在线健康社区信息需求主题分析*

唐晓波[1] 李津[2]
(1.武汉大学信息资源研究中心，武汉 430072；2. 武汉大学信息管理学院，武汉 430072)

摘要：以在线健康社区中高血压问答文本为例，对其信息需求进行主题识别和分析。采用聚类方法分析文本，发现用户最关心疾病的治疗、并发症和生活方式，而且随着时间的推移，并发症和生活方式的关注度有上升趋势。高血压并发症对患者生活造成了较大的影响，同时患者更加注重通过健康生活方式对疾病进行管理。

关键词：在线健康社区；信息需求；主题识别；文本挖掘
中图分类号：G203
DOI：10.3772/j.issn.1673-2286.2019.02.002

随着居民生活水平的提高，健康管理意识的增强，以及互联网的快速发展，人们通过网络获取健康信息的需求越来越大。据统计，截至2016年12月，我国互联网医疗用户规模为1.95亿人，大量用户通过网络平台获取医疗健康的相关服务[1]。对这些在线健康社区的信息进行分析研究，发掘用户的健康信息需求，可以提高社区健康信息服务的质量，促进网络社区平台的建设和发展。《中国心血管病报告》指出，2012年全国18岁及以上成人的高血压患病率为25.2%，根据第六次全国人口普查数据，测算中国高血压患病人数约为2.7亿人[2]。高血压是目前最常见的慢性病，我国每年约200万人的死亡与高血压有关，该病已成为重要公共卫生问题[3]。本文选取在线健康社区中用户的高血压问答为例，利用文本挖掘方法对其进行聚类分析，提取主题，通过对比不同时间段主题分布的变化，了解用户需求的特征和变化趋势，为健康信息服务提供参考。

1 相关研究

1.1 在线健康社区研究现状

在线健康社区是一个包含信息、用户和社区3个要素的复杂系统。信息是用户在社区中反映自身需求、认知和情感的记录；用户是在线健康社区的参与者，不断地产生、搜索、获取和使用健康信息；社区是为用户提供线上信息交流的平台[4]。研究者主要从信息、用户和社区三个维度展开对在线健康社区的研究。

从信息维度，研究者聚焦于在线健康社区中信息的主要内容，对信息的主题和情感进行研究，挖掘社区用户的健康信息需求。Roberts等[5]对美国国立医学图书馆和罕见病信息中心网站上的提问信息进行分析，得到病因、诊断、并发症、临床表现等共13个类别的主题；邓胜利等[6]以百度知道中高血压提问记录作为研究对象，利用文本挖掘软件分析发现用户更关心日常疾病管理、疾病确诊和治疗，并希望在社区中获得情感支持。从用户维度，研究者主要探讨在线健康社区中用户健康信息的获取、搜索和共享等行为。Wong等[7]调查发现在15岁以上的患者中，使用互联网和搜索健康信息频率与年龄呈反比，具社会经济优势的患者在网上获取健康信息的可能性明显高于弱势群体，但是患者的性别、英语水平和地理位置不影响他们搜索健康信息；张克永等[8]构建了网络健康社区用户知识共享的影响因素模型，调查发现自我效能、利他主义、社会信任等因素与用户知识共享行为呈现显著正相关。从社区维度，研究主要集中在社区的价值、运行模式和发展现状等方面。Lee等[9]发现在搜索健康信息时，谷歌导航并不

*本研究得到国家自然科学基金项目“基于文本和Web语义分析的智能咨询服务研究”（编号：71673209）资助。

能满足用户的需求，解决社区网页设计不足的问题非常必要；杨化龙等[10]在分析薄荷网用户的相关数据后，发现社区中用户获得的社会支持和个人目标都对用户的健康有积极影响，且对男性和女性用户影响程度不同，建议在线健康社区的设计者对不同性别的用户开发不同的主页和系统。

1.2 在线健康社区信息需求研究现状

早期对在线健康社区信息需求的研究大多采用问卷调查或者访谈的方式，以社区的用户作为调查对象，统计他们在社区中讨论的热点话题。Armstrong等[11]通过对糖尿病患者的访谈，了解他们在在线健康社区中讨论的热点话题。但这些方法经常受到样本数量的限制以及问卷设计等因素的影响，导致结果具有一定的局限性，难以从整体上反映用户的信息需求。随着在线健康社区的快速发展，用户在社区上发布大量信息，有些研究者开始通过对这些信息文本进行深入分析，以此来反映在线健康社区的用户信息需求。最初研究者普遍采用基于统计分析和人工标注的方法。Zhang[12]统计分析雅虎问答社区的糖尿病患者的问答记录，发现了糖尿病患者关心的12类健康主题；金碧漪等[13]选取Yahoo!Answers网站和Diabetic Connect论坛中糖尿病相关文本，采用人工编码和文本处理等方法，得到8类主题，对比两种网络社区的主题分布情况，大体趋于一致，但在诊断和检查、社会生活主题上各有侧重。郭海红等[14]对寻医问药网站的高血压相关问句进行了人工标注，得到包含诊断、治疗、病情管理、流行病学、健康生活、择医及其他共7个一级主题类目。

近年来，随着自然语言处理研究的快速发展，基于主题识别和文本挖掘的方法也逐渐应用到在线健康社区信息需求分析的研究中。Chen[15]采用K-means方法对3个网络健康社区的发帖文本进行聚类分析，发现不同社区热点主题不同，同时也有如患者经验、治疗、药物和身体管理等相同主题；吕英杰[16]采用EM聚类方法对Medhelp网站的发帖进行主题分析、成员角色分析和情感分析，最终定义了个人详细介绍、情感支持、症状、检查、并发症、用药和治疗共7个热点主题；李重阳等[17]结合LDA和人工标注方法，对百度知道的癌症问答进行分析，发现用户对癌症信息的需求集中在基础病理知识、疾病预防、诊断检查、治疗和其他共5个主题，且各个主题的关注随时间变化而变化。

综上所述，在线健康社区信息需求分析早期采用问卷调查或访谈的方法，两种方法会受到样本数量等因素的影响，难以客观全面地反映在线健康社区的热点话题。以社区中实际发布的文本信息作为研究对象，数据更加真实可信，但依靠人工编码的方式需要消耗大量的人力和时间成本，LDA方法在文本的语义层面也有欠缺。本文以在线健康网站中用户提出的高血压相关问题以及医生回答中的最佳答案作为研究对象，抽取并融合文本的词语特征和词权重特征，采用K-means++方法对文本聚类，提取关键词识别主题并进行分析。

2 研究方案

本文研究方案如图1所示，包括数据采集与预处理、特征抽取与融合、主题识别。

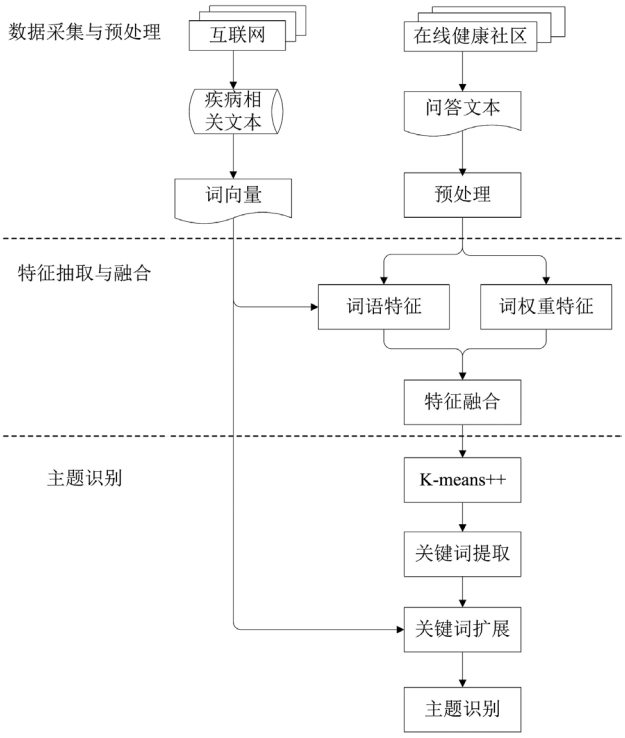

图1 研究方案

2.1 数据采集与预处理

本文利用python爬取在线健康社区中高血压问答文本、高血压常用药品名称及高血压相关文章，搜集常见疾病名称、症状、体征和临床表现等。所有相关文本作为语料库训练词向量，同时高血压常用药品数据用

于构建常用药品词典，疾病名称、症状、体征和临床表现用于构建疾病相关词典。利用jieba对文本进行预处理（包括分词和去停用词），过程中使用了常用药品词典、疾病相关词典和停用词表。药品词典和疾病相关词典用于句子分词时药品名和医学专有词不被划分开，以保证药品和医学词汇表达得完整；停用词表则用于消除句子中没有意义的词，本文以哈尔滨工业大学中文停用词表为基础并作部分修改。

2.2 特征抽取与融合

（1）词语特征。最早由Hinton[18]提出将词映射成多维向量，通过词语向量间的余弦来判断词之间的距离。词向量广泛应用于自然语言处理，它可以很好地表达词语的语义以及词之间存在的相似关系。本文使用Google发布的word2vec词向量计算工具训练词向量，选择skip-gram模型，采用negative sampling训练算法，词向量维度为200，训练窗口为5。

（2）词权重特征。本文使用TF-IDF方法计算文本中词语的权重，该方法主要思想是当某个词语在一篇文档中出现的频率高，且在其他文档中很少出现，则认为该词语具有很好的类别区分能力[19]。其中，TF（Term Frequency）指词频，计算词语在文档中出现的频率，见公式（1）；IDF（Inverse Document Frequency）指逆向文件频率，反应词语在所有文档中出现的频率，见公式（2）；TF-IDF实际是指TF和IDF的乘积，见公式（3）。

$$\mathrm{tf}(\mathrm{t},\ \mathrm{d}) = \frac{f(t,\ d)}{\sum_k f(w_k,\ d)} \tag{1}$$

$$\mathrm{idf}(\mathrm{t},\ \mathrm{d}) = \lg \frac{|D|}{|d \in D : t \in d|} \tag{2}$$

$$\mathrm{tfidf}(\mathrm{t},\ \mathrm{d}) = \mathrm{tf}(\mathrm{t},\ \mathrm{d}) \times \mathrm{idf}(\mathrm{t},\ \mathrm{d}) \tag{3}$$

公式（1）中，$f(t, d)$表示词语t在文档d中出现的次数，$\sum_k f(w_k, d)$表示文档d中所有词语出现的次数之和。公式（2）中，$|D|$表示文档集中文档的总个数，$|d \in D : t \in d|$表示文档集中包含词语t的文档数量，为避免分母为0的情况，一般使用$1+|d \in D : t \in d|$。

（3）特征融合。词语特征揭示了词间的语义关系，词权重特征反映了词语的重要程度，将文档的词语特征和词权重特征采用特征相乘的方式进行融合。文档d的向量可由文档中k个词语的词语特征和词权重特征乘积之和表示，见公式（4）。

$$V_d = \sum_k \mathrm{tfidf}(\mathrm{t},\ \mathrm{d}) \times \mathrm{word2vec}(\mathrm{t}) \tag{4}$$

其中，tfidf（t，d）表示文档d中词语t的TF-IDF值，word2vec（t）表示词语t的词向量。

2.3 主题识别

首先利用K-means++算法对问答文本聚类，然后对每个类别分别提取关键词并扩展关键词，最后识别主题。K-means++算法是基于原始K-means算法，具体算法过程如下[20]。

①从问答文档集中随机选取一个文档作为初始聚类中心c_1；②计算每个文档与当前已有聚类中心之间的最短距离，用$D_{(x)}$表示，接着计算每个文档被选为下一个聚类中心的概率 $\frac{D_{(x)}^{\ 2}}{\sum_{x \in X} D_{(x)}^{\ 2}}$，最后按照轮盘法选择出下一个聚类中心；③重复第2步至选择出k个聚类中心$C = \{c_1, c_2, \cdots, c_k\}$；④计算其他文档到各个聚类中心的距离，并将它们分配到最近的簇中；⑤针对每个类别，重新计算簇的中心$C_i = \frac{1}{C_i} \sum_{x \in C_i} x$；⑥重复第4、第5步至聚类中心位置不再变化。

通过文本聚类后，每个文档分配到距离最近的簇中，即对应一个类别；每个类别包含多个文档，即对应一个文档集。对于这些文档集，我们无法直观地看出每个类别的主题。利用TF-IDF算法，计算多个文档集中词语的权重，选择权重高的词语作为该类别的关键词，然后利用词向量余弦相似度对关键词进行扩展，最后基于关键词识别主题。

3 在线健康社区信息需求结果分析

3.1 数据基本情况

本文利用python爬取39健康网站高血压相关问答对共14 507条、高血压常用药品439种、高血压相关文章共42 396篇。搜集到39健康网站中主要科室的常见标签515个，包含常见疾病名称和症状等。所有文本作为语料库训练词向量，同时高血压常用药品数据用于构建常用药品词典，疾病名称和症状等数据用于构建疾病相关词典。39健康网站问医生专栏用户以患者和医生为主，由用户提问，具有行医资格的医生在线回答。本文以健康网站的问题和最佳答案作为研究对象，问答文本部分示例见表1。

表1 高血压问答文本示例

问 题	最佳答案
我有高血压，一直在服用氨氯地平片降压，听说硝苯地平缓释片也可以降压，想问一下氨氯地平片和硝苯地平缓释片的区别是什么	患者要对症选药，氨氯地平片用于高血压、心绞痛，能显著降低高血压患者肾血管阻力，保护肾脏功能；硝苯地平缓释片可抑制心肌收缩，降低心肌代谢，也可达到降压效果
患有高血压差不多十年，近来出现心慌的现象，想知道这些症状是不是心脏出现毛病	高血压属于慢性疾病，是引发心脑血管疾病的最危险因素；高血压病人出现胸闷，心悸，气短，胸痛的症状，考虑是合并了心脏病；需要做心电图和心脏彩超检查，确诊以后及早治疗
高血压平时吃什么饭菜比较好，在饮食上需要注意什么	饮食注意低盐饮食，忌油腻食物，少吃肉类食品，多吃水果蔬菜，忌烟限酒

根据网站用户发布问题的时间，将2014—2018年问答文本分为5组，数量分布如图2所示，数据量总体呈逐年上升趋势。

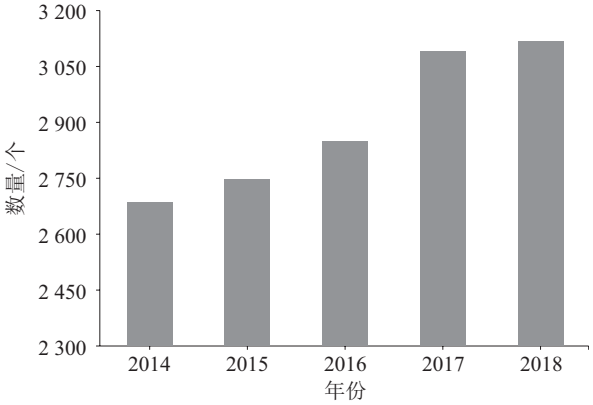

图2 问答文本数量分布

3.2 信息需求主题特征

利用python中sklearn包对文本聚类，计算不同类别个数的误差平方和SSE，并据此确定聚类个数6大类，每个大类再分别重新计算SSE确定小类共16个，依据各个类别提取出来的关键词合并相似类别，最终得到8个子类目，4个主题，主题分布如表2和图3所示。治疗类相关问答记录最多（36.28%），其次是并发症（24.61%）和生活类（23.89%），诊断类最少（15.22%）。

诊断类问答文本最少，说明用户对高血压的诊断标准比较熟悉，高血压以收缩压超过140毫米汞柱和舒张压超过90毫米汞柱作为主要标准，可能伴有头晕呕吐等临床表现。治疗类问答文本最多，高血压治疗以药物治疗为主，西药问答记录远高于中药，用户对西药的服用方法和副作用关注度较高，治疗高血压常用的西药是硝苯地平和氨氯地平等，中药以三七、天麻粉等作为辅助治疗。用户对高血压的并发症关注度较高，并发症以心脑血管疾病、肾脏疾病和眼部疾病为主。高血压是目前最常见的慢性病，疾病所带来的并发症严重影响患者的健康生活。网站上生活类问答文本也较多，说明用户日常生活中自我管理的意识较强，在药物治疗的基础上，通过清淡饮食、增强锻炼和调整心态等方法控制高血压。

表2 信息需求主题及其分布情况

主 题	子类目	关键词	数量/个	占 比
诊断	-	眩晕，呕吐，医院，血压计，测量，检查，舒张压，收缩压，偏高，超过，毫米汞柱	2 208	15.22%
治疗	中药	三七，天麻粉，丹参，灵芝，山楂，泡水	752	5.18%
	西药	硝苯地平，氨氯地平，缬沙坦，一片，缓释片，副作用，依那普利，倍他乐克	4 511	31.10%
并发症	心脑血管	心脏，供血不足，血管，脑部，脑血管，头晕，动脉硬化，冠心病，左心室	1 208	8.33%
	肾脏疾病	肾上腺，肾脏，肾功能，继发性，尿蛋白，肾功能不全，肾衰竭，合并	925	6.38%
	眼部疾病	出血，眼睛，看不清，重影，眼球，充血	1 437	9.90%
生活	饮食	饮食，食物，低盐，低脂，清淡，蔬菜，水果，维生素，芹菜，苦瓜，烟酒，少油	2 081	14.34%
	运动及其他	运动，锻炼，体重，休息，跑步，熬夜，情绪，焦虑	1 385	9.55%

3.3 信息需求主题变化

2014—2018年，39健康网站用户对高血压健康信息的需求呈现一定的变化，见图4。

诊断类信息的需求呈现下降趋势，高血压作为最常见的慢性病，已逐步被大众所认知和熟悉，用户对高血压的诊断标准更加了解。治疗类信息需求一直很高，但也有下降的趋势，与此同时生活类信息需求呈现上

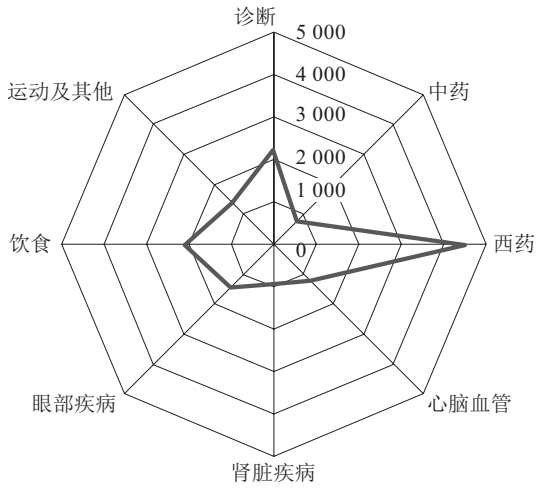

图3 信息需求主题分布

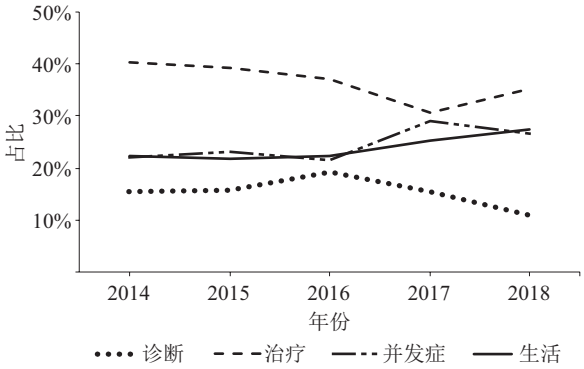

图4 2014—2018年信息主题分布

升趋势。目前高血压的治疗方法比较成熟和完善，需要患者长期服药控制和对自我生活的管理。随着健康意识和知识的增强，患者对高血压的治疗也有更全面和清晰的认识，在药物治疗的基础上，保持健康的生活方式。关于高血压并发症类的信息需求呈现上升趋势，高血压并发症对患者生活的影响日益明显，用户对并发症的危害也越来越重视，并积极地学习相关知识。

4 总结

在线健康社区用户信息需求主题主要包括诊断、治疗、并发症和生活，其中治疗的关注度最高。2014—2018年，用户对诊断和治疗的关注度有下降趋势，对并发症和生活方式的关注度有上升趋势。这说明用户对高血压的基本知识有了一定的了解和掌握，更关心高血压并发症可能带来的更严重伤害；同时在日常生活中注意健康饮食等自我管理和控制，健康意识在逐步提升。

本文采用文本挖掘技术对在线健康社区信息需求展开了主题识别和分析，探讨了其中的现象和原因，为在线健康社区信息服务提供参考。在之后的研究中，还可以从以下两个方面进行改进：①其他疾病的热点主题与高血压可能有差异，有待继续研究；②采用更前沿的技术对文本做深层次、细粒度的挖掘和分析。

参考文献

[1] CNNIC发布第39次《中国互联网络发展状况统计报告》[J]. 中国信息安全，2017(2)：24.

[2] 陈伟伟，王文，隋辉，等. 《中国心血管病报告2016》要点解读 [J]. 中华高血压杂志，2017(7)：605-608.

[3] 黄岚，吕江，王晓慧，等. 基于百度知道平台的网络高血压相关信息现状调查 [J]. 安徽医学，2016(1)：97-100.

[4] 赵栋祥. 国内在线健康社区研究现状综述 [J]. 图书情报工作，2018(9)：134-142.

[5] ROBERTS K，KILICOGLU H，FISZMAN M，et al. Automatically classifying question types for consumer health questions [C]. AMIA Annu Symp Proc，2014：1018-1027.

[6] 邓胜利，刘瑾. 基于文本挖掘的问答社区健康信息行为研究——以“百度知道”为例 [J]. 信息资源管理学报，2016(3)：25-33.

[7] WONG C，HARRISON C，BRITT H，et al. Patient use of the internet for health information [J]. Australian Family Physician，2014，43(12)：875-877.

[8] 张克永，李贺. 网络健康社区知识共享的影响因素研究 [J]. 图书情报工作，2017(5)：109-116.

[9] LEE K，HOTI K，HUGHES J D，et al. Dr Google and the consumer：A qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions [J]. Journal of Medical Internet Research，2014，16(12)：e262.

[10] 杨化龙，鞠晓峰. 社会支持与个人目标对健康状况的影响 [J]. 管理科学，2017(1)：53-61.

[11] ARMSTRONG N，POWELL J. Patient perspectives on health advice posted on internet discussion boards：A qualitative study [J]. Health Expectations，2010，12(3)：313-320.

[12] ZHANG Y. Toward a layered model of context for health information searching：An analysis of consumer-generated questions [J]. Journal of the American Society for Information Science and Technology，2013，64(6)：1158-1172.

[13] 金碧漪，许鑫. 网络健康社区中的主题特征研究 [J]. 图书情报工作，2015(12)：100-105.

[14] 郭海红，李姣，代涛. 中文健康问句分类与语料构建 [J]. 情报工程，2016（6）：39-49.

[15] CHEN A T. Exploring online support spaces：Using cluster analysis to examine breast cancer，diabetes and fibromyalgia support groups [J]. Patient Education & Counseling，2012，87（2）：250-257.

[16] 吕英杰. 网络健康社区中的文本挖掘方法研究 [D]. 上海：上海交通大学，2013.

[17] 李重阳，翟姗姗，郑路. 网络健康社区信息需求特征测度——基于时间和主题视角的实证分析 [J]. 数字图书馆论坛，2016（9）：34-42.

[18] HINTON G E. Learning distributed representations of concepts [C]//Proceedings of the 8th Annual Conference of the Cognitive Science Society. 1986：112-123

[19] 李锐，张谦，刘嘉勇. 基于加权word2vec的微博情感分析 [J]. 通信技术，2017（3）：502-506.

[20] 巩如悦. 基于本体的苹果病虫害垂直搜索引擎研发 [D]. 杨凌：西北农林科技大学，2017.

作者简介

唐晓波，男，1962年生，博士，教授，博士生导师，研究方向：知识组织与情报研究。
李津，女，1994年生，硕士，研究方向：知识组织与情报研究，E-mail：2016201040078@whu.edu.cn。

Analysis on the Topic and Sentiment of Information Needs in Online Health Community

TANG XiaoBo[1] LI Jin[2]
(1. Center for Studies of Information Resources, Wuhan University, Wuhan 430072, China;
2. School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: Taking the text of hypertension-related questions and answers from online health community as an example, this paper explores its topics and sentiment. Using clustering method, result shows that users are more concerned about treatment, complication and lifestyle of the disease. And with time going by, the attention to complication and lifestyle is rising. Hypertension complications have a major impact on the lives of patients. And patients are more focused on managing diseases through healthy lifestyles.

Keywords: Online Health Community; Information Needs; Topic Identification; Text Mining

（收稿日期：2019-01-18）