\_ps}}, f_{j\_\text{GloVe\_ps}})$ 计算公式如下:

$$\text{Dt}_{\text{GloVe\_ps}}(f_{i\_\text{GloVe\_ps}}, f_{j\_\text{GloVe\_ps}}) = \text{Dt}_{\text{COS}}(f_i, f_j) = \frac{\bar{f}_i \cdot \bar{f}_j}{2 \mid f_i \mid \cdot \mid f_j \mid} \tag{7}$$

基于 GloVe 建模和 COS 距离(GloVe-COS)的相似性如算法 2 所示。

算法 2:基于 GloVe-COS 的文本相似性算法

输入:病症文本数据集 $F = \{f_1, f_2, \cdots, f_M\}$,词向量维度 vector_size,上下文窗口大小 window_size
输出:基于 GloVe 的文本相似度 $\text{Dt}_{\text{GloVe\_ps}}(f_i, f_j)$
步骤 1:构造病症文本数据集的词共现矩阵
步骤 2:根据词共现矩阵和 GloVe 建模获得病症词向量集 $W = \{w_1, w_2, \cdots, w_n\}$
步骤 3:for i = 1 to n do
步骤 4:根据词性标注判断 w 是名词还是其它词
步骤 5:根据公式(1)对词向量权重进行标注
步骤 6:end for
步骤 7:根据公式(8)计算基于 GloVe 的文本相似度并输出

2.3　融合 LDA 和 GloVe 模型的相似度结合文本聚类

K-Medoide 算法思想为:在数据集中随机选取 K 个对象作为初始簇,计算其余对象与各个代表对象的距离划分到其所代表的簇;然后反复利用非代表对象替换代表对象,试图找到最优的簇中心;利用代价函数表示聚类质量;当某个代表对象被替换时,除了未被替换的代表对象,其余对象被重新分配。

融合相似度的距离采用平方误差准则 E,公式如下:

$$E = \sum_{e=1}^{K} \sum_{f_i \in c_K} \text{Dt}(f_i, c_K)^2 = \sum_{e=1}^{K} \sum_{f_i \in c_K} \cdot \left[\lambda \cdot \text{Dt}_{\text{LDA}}(f_i, c_K) + (1-\lambda) \cdot \text{Dt}_{\text{GloVe\_ps}}(f_i, c_K)\right]^2 \tag{8}$$

融合 LDA 和 GloVe 模型的聚类算法(LG&K-Medoide)如算法 3 所示。

算法 3:LG&K-Medoide 算法

输入:病症文本数据集 $F = \{f_1, f_2, \cdots, f_M\}$,主题数目 K,融合系数 λ
输出:K 个主题特征词集
步骤 1:从病症文本数据集 $F = \{f_1, f_2, \cdots, f_M\}$ 中随机选取 K 个代表文本为初始簇的中心
步骤 2:利用公式(2)计算相似度融合距离
步骤 3:repeat
步骤 4:计算其余文本 f_i 到 K 个代表文本 c_K 的距离 $\text{Dt}(f_i, c_K)$,并将其划分到最近的簇
步骤 5:随机选择一个其余文本,
步骤 6:计算用其余文本替换代表文本的总代价 S(S 为绝对误差值的差)
步骤 7:if S<0,then 替换代表文本,形成 K 个新簇
步骤 8:until 不再改变
步骤 9:输出 K 个主题特征词集

K-Medoide 算法选择某个到所有点距离之和最小的点作为簇中心,以此能够减轻孤立点对聚类的影响 F1,从而提高聚类效果。

3　实验与结果分析

为了验证本文所提出的融合 LDA 和 GloVe 模型的病症文本聚类算法具有的优势,将其与基于 LDA、LDA+TF-IDF、LDA+Word2Vec 模型同在 K-Medoide 聚类算法中进行比较,在精确率、召回率和 F1 值方面进行测试。

3.1　融合 LDA 和 GloVe 模型的相似度结合文本聚类

本实验在 Windows 7 操作系统下进行,CPU 为 Intel Core I5-4210 M@2.60 GHz,内存 4 GB,编译语言为 Python 3,数据采集软件为八爪鱼 V7.6.4。

本文采用好医生笔记、39 健康、好大夫等咨询平台中患者的病症描述提问作为数据集,经过文本预处理,保留了 13 125 条病症文本。实验数据基本信息如表 1 所示。

表 1 实验数据基本信息

Tab.1 Basic information of experimental data

序号	主题	数量
1	呼吸科	2 397
2	消化科	1 699
3	神经科	1 704
4	骨科	3 016
5	皮肤科	2 029
6	其它	2 280

3.2　聚类评价指标

本实验衡量聚类结果时采用的评价指标包括

精确率(Precision,Pre)、召回率(Recall,Rec)和F1值(F1-Measure)。计算公式如下:

$$\mathrm{Pre}(i,j)=\frac{N_{i,j}}{N_j} \tag{9}$$

$$\mathrm{Rec}(i,j)=\frac{N_{i,j}}{N_i} \tag{10}$$

$$\mathrm{F1}(i,j)=\frac{2\cdot \mathrm{Pre}\cdot \mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}} \tag{11}$$

其中,N_i 表示病症文本数据集中类别 i 的文本数量,N_j 表示病症文本聚类完成后类别 j 的文本数量,$N_{i,j}$ 表示病症文本聚类结果中类别 j 正确划分到类别 i 的文本数量。

3.3 困惑度测试

根据2.2.1所述,本文利用困惑度确定主题数 K 值。实验重复进行10次,不同 K 值对应的困惑度取10次实验结果的平均值,实验结果如图3所示。

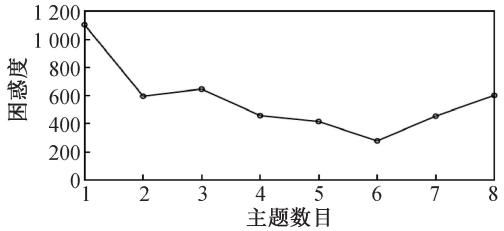

图3 LDA模型在不同主题数目下的困惑度值

Fig. 3 Perplexity value of LDA model under different number of topics

由图3的实验结果可知,当主题取值 $K=6$ 时,困惑度最小,表明此时LDA建模效果最好,故最优主题数 $K=6$。

3.4 融合系数λ值测试

融合系数λ的取值能够根据K-Medoide聚类结果的F1值确定。融合系数取值测试时,分别取 $\lambda=0.1,0.2,\cdots,0.9$,将本实验的算法重复运行10次,不同取值时所对应聚类结果的F1值取10次实验结果的平均值,实验结果如图4所示。

由图4可以看出,随着λ取值的变化,F1值不断提高,当 $\lambda=0.6$ 时,F1达到最高值。由此可得,λ取值为0.6时,病症文本数据的聚类效果最佳。

3.5 准确率测试

确定主题数 $K=6$ 时,将LG&K-Medoide算法与LDA、LDA+TF-IDF、LDA+Word2Vec模型进行准确率对比,结果如图5所示。

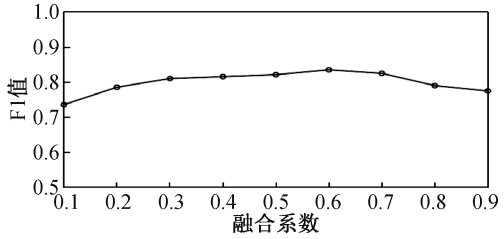

图4 λ在不同取值时所对应的F1值

Fig. 4 λ corresponds to the F1 value at different values

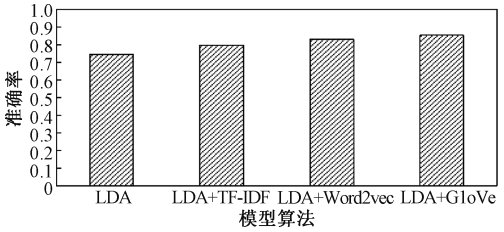

图5 准确率对比结果

Fig. 5 Accuracy comparison results

由图5可知,本文提出的算法在准确率上优于其它模型,较基于LDA+Word2Vec模型的聚类算法相比提高了3%,达到了85%的准确率。由此可见,融合LDA和GloVe模型的病症文本聚类算法在最终结果的准确率上得到了进一步地提升。

3.6 F1值测试

为证明本算法在病症文本聚类方面的优势,实验计算了不同主题下的精确率、召回率以及最终聚类结果的F1值,分别与LDA、LDA+TF-IDF、LDA+Word2Vec模型进行对比。

6个主题所对应的精确率(Precision)比较如图6所示。

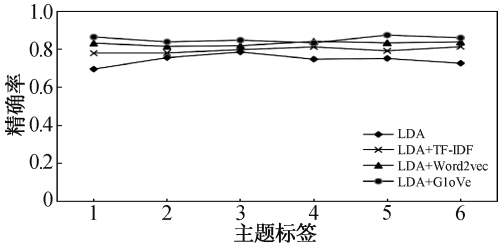

图6 不同主题下精确率对比结果

Fig. 6 Comparison results of accuracy rates under different topics

6个主题所对应的召回率(Recall)比较如图7所示。

从图6、图7中可以看出,分别在6个主题聚类结果下,LG&K-Medoide算法所对应的精确率和召回率均高于LDA、LDA+TF-IDF、LDA+Word2Vec模型算法。在实验中,其它三种算法在不同主题下的精确率和召回率波动较大,而LG&K-Medoide

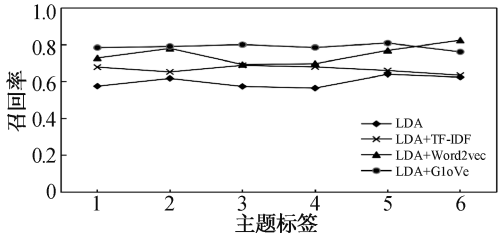

图 7 不同主题下召回率对比结果

Fig. 7 Comparison results of recall rates under different topics

算法处于较平稳的状态,主要是因为数据集为病症文本,不仅可细分为“呼吸科”、“消化科”、“神经科”、“骨科”、“皮肤科”和“其它”6 个主题,还可粗分为“内科”和“外科”2 个主题,所以建模时若未考虑全局语义信息,则在聚类时会出现错误。

为了更直接、准确地比较这四种算法在病症文本上的聚类精度,对最终聚类结果的精确率、召回率和 F1 值进行计算,实验结果如图 8 所示。

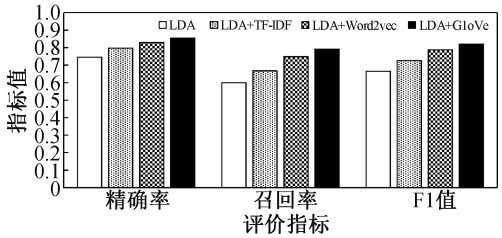

图 8 精确率、召回率、F1 值对比结果

Fig. 8 Accuracy rate, recall rate, F1 value comparison results

从图 8 可看出,四种聚类算法的结果在准确率、召回率和 F1 值上依次逐渐升高,本文所提出的融合 LDA 和 GloVe 模型的聚类算法在三个指标上都处于最优,这是因为该算法在 LDA 模型提取主题分布和词分布的基础上,融合了 GloVe 模型考虑全局语义信息的特点,并且利用了文本集中的统计信息,从而提高病症文本的聚类精度。

最终,根据困惑度和融合系数测试结果,设置聚类算法参数 $K = 6, \lambda = 0.6$。各个主题的特征词集结果如表 2 所示。

表 2 病症主题特征词集

Tab. 2 Illness theme feature word set

主题号	特征词集
1	咳嗽、发烧、有痰、流鼻涕、咽喉、感冒
2	腹泻、厌食、胃胀、发烧、胃痛、恶心
3	手脚麻木、偏头痛、失眠、头晕、多梦、颤抖
4	脚踝扭伤、骨折、椎间盘突出、颈椎、腰椎、腕骨
5	过敏、斑点、瘙痒、糜烂型脚气、疙瘩、脱皮
6	浑身无力、腰酸背痛、浮肿、烦躁、怕冷、出汗

由上表可知,本文提出的聚类算法所得到的实验结果有“呼吸科”、“消化科”、“神经科”、“骨科”、“皮肤科”和“其它”这 6 类主题,与所标注的标签类型一致,且提取的特征词集大都具有代表性,属名词性病症词汇,说明本文提出的基于词性贡献权重的词向量对词语的区分度有一定的提高,使相似度计算更加准确。

4 结论

本文考虑医疗名词蕴含的主要特征,设置了不同词性的贡献权重,突出病症名词的代表性;相似度计算时,结合 LDA 和 GloVe 相似度改进距离函数,提高了 K-Medoide 聚类准确率。实验结果表明,本文提出的 LG&K-Medoide 算法在病症文本数据集上具有更高的聚类精度,且相较与 LDA + Word2Vec 模型在 F1 值上提升了 3%,准确率上提升了 2%。

参考文献:

[1]谭章禄,彭胜男,王兆刚. 基于聚类分析的国内文本挖掘热点与趋势研究[J]. 情报学报, 2019, 38(06): 578-585.

[2]吴宗友,白昆龙,杨林蕊,等. 电子病历文本挖掘研究综述[J]. 计算机研究与发展,2021,58(03):513-527.

[3]POONAM T, RANI P J. Exploring Popular Topic Models [J]. Journal of Physics: Conference Series, 2020, 1706 (1):012171.

[4]KANG J, LEE J, JANG D, et al. A Methodology of Partner Selection for Sustainable Industry-university Cooperation Based on LDA Topic Model[J]. Sustainability, 2019, 11(12):3478.

[5]闫俊伢,马尚才. 基于文本聚类的网络微博舆情话题识别与追踪技术研究[J]. 重庆理工大学学报:自然科学, 2019,33(09):176-181.

[6]KIM S W, GIL J M. Research Paper Classification Systems Based on TF-IDF and LDA Schemes[J]. Human-centric Computing and Information Sciences, 2019, 9(1):1-21.

[7]王少鹏,彭 岩,王 洁. 基于 LDA 的文本聚类在网络舆情分析中的应用研究[J]. 山东大学学报:理学版, 2014,49(09):129-134.

[8]ZHU Z L, LIANG J, LI D, et al. Hot Topic Detection Based on a Refined TF-IDF Algorithm[J]. IEEE Access, 2019: 26996-27007.

[9]马思丹,刘东苏. 基于加权 Word2Vec 的文本分类方法研究[J]. 情报科学,2019,37(11):38-42.

[10]PARK S T, LIU C. A Study on Topic Models Using LDA and Word2Vec in Travel Route Recommendation: Focus on Convergence Travel and Tours Reviews[J]. Personal and Ubiquitous Computing, 2020:1-17.

[11]CHEN W, YU Z, XIAN Y, et al. Mining Keywords from Short Text Based on LDA-based Hierarchical Semantic Graph Model[J]. International Journal of Information Systems in the Service Sector (IJISSS),2020,12(2):76-87.

[12]KIM S, PARK H, LEE J. Word2vec-based Latent Semantic Analysis (W2V-LSA) for Topic Modeling: A Study on Blockchain Technology Trend Analysis[J]. Expert Systems With Applications,2020,152:113401.

[13]郑恒毅,廖城霖,李天柱. 一种面向网络长文本的话题检测方法[J]. 工程科学学报, 2019, 41(09):1208-1214.

[14]王欣研,张向先,张莉曼. 学术APP用户在线评论主题语义关联研究[J]. 情报科学,2020,38(06):25-31.

[15]李少华,李卫疆,余正涛. 基于GV-LDA的微博话题检测研究[J]. 软件导刊,2018,17(02):131-135.

(责任编辑 王利君)

投稿须知

1. 本刊来稿要求:选题新颖,观点明确,逻辑清晰,结构完整,数据真实可靠。

2. 本刊严禁一稿两投、重复内容多次投稿(包括以不同文种分别投稿)以及抄袭他人论文等现象。一旦发现有上述情况,该作者的稿件将作退稿处理。

3. 电子稿件请以WORD(*.doc)文档上传投稿系统。所投稿件如果有照片和曲线图,则需要尽量提供彩色图片,且做到层次分明、清晰,线条粗细均匀、比例合理美观(建议曲线图用Origin、化学结构式等用Chemi Bio Draw制图,再拷贝到文档中)。

4. 稿件的作者必须是直接参与研究工作或对其有重要指导作用的成员(如研究生导师等),协助做实验的人员可放入致谢中。

5. 请务必在稿件首页页脚处依次注明:收稿日期(格式为2015-02-20),基金项目(包括项目来源、项目名称及项目编号),第一和通讯作者简介(包括姓名(出生年—)、性别、民族(汉族可不写)、籍贯、职称、学历及研究方向、E-mail和联系电话)。

6. 通过审查后需要修改和补充实验的稿件,最晚不超过2个月将修改稿返回编辑部,如有困难需及时向编辑部说明情况,逾期按自动撤稿处理。

7. 本刊编辑部对拟用稿有权作技术性和文字性修改,作者若不允许对其文稿作修改,务请在来稿时注明;论文发表后,版权即属编辑部所有。凡在投稿时未作特别声明的,本刊均认为作者已同意将其论文编入有关的数据库并在网上传播。

投稿系统网址:http://xuebao.hebeu.edu.cn/journal.htm

通讯地址:河北省邯郸市经济技术开发区太极路19号 邮编:056038

联系电话:0310-3969121