情报学报 ISSN 1000-0135
第35卷 第1期 55-65, 2016年1月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 35 No. 1, 55-65 January 2016

doi:10.3772/j.issn.1000-0135.2016.001.006

基于改进 *K*-means 聚类的在线新闻评论主题抽取[1)]

夏火松 李保国 杨 培

(武汉纺织大学管理学院,武汉 430073)

摘要 新闻评论反映民众对新闻事件的观点,抽取评论主题,对用户、企业、政府都具有很高的情报分析价值。基于 *K*-means 聚类的主题挖掘算法应用到新闻评论中时,在欧氏距离下,如果使用最大距离法选初始点则会聚成一大类。为解决这个问题,论文首先在预处理阶段增加同义词替换和自动构建领域词典的部分,改善了数据稀疏性和高维性。其次,提出了 *K*-means 改进算法,用隐藏长评论-最大距离法选初始点,解决了初始点多为离群点的问题,用方差拐点确定 *K* 值,解决了预先设定聚类个数的问题,实验发现了先用 BW 权重选初始点,再用新提出的 BW-DF 权重聚类的效果最好。最后,将改进算法与原算法的聚类效果比较,实验结果表明,改进算法准确率高,抽取新闻评论主题的效果明显。

关键词 在线新闻评论 *K*-means 聚类改进 主题抽取 同义词替换 分词领域词典

Topic Extraction in News Comments Based on Improved *K*-means Clustering Algorithm

Xia Huosong, Li Baoguo and Yang Pei

(*School of Management, Wuhan Textile University, Wuhan* 430073)

Abstract News comments on the web express readers' attitudes or opinions about the news events. Opinion topic extraction from news comments is valuable for users, businesses and government. When *K*-means clustering algorithm for topic mining is applied to news comments in the Euclidean distance, it has bad clustering performance through the maximum distance method to select initial centers. To solve this problem, firstly, synonym substitution and field dictionary is introduced in the preprocessing stage to solve the problem of data sparseness and multi dimension. Secondly, the improved *K*-means algorithm is proposed. It selects the initial cluster centers according to maximum distance after the long comments are hidden, which solves the problem that initial centers are outliers. The method of variance inflection is proposed to deal with the problem of the traditional *K*-means algorithm in which k values needs to be input. It is found that the new algorithm has good clustering performance by BW-DF after BW is used to select initial centers. Finally, the effect of improved clustering algorithm is compared with the original one. The results show that the improved algorithm with high accuracy extracts opinion topic effectively.

Keywords online news comments, Improved *K*-means clustering algorithm, topic extraction, synonym substitution, field dictionary

收稿日期:2015年3月26日

作者简介:**夏火松**,男,1964年生,教授,博士,主要研究方向:知识管理、数据挖掘、物流信息管理和电子商务、DSS,E-mail:bxxhs@sina.com。**李保国**,男,1990年生,硕士研究生,主要研究方向:信息管理、数据挖掘。**杨培**,男,1990年生,研究生,主要研究方向:数据挖掘、信息管理。

1) 基金项目:本文系国家自然科学基金项目(71171153)"24小时知识工厂的知识共享活动模型与服务支持系统研究"的研究成果之一。

1 引 言

随着大数据时代的到来,数据分析能够向管理者和政策制定者提供精准的决策支持信息。News 2.0(Journalism 2.0)是指新闻网站利用 Web 2.0 技术鼓励用户参与互动,它实现了用户生成内容和用户至用户的交流[1]。为获得巨大的社会价值和经济效益,News 2.0 网站加强网络新闻的普及并鼓励用户对在线新闻进行评论。这些评论已经成为研究者进行评论意见挖掘(Opinion Mining)的重要情报资源,例如追踪人们对某个特定事件、个人、公司或政府行为的态度变化[2]等。而这项工作的前提是抽取评价主题/对象(Extraction of Opinion Topic or Targets)以及各主题所占评论数量的比例(由于评论的数量可以指示影响程度[3])。通过对评论的研究,民众可以从多角度理解新闻、社会;相关企业可以作危机公关,引导舆论向正面积极的方向发展;政策决策部门可以进行民意询问、公共管理[4]和政府决策[5],有效实现舆情分析和监控。

虽然意见挖掘在产品评论数据上已相对成熟,但遗憾的是它并不集中于新闻评论领域,这里的新闻评论不是新闻文章或专业人士的新闻评论,而是普通民众在交互式网站上对新闻的简短评论。当把针对产品评论的产品特征抽取(Features Extraction)方法直接应用到新闻评论数据时,抽取评价主题(对象)的效果并不明显[6]。这与新闻评论具有不同的特点具有一定关系。第一,产品通常具有一组明确的属性或特征(如尺寸,耐久性)和与之相关的评价词(情感词),可以使用少量的关键词来识别频繁特征词[7],可是新闻评论更加复杂。首先新闻评论有更多的潜在评价主题以及在原新闻上产生漂移的评价主题(如对"占中"的评论中出现一些对香港教育反思的评价主题),其次新闻评论的评价主题与评价词的联系并不太强。第二,从研究目的来讲,产品评论的目的是为了改善各个产品特征的质量或服务,所以更注重各个产品特征的褒贬程度;而新闻评论除了要了解人们对评价主题以及漂移的评价主题的态度倾向外(如对某个人的支持与否),更多的则是要了解各个评价主题以及它所占评论数量的比例,并以此了解舆情动态。

由于评价对象可认为蕴含于评论语料中某个特定的主题,近年来,许多学者将主题挖掘中的主题模型[8]应用到意见挖掘领域来抽取评价主题[9]。从主题挖掘的结果中提取评价主题/对象更适合新闻评论。这是由于基于文本聚类的主题挖掘算法可以得到各评论主题[10]以及各主题所占评论数量的比例,便于从中提取评价对象以及它所占评论数量的比例。而且聚类算法是根据相似度大小分成不同的聚集簇,利于发现热点评论对象和漂移的评论对象。所以这篇文章先使用基于改进 *K*-means 聚类算法的主题挖掘方法得到评论主题及比例,再从中提取评价主题及各评价主题所占评论数量的比例,为后续的意见挖掘作准备。

可是当把 *K*-means 聚类算法直接应用到具有稀疏性和高维性的新闻评论上时,用最大距离法选初始点,在欧氏距离下,则会出现聚成一大类的情况,在余弦距离下则准确率和召回率不高,都不能满足主题抽取的需要。这篇论文针对评论数据及其聚类的特点,提出了相应预处理技术和欧氏距离下 *K*-means 改进算法来解决这个问题。论文组织如下:第一部分是对相关研究工作的梳理;第二部分是研究依据的理论方法;第三部分是根据在线新闻评论的特点,提出评价主题抽取过程与算法改进;第四部分是构造实验并验证改进算法;第五部分是改进前后的算法比较;最后是论文的讨论与结论。

2 相关研究工作

意见挖掘[11]又称为情感分析(Sentiment analysis),有意见信息抽取、情感分类、意见摘要三个任务,意见信息抽取包括评价对象/主题的抽取和评价词(Opinion words or Sentiment Words)抽取。评价对象指评论中讨论的主题,在产品评论中,评价对象是指用户评论的产品或者产品的某一属性/特征,所以评价对象/主题的抽取也叫产品特征抽取。产品特征抽取有三种方法,第一种是基于人工定义的特征种子集的方法;第二种方法比较普遍,为基于名词和名词短语的方法;第三种是基于主题挖掘的主题模型的方法[9]。

文本主题挖掘算法有基于聚类的主题挖掘算法、基于线性代数、基于概率模型、基于主题模型(比如 LDA)算法。这里只对与新闻评论相似的短文本(包括微博、聊天记录、用户评论等)的主题挖掘进行回顾。张晨逸等[12]提出尽管传统文本的主题挖掘已得到了广泛的研究,但对于特殊的短文本,传统的文本挖掘算法不能很好的对它建模。唐晓波等[13]兼顾到微博信息的稀疏性、多维性、海量性等

特点,提出针对此特点的预处理技术,并使用 LDA 进行主题挖掘。唐晓波等[10]把针对短文本的主题挖掘分为文本级主题挖掘、词语级主题挖掘和单句粒度级的主题挖掘。文本级的主题挖掘方法解决了短文本聚类问题。唐晓波等[14]融合了文本聚类和 LDA 来发现微博主题,并在微博聚类中发现,由于微博短文本上的关键词出现次数很少,直接选择常用的 TF-IDF 方法时,得到的相似度与实际情况偏离比较严重,不能达到预期的聚类效果,于是使用重叠文档相似度的方法。对于基于主题挖掘的产品特征抽取的研究,Titov 等[15]使用多粒度的 LDA 模型获取产品特征,并将相似的产品特征聚类到同一主题。Jo 等[16]将评论中每一个句子属于一个主题和一种情感,基于 LDA 模型提出 SLDA 模型来获取产品特征。李芳等[17]应用 LDA 模型挖掘评论文本集合中潜在的评价主题,并利用多特征融合的方法识别评价主题的褒贬倾向性,给用户提供更直接、更方便的评价信息。实验证明,此类方法在产品特征抽取中能够取得一定的效果,但是此方法需要预先定义主题的个数。

Ma 等[18]指出针对产品评论的产品特征抽取方法并不太适用于新闻评论数据的评价主题(对象)的抽取,首次把意见挖掘引入新闻评论(News Comments)领域,并使用中心理论来抽取隐性和显性评价主题。该方法既使用原新闻文章的全局信息,也使用了该新闻的评论信息。Tsagkias 等[19]首次将预测评论数量引入新闻评论领域,并通过评论数量了解新闻的重要性及影响力。

在新闻评论及与之类似的短文本聚类研究中,张立[20]针对新闻评论数据的 K-means 聚类算法的不足,通过构造评论相似度矩阵来选择初始点并对划分类别方法进行改进,并且使用余弦距离。沈幸峰[21]在 LDA 与聚类相融合的新闻评论的话题挖掘中发现,评论中较多的噪声数据对 K-means 的中心点影响较大,一个离群点可能显著扭曲数据的分布,平方误差函数更是严重恶化这一影响,所以使用了 K-medoids 聚类。马莹莹[22]在 BBS 数据的 K-means 聚类中,用 Two-step 算法确定 K 值,用基于复杂网格特性的方法选择初始点,并且使用欧式距离。赵辉等[23]利用结点综合特性来选取 K-means 初始中心点,提高了短文本聚类效果。朱晓峰等[24]提出了聚类算法研究很多,然而将聚类算法应用到微博短文本的舆情监测中,并实现了舆情监测功能的却很少,并分析原因有算法准确度有限、响应时间偏长、对用户介入的要求较高等原因。并使用基于文本平均相似度的 K-means 算法,提高了微博短文本的聚类精度和聚类稳定性。

针对原 K-means 算法应用到新闻评论数据的问题,我们从数据预处理、K-means 算法方面进行改进。首先,在评论数据预处理阶段,增加了其他研究者不常用的同义词替换和自动构建分词领域词典的部分。其次,为选择合适的相似度。这篇文章通过实验比较了欧氏距离与余弦距离在新闻评论数据上的聚类特性区别,为 K-means 改进提供了理论基础;同时,为解决评论的噪声数据、离群点以及高维特性[25]对聚类初始点的影响,使用了隐藏长评论-最大距离法;为解决预先定义主题个数的问题,使用各类评论条数方差拐点的方法来确定 K 值;为解决 TF-IDF 权重函数得到的相似度与实际情况严重偏离的问题,这篇文章测试了多种新提出的权重函数。研究的结果对于未来新闻评论的聚类及主题抽取有着重要的参考价值。

3 相关理论

基于文本聚类的主题挖掘算法,是通过 VSM 将评论文本集转化成结构化向量,再用聚类算法处理,从而得到基于聚类结果的评论主题。聚类有基于划分、基于层次、基于密度、基于网格、基于模型的方法。在文本挖掘中,基于划分的 K-means 算法可以生成质量更好的聚类结果,并有较高的效率和结果可解释性,适于处理大数据集和各类型数据,所以这篇论文选用 K-means 算法[26]。K-means 算法需要解决以下问题:

1)聚类类别数 K 的确定。K 值选择有基于聚类有效性、基于遗传算法的方法。这篇论文在原 K-means 算法中选用人工分类的个数,在欧氏距离下的 K-means 改进算法中提出了各类评论条数方差拐点的方法。

2)初始质心的选择。由于评论向量是高维的,有大量孤立点,所以在 K-means 算法中随机选取初始点很容易陷入局部最优;同时不同的初始点易于得到不同的聚类结果,给比较聚类效果和主题挖掘带来不便;而且还增加了聚类的总迭代次数和聚类总开销。所以初始点选择的改进算法有基于密度、基于优化的方法、最小误差平方和法、层次聚类法等。由于文本聚类的向量都是正数以致欧氏距离和余弦距离都为正数,所以这篇论文在原 K-means 聚

类方法中参考了最大距离法[27]并作适当简化。首先随机选择一个样本点作为初始点,再计算样本集的$(N-1)$个样本点与第一个初始点的距离,并将距离最远的那个样本点作为第二个初始点。在剩余的$(N-2)$个样本点中,选取到前面两个初始点各自距离乘积最大值的那个样本点作为第三个初始点。依此类推找到K个初始点。在欧氏距离K-means改进算法中提出了隐藏长评论-最大距离法。

3)相似度。两点间的相关程度通常用他们的相似度$Sim(D_1,D_2)$来衡量,有欧式距离和余弦距离等。

4)聚类效果度量。在解决欧氏距离聚成一大类的问题时,主要是看哪种方法可以把评论条数大的类聚出来,各种权重以及K值对准确率的影响相对较小,所以主要使用各类评论条数的方差来衡量,方差小的,聚出的大类多,聚类效果好。比较在欧氏距离下改进K-means算法与余弦距离的准确率时,使用外部度量的方法。任意两条评论之间的关系,按照人工分类的标准和自动聚类的标准有表1所示的4种情况。则积极准确率$PP = c/(c + d)$;消极准确率[28]$NP = a/(a + b)$;积极召回率$PR = c/(b + c)$;消极召回率$NR = a/(a + d)$;准确率又称为"精度"、"正确率",而积极与消极是从正面和负面两个方向予以衡量。

表1 评价准确率和召回率时任意两个评论关系的情况

准确率		人工分类中属于同一类	
		是	否
自动聚类中属于同一类	是	c	b
	否	d	a

4 新闻评论主题抽取的过程与K-means算法改进

4.1 新闻评论文本的抓取

即去除网页中的非文本信息,及html标记语言,得到新闻评论文本。抓取方法有通过python、java、c、c++等语言编程的方法和通过火车采集器、GooSeaker、Wget等软件的方法。本文使用python提取评论。

4.2 数据预处理

即对评论文本前期处理,以供有效的文本表示。数据预处理包含以下四个部分。

1)数据清理。即删除大量无用信息和非评论信息以提取出有意义的评论。

2)同义词替换[29]。首先,新闻评论语言口语化严重,还常带有缩写、不规范用语、网络新词等,不利于对评论的理解。把这些词替换为一般的词既有利于对评论的理解又降低了特征项维数。其次,单条新闻评论文本长度短小,但评论集数量庞大,给评论文本表示造成了严重的数据稀疏性以及特征空间高维性的后果,而汉语中有大量的同义词和近义词,可以用同义词替换改善数据稀疏性和特征空间高维性。

3)分词。分词是根据汉语语义的最小单位-词语而非字,来切割句子,以模拟人理解中文的实际过程。分词软件有自带的原分词词典和供用户加进领域词库的用户(领域)词典。新闻评论文本含有大量网络新词、缩写词,不同的新闻又含有特定的领域词。而原词典又不能涵盖所有的词,从而造成原本的一个词语被切开。可当把这个词加入用户词典后,则不会被切开,从而优化对于特定新闻的分词。用户词典的构建有人工和自动两种方式,这篇论文除了根据原新闻构建领域词典外,还根据新闻的评论来自动构建领域词典。分词有基于词典、基于统计和基于语义的分词算法。而这篇论文提出的自动构建领域词典的方法就借鉴了基于统计的算法。步骤为:①获取新闻评论集,然后统计相邻的字词同时出现的次数,次数越多就越可能构成一个词。当达到一定阈值时就构成了一个词。②然后去除在原分词词典已有的词,即可形成领域词典。

4)删除停用词。停用词在所有文本中的频率分布相近,增加了文本之间的相似程度,不利于聚成不同的类。停用词使用频率极高,会占用对聚类有意义的特征词的密度,造成特征词不明显,所以要去掉。

4.3 文本表示

文本表示是文本聚类的基础,即将非结构化文本转换为电脑能够处理的结构化向量。文本表示模型有向量空间模型、语言模型、后缀树模型和基于本体的模型等,这篇论文选用文本聚类经常采用的向量空间模型[30](VSM)。模型为:D_i是第i条评论,D表评论集;$T(T_1,T_2,\cdots,T_n)$为特征项集合,例如可以把评论集D分词后得到的所有词的集合作为T,T_j是第j个特征项,则D_i可表示为$D_i(T_1,T_2,\cdots,$

T_n)；w_{ij}表示特征项 T_j 在第 i 条评论的权重。这样给定一评论 D_i，我们可以把它用向量 $D_i(w_{i1}, w_{i2}, \cdots, w_{in})$表示。VSM 有两个基本问题。

1)特征降维。即降低特征项集合的维度，特征降维有特征选择和特征抽取两种。特征选择有信息增益、期望交叉信息熵、互信息、χ^2 统计法、特征词频方法(TF)、文档频数方法(DF)等。特征抽取有矩阵奇异值分解、主成分分析等。由于聚类前并不知道有几类，这篇论文使用特征词频和文档频数的方法。

2)特征项权重计算。常用权重函数有布尔函数、平方根函数、对数函数、TF 算法、TF-IDF 等。本文用到的：①词频 - 逆向文档频率 TF-IDF：$TF\text{-}IDF = tf_{ij} \times idf_j = (n_{ij}/|D_i|) \times \log(|D|/|\{i:T_j \in D_i\}|)$，其中 n_{ij}是特征项 T_j 在第 i 条评论 D_i 中的出现次数；$|D_i|$：在评论 D_i 中所有字词的个数；$|D|$：评论集 D 中的评论总条数；$|\{i:T_j \in D_i\}|$：评论集中包含特征项 T_j 的评论的数目。②布尔权重(BW)：如果特征项 T_j 出现在评论 D_i 中，则其权重 $w_{ij} = 1$，否则其权重 $w_{ij} = 0$。③原词频权重(RTF)：$w_{ij} = n_{ij}$。④TF 法：$TF = tf_{ij} = n_{ij}/|D_i|$。

因为 TF-IDF 与包含特征词 T_j 的评论数目$|\{i:T_j \in D_i\}|$成反比，所以文档频率低的词语可以产生高权重的 TF-IDF，对聚类结果影响大，利于体现文档频率低但含有重要信息的特征词。但是对于把几百上千条评论聚类成类别数是个位数级的问题来讲，主要类别的评论条数一般较大，所以低文档频率的特征项不应有大的影响。相反，如果文档频率是较大的值，则该词很有可能代表某一个评论条数大的类，具有较好的类别区分能力，应该提高该词权重，增加在相似度计算的影响力。使得含评论条数多的类别更容易在相同的初始值的情况下被聚类出来。所以这篇论文提出 TF-IDF 方法的改进 TF-DF。为寻求最优权重函数，又定义了几种权重作对比实验。⑤词频 - 文档频率权重 TF-DF：$TF - DF = tf_{ij} \times df_j = (n_{ij}/|D_i|) \times |\{i:T_j \in D_i\}|$；⑥布尔 - 文档频率权重 BW - DF：$BW - DF = bw \times df_j$；⑦原词频 - 文档频率权重 RTF - DF：$RTF - DF = n_{ij} \times df_j$；⑧TF - DF^{-2}法：$TF - DF = tf_{ij} \times \sqrt{df_j}$；⑨BW - DF^{-2}法：$BW - DF^{-2} = bw \times \sqrt{df_j}$；⑩RTF - DF^{-2}法：$RTF - DF^{-2} = n_{ij} \times \sqrt{df_j}$；

4.4 采用改进 K-means 聚类算法进行聚类并得到基于聚类结果的评论主题

(1)欧式距离、余弦距离在新闻评论数据下的 K-means 聚类效果分析

为了比较两种距离的聚类效果，人工指定相同的初始点(参考了最大距离的初始点选择结果)，用两种距离聚类，同时，记录迭代到最优过程中的各次迭代结果(以各类的评论个数和评论个数的方差来表示)，截取首次迭代与最后一次迭代的结果如表 2 所示。如果初始点相同，欧式距离和余弦距离首次迭代的聚类结果大致是相同的，这是由于对于稀疏性向量而言，当有共同特征项时，余弦或欧式距离对应的相似度才会增大，从而聚为一类。首次迭代结果有三个共同特点。第一，某一个大类(一般为最大类)聚类精度较低，定义为杂类，其他类聚类精度较高。第二，初次聚类的各类结果与相应初始点的对应性大，即各类与初始点的特征项相近，意思相同。又由于用最大距离选初始点，各类初始点的特征项重叠较少，所以首次迭代的各类结果间特征项重合率也较小(又由于同一意思可以由不同的词语/特征项来表达，所以不同的聚类类别还是可能有相同的意思)。第三，聚类结果中某类的评论条数的多少由初始点选择的好坏决定。

然后在两种距离下都迭代到最优。欧氏距离的平均迭代次数较小，最终聚类结果基本上和第一次迭代的结果相同，所以欧氏距离保持了首次迭代的三个特点。余弦距离的平均迭代次数较大，最终聚

表 2　初始点相同时两种距离的迭代结果

初始点相同时两种距离的迭代结果		方差	各类别的评论个数(条)							
			1	2	3	4	5	6	7	8
欧式距离	第 1 次迭代	0.149	29	46	84	30	30	8	265	31
	第 4 次迭代(最优)	0.148	22	42	84	25	18	8	261	63
余弦距离	第 1 次迭代	0.116	219	49	84	39	30	44	27	31
	第 9 次迭代(最优)	0.047	86	56	93	33	100	31	72	52

类结果和第一次迭代结果相差较大,具有新的特点。第一,各类精度都较低。第二,各类与相应初始点的对应性小,即各类与初始点的特征项差别大,意思不同。而且,各类间特征项出现交叉,进而意思出现交叉。第三,各类评论条数受初始点选择的好坏影响不大,评论条数基本达到平均。以上三个特点在 K 值小时尤为明显。两种距离聚类结果不同的原因如下,第一次迭代后,杂类(一般是评论条数最大的那个)按照聚类规则取平均值并作为本类新的质心,可是杂类中的评论与这个新的质心的相似度对于余弦和欧式距离是不同的。对于欧氏距离,杂类中的评论到这个新的质心的相似度一般比到初始质心的相似度大,所以一般不会散布到其他精度高的类别中而使其他类别的准确率下降。对于余弦距离,相似度一般比到初始质心小,而且可能还会比到其他类的质心的相似度小,所以这些评论会散布到其他精度高的类别中而使得它们准确率下降,使其他精度高的各类与初始点的意思相差较大,进而各类间特征项出现交叉。

可见对于低维稠密向量,迭代有利于优化目标函数,使得聚类效果变好。对于高维稀疏的评论向量,迭代反而会使得各类精度下降,所以对于算法的改进应集中在首次迭代上,而它的关键又在于初始点选择。

(2) K-means 改进算法

欧式距离除那个杂类外准确率较高,所以有利于主题挖掘。可是用最大距离法找初始点时,基本都是离群点(由于评论长度过长或过小),又由于欧式距离各类与初始点的对应性大的特点,出现聚成一大类的情况(这与程序设计为"如果没有更大的相似度,就保持原来的类别"有一定关系)。要解决这个问题,关键是找到不是离群的大类点作为初始点。第一,针对初始点评论长度过长的问题,这篇文章提出了隐藏长评论 - 最大距离法,相应权重和欧氏距离特性又保证了初始点长度基本等于长度阈值限定的适当长度,增大了所选初始点是大类点的概率。第二,适当提高 K 值增大捕获大类点的概率,聚类后再用合并同类的方法弥补 K 值大的弊端。第三,实验发现用 BW 权重可选出更多的大类点,所以用 BW 选初始点。在相同的初始点下,不同权重把与初始点相同类的评论提取出来的能力是不同的,用实验发现效果最好的 BW-DF 聚类。

隐藏长评论 - 最大距离法:即在应用最大距离法之前,先隐藏长度过长的评论。至于长度阈值,第一,一般各类评论条数的方差小的代表选择的初始点更多的为大类点,聚类效果好。第二,当长度阈值设的比较大时,初始点的评论长度较大,则一个初始点一般含有多重意思,从而造成聚类结果中各个类别具有多重含义,不利于提取主题。又由于单重意思的评论可以通过合并来形成多重意思。所以倾向于小值。所以选长度阈值的原则是:方差最小,倾向于小的阈值。

各类评论条数方差的拐点方法:第一,由表3可见随着聚类类别数 K 的增加,新增加的中心点会分最大类(一般为杂类)的评论,而其他类基本不变,从而使各类评论条数的方差减少。当增加的初始点不能有效地抽取最大类的评论时,方差变化不再明显,所以选择拐点。第二,大的 K 值可以捕获更多的大类点,所以倾向于大值。所以选 K 值的原则是:寻找拐点,倾向于大值。Rezaee 等根据经验规律认为最佳的聚类数应该在 2 与 $\sqrt{m}$ 之间,其中 m 为评论集的个数。所以在 2 与 $\sqrt{m}$ 之间寻找拐点。

表3 不同 K 值下的各类评论条数

聚类的 K 值	各类对应的评论个数									
	1	2	3	4	5	6	7	8	9	10
$K=9$	22	42	84	25	17	8	63	19	243	无这一类
$K=10$	22	42	84	25	17	8	63	19	223	20

改进 K-means 步骤为:①用各类评论条数方差的拐点方法取 K 值;②用隐藏长评论 - 最大距离法,在 BW 权重下找初始点;③用 BW-DF 权重函数聚类并得到聚类结果。改进 K-means 聚类算法伪代码如表4所示。

最后,相同主题可能有多种表达方式,它们可能分布在不同类上,由于欧氏距离的聚类结果精度较高,则可以把相同主题合并。然后用 python 提取各类的关键词得到基于聚类结果的评论主题和各主题比例。根据主题抽取的过程和改进算法,研究提出改进 K-means 聚类的新闻评论主题抽取模型如图1所示。

表 4　改进 K-means 聚类算法伪代码

Improved K-means clustering algorithm

K:聚类个数。D_i:评论集 D 中第 i 条评论。$D[BW]$:BW 表示的 D。

$D[BW\text{-}DF]$:BW-DF 表示的 D。S_k:第 k 类,$|S_k|$为其个数。

K = 方差拐点函数($D[BW]$,$D[BW\text{-}DF]$);

$c_1, c_2, \cdots c_K$ = 隐藏长评论_最大距离法函数($D[BW]$,K);

while$c_1, c_2, \cdots c_K \neq \overset{\circ}{c}_1, \overset{\circ}{c}_2, \cdots, \overset{\circ}{c}_K$(判断新旧质心是否相同){

for each D_i in $D[BW\text{-}DF]$){

if $Sim_{ip}(D_i, c_p) = \max_{p=1,2,\cdots K} Sim_{ip}(D_i, c_p)$ {

则将 D_i 划分到 p 类中;}}

for each c_k in $c_1, c_2, \cdots c_K$ {

$\overset{\circ}{c}_k = (\sum_{D \in S_k} D)/|S_k|$}}

return$\overset{\circ}{c}_1, \overset{\circ}{c}_2, \cdots, \overset{\circ}{c}_K$ 及各类的评论;

End

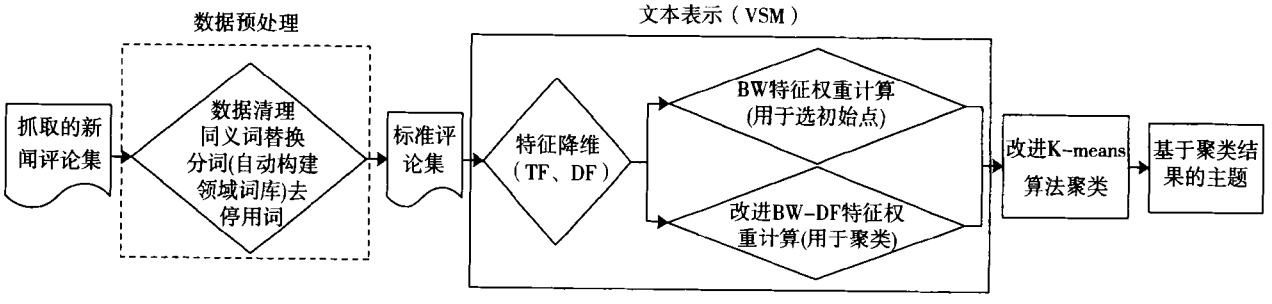

图 1　改进 K-means 聚类的新闻评论主题抽取模型

5　实验结果

5.1　抓取新闻评论文本并转化为结构化数据

相关操作用 python 语言,版本 2.7.10。这篇论文选取 2014 年 10 月 19 日新浪新闻“梁振英:外国势力向来参与香港政治占中也不例外”的 536 条评论。数据清理时,删去小于或等于 3 个字的和回复性质的评论,剩下 523 条。同义词替换时,把一些繁体评论转化为简体后,向原有同义词库(飞达鲁伪原创同义词库)加入网络用语同义词词典和口语化词典,还加入了针对本新闻的同义词集如:“梁 = 梁振英”等。然后进行同义词替换。如表 5 所示,同义词替换后,评论集的总词数减少,词频大于或等于 7 的词数及所有词相加数增加,改善了数据稀疏性和特征空间高维性。而且说明原来低频的同义词通过归并为同一个词而成为高频词。这样评论信息更多的集中于高频词,使得通过词频和文档频率的方法所选取的特征项代表更多的信息。

表 5　同义词替换前后对比

替换前后	所有评论总词数	词频大于或等于 7 的词数	词频大于或等于 7 的所有词相加
同义词替换前	1509	125	1931
同义词替换后	1450	138	2304

在分词时首先根据原新闻和原新闻的评论集,统计相邻的且同时出现次数大于 8 次的词并去除原分词词典已有的词,得到了 72 个词语,然后人工去除了一些毫无意义的词,获得 38 个词语。例如:缩写词“港政府”、“占中者”以及专有名词“梁振英”等。并加入分词领域词典。然后用 python 的结巴分词对评论集进行分词。在删去停用词时,选用四

川大学机器智能实验室停用词库。在特征降维时，将词频小于7的词删除，然后选取6≤文档频数DF≤(523/5)的词，得到134个词。在权重计算时分别用十种权重把评论集转化成结构化数据。以上操作涉及的参数大多根据经验值获取。

5.2 在欧式距离下用原 *K*-means 算法聚类

*K*值使用人工分类的类别数8，把随机选择第一条评论作为第一个初始点，用最大距离法确定其他初始点。初始点选择和聚类使用相同的权重，用十种权重分别聚类得出表6结果。横向表示聚的8个类，“字数”表示某类初始点对应那条评论的字数，“条数”表示某类含有的评论有几条。现在结合初始点的字数对评论条数分析。对于TF-IDF、TF-DF、TF-DF^{-2}、TF权重，由于权重和欧氏距离特性，初始点的字数很小，一来不利于提取有相同特征项的评论，二来增大了离群点的概率，由于欧式距离的聚类特性(各类与初始点的对应性大)，则除第一类由于程序设计原因评论条数多外，其它类的评论条数很小；对于其他6种权重，初始点的字数很大，一来使得评论不是根据意思聚类而是聚集到初始点长度小的的类中，二来增大了离群点的概率。若某类初始点选的好，含有评论数较多。

5.3 在欧式距离下用改进的 *K*-means 聚类

由本文的方法确定隐藏长评论的阈值为去停用词后5个词，发现初始点长度适中，意思明确。*K*取12。

(1)为什么用BW权重函数选择初始点

对于TF-IDF、TF-DF、TF-DF^{-2}、TF权重，选择的中心点的长度本身就很小，而且聚类效果不好，排除。对于BW、RTF、BW-DF、RTF-DF、BW-DF^{-2}、RTF-DF^{-2}中心点评论长度很长，则通过隐藏长评论，再使用最大距离法选择初始点，初始点长度适中，比较聚类结果，即可得出最优方法。图2为6种权重选初始点再分别用十种权重聚类得到结果的各类评论条数的方差。由图2得到用BW权重选初始值，再用十种权重聚类的方差比其他五种方法的情况基本上都小，这说明BW找的初始点更多的是大类点，这些大类点抽取了杂类的评论，使得各类条数的方差较小。另一条新闻验证了BW选初值效果好的鲁棒性。所以用BW权重选初值。其实，由于BW权重各特征项地位相同，不会造成不同初始点的特征项重合而影响聚类效果。由于其他5种权重增大了DF或RTF的影响，各特征项的重要性不再相同，使得DF或RTF大的特征词出现在多个初始点中，增加了各个初始点的相似度，影响了聚类效果。

(2)为何用BW-DF聚类

用BW选初始点，再比较十种权重聚类特性，即可得出最优权重。图3为BW权重选初始点再分别用十种权重聚类得到结果的各类评论条数的方差，其中*K*分别取8和12，并用原*K*-means方法对照。由BW选初值的两条线可见BW-DF权重聚类的方差较小，说明BW-DF更能把与初始点同类别的评论从最大类中抽出来。通过人工判断，结论也是BW-DF的聚类结果最好。

表6 初始中心点的字数统计及各类的评论条数

类别	TF-IDF		BW		RTF		TF-DF		BW-DF		RTF-DF		TF-DF^{-2}		BW-DF^{-2}		RTF-DF^{-2}		TF	
	字数	条数	字数	条数	字数	条数	字数	条数	字数	条数	字数	条数	字数	条数	字数	条数	字数	条数	字数	条数
1	10	500	10	514	10	512	10	426	10	326	10	361	10	403	10	382	10	357	10	447
2	5	1	240	1	163	2	6	1	33	2	70	1	6	50	34	7	69	1	6	6
3	6	6	133	1	237	1	5	6	60	10	170	1	6	6	133	1	163	2	5	1
4	8	4	7	3	91	4	5	5	30	67	31	5	5	1	236	70	89	4	7	1
5	9	7	85	1	69	1	7	1	7	42	90	73	7	30	90	3	24	12	6	36
6	5	2	75	1	75	1	5	53	33	15	33	8	7	24	60	3	49	8	5	5
7	7	1	76	1	49	1	3	24	29	53	59	15	5	2	43	45	85	71	9	21
8	9	2	163	1	55	1	7	7	60	8	87	59	7	7	75	12	59	68	8	6

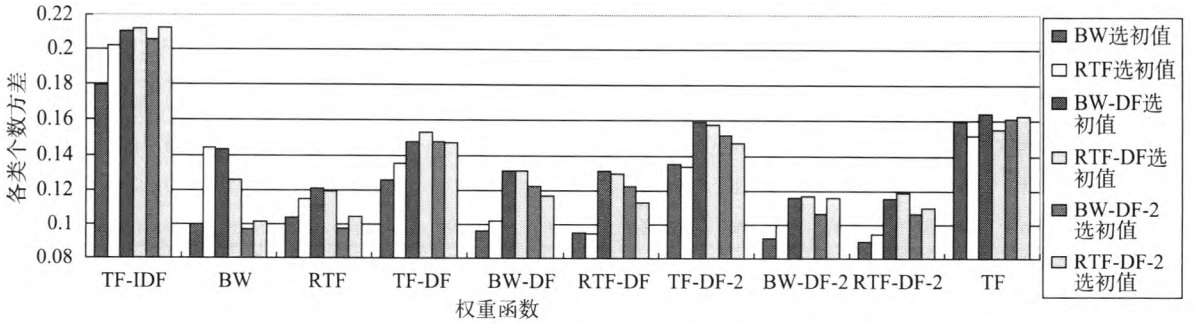

图 2　BW 选初始点与其他权重选初始点的方差比较

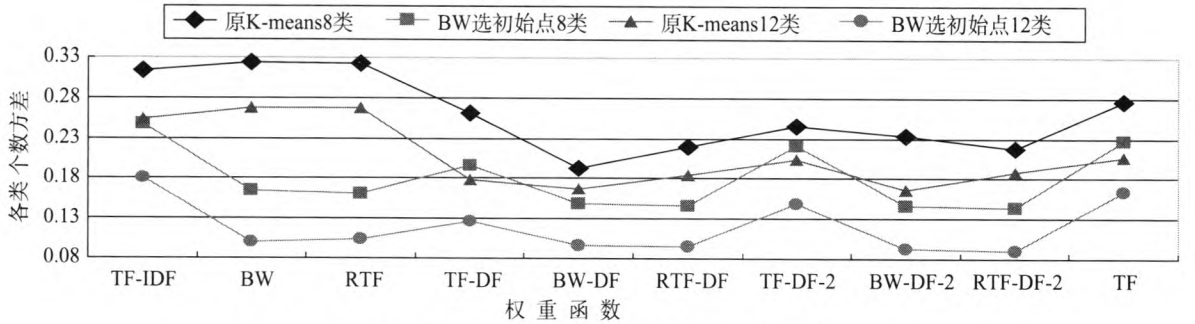

图 3　原方法和 BW 选初始点的方法方差比较

表 7　原 *K*-means 方法和 BW 选初始点方法的方差比较

两种方法	TF-IDF	BW	RTF	TF-DF	BW-DF	RTF-DF	TF-DF^{-2}	BW-DF^{-2}	RTF-DF^{-2}	TF
原 *K*-means	0.24	0.26	0.22	0.19	0.15	0.20	0.19	0.18	0.21	0.22
BW 选初值	0.21	0.17	0.17	0.16	0.13	0.14	0.17	0.15	0.14	0.19

为验证 BW-DF 聚类效果好的鲁棒性，用 2015 年 1 月 2 日新浪新闻“跨年夜英国街头路人醉态”这篇新闻共 473 条评论，由这篇文章提出的方法确定隐藏长评论的长度阈值为 6，$K=13$，在 BW 权重下用最大法距离选初始值，然后得到十种权重聚类的各类评论条数方差如表 7 所示，其中用原 *K*-means 方法对照。可见 BW-DF 权重函数聚类的方差最小。所以用 BW-DF 聚类效果较好。其实，由于 DF 大的词在大类点的概率较大，而 BW-DF 权重增加了 DF 大的词在相似度计算中的影响力，如果某个初始点有 DF 大的词，含有这个词的评论更易于聚类到这个初始点的类中。而 BW-DF^{-2}法中的 DF 由于取了平方根，DF 影响力降低，聚类效果比 BW-DF 差了一点。而 RTF-DF 和 RTF-DF^{-2}法由于 RTF 的影响而使得该方法的鲁棒性不强。

6　改进的 *K*-means 聚类算法与原 *K*-means 聚类算法结果比较

(1)改进算法与在欧式距离下用原 *K*-means 算法比较聚类效果：由图 3、表 7 可见 BW 选初值，BW-DF 聚类的方差比原算法的方差都小。说明改进方法改善了欧氏距离聚成一大类的情况，人工判断也是如此。

(2)在欧氏距离下的改进算法与在余弦距离下用原 *K*-means 算法聚类比较准确率：首先，根据意思把改进算法第 11、4 类合并为一类；第 1、3、10、12、7、6、5 类合并为一类；2、8、9 各为一类，得到改进算法的最后结果共 5 类。然后，在余弦距离下，用原 *K*-means 算法（K = 人工分类数 8，把第一条评论作为第一个初始点，初始点选择和聚类所用权重相同）在十种权重下聚类。为比较准确率，把评论人工分为 8 类，分别为“反对外国势力的扰乱和干预”55 条；“支持香港特区政府、特首、警察应对占中以保证社会秩序”238 条；“香港的稳定与繁荣来之不易，望珍惜”47 条；“从法律角度批评占中行为”21 条；“对占中者受蒙蔽的评价和规劝”36 条；“中国大陆、台湾、香港以及英美各方关系”95 条；“对占中事件应如何应对的建议”21 条；“离群点”10 条。由改进算法和余弦距离十种权重的聚类结果，得出准

表8 欧氏距离的新方法与余弦距离的准确率比较

各种准确率	TF-IDF	BW	RTF	TF-DF	BW-DF	RTF-DF	$TF\text{-}DF^{-2}$	$BW\text{-}DF^{-2}$	$RTF\text{-}DF^{-2}$	TF	改进的算法
PP	0.22	0.26	0.25	0.27	0.28	0.27	0.24	0.27	0.26	0.24	0.70
NP	0.88	0.88	0.88	0.86	0.86	0.86	0.88	0.88	0.88	0.88	0.80
PR	0.41	0.45	0.44	0.42	0.41	0.41	0.42	0.45	0.45	0.42	0.55
NR	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.88

表9 聚类类别在人工分类类别的分布

聚类各类及条数		聚类算法在人工类别中的百分比分布(%)							
		1类	2类	3类	4类	5类	6类	7类	8类
在欧氏距离下的改进算法	第1类共41条(100%)	93	2	0	5	0	0	0	0
	第2类共228条(100%)	5	86	1	1	3	3	1	1
	第3类共42条(100%)	5	0	95	0	0	0	0	0
	第4类共18条(100%)	0	6	0	0	94	0	0	0
	第5类共194条(100%)	2	21	3	9	6	46	9	4
余弦距离用$BW\text{-}DF^{-2}$选初值聚类	第1类共147条(100%)	1	48	1	3	12	26	5	4
	第2类共47条(100%)	4	28	2	9	6	34	15	2
	第3类共93条(100%)	3	92	0	0	1	2	1	0
	第4类共44条(100%)	52	9	0	9	9	11	9	0
	第5类共63条(100%)	10	67	3	2	10	5	2	3
	第6类共54条(100%)	31	6	7	6	7	39	2	2
	第7类共50条(100%)	4	2	76	2	2	14	0	0
	第8类共25条(100%)	0	76	0	12	0	12	0	0

确率与召回率如表8所示。可见改进的方法在PP,PR,NR比余弦距离高,NP相差不大。

由表8可见,在余弦聚类中,用$BW\text{-}DF^{-2}$选初值和聚类的准确率和召回率最高,于是用此方法的结果与改进算法比较聚类类别在人工分类类别分布的百分比情况如表9所示。可见改进算法前四类的聚类精度都达到86%以上。只有第五类聚类聚类精度低,为杂类(一般为合并前的最大类)。而余弦距离只有第3类达到86%以上,而其他的聚类效果除第7类和第8类以外,聚类精度均很低,因此也不能通过合并类别来提高积极准确率。由于改进的算法除了某个杂类的聚类精度低以外,其他各类在包含主要主题的前提下聚类精度仍很大,所以有利于迅速找出各类的主题以及各主题评论条数的比例关系,达到新闻评论主题抽取的目的。最后,由改进算法前4类可迅速得到人工分类中的第1、第2、第3、第5类,然后对第5类(杂类)进行二次聚类来找到人工的第4、第6、第7、第8类,这里就不再累述了。

7 讨论与结论

这篇论文针对新闻评论特殊的文本结构,提出了相应的预处理技术和改进的*K*-means算法,有效地发现了评价主题。在理论上,首先发现了*K*-means聚类的主题抽取算法能较好的应用到新闻评论中。其次,发现了与低维稠密向量的聚类特点不同,对于高维稀疏的评论向量,后续迭代反而会使各类精度下降,这为短文本的*K*-means聚类提供了两个改进方向:第一,首次迭代时的初始点选择;第二,后续迭代的相似度计算架构。实践上,该主题抽取模型可以应用到微博、贴吧、聊天记录、视频弹幕等短文本中,改进算法由于精度高而便于提取出主题和比例,有利于管理者和政策制定者利用评论情报信息进行决策[31]。

本文研究还有许多不足之处和需要进一步改进的地方:尽管本文研究中的数据随机选出,但如果能够选择更多新闻评论类型,将有助于验证本方法的通用性。而且,在对聚类效果不好的那一个杂类进行二次聚类时,会出现效果减弱的情况,需要对此进行进一步研究。

致 谢

感谢国家自然科学基金资助项目(71171153)"24小时知识工厂的知识共享活动模型与服务支持系统研究";感谢湖北省高等学校人文社会科学重

点研究基地－企业决策支持研究中心的支持。

参 考 文 献

[1] Abdul-Mageed M M. Online news sites and journalism 2.0: Reader comments on Al Jazeera Arabic[J]. tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society, 2008, 6(2): 59-76.

[2] 唐晓波,王洪艳.基于潜在狄利克雷分配模型的微博主题演化分析[J].情报学报,2013,32(3):281-287.

[3] Liu Q, Zhou M, Zhao X. Understanding News 2.0: A framework for explaining the number of comments from readers on online news[J]. Information & Management, 2015, 52(7): 764-776.

[4] Walther J B, DeAndrea D, Kim J, et al. The influence of online comments on perceptions of antimarijuana public service announcements on YouTube [J]. Human Communication Research, 2010, 36(4): 469-492.

[5] Houston J B, Hansen G J, Nisbett G S. Influence of user comments on perceptions of media bias and third-person effect in online news[J]. Electronic News, 2011, 5(2): 79-92.

[6] Saha S K. Person Specific Comment Extraction and Classification[D]. Jadavpur University Kolkata, 2012.

[7] Zhuang L, Jing F, Zhu X Y. Movie review mining and summarization [C]//Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006: 43-50.

[8] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. the Journal of Machine Learning Research, 2003, 3: 993-1022.

[9] 王卫平,孟翠翠.基于句法分析与依存分析的评价对象抽取[J].计算机系统应用,2011,20(8):52-57.

[10] 唐晓波,肖璐.基于单句粒度的微博主题挖掘研究[J].情报学报,2013,32(3):281-287.

[11] 姚天昉,程希文,徐飞玉,等.文本意见挖掘综述[J].中文信息学报,2008,22(3):71-79.

[12] 张晨逸,孙建伶,丁轶群.基于 MB-LDA 模型的微博主题挖掘[J].计算机研究与发展,2011,48(10):1795-1802.

[13] 唐晓波,王洪艳.基于潜在语义分析的微博主题挖掘模型研究[J].图书情报工作,2012,56(24):114-119.

[14] 唐晓波,房小可.基于文本聚类与 LDA 相融合的微博主题检索模型研究[J].情报理论与实践,2013,36(8):85-90.

[15] Titov I, McDonald R. Modeling online reviews with multi-grain topic models [C]//Proceedings of the 17th international conference on World Wide Web. ACM, 2008: 111-120.

[16] Jo Y, Oh A H. Aspect and sentiment unification model for online review analysis[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 815-824.

[17] 李芳,何婷婷,宋乐.评价主题挖掘及其倾向性识别[J].计算机科学,2012,39(6):159-162.

[18] Ma T, Wan X. Opinion target extraction in Chinese news comments[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 782-790.

[19] Tsagkias M, Weerkamp W, De Rijke M. Predicting the volume of comments on online news stories [C]// Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009: 1765-1768.

[20] 张立.基于新闻评论数据的 K-means 聚类算法的研究[D].太原:太原理工大学,2011.

[21] 沈幸峰.基于网络评论的话题挖掘[D].杭州:杭州电子科技大学,2013.

[22] 马莹莹.基于聚类的热点主题挖掘[D].西安:西安理工大学,2011.

[23] 赵辉,刘怀亮.面向用户生成内容的短文本聚类算法研究[J].现代图书情报技术,2013,(9):88-92.

[24] 朱晓峰,陈楚楚,尹婵娟.基于微博舆情监测的 K-means 算法改进研究[J].情报理论与实践,2014,37(1):136-140.

[25] Yang C C, Ng T D. Analyzing and visualizing web opinion development and social interactions with density-based clustering[J]. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2011, 41(6): 1144-1155.

[26] Cheung Y M. k*-Means: A new generalized K-means clustering algorithm [J]. Pattern Recognition Letters, 2003, 24(15): 2883-2893.

[27] 翟东海,鱼江,高飞.最大距离法选取初始簇中心的 K-means 文本聚类算法的研究[J].计算机应用研究,2014,31(3):713-719.

[28] Iwayama M, Tokunaga T. Hierarchical Bayesian clustering for automatic text classification[C]//Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2. Morgan Kaufmann Publishers Inc., 1995: 1322-1327.

[29] 邹娟,周经野,邓成,等.特征词提取中同义处理的新方法[J].中文信息学报,2005,19(6):44-49.

[30] 王旭仁,李娜,何发镁,等.基于改进聚类算法的网络舆情分析系统研究[J].情报学报,2014,33(5):530-537.

[31] 丁晟春,孟美任,李霄.面向中文微博的观点句识别研究[J].情报学报,2014,33(2):175-182.

(责任编辑 赵 康)