

基于主题分析的用户评论聚类方法

张会兵¹ 钟 昊¹ 胡晓丽²

(桂林电子科技大学广西可信软件重点实验室 广西 桂林 541004)¹

(桂林电子科技大学教学实践部 广西 桂林 541004)²

摘 要 在社会化商务中对用户评论进行合理的聚类分析有利于商家提供精准服务或推荐信息,文中提出了一种基于主题分析的用户评论聚类方法。根据主题词在用户评论中的互信息强度以及主题词之间的相似度计算主题词权重,并依此构建用户评论主题向量。在此基础上,提出了一种基于用户评论相似度自动选择 canopy 聚类算法初始阈值的自适应 canopy+kmeans 聚类算法,对主题向量进行聚类分析。在亚马逊的评论数据上进行测试,结果表明:该方法充分描述了用户评论中不同主题词对用户观点的突出程度不同,并改善了 K-means 聚类算法易陷入局部最优的缺点,与传统的 LDA+K-means 算法相比,取得了更好的效果。

关键词 用户评论,主题分析,主题向量,自适应聚类

中图法分类号 TP399 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.08.008

User Reviews Clustering Method Based on Topic Analysis

ZHANG Hui-bing¹ ZHONG Hao¹ HU Xiao-li²

(Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China)¹

(Practice and Experiment Station, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China)²

Abstract The rational clustering analysis of user reviews in social business is beneficial to providing accurate service or recommendation information. This paper proposed a user reviews clustering method based on topic analysis. According to the mutual information intensity of topic words in user reviews and the similarity between topic words, the weight of topic words is calculated, and the topic vector of user reviews is constructed. On this basis, an adaptive canopy+kmeans clustering algorithm based on user comment similarity to automatically select the initial threshold of canopy clustering algorithm is proposed, which is used to cluster and analyze the subject vector. On Amazon's review data, the results show that the proposed method makes full use of the weight of different topic words in the user's reviews and improves the disadvantage of the K-means clustering algorithm easily falling into the local optimal. Compared with the traditional LDA+K-means algorithm, the proposed method can achieve better results.

Keywords User reviews, Topic analysis, Topic vector, Adaptive clustering

1 引言

社会化商务中,用户的在线评论不再受到时空条件的约束,用户可以广泛地参与对各种新闻、商品的评论,这些评论代表着用户对目标事物不同粒度的评价,而对用户的评论进行文本挖掘能够表征用户在社会化商务中的行为和偏好,根据用户评论挖掘生成的文本特征对用户进行聚类,能够优化对用户的管理和信息推荐。由于社会化商务中的用户评论文本属于专业性层次不齐、复杂度高的不规则短文本^[1],导致运用机器学习算法进行文本分析时会产生特征稀疏、粒度难以

统一等问题,因此本文主要利用文本分析技术将用户评论映射到同一个向量空间中。

在文本分析技术不断发展的趋势下,主流的文本分析技术有主题发现^[2]、情感分析^[3]和用户意见挖掘^[4]三大类,因此利用文本分析技术对用户评论文本进行分析,提取用户评论文本的特征,并根据特征进行聚类成为了研究的热点。传统的聚类方法^[5-6]仅考虑了评论文本中主题词的分布情况,未考虑不同主题词对于用户观点具有的更深层次的表征;再者, K-means 聚类算法在确定聚类数和初始聚类中心时存在随机性^[7],导致聚类结果易陷入局部最优解。鉴于此,本文针对传

到稿日期:2018-07-19 返修日期:2018-11-23 本文受国家自然科学基金项目(61662013, U1501252, U1711263, 61662015, 61562014), 广西科技重大专项(AA17202024), 广西自然科学基金项目(2017GXNSFAA198372, 2016GXNSFAA380149), 广西师范大学教育发展基金会第四批“教师成长基金”项目(EDF2015005)资助。

张会兵(1976—),男,博士,副教授,主要研究方向为物联网/移动互联网/嵌入式与移动计算;钟 昊(1993—),男,硕士生,主要研究方向为农业物联网、社交网络计算与可信计算, E-mail: 401924605@qq.com;胡晓丽(1978—),女,硕士,讲师,主要研究方向为计算机应用技术、物联网, E-mail: huxiaoli@guet.edu.cn(通信作者)。

统的基于 LDA 和 K-means 文本聚类方法的不足,构建主题词之间的层次关系^[8],生成主题词层次结构,并提出了一种自适应的 Canopy+K-means 聚类算法 KHS(Keyword Hierarchy Structure)+ACK-means(Adaptive Canopy K-means)。本文的主要贡献如下:

(1)基于主题词的互信息强度以及主题词之间的相似度生成主题词层次结构,使主题词所在的层次能够表征其对用户观点的突出程度。根据用户评论中包含的主题词数以及主题词所在的层次,定义了用户偏好向量的生成方法,使所有的用户评论均映射到同一个向量空间中。

(2)根据任意两个评论文本之间的相似度设定 canopy 聚类算法中的阈值 T_1 和 T_2 ,从而得到初步聚类的个数 K 和聚类中心集合 C ,初始化 K-means 算法的初始 K 值和初始聚类中心,避免了 K-means 算法随机获取聚类中心导致的聚类结果易陷入局部最优解的问题,使聚类结果更趋近于全局最优解。

本文第 2 节介绍了相关的研究工作;第 3 节描述了构建主题词之间的层次关系和结合 canopy 聚类对 K-means 算法进行改进的意义,对模型和算法的流程做了概括;第 4 节给出了主题词之间的层次关系的构建、用户评论的主题向量的相关计算以及自适应的 canopy+K-means 算法的过程;第 5 节在真实数据集上进行了实验和比较;最后总结全文。

2 相关工作

基于主题发现的文本聚类方法与本文提出的基于主题分析的用户评论聚类方法相似。传统的基于主题发现的文本聚类方法结合 LDA 与 K-means 对文本进行聚类。吴江等^[9]将 LDA 技术应用于在线医疗社区的用户评论文本,对用户评论文本进行特征提取,可以降低文本表示向量的维度,并基于文本表示向量采用 K-means 算法对用户评论文本进行聚类,最后根据用户评论文本的聚类结果对用户进行聚类。

据此,不少科研工作者在 LDA 模型的基础上对其进行了改进,如 Alharbi 等^[10]针对文本冗余、同义、多义、噪声和高维等性质,提出了一个基于 LDA 模型的特征选择模型用于文本聚类,该模型提出了一种新的扩展随机集理论,该理论根据

LDA 模型生成的主题词,为文本中的所有词汇都赋予一定的权重,从而提升文本聚类的效果。Cai 等^[11]提出了一种特征分组的方法用于文本聚类。在该方法中,每组文档用向量空间模型表示。将 LDA 算法应用到文本数据中生成特征组作为主题,将这些主题作为每组文本的特征值输入到 K-means 算法中进行文本聚类,并在 LDA 生成特征组时,提出一种基于熵的词过滤方法,用于去除相应主题的主题词分布中出现概率较低的词。Shi 等^[12]在 LDA 模型的基础上,结合了 Word2vec 词向量模型对其进行改进,使用 K-means++ 算法对 Word2vec 模型得到的词向量进行聚类,从而优化 LDA 模型的训练过程,最终得到更好的主题分布。Tang 等^[13]综合考虑了微博文本的特征和短文本聚类的效率问题,提出了一种结合频繁词集和 LDA 主题模型的文本聚类方法,该方法不仅能有效地划分文本,并且能更好地挖掘类簇中的潜在主题。Suadaa 等^[14]提出了利用 LDA 生成主题以及主题词,并用主题词在文本中的 TF-IDF 值作为文本的特征值,各个主题词的 TF-IDF 值构成文本的特征向量,最后采用 K-means 算法对文本的特征向量进行聚类。Li 等^[15]提出了一种 LDA 模型和 VSM 模型相结合的聚类方法,采用包含文本、主题和主题特征词的三层架构,通过整合 VSM 和 LDA 两个模型,设定阈值,对文本通过 VSM 和 LDA 两个模型生成的两个向量赋予不同的权重,并根据两个权重不同的向量对文本相似度进行计算,最后根据评论文本两两之间的相似度采用 K-means 算法对文本进行聚类。

上述研究存在两个不足之处:1)均未在主题词层面分析各个主题词对用户观点突出程度;2)在研究如何设置 K-means 初始聚类个数和聚类中心时,聚类的结果容易陷入局部最优解的缺陷仍没有得到有效的改善。因此,基于文本主题分析的用户聚类方法仍有较大的改进之处,这正是论文研究的重点与创新。

3 模型和算法

图 1 给出了基于主题分析的用户评论聚类模型的流程图。为了表述方便,相关的符号定义如表 1 所列。

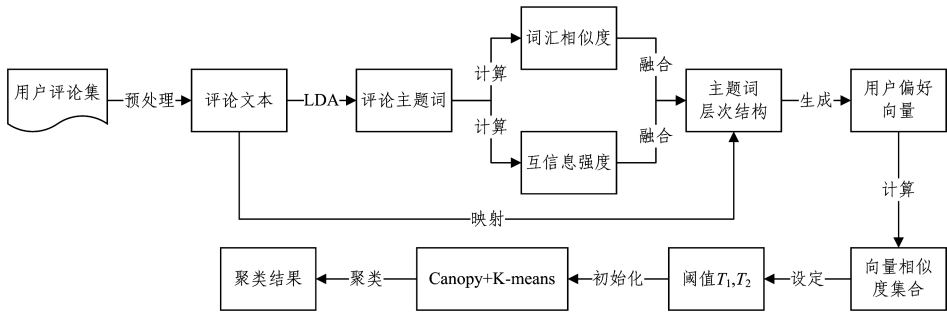


图 1 用户评论聚类流程图
Fig. 1 Process of user reviews clustering

首先,鉴于不同主题词对用户观点的突出程度不同,使用 LDA 处理用户评论集 $R = \{R_1, R_2, \dots, R_m\}$,生成主题集 $T = \{T_1, T_2, \dots, T_n\}$ 和各主题下的主题词分布 $W_i = \{W_{i1}, W_{i2}, \dots, W_{iN}\}$ 。融合主题词之间的相似度与互信息强度,生成主题词层次结构,其中主题词 $W_i = \{W_{i1}, W_{i2}, \dots, W_{iN}\}$ 对

应的层次为 $H_i = \{H_{i1}, H_{i2}, \dots, H_{iN}\}$,以主题词所在层次赋予其不同的权重,以表征主题词对用户观点的突出程度,即层次越深对用户观点的突出程度越大,并将用户评论 R_m 映射于主题词层次结构得到主题词数及平均深度,从而得到用户评论在每个主题上的表现值 U_m^n ,由用户评论在每个主题上的

表现值构成主题向量 $\mathbf{U}^m = \{U_1^m, U_2^m, \dots, U_n^m\}$ 。然后,考虑 K-means 算法易陷入局部最优解的问题,在用户评论聚类阶段,用户评论间的相似度通过计算用户评论对应的主题向量之间的欧氏距离得出。获取相似度的最大值 $d_{\max}$ 、最小值 $d_{\min}$ 以及 $d_{\min}$ 的位数 d_w ,以 d_w 为步长,自动设置 Canopy 聚类算法

的阈值 T_1 和 T_2 ,根据步长 d_w 不断更新 T_1 和 T_2 ,获取聚类个数 K 以及聚类中心集合 $C = \{C_1, C_2, \dots, C_K\}$,用于初始化 K-means 算法;对主题向量 $\mathbf{U}^m = \{U_1^m, U_2^m, \dots, U_n^m\}$ 进行聚类,并采取常用的 Silhouette Coefficient 和 Calinski-Harabaz Index 指标评估聚类结果,最后取不同 K 值下聚类效果最好的值。

表 1 本文词汇表
Table 1 Glossary used in this paper

名称	符号表示	说明
用户评论集	$R = \{R_1, R_2, \dots, R_m\}$	评论集合共有 m 条记录
主题集	$T = \{T_1, T_2, \dots, T_n\}$	主题集共有 n 个主题
主题 i 下的主题词集	$W_i = \{W_{i1}, W_{i2}, \dots, W_{iN}\}$	主题词集中共有 N 个主题词
主题 i 下的主题词层次结构	$H_i = \{H_{i1}, H_{i2}, \dots, H_{iN}\}$	主题词 W_{iN} 对应层次 H_{iN}
第 m 条评论对应的主题向量	$\mathbf{U}^m = \{U_1^m, U_2^m, \dots, U_n^m\}$	R_m 在每个主题上的表现值为 U_n^m
用户评论两两相似度最大值	$d_{\max}$	评论集中任意两条评论的相似度
用户评论两两相似度最小值	$d_{\min}$	评论集中任意两条评论的相似度
$d_{\min}$ 的位数	d_w	相似度最小值 $d_{\min}$ 的位数
聚类个数	K	初始化 K-means 算法的聚类个数
初始聚类中心集合	$C = \{C_1, C_2, \dots, C_K\}$	有 K 个类别对应 K 个聚类中心

4 模型和算法流程

4.1 评论主题词层次分析

每个主题词的概率值不同,可以通过计算主题词之间的相似度以及主题词的互信息强度来确定主题词之间的层次关系,生成主题词层次结构。依据主题词在层次结构中的层次来表征其对用户观点的突出程度。主题词之间的层次关系识别原则为主题词的概率值越大就越可能成为层次树的上层概念。采用互信息值强度作为主题词的度量方式,比较每个主题词的互信息强度来对主题词进行上下位关系的判定。同时,以主题词之间的相似度为构建结构的约束条件,令相似度高的主题词分布在层次结构的同一分支中,而相似度低的主题词分布在层次结构的不同分支中。

首先计算主题词互信息强度并进行降序排列,在每个主题下均得到有序的主题词序列集合 $W_i' = \{W_{i1}': MI(W_{i1}'), W_{i2}': MI(W_{i2}'), \dots, W_{iN}': MI(W_{iN}')\}$,且 $MI(W_{i1}') > MI(W_{i2}') > \dots > MI(W_{iN}')$,选择其中互信息值强度最大的主题词 W_{i1}' 作为层次树的上位概念词并从集合 W_i' 中删除 W_{i1}' 。此时,选择 W_{i2}' 作为层次结构待判定主题词,若主题词 W_{i2}' 与上位概念词 W_{i1}' 之间的关系满足定义 1 的要求,则将主题词 W_{i2}' 作为上位概念词 W_{i1}' 的下位概念词加入到层次结构中,并从集合 W_i' 中删除 W_{i2}' ;若不满足定义 1 的要求,则在集合 W_i' 中保留 W_{i2}' 。

定义 1 主题词 W_{ia} 和 W_{ib} 的层次关系判定:

- 1) 如式(1)所示,满足 $SIM(W_{ia}, W_{ib}) > \partial$, ∂ 是调节参数;
- 2) 如式(3)所示,满足 $MI(W_{ia}) < MI(W_{ib})$ 。

同理,按序依次对待识别层次的主题词进行判定,直至所有的主题词都标识结束,则该主题下的层次结构构建完成。采用同样的方法对每个主题下的主题词进行层次结构的构建,最终生成 n 个主题词层次结构,主题 i 下的主题词构成的层次树为 $H_i = \{H_{i1}, H_{i2}, \dots, H_{iN}\}$,其中 $H_{i1} \neq H_{i2} \neq \dots \neq H_{iN}$ 并且 $H_{iN} \neq H_{jN}$,因此主题词在层次结构中所处的层次各不相同。

主题 i 中的两个主题词 W_{ia} 和 W_{ib} 的相似度计算公式如下:

$$SIM(W_{ia}, W_{ib}) = \frac{(E_{W_{ia}} E_{W_{ib}})}{\sqrt{(E_{W_{ia}})^2} \sqrt{(E_{W_{ib}})^2}} \tag{1}$$

其中, $E_{W_{ia}}$ 表示在用户评论集 R 内,根据主题词 W_{ia} 在每条用户评论中的 TF-IDF 值构成的空间向量, $E_{W_{ia}} = \{E_{W_{ia},1}, E_{W_{ia},2}, \dots, E_{W_{ia},m}\}$ 。向量的元素 $E_{W_{ia},m}$ 表示主题词 W_{ia} 在用户评论集 R 中第 m 条评论中的 TF-IDF 值,如式(2)所示:

$$E_{W_{ia},m} = \frac{F_{W_{ia},m}}{\sum_{k=1}^N F_{W_{ik},m}} \times \log \frac{|R|}{|\{j: W_{ia} \in R_j\}|} \tag{2}$$

其中, $F_{W_{ia},m}$ 表示主题词 W_{ia} 在用户评论集 R 中出现的次数, $|R|$ 表示评论文本的总数, $|\{j: W_{ia} \in R_j\}|$ 表示包含词汇 W_{ia} 的文本总数。

在主题 i 下,主题词 W_{ia} 的互信息强度指主题词 W_{ia} 与其他主题词的点互信息的累加和:

$$MI(W_{ia}) = \sum_{k=1}^N PMI(W_{ia}, W_{ik}) \tag{3}$$

其中,两个主题词的点互信息计算公式如下:

$$PMI(W_{ia}, W_{ib}) = \log \frac{P_{(W_{ia}, W_{ib})}}{P_{W_{ia}} P_{W_{ib}}} \tag{4}$$

其中, $P_{W_{ia}}$ 表示主题词 W_{ia} 出现的概率,主题词 W_{ia} 和 W_{ib} 同时出现的概率则用 $P_{(W_{ia}, W_{ib})}$ 表示。

4.2 评论主题向量构建

主题 i 下的主题词集合 $W_i = \{W_{i1}, W_{i2}, \dots, W_{iN}\}$ 中的每个主题词在层次结构中所对应的层次为 $H_i = \{H_{i1}, H_{i2}, \dots, H_{iN}\}$,利用层次 H_{iN} 赋予主题词 W_{iN} 的权重。对于用户评论 R_m ,其包含主题 i 下的主题词数 S_i 的计算方式如下:

$$S_i = \begin{cases} S_i + 1, & \exists W_{iN} \in R_m \\ S_i, & \exists W_{iN} \notin R_m \end{cases} \tag{5}$$

遍历 n 个主题,从而从而得到用户评论中包含的每个主题下的主题词数 $S = \{S_1, S_2, \dots, S_n\}$ 。

根据用户评论 R_m 中包含主题词集 W_i 中的主题词以及主题词在层次结构 H_i 中的对应层次,计算每条用户评论在

主题层次树上的平均深度,计算公式如下:

$$L_i = \frac{\sum_{j=1}^N H_{ij} (\exists W_{ij} \in R_m)}{S_i} \quad (6)$$

遍历 n 个主题,从而得到用户评论在主题层次结构下的平均深度 $L = \{L_1, L_2, \dots, L_n\}$,其中 H_{ij} 表示主题 i 下主题词 W_{ij} 所在的层次, L_i 表示用户评论 R_m 在主题 i 的主题词层次树下的平均深度。

根据用户评论在每个主题下包含的主题词数 $S = \{S_1, S_2, \dots, S_n\}$ 以及用户评论在层次结构下的平均深度 $L = \{L_1, L_2, \dots, L_n\}$,采用式(7)的类似指数式函数来计算用户评论 R_m 对应于用户对主题 n 的偏好程度 U_n^m :

$$U_n^m = e^{-(2/L_n + 2/S_n)} \quad (7)$$

逐一计算 U_n^m ,从而得到对应的用户偏好向量 $U^m = \{U_1^m, U_2^m, \dots, U_n^m\}$ 。该方法充分考虑了用户评论 R_m 包含每个主题的主题词数 S 以及在主题词层次树下的平均深度 L 对于用户观点突出程度的不同。

4.3 自适应 canopy+K-means 聚类算法

鉴于 k-means 算法中的聚类数 K 需要事先指定, K 值的选定在事先不确定聚类数的聚类中难以估计,并且初始聚类中心的选择对 K-means 聚类算法有很大的影响,聚类结果易陷入局部最优解,而 Canopy 聚类算法最大的特点在于不需要指定聚类的个数 K ,因此,本文利用 Canopy 聚类对数据进行粗聚类,从而得到聚类的个数 K 以及对应的聚类中心集合 C ,将其作为 K-means 聚类算法的初始化,再利用 K-means 算法对数据进行细聚类,从而避免 K-means 随机选择聚类中心导致聚类结果浮动的问题,以及在一定程度上使聚类结果接近于全局最优解。而 Canopy 聚类算法的缺点在于在初始化两个阈值 T_1 和 T_2 时,阈值 T_1 和 T_2 的选择决定着聚类个数和聚类中心集合的生成。由于 T_1 和 T_2 大多由人工指定,具有不确定性和随机性,因此本文结合了用户文本之间的相似度,用文本之间的相似度为 T_1 和 T_2 赋予不同的值,由于不同数据集中文本之间的相似度不同,因此 T_1 和 T_2 随着数据集中的文本相似度不断变化,直至得到聚类最优解。本文提出的自适应 Canopy+K-means 聚类算法的具体过程如下:

- (1)用户评论集合对应的向量集合为 U^m ,计算每个向量与其他向量的距离,得到距离集合 D 。
- (2)获取距离集合 D 中的最大值 $d_{\max}$ 、最小值 $d_{\min}$ 、最小值 $d_{\min}$ 的位数 d_w 。
- (3)以最大值 $d_{\max}$ 为终止值,以最小值 $d_{\min}$ 为阈值 T_2 的初始值,以最小值 $d_{\min}$ 的位数 d_w 为步长,以 d_w 与阈值 T_2 的和为阈值 T_1 的初始值, $T_1 > T_2$ 。
- (4)从向量集合 U^m 中任取一个样本 p 作为一个簇的聚类中心,将其加入聚类中心集合 C ,从 U^m 中移除 p 。
- (5)计算 U^m 中全部点到 p 的距离 $dist$ 。
- (6)若 $dist < T_1$,则该点与 p 可能同类,该点作为弱关联。
- (7)若 $dist < T_2$,则该点与 p 同类,将点移出 U^m ,作为强关联。
- (8)反复执行步骤(4)一步骤(7),直至 U^m 为空。
- (9)当 U^m 为空时,得到样本簇数 K 以及类簇中心集合 C 。

(10)以簇数 K 和类簇中心集合 C 作为 K-means 算法的初始值,K-means 算法执行完毕后得到一个聚类结果,将聚类结果存入聚类结果集合 $Result$ 。

(11) T_2 不变, T_1 根据步长逐渐增大,满足 $T_1 > T_2$,反复执行步骤(4)一步骤(10),直至 T_1 到达最大值 $d_{\max}$ 。

(12)将最小值 $d_{\min}$ 增加 d_w ,最小值 $d_{\min}$ 为阈值 T_2 的初始值,以 d_w 与阈值 T_2 的和为阈值 T_1 的初始值, $T_1 > T_2$,反复执行步骤(4)一步骤(11),直至 T_2 到达最大值 $d_{\max}$ 。

(13)取不同聚类个数 K 的聚类结果集合 $Result$ 的最优结果。

该算法根据用户评论文本之间的相似度自动选择 Canopy 聚类算法的最优初始阈值 T_1 和 T_2 ,通过 Canopy 聚类得到聚类个数 K 和聚类中心集合 C ,避免了 K-means 算法中人工设定阈值以及初始聚类中心选择的随机性。以文本相似度的位数作为步长,在不断更新 T_1 和 T_2 的过程中,每个聚类个数 K 值所对应的聚类中心集合不同,因此聚类结果也不同,选择最好的聚类结果可以使聚类结果较 K-means 更加稳定,并在一定程度上趋近于全局最优。

5 实验和结果

5.1 数据集和评价标准

本文选取亚马逊上真实的用户评论作为数据集进行数据处理和用户评论聚类的实验,共选取了 8 类商品作为实验对象,具体如表 2 所列。

表 2 商品名称与编号
Table 2 Name and number of goods

商品名	商品编号
Tent(帐篷)	B00CDFVK0O
Belt(腰带)	B009A116Z8
Microphone(话筒)	B00XBQ8UGG
Zipper(拉链)	B005SG3ZAI
Christmas-Ornament(圣诞装饰)	B01BE8VXVG
Floating-Lantern(灯笼)	B00C1UXXC4
Waterproof(雨具)	B01EWBZW0A
Package(包装盒)	B004UWZ090

采用 Silhouette Coefficient(简称 SC 指数)和 Calinski-Harabaz Index(简称 CI 指数)两个指标评估用户评论聚类的效果。SC 指数适用于在未知聚类个数的情况下评价聚类效果,其结合了内聚度和分离度两种因素,一个数据集合的 SC 指数是所有类的 SC 指数之和的平均值。其计算公式如下:

$$SC = (\sum_{i=1}^m \frac{avg(i) - avg(i)}{\max(avg(i), avg(i))}) / m \quad (8)$$

其中, m 表示用户评论的个数; $avg(i)$ 表示数据元素 i 与其同类的其他元素的平均距离,即内聚度; $avg(i)$ 表示数据元素 i 与其不同类的其他元素的平均距离,即分离度。SC 指数的取值范围为 $[-1, 1]$,若同类元素之间的距离越近,不同类元素之间的距离越远,则 SC 指数越大,聚类效果就越好。

CI 指数亦是在实际聚类个数未知的情况下,结合类别内部之间的协方差以及类别之间的协方差对聚类结果进行评估,计算公式如下:

$$CI = \frac{tr(Dif_K) m - K}{tr(Sim_K) K - 1} \quad (9)$$

其中, m 表示用户评论的个数, K 表示聚类的个数, Dif_K 表示不同类别之间的协方差矩阵, Sim_K 表示同类内部元素之间的协方差矩阵, tr 表示计算矩阵的迹。因此, 当类别之间的协方差越大, 同类内部的协方差越小, 则 CI 指数越大, 聚类效果就越好。

5.2 测试分析

设置主题数 n 为 5, 主题词数 N 为 50, 对用户评论进行聚类, 聚类之前无法预知聚类的个数 K , 因此, 本文设置聚类个数 K 的值为 2~10 并对聚类效果进行比较。同时, 将本文方法的效果与传统的采用 LDA+kmeans 方法对用户评论进行聚类的效果进行比较。以编号为 B009A116Z8 的商品和编号为 B01EWBZW0A 的商品为例, 设置聚类个数 K 为不同的值时, B009A116Z8 商品和 00CDFVK0O 商品的用户评论聚类结果评估指标 SC 指数和 CI 指数的实验结果分别如图 2 和图 3 所示。

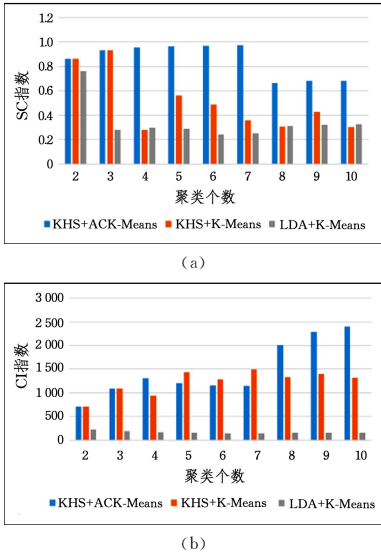


图 2 B009A116Z8 商品的用户评论聚类结果对比(电子版为彩色)
Fig. 2 Results comparison of B009A116Z8 commodity's user comment clustering

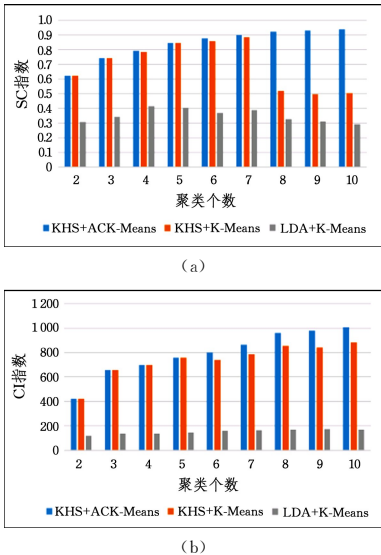


图 3 B01EWBZW0A 商品的用户评论聚类结果对比(电子版为彩色)
Fig. 3 Results comparison of B01EWBZW0A commodity's user comment clustering

如图 2 和图 3 所示, 设置不同的聚类个数, 聚类的结果有明显差异, 对于 B01EWBZW0A 商品, 当聚类个数增多时, 用户评论的聚类效果较好。而对于 B009A116Z8 商品, 当聚类个数增多时, 用户评论的效果却不是呈递增或者递减的趋势, 因此, 对于不同的商品, 应根据实际需求选择不同的用户评论聚类个数。本文提出的基于主题分析的用户评论聚类方法的聚类效果在 SC 指数和 CI 指数两个指标上均优于传统的 LDA+K-means 方法。如在测量 B009A116Z8 商品的用户评论聚类结果的 SC 指数时, 当聚类个数 K 为 4 和 10 时, 对主题词进行层次分析并使用 K-means 聚类, 结果陷入了局部最优解, 导致聚类结果在 SC 指数上不如传统 LDA+K-means 方法, 而本文提出的自适应 Canopy+K-means 聚类方法则有效地缓解了聚类结果陷入局部最优解的不足。由于自适应 Canopy+kmeans 聚类方法在聚类时将数据集样本之间相似度的位数作为寻找最优解的步长, 因此不能完全解决聚类结果陷入局部最优解的问题, 如在测量 B009A116Z8 商品的用户评论聚类结果的 CI 指数时, 当聚类个数 K 分别为 5, 6, 7 时, 自适应 Canopy+kmeans 聚类方法则不如传统的 K-means 方法, 但聚类结果仍优于传统的 LDA+K-means 方法。而如图 3 所示, 当聚类个数分别为 2, 3, 4, 5 时, 提出的自适应的 Canopy+K-means 聚类方法与传统的 K-means 算法的聚类结果相同; 而当聚类个数 K 值大于 5 时, 提出的聚类算法则优于传统的 K-means 算法。这是由于当聚类个数 K 值增大时, 聚类中心的选择随机性增大, 因此传统的 K-means 算法会使聚类结果更容易陷入局部最优解。

最后, 对所选取数据集集中的商品的用户评论进行聚类, 对于每种商品, 选取采用本文算法得到的最好聚类结果与传统的 LDA+kmeans 算法得到的最好聚类结果进行比较, 其评估指标 SC 指数和 CI 指数的结果如图 4 所示。

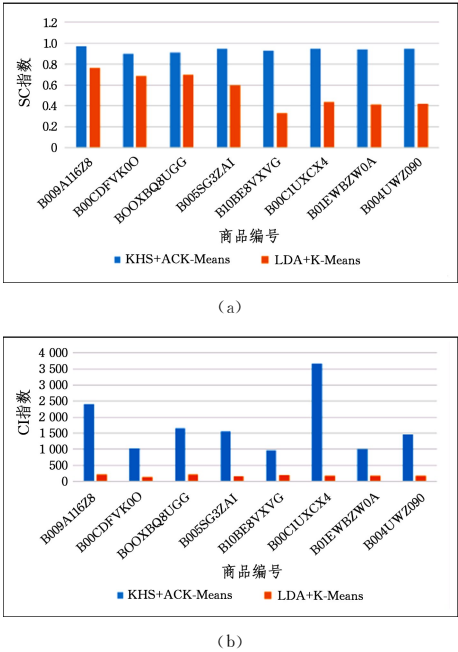


图 4 用户评论聚类结果(电子版为彩色)
Fig. 4 Results of user comment clustering

由图 4 可以看出, 提出的基于主题分析的用户评论聚类

方法在评估聚类结果的 SC 指数和 CI 指数两个指标上均优于传统的 LDA+K-means 算法。而由于各类商品的用户评论长短不一,评论质量无法统一,因此用户评论聚类效果的优化程度不同。对于用户评论主题词层次关系明显的商品如 B01BE8VXVG,B00C1UXCX4,B01EWBZW0A 和 B004UWZ-090,聚类效果的优化程度明显较大;而对于用户评论主题词之间层次关系不明显的商品,则优化程度较小。

结束语 本文提出的基于用户评论主题分析的用户评论聚类方法,从用户评论的主题分析出发,考虑了不同主题词对用户观点的突出程度不同,对主题词进行层次分析,并构建主题词层次结构,使不同主题词能够更好地表征用户的观点。同时设计用户评论到主题词层次结构的映射规则,每条用户评论生成对应的主题向量。而为了解决 K-means 聚类易陷入局部最优解的缺点,本文提出了一种自适应的 canopy+k-means 聚类方法,从而提高了用户评论聚类的效果,使聚类在一定程度上趋近于全局最优解。

在未来的工作中,将把重点放在使聚类结果更趋近于聚类算法的全局最优解方面。在本文实验中,提出的自适应 Canopy+k-means 算法在数据集上的某些指标还存在陷入局部最优解的问题,但与原始的 k-means 算法相比,陷入局部最优解的问题已有改善,未来工作将逐步分析陷入局部最优解的原因,使聚类结果更加趋近于全局最优解。

参 考 文 献

[1] QIAO Z,ZHANG X,ZHOU M,et al. A Domain Oriented LDA Model for Mining Product Defects from Online Customer Reviews[C]// The, Hawaii International Conference on System Sciences. 2017.

[2] JO Y,OH A H. Aspect and sentiment unification model for on-line review analysis[C]// ACM International Conference on Web Search and Data Mining. ACM,2011:815-824.

[3] IVAN T,MCDONALD R. A joint model of text and aspect ratings for sentiment summarization[C]//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. ACL,2008:308-316.

[4] ZHANG W,XU M,JIANG Q. Opinion Mining and Sentiment Analysis in Social Media:Challenges and Applications[C]// International Conference on HCI in Business, Government, and Organizations. Springer,Cham,2018.

[5] XIAO R,KONG L,ZHANG Y. A text clustering model for diverse versions discovery[J]. Caa Transactions on Intelligent

Systems,2012,2(4):1113-1117.

[6] YANG Y,WANG J,HUANG W,et al. TopicPie: An Interactive Visualization for LDA-Based Topic Analysis[C]// IEEE Second International Conference on Multimedia Big Data. IEEE,2016: 25-32.

[7] KHANMOHAMMADI S,ADIBEIG N,SHANEHBANDY S. An Improved Overlapping k-Means Clustering Method for Medical Applications[J]. Expert Systems with Applications,2017, 67(1):12-18.

[8] FISH L S. Hierarchical Relationship Development:Parents and Children[J]. Journal of Marital & Family Therapy, 2010, 26(4):501-510.

[9] WU J,HOU S X,JIN M M,et al. LDA Feature Selection Based Text Classification and User Clustering in Chinese Online Health Community[J]. Journal of the China Society for Scientific and Technical Information,2017(11):1183-1191. (in Chinese)

吴江,侯绍新,靳萌萌,等. 基于 LDA 模型特征选择的在线医疗社区文本分类及用户聚类研究[J]. 情报学报,2017(11):1183-1191.

[10] ALHARBI A S,LI Y,XU Y. Integrating LDA with Clustering Technique for Relevance Feature Selection[C]// Australasian Joint Conference on Artificial Intelligence. Springer, Cham, 2017:274-286.

[11] CAI Y,CHEN X,PENG P X,et al. A LDA Feature Grouping Method for Subspace Clustering of Text Data[C]//Pacific Asia Workshop on Intelligence and Security Informatics. Springer International Publishing,2014:78-90.

[12] SHI M,LIU J,ZHOU D,et al. WE-LDA: A Word Embeddings Augmented LDA Model for Web Services Clustering[C]// IEEE International Conference on Web Services. IEEE Computer Society,2017:9-16.

[13] XIAOBO T. Research on Micro-blog Topic Retrieval Model Based on the Integration of Text Clustering with LDA[J]. Information Studies:Theory & Application,2013,36(8):85-90.

[14] SUADAA L H,PURWARIANTI A. Combination of Latent Dirichlet Allocation (LDA)and Term Frequency-Inverse Cluster Frequency (TFxICF)in Indonesian text clustering with labeling [C]//International Conference on Information and Communication Technology. IEEE,2016:1-6.

[15] LI C,YANG C,JIANG Q. The research on text clustering based on LDA joint model[J]. Journal of Intelligent & Fuzzy Systems,2017,32(5):3655-3667.