doi:10.11887/j.cn.201703014　　http://journal.nudt.edu.cn

多特征融合文本聚类的新闻话题发现模型*

车　蕾[1,2],杨小平[1]

(1. 中国人民大学 信息学院, 北京　100872; 2. 北京信息科技大学 信息管理学院, 北京　100192)

摘　要:融合新闻命名实体、新闻标题、新闻重要段落、文本语义等多特征影响,提出基于多特征融合文本聚类的新闻话题发现模型。模型根据新闻的多特征影响,提出一种多特征融合文本聚类方法。该方法针对新闻标题、新闻重要段落等特征因素构建向量空间模型及相似度算法,基于潜在狄利克雷分配模型构建主题空间模型及相似度算法,针对命名实体构建命名实体模型及相似度算法,并将三种相似度算法形成最优融合。基于多特征融合文本聚类方法,模型改进了用于新闻话题发现的 Single-Pass 算法。实验是在真实新闻数据集上开展的,实验结果表明:该模型有效地提高了新闻话题发现的准确率、召回率和综合评价指标,并具有一定的自适应能力。

关键词:新闻话题;多特征融合;潜在狄利克雷分配;向量空间模型;主题空间模型

中图分类号:TP391　　**文献标志码**:A　　**文章编号**:1001-2486(2017)03-085-06

News topic discovery model of multi feature fusion text clustering

CHE Lei[1,2], *YANG Xiaoping*[1]

(1. School of Information, Renmin University of China, Beijing 100872, China;
2. School of Information Management, Beijing Information Science & Technology University, Beijing 100192, China)

Abstract: The news topic discovery model based on multi feature fusion text clustering was proposed fusing multi features of news, such as named entities, news headlines, important paragraphs, text semantics and so on. Based on the multi feature influence of news, a multi feature fusion text clustering method was put forward in this model. In this way, vector space model and similarity algorithm based on feature words, news headlines, important paragraphs were constructed, subject space model and similarity algorithm based on latent Dirichlet allocation were constructed, named entity model and similarity algorithm based on named entities were constructed, and those three similarity algorithms were fused optimally. Based on multi feature fusion text clustering method, the Single-Pass algorithm used in the news topic discovery was improved. Experiments were carried out on the real news data set, and the experimental results show that the model can improve the accuracy rate, recall rate and comprehensive evaluation index of the news topic discovery, and have some ability of self-adaption.

Key words: news topic; multi feature fusion; latent Dirichlet allocation; vector space model; subject space model

随着信息化的发展,互联网逐渐成为人们获取信息的一个主要途径,突发新闻事件可以在互联网上瞬间传播。如何发现新闻话题、如何追踪新闻事件的发展过程,是迫切需要解决的问题。新闻话题的内容会伴随时间发展而发生变化,新闻话题的强度也会伴随时间发展经历一个从高潮到低潮的过程。如何按时间顺序挖掘新闻集合中话题的演化过程,从而帮助用户追踪感兴趣的话题,具有实际意义。因此,话题演化研究具有现实的应用背景。话题发现是话题演化研究中的关键环节。

当前话题发现的研究主要集中在建立更好的文本表示形式和充分利用新闻语料特征两个方面。Allan 等[1]最先采用信息检索领域的向量空间模型(Vecor Space Model,VSM)构建话题模型。Yang 等[2]运用 Rocchio 算法对基于 VSM 的话题模型进行扩展。Nallapati[3]提出语义语言模型。Lee 等[4]利用增量型的方法在话题发现过程中不断提炼基于话题的特征词,给予这些特征词更大的权重,从而提高话题区分能力。王少鹏[5]利用词频-反文档频率(Term Frequency-Inverse Document Frequency,TF-IDF)算法和潜在狄利克雷分配(Latent Dirichlet Allocation,LDA)主题模型分别计算文本的相似度,然后使用 K-means 算法

* **收稿日期**:2016-02-10
基金项目:国家自然科学基金资助项目(61272513);北京市教育委员会科技计划面上资助项目(KM201511232016,SM201511232004);北京高等学校青年英才计划资助项目(YETP1503)
作者简介:车蕾(1979—),女,河南洛阳人,副教授,博士研究生,E-mail:chelei@bistu.edu.cn

进行聚类分析。马晓姝[6]将基于 VSM 命名实体特征词的文本相似度和基于潜在主题的文本相似度进行线性结合,使用文本聚类算法进行聚类。但是这些研究缺乏新闻特征词、命名实体、新闻标题、新闻重要段落等特征对话题发现的影响。

1 LDA 模型及相关概念

话题模型本质上就是一种文本的降维表示。TF-IDF 是最早的文本降维模型,该模型的不足是无法从语义层面表示文本。随后 Deerwester 等[7]提出了隐性语义分析(Latent Semantic Analysis, LSA)模型,采用矩阵的奇异值分解技术对文本进行降维,以从文本中发现隐含的语义维度,该模型的不足是没有能力处理一词多义和一义多词的问题。Hofmann[8]在 LSA 基础上提出了概率隐性语义分析(Probabilistic Latent Semantic Analysis, PLSA)模型,并使用期望最大化(Expectation Maximization, EM)算法学习模型参数,该模型的不足是参数数量会随着文集增长而线性增长,并且产生过拟合的问题。Blei 等[9]提出了 LDA 模型,LDA 模型既是一个概率生成模型,又是一个话题模型,该模型很好地解决了 PLSA 模型出现的问题,具有很好的泛化能力。LDA 模型在文本挖掘、机器学习等领域发挥着重要作用。

1.1 LDA 基本概念

LDA 模型是一个经典主题模型,它可以计算出每篇文档的主题概率分布。在此基础上可以利用 LDA 模型获得各个节点文本内容的主题分布信息,运用协作演化的思想,将内容相似度和结构信息相融合,以提升节点相似性评定的效果[10]。在 LDA 模型中假设文档是多个隐含主题上的混合分布,各个主题是一个固定词表上的混合分布。LDA 文档生成过程的概率模型如图 1 所示,表 1 说明了该模型中各符号的含义。

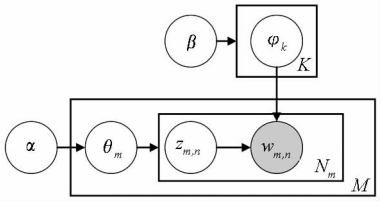

图 1 LDA 文档模型生成图

Fig. 1 LDA document model generating

LDA 模型采用 Dirichlet 分布作为概率主题模型中多项分布的先验分布。图 1 代表的概率模型如式(1)所示。

$$p(\theta,z,w\mid\alpha,\beta)=p(\theta\mid\alpha)\prod_{n=1}^{N}p(z_n\mid\theta)p(w_n\mid z_n,\beta) \tag{1}$$

表 1 LDA 模型中各符号的含义

Tab. 1 Meaning of symbols in LDA model

符号	含义
α	θ 的超参数(先验分布)
β	φ 的超参数(先验分布)
θ	文档 - 主题概率分布
φ	主题 - 词概率分布
M	文档数
N_m	第 m 个文档的单词数
K	主题数
$z_{m,n}$	第 m 个文档中第 n 个词的主题
$w_{m,n}$	第 m 个文档中的第 n 个词
空心圆圈	隐藏变量
实心圆圈	可观测的变量
矩形框	重复抽样过程

在 LDA 中,参数 α 和 β 是固定值,由用户事先制定。文档中的各个单词 $w_{m,n}$ 是可观测的数据。文档 - 主题概率分布以及主题 - 词概率分布,即 θ 和 φ 是隐式参数,需要通过概率推导求解。LDA 模型的参数个数只与主题数和词数有关。使用 Gibbs[11]采样间接估算 θ 和 φ。

1.2 存在的问题及解决方案

LDA 主题模型可以挖掘文本内容中的潜在语义信息,但是维度较低,很难保证信息的完整性,对文本类别的区分能力不够完全。另外,对于新闻文本而言,新闻的标题和第一段,对文本类别的区分度也占有一定分量。基于 TF-IDF 的 VSM 可以发挥从词语层面对文本信息进行充分挖掘的优势。通过把 LDA 主题模型和传统的基于 TF-IDF 的 VSM 相结合,从特征词和主题两方面对文本进行聚类分析,结合两种模型的优点,弥补两种方法的缺陷。另外,也将新闻的标题、新闻重要段落、命名实体等具有明显主题特征的信息纳入模型中。该模型可以对文本信息进行充分挖掘,保证了新闻话题发现的准确度。

2 话题发现模型的构建

本模型的实施流程如图 2 所示,包括数据抓取、数据预处理、建模、计算相似度、基于改进的 Single-Pass 算法进行聚类、挖掘出相应的新闻话

题。其中,建模过程包括:VSM 特征词建模、LDA 主题建模和命名实体。计算相似度包括:计算基于特征值的文本相似度、计算基于主题的文本相似度和计算命名实体的相似度。

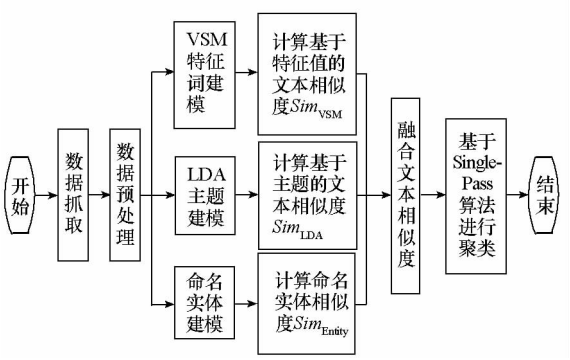

图 2　多特征融合文本聚类分析流程图

Fig. 2　Flow chart of multi feature fusion text clustering

2.1　多特征融合文本相似度

确定文本的相似度是进行聚类分析非常关键的一步。把基于 TF-IDF 权值策略和基于 LDA 主题模型以及基于命名实体模型的相似度进行最优线性结合,得到文本的相似度,即多特征融合文本相似度。根据计算得到的文本相似度矩阵,使用 Single-Pass 算法进行聚类分析。提取人名、地名、机构名等命名实体构建命名实体模型。

线性结合的如式(2)所示。

$$\begin{cases} sim(\boldsymbol{d}_1,\boldsymbol{d}_2) = \alpha sim_{\mathrm{VSM}}(\boldsymbol{d}_1,\boldsymbol{d}_2) + \\ \qquad\qquad \beta sim_{\mathrm{LDA}}(\boldsymbol{d}_1,\boldsymbol{d}_2) + \\ \qquad\qquad \lambda sim_{\mathrm{Entity}}(\boldsymbol{d}_1,\boldsymbol{d}_2) \\ \alpha+\beta+\lambda = 1 \end{cases} \tag{2}$$

其中:α,β,λ 分别为向量空间模型的文本相似度系数、主题空间模型的文本相似度系数、命名实体模型的相似度系数。$sim_{\mathrm{VSM}}(\boldsymbol{d}_1,\boldsymbol{d}_2)$ 为两文本间向量空间模型的文本相似度;$sim_{\mathrm{LDA}}(\boldsymbol{d}_1,\boldsymbol{d}_2)$ 为两文本间主题空间模型的文本相似度;$sim_{\mathrm{Entity}}(\boldsymbol{d}_1,\boldsymbol{d}_2)$ 为两文本间命名实体模型的相似度。

2.1.1　向量空间模型文本相似度

不同模型对应不同的相似度计算方法,采用关键词的标准化 TF-IDF 值来衡量向量空间模型中的文本,采用余弦相似度来计算文本相似度。余弦相似度计算如式(3)所示。

$$sim_{\mathrm{VSM}}(\boldsymbol{d}_1,\boldsymbol{d}_2) = \frac{\boldsymbol{d}_1 \cdot \boldsymbol{d}_2}{|\boldsymbol{d}_1| \times |\boldsymbol{d}_2|} = \frac{\sum_{i=1}^{n}(\boldsymbol{a}_i \times \boldsymbol{b}_i)}{\sqrt{\sum_{i=1}^{n}\boldsymbol{a}_i^2 \times \sum_{i=1}^{n}\boldsymbol{b}_i^2}} \tag{3}$$

其中,$\boldsymbol{d}_1$ 和 $\boldsymbol{d}_2$ 表示两个文档,$\boldsymbol{a}_i$ 表示文档 $\boldsymbol{d}_1$ 中第 i 个向量,$\boldsymbol{b}_i$ 表示文档 $\boldsymbol{d}_2$ 中第 i 个向量。

新闻模型如下:News(NewsID, NewsTitle, Content, FirstParagraph, NewsURL, Source, Date, TopicID)。新闻文本模型 New_i 中每个特征向量 $\boldsymbol{w}$ 的权重计算如式(4)所示。

$$\begin{cases} Weight(\boldsymbol{w},New_i) = \chi TFIDF(\boldsymbol{w},New_i) + \\ \qquad\qquad \lambda Count(\boldsymbol{w},NewsTitle_i) + \\ \qquad\qquad \kappa Count(\boldsymbol{w},firstParagraph_i) \\ \chi+\lambda+\kappa = 1 \end{cases} \tag{4}$$

其中:χ,γ,κ 分别为 TF-IDF 值、向量 $\boldsymbol{w}$ 在新闻标题中出现的次数、向量 $\boldsymbol{w}$ 在新闻第一段出现的次数的系数。

VSM 模型中,话题向量的每个特征向量(词语)$\boldsymbol{w}$ 的计算如式(5)所示。

$$Weight(\boldsymbol{w}) = \sum_{i=1}^{NewsCount}\left[\frac{M}{TimeDist(New_i+M)}\right] \times TFIDF(\boldsymbol{w},New_i) \times \frac{Count(\boldsymbol{w} \in title)}{NewsCount} \tag{5}$$

其中:$NewsCount$ 表示该话题下的新闻数目;$TimeDist(New_i+M)$ 表示第 i 个新闻的开始时间与话题的开始时间的距离,M 为调整参数;$TFIDF(\boldsymbol{w},New_i)$ 表示第 i 个新闻中出现词语 $\boldsymbol{w}$ 的权重。

考虑到文本长度对权值的影响,需要对特征权值公式做归一处理,将各权值规范到[0,1]之间,如式(6)所示。

$$W_{ik} = \frac{tf_{ik} \times \ln\left(\frac{N}{n_k}+0.01\right)}{\sqrt{\sum_{k=1}^{n}\left[(tf_{ik}) \times \ln\left(\frac{N}{n_k}+0.01\right)\right]^2}} \tag{6}$$

2.1.2　主题空间模型文本相似度

采用服从 Dirichlet 分布的主题概率向量来衡量 LDA 主题模型中的文本,并用延森 - 香农(Jensen-Shannon,JS)距离函数来计算文本概率向量的相似度。

$\boldsymbol{p}=(p_1,p_2,\cdots,p_k)$ 到 $\boldsymbol{q}=(q_1,q_2,\cdots,q_k)$ 的距离定义如式(7)所示。

$$D_{\mathrm{js}}(\boldsymbol{p},\boldsymbol{q}) = \frac{1}{2}\left(\sum_{j=1}^{k} p_j \ln\frac{2p_j}{p_j+q_j} + \sum_{j=1}^{k} q_j \ln\frac{2q_j}{p_j+q_j}\right) \tag{7}$$

其中,$\boldsymbol{p},\boldsymbol{q}$ 为主题概率分布。

采用 LDA 模型对文本向量进行建模,并使用 Gibbs 抽样法对建模后的文本向量矩阵进行求解,得到文本 - 主题矩阵和主题 - 词矩阵,从而获

取每个文本的概率向量。主题空间模型中,每个主题的概率向量计算如式(8)所示。

$$\text{主题的概率向量} = \left(\frac{\sum_{i=1}^{NewsCount} d_{11}}{NewsCount}, \frac{\sum_{i=1}^{NewsCount} d_{12}}{NewsCount}, \cdots, \frac{\sum_{i=1}^{NewsCount} d_{1t}}{NewsCount} \right) \tag{8}$$

2.1.3 命名实体相似度

命名实体部分使用的是 Jaccard 来计算其相似度,如式(9)所示。

$$sim_{\text{Entity}}(d_1, d_2) = \frac{\text{count}(\{e \in d_1\} \cap \{e \in d_2\})}{\text{count}(\{e \in d_1\} \cup \{e \in d_2\})} \tag{9}$$

其中,d_1、d_2 分别表示两个实体,e 表示实体里面的每个对象。

每个话题下的命名实体就是该话题下所有新闻的命名实体的并集。

2.2 多特征融合文本聚类算法

文本聚类的主要方法可分为 5 类[7]:划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法。基于划分的聚类算法需要对 K 个分组进行初始化,并通过迭代方法将数据聚合到分组中,使得每次得到的分组方案较前一次好;基于层次的聚类方法通过对数据集合按照某种指定的方法进行层次划分,直到满足收敛或者满足某种条件时停止;基于密度的聚类方法是基于密度的,而其他的算法基于各种不同的距离计算方式,其克服了基于距离的算法只能发现一定距离内的类簇的局限性;基于网格的聚类方法基于网格结构划分文档集合,然后在网格上进行聚类;基于模型的聚类方法依据建立的数学模型,对给定的文档数据与该数学模型进行拟合。

Single-Pass 算法是话题检测的一种著名算法,实际是 K 最近邻(K-Nearest Neighbor, KNN)的一种应用,为 KNN 算法[12]。Single-Pass 算法常被用于 Web 新闻话题发现,其算法的优劣与新闻的报道时间有关,并且该算法是一种增量的聚类算法,因此适合于 Web 话题检测。基于多特征融合相似度改进了新闻话题发现算法 Single-pass,改进算法的步骤如图 3 所示。提出的多特征融合文本聚类算法可以应对实时新闻报道,并对其进行动态聚类。

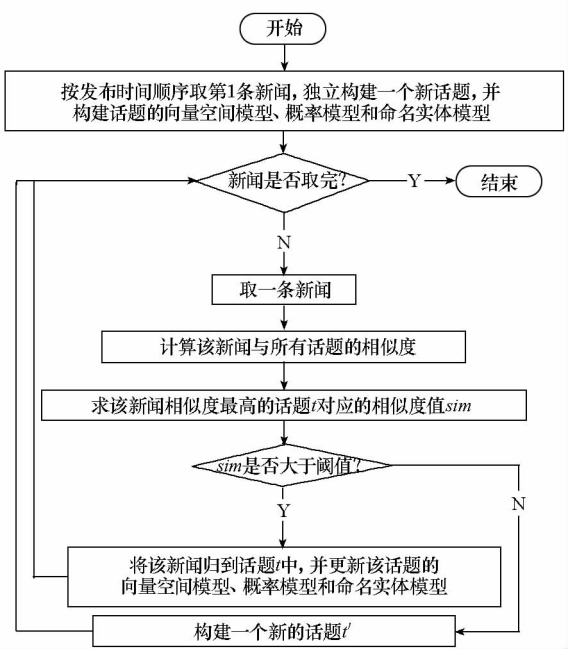

图 3 基于多特征融合文本聚类的新闻话题发现步骤

Fig. 3 Steps of new topics discovering based on multi feature fusion text clustering algorithm

3 实验设计及结果分析

3.1 实验环境及数据

实验是在 CPU 为 Intel(R) Core(TM) i5 - 3320M CPU @2.60 GHz、内存为 8 GB、操作系统为 Windows 8.1 专业版、处理器基于 X64 的 PC 机上运行的。算法基于 Java 语言实现。

实验数据是 2014 年 2 月至 2015 年 4 月从新闻网站的专题版块抓取的 20 个专题的相应新闻,共 3009 条新闻。首先对这些新闻数据进行预处理,包括分词、去停用词、识别命名实体。然后构建新闻的特征向量表,并结合词频、新闻标题、新闻第一段等多特征计算特征向量的权重。为下一步的实验分析做好准备。

3.2 性能评价指标

为了能够对话题检测的效果进行有效的评价,话题检测与跟踪(Topic Detection and Tracking, TDT)提出了评价标准,包括:准确率、召回率、漏报率、误报率和综合评价指标(F1-Measure)。准确率、召回率、综合评价指标越高越好,漏报率和误报率越低越好,预测主题数目越接近实际主题数目越好。其中:

$$\text{准确率} = \frac{\text{识别出的新闻数目}}{\text{新闻总数目}}$$

$$\text{召回率} = \frac{\text{识别出的关于某个话题的新闻数目}}{\text{语料库中描述该话题的新闻数目}}$$

$$漏报率=\frac{没有识别出的与某话题相关的新闻数目}{语料库中描述该话题的所有新闻数目}$$

$$误报率=\frac{对某话题识别错误的新闻数目}{语料库中与该话题不相关的新闻数目}$$

$$综合评价指标=\frac{2\times准确率\times召回率}{准确率+召回率}$$

3.3　实验结果与分析

实验分析主要包括如下内容:通过多次实验结果的比对,确定令性能评价指标达到最优的相似度阈值;当阈值确定后,通过多次实验结果的比对,确定令性能评价指标达到最优的特征向量权重各因素影响因子的取值;通过 LDA 实验,确定 LDA 最优主题数目;最后,3.3.4 节基于确定好的各个参数以及相同实验数据,分别开展仅采用 VSM 方法、VSM-多特征(Multi Feature, MF)方法、LDA 方法、VSM-LDA 和多特征融合(Multi Feature Fusion,MFF)方法的实验,通过比对各实验结果,证明基于多特征融合方法的文本聚类模型的有效性和优势。所有最优参数值或阈值都是由系统自动根据具体数据源计算获得。

3.3.1　相似度阈值的选择

将实验结果数据首先按|预测主题数目 - 实际主题数目|(即差的绝对值)的升序排序,再按 F1-Measure 的降序排序,排在第一位的数据就是综合评价最优的。图 4 显示了不同相似度阈值下的实验对比结果,从中可以分析出,阈值取 0.4 时综合评价最优的。

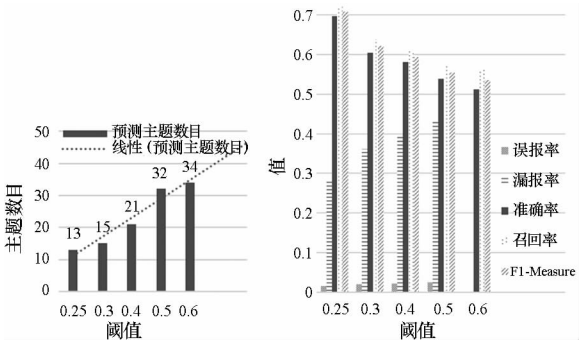

图 4　不同阈值下的主题数目估计情况、误报率、漏报率、召回率、准确率和 F1-Measure

Fig. 4　Subject number estimation, false positive rate, false negative rate, recall rate, accuracy rate and F1-Measure under different thresholds

3.3.2　特征向量权重各影响因素权重的选择

影响因子的特征向量权重如式(4)所示,为找到最优特征向量权重各影响因素的权重,实验遍历 3 个影响因素的所有组合的可能。图 5 显示了特征向量权重各影响因素不同权重组合的性能评价指标,该图显示的仅是综合评价较好(即预测主题数目接近实际主题数目,F1-Measure 较高)的一些影响因素权重组合。综合 F1-Measure 和主题数目考虑,当 $\chi=0.6$,$\gamma=0.3$,$\kappa=0.1$ 时性能评价指标最好。从该实验可以看出,考虑标题和第一段对主题的影响相比仅考虑传统 TF-IDF 对主题的影响的评价效果要好。

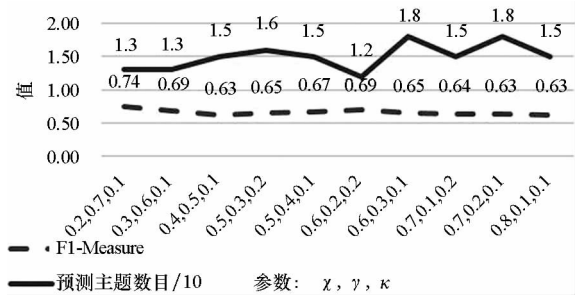

图 5　特征向量权重各影响因素不同权重组合的性能

Fig. 5　Performance of different weight combinations of each effected factors of feature vector weight

3.3.3　LDA 最优主题数 T 的选择

为确定使 LDA 算法达到最优性能评价指标所对应的相似度阈值和主题数,实验仅用 LDA 算法计算相似度,遍历了 0.001 至 0.5 范围的相似度阈值和 10 至 150 之间的 LDA 主题数目,将实验结果数据首先按|预测主题数目 - 实际主题数目|(即差的绝对值)的升序排序,再按 F1-Measure 的降序排序,筛选出前 n 条实验结果,筛选结果如图 6 所示。综合 F1-Measure 和主题数目,LDA 主题数目为 50 时,预测主题数目为 20,F1-Measure 为 0.807 180,性能综合评价最好。

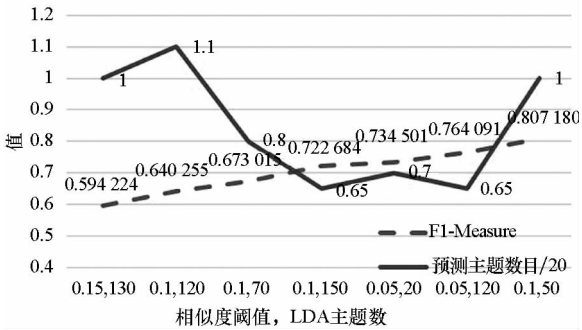

图 6　不同相似度阈值不同 LDA 主题数的性能

Fig. 6　Performance of different similarity thresholds and different LDA subject

3.3.4　多特征融合模型

为确定使多特征融合模型达到最优性能评价指标,实验遍历了式(2)中的 α、β 和 λ 所有可能组合,从实验结果中筛选出预测主题数在 18 至 30 之间的记录,筛选结果如图 7 所示。综合 F1-

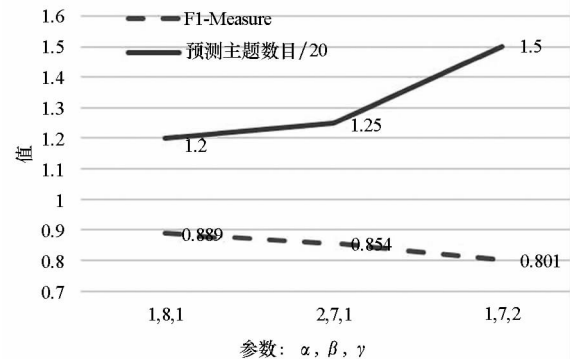

图 7 多特征融合模型中各因素影响因子的选择

Fig. 7 Choice of effected factors in multi feature fusion model

Measure 和主题数目考虑，α、β 和 λ 分别取 1，8，1 的时模型性能评价指标最好。表 2 显示了多特征融合模型（MFF）与其他模型（VSM、VSM-MF、LDA 和 VSM-LDA）的评价指标比对情况。其中，VSM、LDA、VSM-LDA 模型都有同行做过相关研究。从表 2 中可以看出，结合了多特征的 VSM-MF 比单纯的 VSM 的性能要优；虽然 LDA 的性能要优于 VSM 和 VSM-MF，VSM-LDA 的性能优于前 3 个，但是提出的将 VSM、LDA 和多特征因素结合起来的多特征融合模型 MFF 的性能相比之下是最优的。

表 2 多特征融合模型与其他模型的评价指标比对

Tab. 2 Comparison of evaluating index of multi feature fusion model and other models

模型	准确率	召回率	漏报率	误报率	F1-Measure
VSM	0.581	0.608	0.392	0.022	0.594
VSM-MF	0.643	0.664	0.336	0.019	0.654
LDA	0.830	0.785	0.215	0.009	0.807
VSM-LDA	0.816	0.828	0.172	0.010	0.822
MFF	0.901	0.878	0.122	0.005	0.889

4 结论

本文基于 LDA 模型提出了一种改进的多特征融合文本聚类的新闻话题发现模型，该综合考虑新闻的各组成部分、新闻的词频、新闻的语义等多特征，提高了 Web 新闻话题发现的准确性，并能更好地发现新主题。实验表明，改进模型具有更优的主题检测效果。由于实验中的所有最优参数值或阈值都是由系统自动根据具体数据源计算获得，所以模型具有很好的自适应能力。

在 Web 新闻话题发现的研究方面还存在很多问题需要进一步探讨。例如：目前多特征是按照线性方式进行结合的，非线性的结合模型是否会更好地提高话题发现的准确度有待研究。

参考文献（References）

[1] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking [C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998: 37 – 75.

[2] Yang Y, Ault T, Pierce T, et al. Improving text categorization method for event tracking[C]//Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000: 65 – 72.

[3] Nallapati R. Semantic language models for topic detection and tracking[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003: 71 – 81.

[4] Lee C H, Wu C H, Chien T F. BursT: a dynamic term weighting scheme for mining microblogging messages[C]// Proceedings of the 8th International Symposium on Neural Network, 2011: 548 – 557.

[5] 王少鹏. 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(9)：129 – 134.
WANG Shaopeng. Research of the text clustering based on LDA using in network public opinion analysis[J]. Journal of Shandong University: Natural Science, 2014, 49(9): 129 – 134. (in Chinese)

[6] 马晓姝. 基于 LDA 模型的新闻话题发现研究[D]. 长春：东北师范大学，2014.
MA Xiaoshu. News topic discovery research based on the LDA model[D]. Changchun: Northeast Normal University, 2014. (in Chinese)

[7] Deerwester S, Dumais S, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society of Information Science, 1990, 41(6): 391 – 407.

[8] Hofmann T. Probabilistic latent semantic indexing [C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1990: 50 – 57.

[9] Blei D M, Ng A Y, Jordan M I, et al. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993 – 1022.

[10] 胡艳丽，白亮，张维明. 网络舆情中一种基于 OLDA 的在线话题演化方法[J]. 国防科技大学学报，2012，34(1)：150 – 154.
HU Yanli, BAI Liang, ZHANG Weiming. OLDA-based method for online topic evolution in network public opinion analysis[J]. Journal of National University of Defense Technology, 2012, 34(1): 150 – 154. (in Chinese)

[11] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(S1): 5228 – 5235.

[12] 吴少凯. 基于桶的二次聚类新闻热点话题挖掘及应用[D]. 广州：华南理工大学，2013.
WU Shaokai. Mining and application of hot news topics by bucket-based quadratic clustering [D]. Guangzhou: South China University of Technology, 2013. (in Chinese)