情报学报　2017 年 11 月　第 36 卷　第 11 期
Journal of the China Society for Scientific and Technical Information, Nov. 2017, 36(11): 1183-1191

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

吴　江，侯绍新，靳萌萌，胡忠义

（武汉大学信息管理学院，武汉　430072）

摘　要　随着互联网时代的快速发展，在线医疗社区的出现打破了时空限制，为用户提供了丰富的医疗信息和情感帮助，已经成为社会支持的重要来源，受到用户的广泛关注和参与。对在线医疗社区进行用户文本挖掘能够揭示社区中用户的参与行为，从而优化其用户管理和信息推荐。已有的研究对象主要集中在英文在线医疗社区，鲜有文献对中文在线医疗社区进行研究。基于社会支持理论，本文设计了一个中文用户文本挖掘流程来研究中文在线医疗社区中的社会支持类型和用户参与。利用中文文本挖掘及机器学习方法，对中文糖尿病社区“甜蜜家园”进行研究。本文利用 LDA（Latent Dirichlet Allocation）模型进行特征提取来构建低维度文本表示向量，采用二元分类法将用户文本分为不同的社会支持类型。最后，基于分类结果使用 *K*-means 算法进行用户聚类来识别用户角色。相比传统的特征提取方法，利用 LDA 进行特征提取能显著地降低数据维度，优化分类模型，提高分类准确率和分类效率。结果表明，本文提出的中文用户文本挖掘流程在文本分类与用户聚类中效果显著。

关键词　在线医疗社区；LDA 模型；特征提取；文本分类；用户聚类

LDA Feature Selection Based Text Classification and User Clustering in Chinese Online Health Community

Wu Jiang, Hou Shaoxin, Jin Mengmeng and Hu Zhongyi

(*School of Information and Management, Wuhan University, Wuhan*　430072)

Abstract: The emerging online health communities (OHCs) provide abundant medical information and emotional connection for users in today's rapidly developing Internet era, without the limitation of time and space. OHCs, which have been regarded as one of the major sources of social support, have become increasingly popular among people with health issues in China. The user text mining of OHCs can reveal a user's behavior, and hence can be used to optimize user management and information recommendation. Most studies used English OHCs as their research objects, while few focused on Chinese OHCs. Based on the social support theory, we designed a Chinese content analysis process to reveal the social support and user engagement in OHCs. Using a case study of an OHC among diabetics, we first extracted the features using an LDA model to construct low-dimensional text representation vectors, and then used binary classification to divide users' posts and replies into different types of social support. Finally, we used the *K*-means algorithm to cluster the users based on the classification results to identify user roles. Compared with the traditional vector space model, the LDA feature extraction can not only significantly reduce the data dimension and the amount of human annotation data, but also improve the classification accuracy and efficiency. Results showed that

收稿日期：2017-03-28；**修回日期**：2017-08-12

基金项目：国家自然科学基金项目“内容关系互动下的在线医疗社区用户行为演化研究”（71573197）。

作者简介：吴江，男，1978 年生，博士，教授，博士生导师，E-mail: jiangw@whu.edu.cn；侯绍新，男，1993 年生，硕士研究生，主要研究在线医疗社区文本挖掘；靳萌萌，女，1995 年生，硕士研究生，主要研究在线医疗社区及共享经济；胡忠义，男，1987 年生，博士，讲师，主要从事数据挖掘及数据分析。

the process performed well in text classification and user clustering.

Key words: OHCs; LDA model; feature extraction; text classification; user clustering

1 引 言

在线医疗社区是指能够打破时间地域限制、将患者和医生聚集在一起的互联网平台，能够最大程度地整合各种医疗资源，提高医疗服务水平，对于提高用户的健康意识和健康水平具有重大意义[1-2]。在线医疗社区的出现改变了人们获取健康信息以及疾病信息的方式[3]。近年来，越来越多的人使用网络来满足自身的健康需求[4]。与传统健康相关网站只允许用户获取信息的方式不同，在线医疗社区允许具有相同疾病或治疗经历的用户进行及时交流，满足他们迫切的信息或情感需求。

在线医疗社区能够为参与者提供有效的社会支持[2;5-6]，并对用户的健康行为产生积极的影响[7]。在线医疗社区主要包含三种社会支持类型：信息支持、情感支持以及陪伴[8-9]。不同社会支持类型是否会影响用户社区参与行为？Wang 等[5]依据社会支持相关理论，利用文本挖掘方法对英文乳腺癌患者社区 Breastcancer.org 进行用户文本分类，研究用户文本的社会支持类型，并根据用户文本类型进行用户聚类来研究不同群体的用户特征。在线医疗社区文本属于专业性较强、复杂度高的不规则短文本，这些特点增加了机器学习文本分析难度，进而导致在线医疗社区文本分析的主要方法是内容分析法，通过对社区上大量的文本进行人工标注和统计分析来发现用户行为。Rodgers 等[10]对某个乳腺癌医疗社区的文本信息进行内容分析后发现该社区的患者在社区中得到自己需要的信息并且与其他患者进行情感交流以及和其他患者相互鼓励后，负面情绪明显改善，更加有信心来接受接下来的治疗。现有研究的不足包括以下两个方面：①研究对象主要是国外在线医疗社区中的英文文本，而对于中文文本挖掘的研究较少。由于中文与英文在语义、语法等方面存在差异，无法完全采用国外的研究方法进行中文文本挖掘研究。②主要采用人工分类的内容分析法，虽然准确率较高，但是时间成本和人工成本大，并且可能存在人为因素的影响造成处理结果存在差异。

针对以上两点不足，本文提出了适用于中文在线医疗社区的文本挖掘方法，并采用机器学习方法来代替人工分类，减少了时间和人工成本，提高了工作效率。结果表明，利用本文提出的基于 LDA 模型进行特征提取的方法能显著降低数据表示维度及减少人工标注数据量，同时提高了分类准确率。利用最大距离原则和肘部法则改进聚类算法能够显著提高聚类效果，不同聚类中用户特征显著。

2 相关方法研究

文本挖掘[11]（Text Mining，TM），又称为文本数据挖掘（Text Data Mining, TDM）或文本知识发现（Knowledge Discovery in Texts, KDT），是指为了发现知识，从大规模文本库中抽取隐含的、以前未知的、潜在有用的知识的过程。从 Feldman 等[12]在 1995 年首次提出文本挖掘概念至今，文本挖掘在国外特别是拉丁语系国家发展迅速。

文本数据表示对于文本分析至关重要，数据表示质量直接影响分析结果。通过分析用户文本，Wang 等[5]提取了 4 个维度 19 个特征作为文本分类特征，包括基本特征、词汇特征、情感特征和主题特征。这种特征提取方法不仅极大程度地保留了文本信息，同时降低了数据表示维度和人工标注数据数量。然而，在汉语文本处理领域，这种方法并不适应汉语本身特点。汉语是一种语义型语言，重“意合”，轻形式，而且语形、语法、语义等各层面歧义现象非常严重[13]。同时，这种特征提取方法还必须有基础资源的支持，包括英语词库、义类词典、领域专业词典、语义语法规则库、常识知识库等。相比于国外，国内对这些支持资源的建设还不够完善。

传统的文本数据通常基于向量空间模型（Vector Space Model，VSM），利用特征词和权值构成向量表示[14]。向量空间模型得到的特征向量的维数往往会达到数十万维，过高的维度不仅会影响文本分析效果，同时会大大增加机器学习的时间以及人工标注的数据量。为了降低特征维数，Hotho 等[15]和 Pinto 等[16]分别引入 WordNet、MeSH，实现了文本特征提取；Banerjee 等[17]采用搜索引擎引入了维基百科外部语料库。这种特征提取方式往往受外部语料库限制，尤其是一些专业性较强的领域很难通过外部语料库进行特征提取。随着概率主题模型在文本特征提取中的成功应用，将概率主题模型应用到文本挖掘成为一个新的趋势。目前主要的主题模型有 LSI、PLSI 以及 LDA。在主题模型中 LDA 是一个完全概率生成模型，获得广泛应用。研究者大多直接将 LDA

应用到文本分类中，其中，Rubin 等[18]在文档分类中建立了实现一种多标签分类的主题模型，并系统地比较了统计主题模型和判别模型技术的优劣。另一方面，研究者也开始采用主题模型对文本特征进行扩展。其中，Phan 等[19]在基于外部语料库的文本特征扩展中采用了隐主题，Chen 等[20]在短文本分类中通过多粒度主题模型（multi-granularity topics）实现了文本特征扩展和基于不同标签的最优主题选择技术。

本文以中文糖尿病社区“甜蜜家园”为研究对象，结合汉语文本分析的特点，针对在线医疗社区进行了用户文本分类及用户聚类，主要创新包括：①改进特征提取方式，利用 LDA 模型进行主题聚类，从中提取出主题特征词库，利用主题特征词库构建数据的低维特征向量。②采用二元分类法，针对每个社会支持类型利用不同的分类算法建立分类器，从中选择最佳分类算法用于构建分类器。③优化聚类算法，利用最大距离原则选择聚类中心以及利用肘部法则确定最佳聚类数。

3　数据分析流程及关键技术

为了实现在线医疗社区的用户文本分类及用户聚类，本文设计了一个基于 LDA 模型特征提取的文本分析流程。图 1 详细描述了该流程，主要包括数据采集与数据预处理、利用 LDA 进行特征提取、构建分类模型进行文本分类、特征提取效果评价、构建聚类模型及用户聚类分析 5 个部分，各部分流程与涉及的关键技术描述如下。

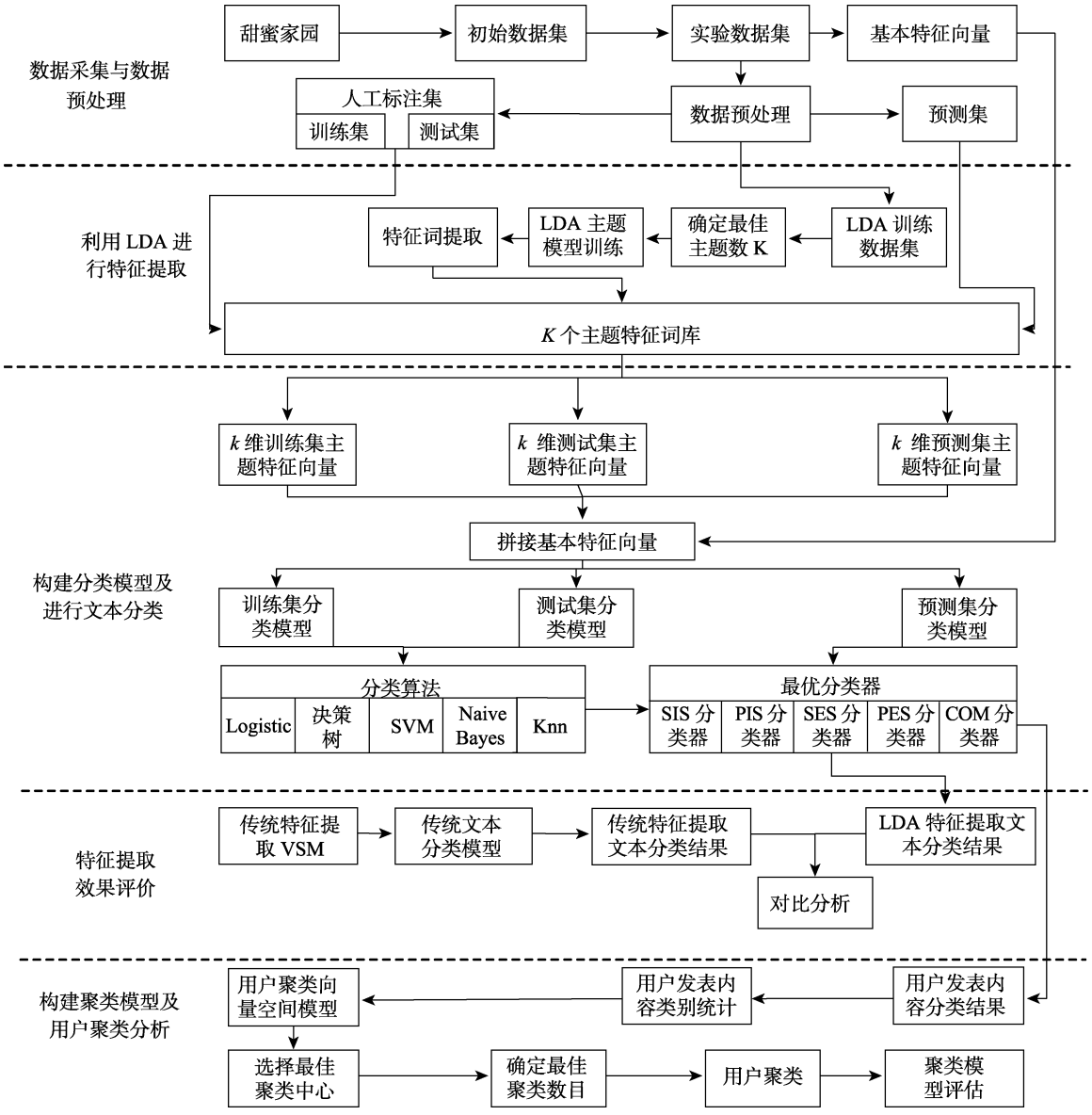

图 1　数据分析流程图

3.1 数据采集及预处理

本文使用的数据来源于糖尿病在线医疗社区“甜蜜家园”。我们基于 Python 语言设计了多线程爬虫工具，采集了该论坛上用户发帖或回帖文本内容，形成初始的实验数据集；对初始实验数据集进行清理、筛选和抽取，得到实验数据集。

文本的一些基本特征对文本的社会支持类型分类具有重要的作用，寻求帮助的文本长度一般要小于提供帮助的文本长度，包含“!”的文本感情色彩更加强烈，包含“？”的文本属于寻求信息帮助的概率更大，包含“url”的文本多数属于提供信息帮助[5]。因此，本文对实验数据集进行基本特征的提取，并形成基本特征向量。

将实验数据集拆分成训练集、测试集、预测集及 LDA 模型训练集 4 个数据集，将糖尿病专有名词作为用户字典以及常用中文停词表作为去停词表加入分词包，利用 JiebaR 分词[21]对 4 个数据集进行分词处理，最终完成数据采集和数据预处理工作。

3.2 基于 LDA 模型的特征提取

LDA 是一种文档主题生成模型，也称为三层贝叶斯概率模型，包含词、主题和文档三层结构[22]。其核心思想是文档可以表示为一系列潜在主题的随机混合，其中每一个主题都代表了在所对应的文档集中全部词的概率分布，与潜在主题相关的词的概率分布较高。

在使用数据集训练 LDA 模型时，为了得到最佳的主题模型，需要确定最佳的主题数 k。Blei 等[22]在其研究中详细阐述了使用困惑度（Perplexity）作为确定主题数的标准，通过绘制 Perplexity-topic number 曲线，寻找最佳主题数。

困惑度是一种信息理论的测量方法，可以理解为，对于一篇文档 d，我们的模型对文档 d 属于哪个 topic 有多不确定，这个不确定程度就是 Perplexity。其他条件固定的情况下，topic 越多，则 Perplexity 越小，但是容易过拟合。

使用 LDA 训练数据集进行 LDA 模型的训练。通过 Perplexity-topic number 曲线确定最佳的主题数 k，然后对数据集进行主题聚类。主题词与主题的相关性通过公式（1）计算得出：

$$\text{relevance}(\text{term } w \mid \text{topic } t) = \lambda \times p(w \mid t) + (1-\lambda) \times p(w \mid t) / p(w) \quad (1)$$

式中，$p(w \mid t)$ 表示单词 w 在主题 t 中的概率，λ 是可调节参数。如果 λ 接近 1，那么在该主题下更频繁出现的词，主题相关性更高；如果 λ 越接近 0，那么该主题下更特殊、更独有的词，主题相关性更高。为了确保提取的主题特征词彼此之间独立以及主题下的单词能更好地突出该主题的特征，本文将 λ 调节为 0 来计算主题-单词相关性。对每个主题下的词根据相关性进行排序，取前 300 个词作为该主题特征的特征词。最终，形成 k 个主题特征词库，完成特征提取。

3.3 用户文本分类

文本分类属于有监督的学习算法，需要使用训练集来训练模型，使用测试集来评价模型质量，以及使用构建好的模型进行预测。

1）构建文本分类模型

首先，使用主题特征词库对训练集、测试集和预测集进行特征提取，形成三个 k 维的主题特征向量，结合基本特征向量和人工标注结果形成分类模型。模型中除评论长度这一特征外都是用 0,1 表示，因此需要将评论长度这一维度进行数据转换。转换的原则是根据分位数，0~25%设置为 0，25%~50%设置为 1，50%~75%设置为 2，75%~100%设置为 3。最终形成用于分类的基本特征与主题特征结合的文本分类模型。

2）最佳分类算法选择

根据社会支持的理论[8-9]，本文将在线医疗社区用户文本分为 5 类，分别为寻求信息支持（SIS）、提供信息支持（PIS）、寻求情感支持（SES）、提供情感支持（PES）以及陪伴（COM），分类示例如表 1 所示。从表 1 中可以看出，一个文本可以同时属于多个类型，如文本“饮食 运动 主要是运动，你的情况和正常人一样，要是吃了不动，以后可真的悲催了，少吃糖，没事的”同时属于 PIS 和 PES。如果使用一个多分类器进行分类，上述一个文本属于多个分类的真实情况就会被掩盖。因此，本文使用二元分类法，为每一个社会支持类型的判别建立一个 0~1 分类器，最后的分类结果是一个长度为 5 的数组，数组中每个元素的取值为 0 或 1。如某个文本的分类结果为[1,0,0,0,1]，即第一个分类器（SIS 分类器）和第五个分类器（COM 分类器）判定该文本属于这两类，五个分类器之间彼此独立。

不同分类算法的分类效果存在差异，为了找到最佳分类算法来构建分类器，分别采用 Logistic 算法、决策树算法、朴素贝叶斯算法、KNN 算法以及

表 1　文本分类示例

社会支持类型	文本
SIS	**1. 吃了一年的二甲双胍目前记忆力明显下降。这是怎么回事？我才 28 岁，这样发展下去没几年还不先变成傻子了？！** 2. “带泵 20 天左右就发现，部位有时吸收好，有时不好，”请问为什么会吸收不好啊？才 20 天就这样？ 3. 是不是将要从冰箱里拿出来没有等到温度和室内平衡直接着就灌了？
PIS	1. 的确是高了，看来光靠二甲是不够了，让医生调下。 **2. 饮食 运动 主要是运动，你的情况和正常人一样，要是吃了不动，以后可真的悲催了，少吃糖，没事的** 3. 河南省：省级医院：102 元/天、市级医院：87 元/天、县级医院：71 元/天
SES	**1. 吃了一年的二甲双胍目前记忆力明显下降。这是怎么回事？我才 28 岁，这样发展下去没几年还不先变成傻子了？！** **2. 我结婚半年，查出糖前，想要宝宝，家里姥爷是，很瘦，所以不好控制，心里也很难受！我才 25 岁，但日子还要过！心情好点，一起努力！** 3. 病情在进一步加重，我该何去何从
PES	**1. 饮食 运动 主要是运动，你的情况和正常人一样，要是吃了不动，以后可真的悲催了，少吃糖，没事的** **2. 我结婚半年，查出糖前，想要宝宝，家里姥爷是，很瘦，所以不好控制，心里也很难受！我才 25 岁，但日子还要过！心情好点，一起努力！** 3. 32 岁了都，你还有什么郁闷的，很多几岁甚至几个月的孩子得了又能怎么样呢？照样打胰岛素，你知足吧，你要感到庆幸
COM	1. 有同感，我们可是安徽同乡啊，欢迎多多交流！ 2. 回复 2#半糖葫芦多谢多谢 3. 今天早上堵管啦，我用盐水冲开因为针头还没钝继续使用。

不同内核函数的SVM算法为每个社会支持类型的判别建立分类器。然后，使用准确率和 F1 值作为模型评价标准，从中为每一个社会支持类型找到最佳的分类算法进行建模。准确率和 F1 计算公式如下：

$$\text{accuracy}=(\text{TP}+\text{FN})/(\text{TP}+\text{FN}+\text{FP}+\text{TN}) \quad (2)$$

$$\text{F1}=2\times P\times R/(P+R) \quad (3)$$

$$\text{Precision}=\text{TP}/(\text{TP}+\text{FP}) \quad (4)$$

$$\text{Recall}=\text{TP}/(\text{TP}+\text{FN}) \quad (5)$$

其中，TP 是正确的肯定（true positive），FP 是错误的肯定（false positive），TN 是正确的否定（true negative），FN 是错误的否定（false negative）。准确率是对于给定的测试数据集，分类器正确分类的样本数与总样本数之比。Precise 和 Recall 分别是精确率和召回率，精确率是所有被检索到的（TP+FP）中，“应该被检索到的（TP）”占的比例，召回率是所有检索到的（TP）占所有“应该被检索到的（TP+FN）”的比例。从公式（4）和公式（5）可以看出，精确率和召回率一般是成反比的关系，F1 综合两个评价指标，可以对分类器进行整体评价[23]。

3）使用分类模型进行分类

使用最佳分类算法为每一个社会支持类型建立分类器，使用训练集训练分类器构建分类模型。然后，使用训练好的模型对预测集进行预测，将预测结果写回原文件，和用户名一一对应。

使用传统特征提取方式构建用户文本分类模型，同样根据二元分类原则对每个类型利用最佳分类算法构建二分类器，利用构建的分类器对用户文本进行分类。通过对比两种不同分类方式的准确率来评价本文提出的特征提取方式。

3.4　特征提取效果评价

使用传统特征提取方式构建用户文本分类模型，同样根据二元分类原则对每个类型利用最佳分类算法构建二分类器，利用构建的分类器对用户文本进行分类。通过对比两种不同分类方式的准确率来评价本文提出的特征提取方式。

3.5　用户聚类

1）用户聚类模型

将分类结果按用户分组，统计用户的不同类型文本数量，形成用户文本类型组成向量，如（uname,中 10,4,1,8,5），“uname”是用户的注册名，“10,4,1,8,5”分别表示“SIS”、“PIS”、“SES”、“PES”和“COM”5 种社会支持类型的数量。为了消除不同用户文本数量的差异，对用户向量进行比例转换，最终形成如

（uname,0.2,0.3,0.1,0.4,0）的用户聚类向量。

2）*K*-means 聚类算法原理

K-means算法的基本思想是以空间中*k*个点为中心进行聚类，对最靠近它们的对象归类，通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果[24]。*K*-means 的参数是类的中心位置和其内部观测值的位置。与广义线性模型和决策树类似，*K*-means 参数的最优解也是以成本函数最小化为目标。*K*-means 成本函数公式如下：

$$J=\sum_{k=1}^{k}\sum_{i\in C_k}\left|x_i-u_k\right|^2 \qquad (6)$$

其中，u_k 是第 k 个类的中心位置。成本函数是各个类畸变程度（distortions）之和。每个类的畸变程度 J 等于该类中心与其内部成员位置距离的平方和。求解成本函数最小化的参数就是一个重复配置每个类包含的观测值、并不断移动类中心的过程。

3）确定最佳聚类数 k

聚类数对聚类效果具有至关重要的影响，本文通过肘部法则[25]来确定最优聚类数。利用 *K*-means 算法的成本函数，通过计算不同 k 值下的成本函数值，绘出 *k*-平均畸变程度（成本函数值/*k*）曲线。曲线中，随着 k 值的增大，平均畸变程度会减小，每个类包含的样本数会减少，于是样本离其重心会更近。但是，随着 k 值继续增大，平均畸变程度的改善效果会不断降低。k 值增大过程中，畸变程度的改善效果下降幅度最大的位置对应的 k 值就是肘部。

4）选择最佳聚类中心

K-means 的初始中心位置是随机选择的。如果选择的中心点彼此相距不远，这种随机选择的中心会导致 *K*-means 陷入局部最优解。本文采用距离最大化原则，利用公式（7）来选择初始中心位置：

$$\text{distance}_i=\sum_{j=1}^{j<K}\left|x_i-u_j\right|^2 \qquad (7)$$

其中，x_i 表示数据 i 的位置，u_j 表示第 j 个中心的位置，k 表示聚类数，distance_i 表示第 i 个数据与第 j 个中心的距离和。首先，从数据集中随机选取一个数据作为第一个初始中心位置，然后利用公式（7）寻找与第一个中心距离最大的点作为第二个中心，再寻找第三个点与前两个中心距离和最大的点作为第三个中心，以此类推，直到找到 k 个初始聚类中心。

5）聚类效果评价

K-means 是一种非监督学习，没有标签和其他信息来比较聚类结果。可以使用轮廓系数（Silhouette Coefficient）来评价聚类效果。轮廓系数是类的密集与分散程度的评价指标。它会随着类的规模的增大而增大。彼此相距很远、本身很密集的类，其轮廓系数较大；彼此集中、本身很大的类，其轮廓系数较小[26]。

4 实验结果

4.1 实验数据

经过对“甜蜜家园”数据的采集和预处理，得到的实验数据包含 1356417 条文本数据，共涉及 39675 名用户。其中主题文本数量为 31094，回复文本数量为 1325323。利用分层抽样的方法将实验数据分为三个数据集，数据组成如表 2 所示。

表 2 数据组成

	回复文本	主题文本	总计
训练集	1349	953	2302
测试集	319	141	460
预测集	1323655	30000	1353655
总计	1325323	31094	1356417

4.2 基于 LDA 模型特征选择结果

为了确定最佳主题数，本文计算了不同 k 值对应的困惑度，困惑度-主题数曲线如图 2 所示。从图中可以明显看出，随着主题数增加，困惑度随之减小，并且困惑度下降速度变慢。当主题数为 50 时，困惑度趋于平稳，继续增加主题数所得到的收益要小于增加主题数的投入。因此，确定最佳主题数为 50。

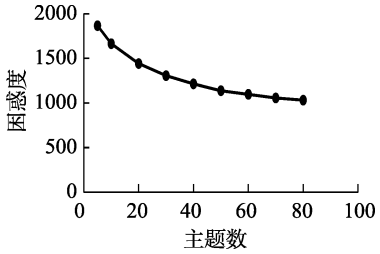

图 2 困惑度-主题数曲线

将数据集和 k=50 输入 LDA 模型中进行主题聚类，将 λ 设置为 0，利用公式（1）计算主题词与该主题的相关性，根据相关性排序从每个主题中选出 300 个最能代表该主题特征的词语[27]，最终形成 50 个主题特征词库，完成特征提取。

4.3 用户文本分类结果

不同社会支持类型的算法表现如表 3 所示。其

中，SVM 表示支持向量机算法，SVM_1 指其内核函数为线性函数，SVM_2 内核函数为多项式函数，SVM_3 内核函数为径向基函数，SVM_4 内核函数为 Sigmoid 函数[28]；Logistic 算法[29]是一种广义的线性回归分析算法；KNN 算法[30]是 K 最近邻算法，根据 K 个最近邻居来预测本身；Bayesian 算法[31]是一种利用概率统计知识进行分类的算法；Tree 算法[32]是在已知各种情况发生的概率基础上，通过构成决策树来进行预测的分类算法。

从表 3 中可以看出，不同的社会支持类型最佳分类算法不同，利用准确率和 F1 值来确定最佳分类算法："PIS"、"PES"、"SES" 和 "COM" 类型的最佳分类算法都是 Logistic 算法，该算法的 F1 值要优于其他算法，准确率方面也较优秀；对于 "SIS" 类型，最佳分类算法为线性核函数的 SVM_1，其准确率和 F1 值都大于其他算法。不同类型最优算法如表 4 所示。对不同的社会支持类型使用最佳分类算法构建分类器对预测集进行预测，预测结果如表 5 所示。

表 3　算法评价表

算法	评价指标	SIS	PIS	SES	PES	COM
Logistic	准确率	0.8711	0.7781	0.9811	0.9611	0.7004
	F1	0.7232	0.7195	0.7698	0.7948	0.7934
SVM_1	准确率	0.8753	0.7768	0.9826	0.9579	0.7004
	F1	0.7679	0.7145	0.7648	0.7878	0.7933
SVM_2	准确率	0.7818	0.7152	0.9826	0.9579	0.5414
	F1	0.4713	0.5440	0.6635	0.7259	0.6895
SVM_4	准确率	0.8505	0.7690	0.9826	0.9579	0.7124
	F1	0.7487	0.6769	0.7414	0.7730	0.7918
SVM_4	准确率	0.8497	0.7469	0.9826	0.9579	0.7078
	F1	0.7443	0.6399	0.7195	0.7593	0.7831
KNN	准确率	0.8247	0.7321	0.9833	0.9590	0.6622
	F1	0.6789	0.6738	0.7384	0.7722	0.7898
Bayesian	准确率	0.8666	0.7753	0.9805	0.9525	0.6961
	F1	0.7384	0.7174	0.7661	0.7939	0.7950
Tree	准确率	0.8228	0.7026	0.9599	0.9236	0.7000
	F1	0.7399	0.7144	0.7635	0.7862	0.7777

表 4　最优算法

类型	SIS	PIS	SES	PES	COM
算法	SVM_1	Logistic	Logistic	Logistic	Logistic
准确率	0.8753	0.7781	0.9811	0.9611	0.7004

表 5　预测结果表

社会支持类型	SIS	PIS	SES	PES	COM
数量	98277	279973	9383	23441	804329

从表 5 中可以看出，"COM" 类别的文本最多，"PES" 类别的文本最少。在线医疗社区中，用户主要目的是寻求信息以及分享信息，而倾述情感以及给予安慰和鼓励表现不是那么强烈；在信息和情感之外，用户大部分处于社区中的时间都是和其他用户以相互陪伴的方式度过[5]。

4.4　特征提取效果评价

利用传统特征提取方式进行分类，不同类型对应的最优算法以及准确率如表 6 所示。

表 6　传统特征提取方式分类评价表

类型	SIS	PIS	SES	PES	COM
算法	SVM_1	SVM_1	Logistic	Logistic	KNN
准确率	0.712	0.632	0.601	0.877	0.534

对比表 4 和表 6 可以得出结论，相比于传统的特征提取方式，本文提出的基于 LDA 模型特征提取的文本分类结果准确率更高。同时，传统的 VSM 模型维度高达 75422，通过降维保留了 540 个特征；利用本文提出的特征提取方式得到的分类模型只有 50 个主题特征和 5 个基本特征，较低的维数提升了分类效率。

4.5　用户聚类结果

为了确定最佳聚类数 k，本文分别计算了聚类数为 1~9 的平均畸变程度，k-平均畸变程度（成本函数值/k）曲线如图 3 所示。

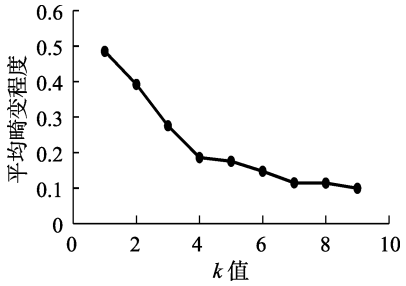

图 3　k-平均畸变程度曲线

从图 3 可以明显看出，k=4 是畸变程度的改善效果下降幅度最大的位置。根据肘部法则原理，k=4 即是该曲线的肘部，也就是本文需要的最佳聚类数。利用距离最大化原则确定的聚类中心如表 7 所示。

第一个聚类中心是随机选择数据集中序号=11833 的用户向量，根据距离最大化原则，依次选取 Index=411246 的用户向量为聚类中心。利用最佳聚类数和

最佳聚类中心进行用户聚类，得到的用户聚类结果如表 8 所示。

表 7 聚类中心表

序号	聚类中心	欧氏距离
11833	[1,0,0,0,0]	0
4	[0,0,0,0,1]	1.41
11	[0,1,0,0,0]	2.83
246	[0,0,0,1,0]	4.23

表 8 聚类结果

	SIS	PIS	SES	PES	COM	Total
Cluster1	0.7874	0.0577	0.0008	0.0018	0.1520	6447
Cluster2	0.0155	0.0190	0.0032	0.0062	0.9558	15250
Cluster3	0.0163	0.9255	0.0004	0.0037	0.0538	4153
Cluster4	0.2226	0.2624	0.0669	0.0872	0.3606	13825

表 8 中第 2~6 列分别表示各分类文本在每一个聚类中的占比，Total 表示该聚类中用户的数量。从表中可以看出，Cluster1 用户“SIS”类评论的比例远远大于其他类别，该聚类用户的发表内容多数是为了寻求信息支持。根据这一特点，该类用户可以归纳为“信息需求者”。Cluster3 用户的“PIS”类发表内容比例为 92.55%，该类用户以分享信息、提供信息帮助为主。因此，该类用户可以归纳为“信息分享者”。Cluster2 用户发表内容几乎全部是“COM”类，该类用户经常在社区中活动，他们区别于信息驱动型用户，以分享生活、娱乐聊天以及交友陪伴为主，这类用户可以归纳为“社区陪伴者”。Cluster4 用户发表内容相比前三类用户比较均衡，情感类“SES”、“PES”比例明显高于前三类，信息类“SIS”、“PIS”比例相比前三类较为均衡，这类用户在社区中没有明确的目的，以“散步”的心态参与社区交流，这类用户可以归纳为“社区散步者”。

使用平均轮廓系数来评价聚类模型效果，得到的 *k*-平均轮廓系数曲线如图 4 所示。从图中可以看出，当 *k*=4 时，轮廓系数最大，说明 *k*=4 时聚类模型效果最好。

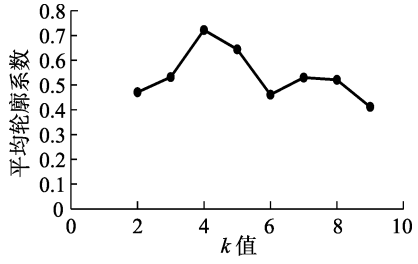

图 4 *k*-平均轮廓系数曲线

5 结 语

在线医疗社区以其方便快捷、无地域差异及高度整合医疗资源的优势，广泛成为人们进行自我健康管理的平台。本文总结前人对在线医疗社区的研究理论，从在线医疗社区社会支持的角度切入，利用汉语文本挖掘以及机器学习的方法对用户文本进行分类，并在分类结果的基础上对用户进行聚类分析。本文在总结 LDA 在文本分类领域的研究基础上，提出了基于 LDA 模型进行主题特征词库提取的方法。相比于传统方法，该方法能显著减少人工标注数据量，同时在提高分类准确率和分类效率上效果显著。在用户聚类中，本文使用最大距离确定聚类中心以及利用肘部法则确定聚类数，优化了聚类效果，聚类结果中不同聚类用户特征显著。利用本文的研究方法，可以帮助在线医疗社区对用户进行精准划分，实现对用户的精细管理。同时，在此基础上还可以对用户社群行为进行分析以及对用户进行个性化的信息推荐服务等。

参 考 文 献

[1] van der Eijk M, Faber M J, Aarts J W, et al. Using online health communities to deliver patient-centered care to people with chronic conditions[J]. Journal of Medical Internet Research, 2013, 15(6): e115.

[2] Wang X, Zuo Z Y, Zhao K. The evolution of user roles in online health communities – a social support perspective[C]// Proceedings of Pacific Asia Conference on Information Systems. IEEE Computer Society, 2015: 48-56.

[3] Ziebland S, Chapple A, Dumelow C, et al. How the Internet affects patients' experience of cancer: a qualitative study[J]. BMJ, 2004, 328(7439): 564.

[4] Guarino L, Scremín F, Borrás S. The social life of health information[J]. Pew Internet, 2009, 1(1): 13-21.

[5] Wang X, Zhao K, Street N. Social support and user engagement in online health communities[C]// Proceedings of International Conference on Smart Health. Cham: Springer, 2014: 97-110.

[6] Beaudoin C E, Tao C C. The impact of online cancer resources on the supporters of cancer patients[J]. New Media & Society, 2008, 10(2): 321-344.

[7] Ba S, Wang L. Digital health communities: The effect of their motivation mechanisms[J]. Decision Support Systems, 2013, 55(4): 941-947.

[8] Keating D M. Spirituality and support: a descriptive analysis of online social support for depression[J]. Journal of Religion and Health, 2013, 52(3): 1014-1028.

[9] Bambina A. Online social support : the interplay of social net-

works and computer-mediated communication[M]. Youngstown: Cambria Press, 2007.

[10] Rodgers S, Chen Q M. Internet community group participation: psychosocial benefits for women with breast cancer[J]. Journal of Computer-Mediated Communication, 2005, 10(4).

[11] 谌志群, 张国煊. 文本挖掘研究进展[J]. 模式识别与人工智能, 2005, 18(1): 65-74.

[12] Feldman R, Dagan I. Knowledge discovery in textual databases (KDT)[C]// Proceedings of the First International Conference on Knowledge Discovery and Data Mining. Palo Alto: AAAI Press, 1995: 112-117.

[13] 鲁川. 汉语语法的意合网络[M]. 商务印书馆, 2001.

[14] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[15] Hotho A, Staab S, Stumme G. Ontologies improve text document clustering[C]// Proceedings of the IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2003: 541-544.

[16] Pinto D, Rosso P, Benajiba Y, et al. Word sense induction in the arabic language: A self-term expansion based approach[C]// Proceedings of the 7th Conference on Language Engineering of the Egyptian Society of Language Engineering, 2007: 235-245.

[17] Banerjee S, Ramanathan K, Gupta A. Clustering short texts using wikipedia[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2007: 787-788.

[18] Rubin T N, Chambers A, Smyth P, et al. Statistical topic models for multi-label document classification[J]. Machine Learning, 2012, 88(1-2): 157-208.

[19] Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]// Proceedings of the 17th International Conference on World Wide Web. New York, 2008: 91-100.

[20] Chen M G, Jin X M, Shen D. Short text classification improved by learning multi-granularity topics[C]// Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press. 2011: 1776-1781.

[21] Wu Z M, Tseng G. Chinese text segmentation for text retrieval: Achievements and problems[J]. Journal of the American Society for Information Science., 1993, 44(9): 532-542.

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

[23] Ueno M. Data mining and text mining technologies for collaborative learning in an ILMS "*Samurai*"[C]// Proceedings of the IEEE International Conference on Advanced Learning Technologies. Washington DC: IEEE Computer Society, 2004: 1052-1053.

[24] MacKay D. An example inference task: clustering[M]// Information Theory, Inference and Learning Algorithms. Cambridge: Cambridge University Press, 2003: 284-292.

[25] 杨善林, 李永森, 胡笑旋, 等. *K*-means 算法中的 *k* 值优化问题研究[J]. 系统工程理论与实践, 2006, 26(2): 97-101.

[26] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics, 1987, 20: 53-65.

[27] Sievert C, Shirley K E. LDAvis: A method for visualizing and interpreting topics[C]// Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Association for Computational Linguistics, 2014: 63-70.

[28] Osuna E, Freund R, Girosi F. Training support vector machines: an application to face detection[C]// Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 1997: 130-136.

[29] Steyerberg E W, Harrell Jr F E, Borsboom G J J M, et al. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis[J]. Journal of Clinical Epidemiology, 2001, 54(8): 774-781.

[30] Guo G D, Wang H, Bell D, et al. KNN model-based approach in classification[C]// Proceedings of the Conference: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Heidelberg: Springer, 2003, 2888: 986-996.

[31] Ronquist F, Huelsenbeck J P. MrBayes 3: Bayesian phylogenetic inference under mixed models[J]. Bioinformatics, 2003, 19(12): 1572-1574.

[32] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

（责任编辑　马　兰）