软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(4):1143−1161 [doi: 10.13328/j.cnki.jos.005987]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

中文软件问答社区主题分析研究*

蒋 竞, 吕江枫, 张 莉

(北京航空航天大学 计算机学院,北京 100191)

通讯作者: 张莉, E-mail: lily@buaa.edu.cn

摘 要: 软件问答社区是软件开发者通过问答方式进行技术交流的网络平台.近年来,软件问答社区积累了大量用户讨论的技术问答内容.一些研究者对 Stack Overflow 等英文问答社区进行主题分析研究,但是缺少对于中文软件问答社区的分析.通过对中文软件回答社区开展主题分析研究,不仅可以指导开发者更好地了解技术动向,而且可以帮助管理者改进社区、吸引更多用户参与.“开源中国”是中国最大的技术社区之一.对“开源中国”开展了开发者问题主题分析研究.收集“开源中国”的 92 383 个开发者问题,采用隐狄利克雷分配模型的主题分析方法,分析开发者问题的主题分布、热度趋势、回答情况和关键技术热度等.发现:(1) 开发者讨论的技术主题分为前端开发、后端开发、数据库、操作系统、通用技术和其他 6 个类别.其中,前端开发讨论占比最大.(2) 后端开发下的主题中用户的关注重点从传统的项目部署、服务器配置转移到较新的分布式系统等主题.(3) 数据展示主题的零回答问题比例最高,数据类型主题下的零回答问题比例最低.(4) 在技术学习主题下,用户对于 Java 的讨论明显多于对 Python 的讨论.

关键词: 软件问答社区;主题模型;经验研究;隐狄利克雷分配模型;开源中国

中图法分类号: TP311

中文引用格式: 蒋竞,吕江枫,张莉.中文软件问答社区主题分析研究.软件学报,2020,31(4):1143−1161. http://www.jos.org.cn/1000-9825/5987.htm

英文引用格式: Jiang J, Lü JF, Zhang L. Topic analysis on Chinese programming question and answer websites. Ruan Jian Xue Bao/Journal of Software, 2020,31(4):1143−1161 (in Chinese). http://www.jos.org.cn/1000-9825/5987.htm

Topic Analysis on Chinese Programming Question and Answer Websites

JIANG Jing, LÜ Jiang-Feng, ZHANG Li

(School of Computer Science and Engineering, Beihang University, Beijing 100191)

Abstract: Programming question and answer website is a network platform where software developers can exchange technical knowledge by posting and answering questions. With the development of Internet and growth in the number of software developers, programming question and answer websites accumulate extensive discussion contents of software engineering knowledge. Researchers have applied topic analysis on English question and answer websites in recent years, yet there are few similar studies on Chinese programming question and answer websites. Analyzing these contents can help developers know more about the trends of techniques. It also benefits website administrator to improve the forum for better user experience, etc. This study applies latent Dirichlet allocation (LDA) to automatically cluster the main topics in 92 383 questions on OSCHINA. Then, several analyses are applied to these topics, including trend analysis, difficulty analysis, and keyword analysis. Several findings are as follow: (1) Topics concluded from user discussion can be divided into 6 categories, including front-end development, back-end development, databases, operating systems, general techniques, and others. Within those categories, front-end development contains the most question posts. (2) Using trend analysis, it is found that in back-end development, developers are paying more attention to more up-to-date and advanced topics (distributed systems, system design & Web interfaces) rather than basic topics (project deployment, server configuration). (3) It is also found that data

* 基金项目: 国家重点研发计划(2018YFB1004202); 国家自然科学基金(61672078)

Foundation item: National Key Research and Development Program of China (2018YFB1004202); National Natural Science Foundation of China (61672078)

收稿时间: 2019-07-21; 修改时间: 2019-10-09; 采用时间: 2019-12-02

presentation is the most difficult topic, as it has the highest ratio of questions which are never answered while its popularity is above average. (4) The trend of different specific techniques is analyzed in one topic. For instance, the popularity of Java in the technique learning topic is obviously higher than the popularity of Python.

Key words: programming question and answer websites; topic model; empirical study; latent Dirichlet allocation; OSCHINA

随着软件开发领域的蓬勃发展,新的技术、工具、平台和编程语言不断出现.Stack Overflow、开源中国等软件问答社区积累了大量用户对于这些技术内容的讨论.近年来,一些学者对 Stack Overflow 开发人员问答内容进行主题分析[1−5],了解开发者关心的技术内容及其发展趋势.现有工作主要研究 Stack Overflow 等英文软件问答社区[1−5],缺少对于中文软件问答社区的主题分析,不清楚中国开发者关注的问答内容.

开源中国是中国最大的技术社区之一.用户可以在开源中国上查找、发表各种开源软件,参与技术问答.通过对开源中国问答社区进行主题分析,了解中国开发者关心的技术内容及其发展趋势,具有重要的研究意义.通过分析软件问答社区主要讨论的技术主题,可以帮助开发者更好地了解技术动向,有助于他们在开发过程中对技术、平台、工具的选择.此外,问答社区分析结果帮助社区管理者了解开发者关注的技术内容,采用热点推荐等方式吸引更多的用户参与社区讨论.

本文对开源中国开展了开发者问题主题分析研究.本文收集开源中国的 92 383 个开发者问题,然后采用隐狄利克雷分配模型(LDA)的主题分析方法,获取用户问答中讨论的主要主题,并对这些主题进行了热度趋势分析、回答情况分析和关键技术热度趋势分析.本文发现:(1) 开发者在开源中国问答社区主要讨论的主题涉及范围很广,可以被归类为前端开发、后端开发、数据库、操作系统、通用技术和其他 6 个类别.其中,前端开发讨论占比最大.(2) 后端开发下的主题中用户的关注重点从传统的项目部署、服务器配置转移到较新的分布式系统等主题.(3) 数据展示主题的平均浏览数较高,但是零回答问题比例最高;数据类型主题的平均浏览数较低,但是零回答问题比例最低.(4) 分析了同一主题下,不同技术手段、技术工具的使用热度对比.例如,技术学习主题下用户对于 Java 的讨论明显多于对 Python 的讨论.本文将实验数据集以及主题分析的实时结果在网页上进行展示,以供研究者参考(http://developertopic.cn/).

本文第 1 节分别对主题模型和软件问答社区的相关研究现状进行总结.第 2 节介绍本文研究方法,分别说明每一个研究问题、实验整体设计、数据的获取和预处理以及主题提取的方法与实现.第 3 节分别针对每一个研究问题,提出有针对性的分析方法和结论.第 4 节对重要结论、启示进行总结,并进行有效性分析.第 5 节对全文的工作内容进行总结.

1 相关研究

1.1 关于主题模型的研究

主题模型是一种通过无监督学习的方式,对文本中的一些隐含语义结构进行聚类的统计模型[6].主题模型的概念最早由潜在语义索引(latent semantic indexing,简称 LSI)[6]提出,之后由 Hoffmann 等人[7]提出 pLSI (probabilistic latent semantic indexing),形成较为完善的主题模型.2003 年,Blei 等人[8]提出隐狄利克雷分配模型(latent Dirichlet allocation,简称 LDA),LDA 能在无标注文集中提取出文档-主题的概率分布和主题-词的概率分布.从而知道这个文集中每一个文档所属的主题和每一个主题所包含的主题词.这一主题模型被大量应用于自然语言处理问题之中,如文本分类问题[9,10]、源代码分析问题[11]、主题发现与提取问题[1−5,12]以及标签推荐、更正问题等[13,14].

1.2 软件问答社区主题分析的经验研究

国外有很多成熟的关于软件开发问答社区上主题的挖掘和分析研究.这些研究的整体思路采用经验研究,经验研究是一种直接或间接地通过者实验获得知识的方法.经验研究在软件工程领域已经得到广泛的应用并备受关注[15].基于主题模型(topic model)的研究主要通过 LDA[8]对用户问题进行主题提取,分析单一问题下或所有问题的主题及分类.如 Barua 等人[1]通过对 Stack Overflow 上所有问题和回答数据进行 LDA 分析,之后对

于得到的主题进行时间趋势分析以及内容中关键技术影响力的分析.该论文以 Stack Overflow 上的主要讨论内容和这些内容的变化趋势等作为结论.Yang 等人[2]通过对所有 Stack Overflow 上的安全问题进行 LDA 分析得出用户关注的安全方面的主题,并将这些主题分为手机安全、网页安全、软件安全、系统安全与加密五大类.之后对得到的主题进行了难易度分析.Wang 等人[3]也使用了 LDA 分析了 Stack Overflow 上的所有分类下的问题,将这些问题分为用户界面、报错、大型代码段分析、网络文档和其他 5 个大种类,之后分析了软件问答社区用户的互动和分类.Treude 等人[16]通过主题模型分析了 Stack Overflow 上的问题并将主题分为 10 个大类别.文献[4,5,17−19]都是针对 Stack Overflow 上的某一类别问题进行进一步的主题划分.其中,文献[5]通过统计每个主题的热度和难易度的影响因素,利用统计学方法,证明在问答社区上,问题的难度和热度呈负相关关系.文献[18]分析了 Stack Overflow 上关于大数据问题的主题、热度和难度.文献[19]分析了 Stack Overflow 上关于机器学习问题的主题,分析了这些主题的热度和难度以及热度难度的关系.

国内并没有针对中文软件问答社区的主题内容分析.因此缺少对中国用户关注的技术问题的了解,也缺少对于中文问答社区的社区生态、用户模型等分析的基础.

2 研究方法

在本文的这一部分,依据经验研究的思路,首先提出了研究问题并分别说明研究问题的研究意义.之后说明实验的总体设计,宏观上介绍了实验的每一步流程和逻辑关系.之后详细介绍了数据获取与预处理,以及主题提取的详细实现,为后续主题分析作铺垫.

2.1 研究问题

根据软件问答社区内容分析的研究目标,本文将研究工作具体分为 4 个研究问题.

研究问题 1. 用户在开源中国问答社区主要讨论的主题有哪些?

现如今,开发者在开发过程中需要解决各种各样的问题,也会面临各种各样技术之间的选择.在遇到这些问题时,开发者往往会前往软件问答社区去寻找答案.想要分析这些软件问答社区上的内容,首先要找出该问答社区中用户主要讨论的主题都有哪些,对于这些主题的内容进行基本分析、归类,形成对其他研究问题进行研究的基础.同时,找出这些主题、对这些主题进行基本内容的分析介绍和归类也能对开发者日常开发中关注什么样的技术问题形成一个整体上的认识.

研究问题 2. 这些讨论的主题其热度随时间发生变化的趋势是什么样的?

当前技术领域迭代更新迅速,新技术、新方案层出不穷.在得出了问答社区上讨论的主题都有哪些的条件下,通过分析这些技术主题随时间的变化趋势,开发者、企业管理者、研究者等用户群体就能知道哪些技术内容的热度正在上升,而哪些技术内容的热度正在下降,从而对这些群体的决策起到帮助作用.

研究问题 3. 这些讨论主题的回答情况如何?

在软件问答社区众多的问题中,存在回答数量多、回答数量少的问题,也存在完全没有回答的问题.通过对每一个主题的回答情况进行统计,可以推测一些主题下问题的回答难度.例如,当某一主题热度足够大,但无回答问题的比例也很多时,可以推测这个主题下的问题普遍较难回答.反之,当一个主题无回答问题的比例很小时,那么该主题下问题的平均难度较低.这些结论将对社区用户起到启示作用.

研究问题 4. 用户对特定技术的兴趣是如何随时间发生变化的?

尽管通过研究问题 1 得到了软件问答社区中用户讨论的主题,但是这些主题往往是一些宏观的技术领域而不是某一个具体的技术或工具,例如,在一个与数据库相关的主题下,包含着 MySQL、SQL server 等具体的工具,但是单独依靠之前研究问题的分析无法分析这些具体的技术热度与走势.通过分析、对比这些具体技术的走势,可以对一些关心这些技术的用户起到指导作用.

2.2 研究总体设计

本文要从开源中国问答社区获取用户的讨论主题并加以分析.本文采用主题模型的方法来发现主题.主题

模型是一种信息检索技术,它能够自动地从指定的文本文集中查找主题,而不需要训练数据或预定义分类[20−22].主题模型在原始数据的基础上,能够通过每一个文档中的词频和词共现频率构建出一种相关词组成的模型.本文从数据获取到分析展示的流程如图 1 所示.首先进行数据采集得到原始数据,之后进行预处理将原始数据进行删减、加工.在数据获取和预处理环节之后,通过主题模型处理数据来提取文本中的主题,但是得到的主题并不能直接用于解答研究问题.所以在获取主题之后,通过每个主题下问题的其他信息,如时间戳、访问量等,对主题进行分析.

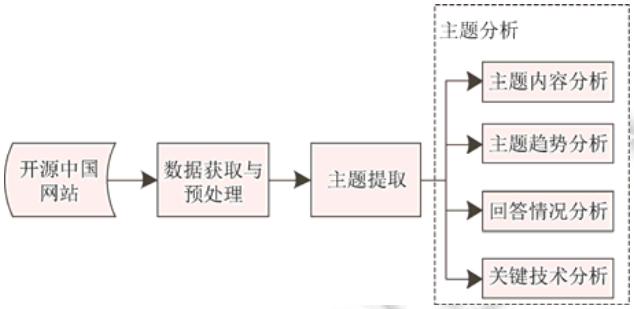


Fig.1 An overview of our research method

图 1 分析流程图

2.3 数据获取与预处理

关于研究数据,本文针对开源中国问答社区,通过数据采集过程,从开源中国问答社区获取 92 383 个有效问题.收集的问题中包含少量的 2011 年和 2019 年的数据,由于这两年的数据过少,在实际应用中没有使用.论文使用的数据时间跨度为 2012 年 1 月~2018 年 12 月,共计 90 204 条有效数据.

收集到的问题详细信息见表 1,每一个问题包含用于展示问题主要内容的问题题目(title)、用于描述问题具体信息的主干(body)、用于给问题归类的标签(tag),这里的标签是用户自行设定的.以上 3 个数据属性代表了这个问题的内容,这些内容可用于提取主题和分析主题,从而得到有意义的结论.开源中国问答社区上每一个问题除了有具体内容,还有一些记录其他信息的属性,如发布时间(pubDate)记录问题的提出时间,回答数量(answerCount)记录问题所拥有的回答数量、浏览次数(viewCount)记录问题被用户浏览的次数.以上这些元数据对主题分析起着至关重要的作用.

Table 1 Detailed information of a question

表 1 数据项明细

字段名	类型	说明
id	int	问题 ID
pubDate	datetime	发布时间
author	string	作者
authorid	long	作者 ID
body	string	问题内容
title	string	问题标题
answerCount	int	回复数
viewCount	int	浏览数
url	string	问题链接
Tag	string	问题标签

数据的预处理针对数据中的问题内容(body)和问题标题(title),整体上分为 3 个步骤.首先,对于每个数据项,通过正则表达式去除原始数据问题内容中的 html 标签.其次,由于问题中的不同代码段往往含有相似的内容,所以这一类代码段对于提取主题并没有帮助[23],因此在预处理的第 2 步通过正则表达式去除掉原始数据问题内容中的代码段.最后,通过 jieba 分词工具,对原始数据中的问题题目和问题内容进行分词,将分词后的结果替换掉原本的问题题目和问题内容,同时删除预处理后题目或内容为空的问题,并将已处理的结果保存为文件,以便

下一步分析.

2.4 主题提取

本文基于LDA方法从问题数据中提取主题.LDA是一种成熟的主题模型,被广泛应用于自然语言处理中主题发现、主题分析、文本聚类等问题.LDA 是一种 3 层贝叶斯模型,用于离散的数据集(如文本),其中认为数据集中的每一个数据项都服从一个隐含的主题集合上的概率分布[8].也就是说,每一个文档项都有一定概率属于一个主题集中的主题,同时文本集中每个单词也同样有一个主题分布.对于文档,认为某文档的主题是其主题分布中概率最大的主题.对于主题,认为其词分布中概率最大的几个词能够概括主题的内容.例如,对于一个与移动端开发相关的问题,可能归类为以“android”“手机”“ios”等词为主题词的主题.

LDA 算法从原理上说,假设在原始数据语料库 C 中,对于任一文档 d,存在先验的主题分布 θ_d,其中,θ_d 服从参数为 α 的 Dirichlet 分布.同样,假设主题中词的先验分布 φ_K 也服从参数为 β 的 Dirichlet 分布.通过两个先验分布和对于原始文档进行统计,可以得到文档中词-主题的多项分布和全部文本中主题-词的多项分布.利用多项分布与 Dirichlet 分布共轭,可以更新 θ_d 和 φ_K 的后验分布.亦即确定了每一个文档属于哪个主题、每一个主题中哪些词出现的概率更大.

LDA 算法的基本流程为(基于 Gibbs 采样):

(1) 选择合适的主题数 k 以及参数 α 和 β.

(2) 对于每一个文档的每一个词,随机设置一个主题编号 z.

(3) 对于文本库中的每一个词,利用 Gibbs 采样公式计算这个词属于每一个主题的概率,并更新语料库中该词的主题编号.

(4) 重复步骤(3),直到结果收敛,并根据文档词的主题分布计算文档的主题分布.

本文参考文献[1,3,4]中的实验方法,利用默认值设置 k=20,α=50/k 和 β=0.01.本文通过 LDA 方法分析预处理后的数据.得到每个主题的词分布、文档-主题分布,将这两个分布记录为文件,以用于之后的主题分析.

3 软件问答社区主题结果分析

本节以数据采集和主题提取的结果为原始资料,针对每一研究问题进行实验得出实验结果,并分别对实验结果进行分析,从而解答每一个研究问题.在本节最后,总结了分析 4 个研究问题所得出的重要结论,并讨论了这些结论的有效性.

3.1 研究问题1:软件问答社区的讨论主题

在这一研究问题中,本文要研究并解答用户在开源中国问答社区主要讨论的主题有哪些,并对这些主题进行归类以及具体介绍.

3.1.1 分析方法

在上文提到的主题抽取过程中,本文通过使用主题数 k=20 的 LDA 算法分析了预处理后数据集中的全部问题数据,得到了 20 个社区中用户讨论的主题.对于每一个主题 Z_k 都有一个词分布 φ_K,对于每一个文档 d 都有一个主题分布 θ_d.

根据前一步中 LDA 算法的结果,通过每个主题的词分布 φ_K 以及词袋模型从原始数据中提取出的特征词,获取这个主题中出现概率较大的 10 个词作为主题内容的概括.根据主题词人工地为主题标注名称.同时计算出每个主题 Z_k 的占比 $Share(Z_k)$ 并加以展示.其中,$Share(Z_k)$ 的计算方法如下:

$$Share(Z_K)=\frac{Num(Z_K)}{N},$$

上式中,N 为全部问题数量.$Num(Z_k)$ 是主题为 Z_k 的问题数量.

之后,通过对每一个主题下的文档-主题分布 θ_d 进行排序,选择概率最大的问题作为这一主题的示例问题加以展示和分析.最后将得到的全部主题的内容、占比作为用户在开源中国问答社区讨论的所有主题加以展示.

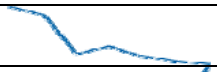
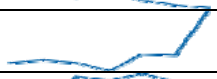
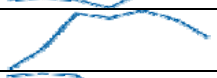
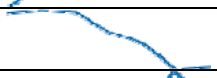
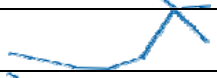
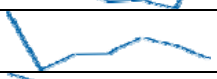
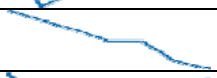
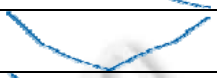
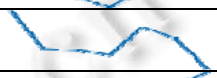
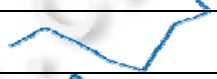
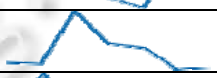
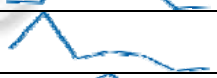
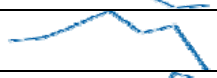
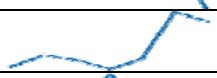
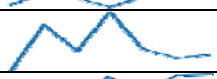
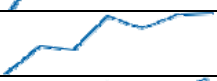
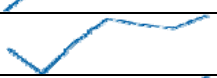
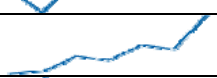
3.1.2 从整体分析

根据分析方法,得出每一主题的关键词、占比、主题名称等数据,将结果按占比排序并进行展示.之后,根据研究问题 2 中的热度趋势分析,将每一个主题的热度趋势展示在表中,具体主题的列表见表 2.

从 20 个主题的总体分析结果上看,用户在开源中国问答社区上关心众多不同内容的技术主题,从日常开发中会经常遇见的操作系统与软件开发问题,到分布式系统这样较为专业的主题.在这 20 个主题中,操作系统与软件安装主题的讨论占比最大.

Table 2 List of all topics

表 2 主题列表

主题名	主题词	主题占比	趋势
操作系统与软件安装	安装 inux 运行 版本 编译 系统 命令 提示 python windows	0.075 8	
数据库与 SQL 语句	数据 数据库 mysql 查询 sql id 字段 语句 oracle 插入	0.070 3	
项目部署	项目 信息 eclipse tomcat maven 报错 jar 启动 运行 java	0.063 1	
页面与界面	点击 显示 按钮 事件 页面 android 界面 添加 选择 设置	0.062	
技术学习	开发 学习 java 技术 开源 框架 项目 python 推荐 工作	0.057 1	
文件操作	文件 图片 上传 下载 生成 目录 java 路径 读取 文件夹	0.054 6	
服务器配置	服务器 访问 nginx 配置 连接 ip tomcat 端口 地址 apache	0.052 5	
移动端开发	android 客户端 手机 app ios 视频 服务端 安卓 开发 播放	0.051 7	
数据类型	代码 乱码 php python 输出 字符串 java 数组 中文 输入	0.051	
前端编程	页面 js 浏览器 html 标签 代码 显示 加载 jsp jquery	0.049 1	
函数与方法	方法 对象 函数 调用 java 代码 参数 属性 类型 定义	0.047 5	
数据展示	显示 echarts 图片 设置 地图 数据 效果 林峰 位置 颜色	0.047 2	
数据库配置	方法 异常 数据库 jfinal 调用 代码 连接 报错 事务 配置	0.045 5	
后端数据	数据 请求 返回 获取 json 后台 参数 提交 java 页面	0.043 3	
操作系统调度	执行 线程 时间 内存 日志 运行 程序 测试 java 进程	0.042 9	
网页应用	网站 php 网页 微信支付 链接 邮件 获取 功能 分享	0.042 3	
Spring	Spring 配置 项目 git 代码 mybatis 配置文件 hibernate 注解 xml	0.041 4	
分布式系统	数据 redis 缓存 节点 集群 消息 hadoop 同步 配置 启动	0.034 4	
系统设计	系统 服务 功能 用户 设计 项目 接口 业务 提供 公司	0.034 4	
用户安全	用户 登录 权限 密码 session 登陆 验证 加密 信息 页面	0.034	

3.1.3 从类别分析

在得到 20 个主题之后,本文研究发现,有很多主题可以被划分为同一个大类别.为了得到更具普适性、更方便读者建立宏观认识的结论,本文将这些主题分为了 6 大类别,分别为:前端开发、后端开发、数据库、操作系统、通用技术和其他.类别与主题的关系见表 3.其中,每一类别按占比排序,每一类中的所有主题按占比排序.

Table 3 Categories of topics

表 3 类别与主题关系表

类型	问题比例	主题名
前端开发	0.252 3	页面与界面 移动端开发 前端编程 数据展示 网页应用
通用技术	0.244 2	技术学习 文件操作 数据类型 函数与方法 用户安全
后端开发	0.227 7	项目部署 服务器配置 后端数据 分布式系统 系统设计
操作系统	0.118 7	操作系统与软件安装 操作系统调度
数据库	0.115 8	数据库与 SQL 数据库配置
其他	0.041 4	Spring

本文给予每一个类别相对明确的定义,下面给出每一类别的定义及具体内容介绍.

前端开发类别.

前端开发的主题主要讨论呈现给用户的网页页面或手机 app 界面相关的问题.根据表 3,前端相关的问题在开源中国问答社区上占比为 25.23%,被问及的问题数量最多.这一部分包括页面与界面、移动端开发、前端编程、数据展示以及网页应用 5 个主题.

页面与界面主题在前端开发分类中占比最大,这一主题主要包括关于页面、界面事件和页面、界面设计的问题.例如,通过文档-主题分布,发现该主题下概率最大的问题为关于实现页面滑动菜单的问题.

移动端开发主要是指移动端应用在手机一端的开发,实现和应用用户交互、展示的功能.

前端编程主要是指与前端编程关系很大的问题,一般是在问前端的某一类功能如何编程实现或某前端代码为何出错.

数据展示主要涉及结构化的数据如何在前端展示的问题,一般是关于网页图表、表格的问题.

网页应用主要是指一些网页相关应用的综合讨论.从这个主题的主题词可知,网页应用包括与微信相关的应用、与支付相关的应用、与邮件相关的应用等.

综上分析可知,在开源中国问答社区上,用户在前端开发中讨论页面与界面的设计、实现相关的问题最多.用户在前端开发类别中,也同时关心移动端开发、前端编程、数据展示和网页应用相关的问题.

通用技术类别.

通用技术的问答占比为 24.42%,在开源中国问答社区也占有很大的比例.这些技术涉及范围很广,基本上,无论是在前端、后端、数据库,还是在操作系统下,都能用到这些通用技术.这一分类下包括技术学习、文件操作、数据类型、函数与方法和用户安全这 5 个主题.

首先,占比最大的是技术学习,这一类问题的主要内容是新技术、新概念的学习与推荐.主要包括两部分.(1) 新手在问答社区提问,请求学习帮助或技术推荐.(2) 由一些用户在问答社区分享的技术推荐文章,用来介绍

一种技术.例如,通过技术学习主题的文档-主题分布,提取出主题下概率最大的问题为“如何用最短时间高效学习python”.

文件操作主题,包括文件的上传与下载、文件的读写、文件的创建、文件的路径问题等一系列与文件相关的问题,统称为文件操作.文件操作问题不光可以在后端中遇见,也可以在前端、数据库、操作系统问题中出现,例如怎么上传文件到网页、怎么导出生成 sql 文件等.所以将文件操作归类于通用技术类.

数据类型主题,顾名思义,是关于各种编程语言中数据类型的汇总问题,包括数字、字符、字符串、数组等.这类问题也属于开发者应该掌握的通用技术,能被应用在前端、后端、数据库等多种场景下,所以被归类为通用技术类.

函数与方法主题,是关于各种与函数或方法相关的问题的主题,同时,这一类别涉及了部分面向对象编程的内容和概念(基本上都与方法相关),由于这一主题同样可以应用在前端、后端等众多场景,属于开发者都应掌握的通用技术,所以放在这一分类当中.

最后,用户安全主题,是关于用户身份验证、用户安全、加密等问题的汇总,但是也包括一小部分单纯与用户登录或安全相关的问题.这一主题同样属于可以被应用在多个场景下的通用技术.

综上,在通用技术方面,用户主要讨论技术学习、文件操作、数据类型、函数与方法和用户安全.

后端开发类别.

在 20 个主题中,本文将项目部署、服务器配置、后端数据、分布式系统以及系统设计主题归类为后端开发类别.分别介绍如下.

在后端开发中,项目部署相关的问题占比最高.这一主题主要包括和后端项目部署相关的问题,如项目在服务器(主要为 tomcat)上的部署,主要包括但不限于 Java 项目和 Spring 项目.

服务器配置主题在后端开发类别中占比较高,这一主题主要包含各种服务器的设置、搭建问题.

后端数据主题,主要是关于两种前后端数据交流的问题,如 HTTP GET 和 HTTP POST 以及 URL 传参数等.

分布式系统主题,顾名思义,包括绝大部分与分布式相关的问题.例如分布式计算、数据的分布式存储、服务器集群等.

系统设计主题主要包含系统宏观设计、功能设计、接口设计等相关问题.但也包括一小部分单纯的规则设计、流程设计等设计问题.根据系统设计主题的文档-主题分布,提取出属于该主题概率最大的问题为“类电商系统如何进行前后台分离”.

操作系统类别.

操作系统类别包括操作系统与软件安装以及操作系统调度这两个类别.这两个主题都与操作系统概念或具体操作系统型号关系较大.其中,操作系统主题主要是关于操作系统下的软件安装、与操作系统相关的 bug 以及操作系统命令的问题,从主题词上分析,应该还包含一些 Python 程序包安装的问题.而操作系统调度这一主题主要是关于进程、线程的调度问题,存在少量关于系统内存、系统时间的讨论.

数据库类别.

数据库类别包括数据库与 SQL,数据库配置两个主题.这一类主题的特点是明显围绕数据库讨论.其中,数据库与 SQL 主题占比最大,主要内容就是与数据库,尤其是 SQL 语句使用相关的问题.数据库配置主题和上一主题的区别是更加突出数据库本身的配置、链接、备份等内容.例如,根据数据库配置主题的文档-主题分布,提取出该主题下概率最大的问题为 SQLserver 连接报错的解决方案.

其他类别.

这一类别只有 Spring,是因为 Spring 作为开发工具,并不是归类于上述前端、后端、通用技术等任何主题.在此次结果中,只有 Spring 这个技术工具形成了独立主题,这说明:(1) Spring 是一个很新的工具,它在 2003 年才发布最初版本,近年来才在实际开发中占比激增.用户普遍对于 Spring 不熟悉,网络上的教程也不如其他技术工具多,所以对于 Spring 本身语法、特性等属性的讨论较多,从而能够形成一个主题.(2) Spring 不同于 Java、Python 等传统的语言或工具,具有太多新的特性.例如前 10 个主题词中的 MyBatis(提供持久层框架)、Hibernate(一种数据映射框架)、注解(用于实现 AOP 功能的语法结构)都和 Spring 密切相关,这些特性成为了 Spring 项目的特

征词,使得 Spring 相关的问题特征明显,更容易被 LDA 算法聚为一类.

Spring 主题下概率最高的问题都是关于 Spring 自身配置、自身特性的讨论.例如,“Spring boot 怎么取消自动将 Filter 加入到容器的 Filter 链中”“为什么要在 Spring 中进行 bean 的配置”.这证实了上文中关于 Spring 单独形成主题的原因的正确性.综上,Spring 成为唯一的开发工具相关主题并非因为它的热度已经超过了其他开发工具,而是因为针对 Spring 自身特性的讨论问题,比针对其他某一种开发工具自身特性的讨论要多.也因为 Spring 相对于其他传统开发工具,拥有更多独有的特性和功能,从而使 Spring 的问题更具有特征,更容易形成独立主题.

3.1.4 小 结

在这一研究问题中,通过主题模型得到了用户主要讨论的 20 个主题,并将这些主题分成了前端、后端、数据库、操作系统、通用技术以及其他 6 个类别,并对于每一类别的内容进行了详细讨论.在这些类别中,前端类别主题的讨论占比最多.在单个主题中,操作系统与软件安装的讨论最多.

3.2 研究问题2:主题热度变化趋势

在这一研究问题中,本文首先统计全部问题的时间趋势情况,对比每一类别整体的热度趋势.之后按类别分析、对比每一个主题的热度趋势.从而对技术热点及动向有一定了解.

3.2.1 分析方法

通过分析主题下每个问题的时间戳,统计主题热度随时间的变化并输出展示.本问题中使用平均浏览量来评估某个主题的热度.定义在某一时间段 T 下,主题 Z_K 的热度为 $TopicImpact(Z_K,T)$,其计算公式如下:

$$TopicImpact(Z_K,T)=\frac{\sum_{d_i\in D_{T,K}} View(d_i)}{\sum_{d_i\in D_T} View(d_i)},$$

其中,$D_{T,K}$ 是主题 Z_k 在时间段 T 内的问题集合,D_T 为时间段 T 内的全部问题集.$View(d_i)$是问题 d_i 的浏览数.根据主题的热度,定义类别 C_K 在时间段 T 的热度 $CategoryImpact(C_K,T)$,其计算公式如下:

$$CategoryImpact(C_K,T)=\sum_{Z_K\in C_K} TopicImpact(Z_K,T).$$

3.2.2 从整体分析

本文采用上文中的类别热度计算方法,针对全体问题,统计了各个类别的热度趋势,如图 2 所示.由图 2 可知,前端开发、后端开发和通用技术类别热度较高,其中前端开发热度自 2014 年起呈下降趋势,后端开发类别自 2013 年~2016 年热度呈上升趋势.

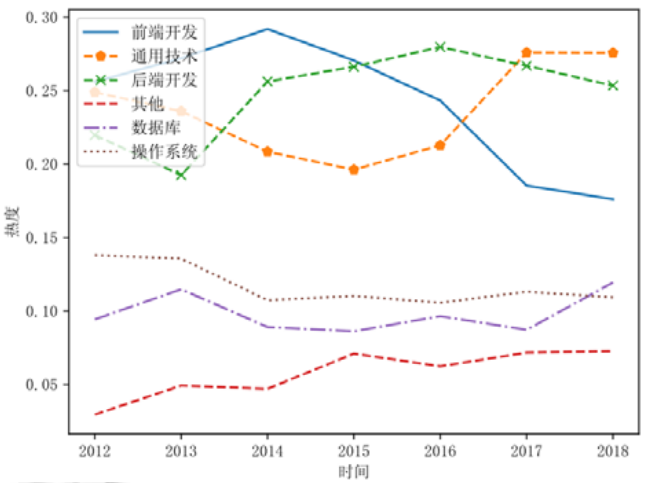


Fig.2 Popularity of categories

图 2 类别热度趋势

3.2.3 从类别分析

首先,根据前文研究问题 2 中的方法,分析所有主题的变化趋势,见表 2.之后,根据研究问题 1 中的 6 个类别,分别对每一类别中的主题热度趋势进行分析对比.

前端开发类别.

针对所有在前端开发类别下的主题进行了趋势分析.具体结果如图 3 所示.

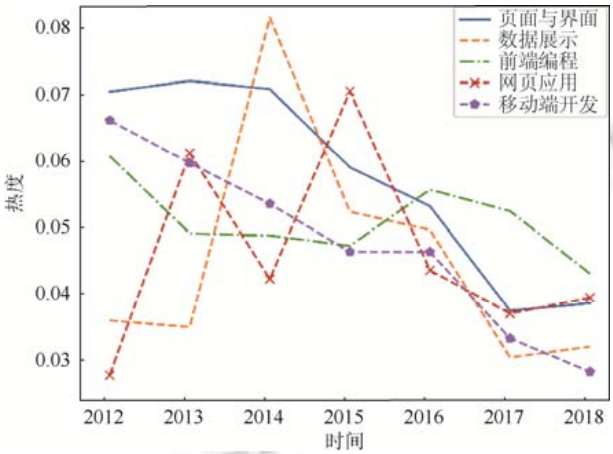


Fig.3 Popularity of front-end topics

图 3 前端类别主题热度趋势

由分析结果可知,前端开发的众多主题虽然在 2015 年之前趋势波动较大,无法分析,但在 2015 年之后,所有前端主题热度一致下降.这种整体趋势代表了整个前端领域的关注度的下降.

通用技术类别.

针对所有在通用技术类别下的主题进行趋势分析,如图 4 所示.

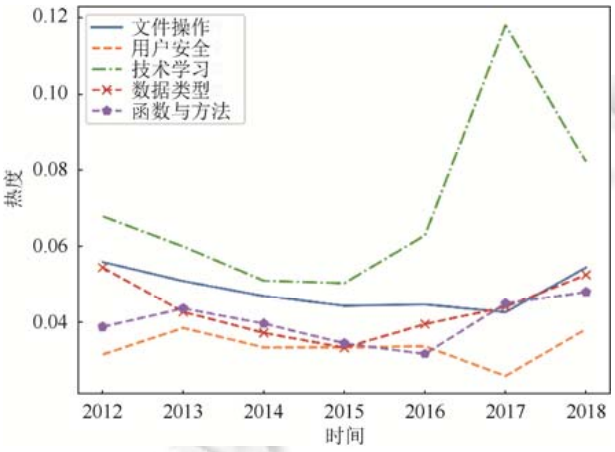


Fig.4 Popularity of general techniques topics

图 4 通用技术类别主题热度趋势

由分析结果可以看出,在开源中国问答社区上除技术学习主题之外,其他讨论主题的热度趋势变化基本不大.这与通用技术的定义是相符的,通用技术类别下的文件操作、用户安全、数据类型和函数与方法这 4 类问题由于通用性强,所以受到技术热点变化的影响较小.

后端开发类别.

针对所有在后端开发类别下的主题进行了趋势分析.具体结果如图 5 所示.

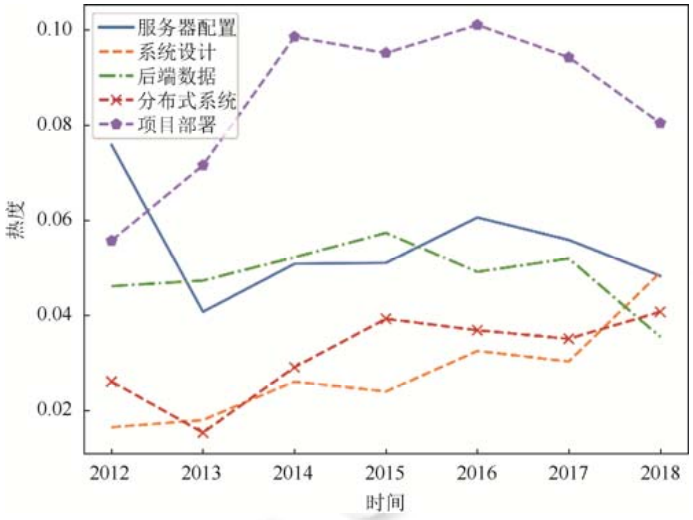


Fig.5 Popularity of back-end topics

图 5 后端类别主题热度趋势

由分析结果可知,后端开发类别中,项目部署的热度保持最高,该主题在 2012 年～2014 年热度快速上涨,这与前端大部分主题热度的上涨是同步的.但是从 2016 年来出现下降趋势,也与前端大部分主题热度趋势相同.服务器配置主题的热度除了在 2012 年～2013 年之间快速下降之外,其他时候热度先小幅上升,从 2016 年又小幅下降,总体上基本保持持平.说明开发者一直对服务器配置问题保持一定的关注,因为这基本上是后端开发要解决的基本问题.服务器配置主题与项目部署主题一样,从 2016 年以来出现下降趋势.后台数据主题与服务器配置主题趋势基本相同,从 2016 年以来出现下降趋势.

值得关注的是分布式系统主题以及系统设计主题.自 2012 年以来基本呈热度上升趋势.说明开发者的关注点逐渐从基础的后端开发(项目部署、服务器配置、后端数据传递)内容转移到了系统设计以及分布式系统这样的进阶内容.

数据库类别.

针对所有在数据库类别下的主题进行了趋势分析.具体结果如图 6 所示.

由分析结果可知,数据库类别下,数据库与 SQL 主题的热度在年线上呈上升趋势.关于数据库配置的热度从 2013 年起下降趋势明显.结果表明,在数据库方面,用户更加关心 SQL 语言的使用,而不是数据库本身的配置、连接问题.同时,也因为主流数据库版本更新速度较慢,几年前的问题很可能足够解决开发者们当前遇到的大部分数据库配置问题,所以数据库配置主题的热度总体上呈下降趋势.

操作系统类别.

针对所有开发工具类别下的主题进行了趋势分析,具体结果如图 7 所示.

由分析结果可知,操作系统类别下,操作系统与软件安装主题的热度在年线上呈下降趋势.操作系统调度主题的热度在 2015 年~2017 年稍有提升.

其他类别.

由于其他类别中只有 Spring,对其单独进行热度趋势分析如下.

由分析结果可以看出,Spring 类别的热度在近几年快速上升,从 2012 年~2018 年,Spring 的热度上升超过 1 倍.项目开发者、技术学习者都可以考虑多在 Spring 相关技术上投入更多时间来学习.

Fig.6 Popularity of database topics

图 6 数据库类别主题热度趋势

Fig.7 Popularity of operating system topics

图 7 操作系统类别主题热度趋势

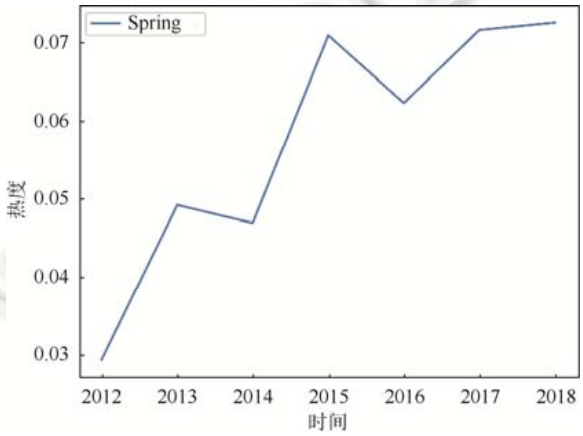


Fig.8 Popularity of the other topic

图 8 其他类别主题热度趋势

3.2.4 小 结

通过这一研究问题,本文分析了不同类别下每一主题的热度变化趋势.总体上说,前端开发类别下主题的热度逐年下降,通用技术类别下主题的热度基本维持不变,后端开发类别下主题的热度变化不同,传统的项目部署、服务器配置热度逐年下降,而系统设计、分布式系统这样更新、更复杂的主题热度在逐年上升.

3.3 研究问题3:主题回答情况分析

在这个研究问题中,希望通过分析每一个主题的回答情况,对主题下问题的回答难度有一定评估.

3.3.1 分析方法

本文通过以下两个维度来分析主题的回答情况:(1) 统计了主题下没有回答的问题所占的比例(零回答问题比例 NP).(2) 在统计零回答问题比例的基础上,统计了主题下问题的平均浏览量.主题 Z_k 的无回答问题比例 NP 的计算公式定义如下:

$$NP(Z_K)=\frac{No\_Ans\_Count(DZ_K)}{Count(DZ_K)}\times 100\%,$$

其中,DZ_k 是主题 Z_k 的问题集合.$Count(DZ_k)$是 DZ_k 中的全部问题个数.$No_Ans_Count(DZ_k)$是 DZ_k 中的没有任何

回答的问题个数.本文结合两种回答情况的统计一起进行分析对比.如果某个主题下的问题平均浏览量较高,问题受到用户大量关注,但是无回答的问题占比很高,那么这类问题较难回答.

针对这一研究问题,通过 LDA 算法提取的主题结果,结合原始数据统计每一主题的平均浏览数以及没有回答的问题占比.将这些统计数据进行排序并加以展示.

3.3.2 从整体分析

根据上文的分析方法,得出每个问题的平均浏览数以及无回答的问题比例.将数据项按零回答问题比例加以排序的结果见表 4.

Table 4 Statistics for each topic of answers

表 4 主题回答情况分析表

主题名称	类别	平均浏览数	零回答问题比例
数据展示	**前端开发**	**974.74**	**0.426 1**
分布式系统	后端开发	847.11	0.336 1
移动端开发	前端开发	845.19	0.333
页面与界面	前端开发	863.85	0.328 6
文件操作	通用技术	761.05	0.313 3
操作系统与软件安装	操作系统	905.56	0.299 2
网页应用	前端开发	105 9.71	0.269 3
前端编程	前端开发	905.41	0.265 4
操作系统调度	操作系统	777.91	0.263 5
服务器配置	后端开发	923.95	0.258 6
系统设计	后端开发	664.7	0.255 1
用户安全	通用技术	882.87	0.252 2
项目部署	后端开发	125 9.62	0.250 6
技术学习	通用技术	991.46	0.227 7
Spring	其他	1254.4	0.223 8
函数与方法	通用技术	701.58	0.216 9
数据库与 SQL	数据库	674.03	0.215 6
数据库配置	数据库	780.79	0.206 2
后端数据	后端开发	106 1.09	0.203 5
数据类型	通用技术	690.28	0.201 4

由分析结果可知:数据展示(在表 4 中标粗展示)相关的问题拥有平均浏览数为 974.74 次,大于所有主题问题的平均浏览数 891.25 次,说明数据展示主题相对较热.同时,这类问题的没有回答的问题比例最大,有 42.6%的问题没有回答.在数据展示话题相对较热的条件下,无回答的问题比例最大,可以认为这个主题的问题较难回答.

根据主题无回答的问题比例,本文发现数据展示、分布式系统、移动端开发、页面与界面和文件操作这 5 个主题有超过 30%的问题没有得到回答.对于零回答问题比例较高这一问题,未来可应用自动推荐回答的方法[24]为提问者提供参考答案.

对于数据类型主题,该主题的平均浏览量较低,但是零回答问题比例最低.该主题在用户关注不大的情况下,未问答的问题比例最少,说明这一主题的问题容易回答.

3.3.3 从类别分析

从类别角度上看,前端开发类别下主题都相对较难.根据前文分析,前端开发类别包括数据展示、移动端开发、页面与界面、网页应用和前端编程这 5 个主题.在回答情况分析中,根据表 4,这 5 个主题的平均浏览数都接近或高于平均值,说明这些主题都具有足够的用户浏览量.同时,这些主题的零回答问题比例都大于平均值,结合它们的浏览数,认为前端开发类别下的 5 个主题都相对较难.

后端开发、通用技术、操作系统类别下的主题有的较难回答,有的容易回答,并不具备一致的特征.

3.3.4 小 结

在这一研究问题中,本文通过对于每一个主题的回答情况进行分析和对比.数据展示主题的平均浏览数高,但是零回答问题比例最高;数据类型主题的平均浏览数较低,但是零回答问题比例最低.

3.4 研究问题4:一些关键技术的趋势分析

对比主题的热度能够提供对于技术动向的宏观认识,但是无法体现具体技术手段、技术工具的热度关系.所以这一部分根据前文研究问题 2 的分析方法,挑选研究问题 1 中由多个关键技术作为主题词的主题,对这些主题进行主题词中关键词的趋势分析.

3.4.1 分析方法

在这一部分,本文希望对主题下的关键技术进行趋势分析与对比.首先在关键字选取上,本文根据主题的词分布,提取其中条件概率最大的 10 个词,从中选取代表某一特定技术或一个编程语言的关键字,使选择的词与主题更相关,同时保证客观性.例如,对于技术学习主题,其排序在前 10 的主题词为:“开发 学习 java 技术 开源 框架 项目 python 推荐 工作”,其中,Java 和 Python 代表编程语言,因此选取 Java 和 Python 进行分析.而对于服务器配置主题,由关键词:服务器、访问、nginx、配置、连接、ip、tomcat、端口、地址、apache 组成,其中,nginx、tomcat、apache 都是服务器的具体解决方案,可以用于关键字分析.在统计方法上,相对于分析标签,本文直接分析问题文本内容(包括正文、题目、标签),扩大搜索范围.在关键技术热度的统计方法上,本文统计出现关键词的问题的浏览数占比,作为关键词的热度.设关键词 k 对于在第 y 年对主题 t 的影响力 Impact 为

$$Impact(k,t,y)=\frac{\sum_{D_i\in Q(y,t,k)}View(D_i)}{\sum_{D_i\in Q(y,t)}View(D_i)},$$

其中,$Q(y,t,k)$指的是 y 年内含有关键词 k(包括 k 的大小写形式)的 t 主题下的问题.$Q(y,t)$指的是 y 年内 t 主题下的所有问题.$View(D_i)$是该问题的浏览量.

3.4.2 从整体分析

对于全部 20 个主题,根据主题词,挑选其中技术学习、服务器配置、移动端开发、数据库与 SQL、操作系统与软件安装和前端编程主题,对它们进行了关键技术分析,具体每一个主题的分析结果如图 9 所示.

根据以上分析结果,可以得出如下结论.

(1) 在技术学习主题中,Java 的讨论热度占比仍高于 Python.说明对于初学者来说,讨论 Java 语言的热度要大于讨论 Python 的热度.这与编程语言热度排行榜是相符的,根据 Tiobe 的编程语言排行榜,Java 语言的热度高于 Python 语言的热度(TIOBE,https://www.tiobe.com/tiobe-index).另外,越来越多的学校选择 Java 作为教学语言[25],这可能也造成 Java 讨论热度更高.

(2) 在服务器配置主题中,关于 nginx 服务器的讨论逐年上升而关于 apache 服务器的讨论逐年下降,尽管 apache 在全球网站中的使用比例超过 nginx 服务器的使用比例.但是根据 builtwith 上的统计数据,使用 nginx 的中国网站数量约是使用 apache 的中国网站数量的 1.25 倍(BUILDWITH,https://trends.builtwith.com/Web-Server/nginx).结合关键技术趋势分析结果可见,在中国开发者中,nginx 服务器热度大于 apache 服务器热度.

(3) 在移动端开发的问题中,安卓开发的讨论热度遥遥领先于 ios 开发的讨论热度,这与安卓手机的市场占有比例远远大于 ios 系统手机,同时安卓开发岗位多于 ios 开发岗位这两个事实相符.

(4) 对于数据库与 SQL 主题,mysql 的热度明显高于 oracle 数据库的讨论热度.这与相关工作中对 Stack Overflow 社区关键技术趋势对比的结论是一致的[1].

(5) 在操作系统与软件安装方面,虽然 Linux 用户少于 Windows 用户,但在软件问答社区,关于 Linux 的问题讨论热度仍多于关于 Windows 问题的讨论热度.

(6) 最后,对于前端编程主题,用户讨论更多的是 JavaScript 而非 HTML.

3.4.3 小 结

本文通过对主题的关键词进行热度趋势分析,得出特定主题下关键技术的热度趋势对比情况.例如,在服务器配置主题中,关于 nginx 服务器的讨论热度最高.对于开发者和学习者,这些分析可以在具体技术的选择上起到指导作用.

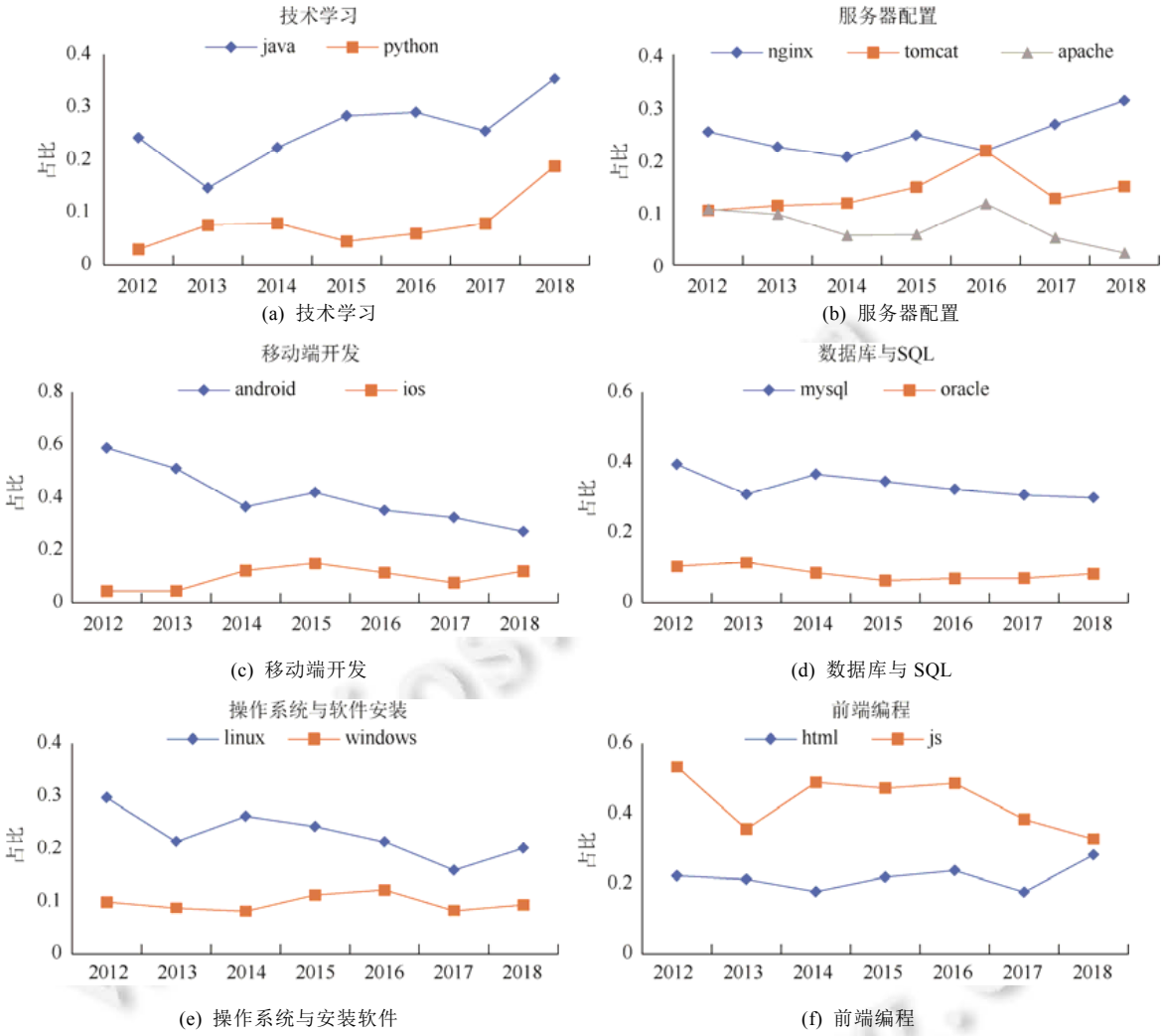


Fig.9 Popularity of keywords

图 9 关键词趋势图

4 讨 论

4.1 重要结论总结

本文通过对开源中国问答社区的主题进行分析,从而对中国的软件开发行业发展现状可以有更深入的了解.本节总结本文研究得出的重要结果.下面是总结的 4 个重要结论.

前端开发进入缓慢发展时期.

本文通过 LDA 主题模型从开源中国问答社区的数据中提取了 20 个用户主要讨论的主题,其中本文将网页与页面、前端编程、移动端开发、数据展示以及网页应用这 5 个主题归类为前端开发类别.这一类别在开源中国问答社区的讨论占比最高.但是,在之后的热度趋势分析中,发现前端开发类别下的主题热度在 2015 年之后均呈现下降趋势.在之后的回答情况分析中,发现前端类别的主题回答难度都相对较大.

越来越少的关注热度可能意味着前端开发技术的迭代放缓,使新问题减少;也可能意味着关注前端技术的人群在逐渐减少.无论是上述哪种可能,都意味着前端开发本身从之前快速发展进入了当下缓慢发展时期.

后端开发重点的转移.

在 20 个主题中,项目部署、服务器配置、后端数据、分布式系统以及系统设计这 5 个主题被归类为后端开发类别.在主题热度趋势分析中,发现项目部署、服务器配置这两个传统的后端开发话题热度在逐年下降.同时,分布式系统、系统设计这两个相对更新、更复杂的主题热度在逐年上升.可以认为,在后端开发方面,技术的讨论热点正在从传统的项目部署、服务器配置转移到更新、更复杂的分布式系统和系统设计主题上.这对于企业或项目开发者来说,可以考虑加强后端开发中分布式系统以及系统设计相关任务所占的比重,以提升自己软件的设计与性能.对于后端开发技术学习者来说,在学习项目部署与服务器配置的基础上,应该尝试投入时间和精力学习分布式系统和系统设计的内容,以保持在后端开发中的竞争力优势.

通用技术值得重视.

在本文中,通用技术类别代表开发者在日常开发中往往需要掌握的通用技术,其中每一项技术都可以在前端、后端、数据库、操作系统等多个应用场景中使用到.通用技术类别的讨论在开源中国问答社区上占有第二多的占比.同时,从热度趋势分析中可以看出,通用技术类别下主题的热度(除技术学习主题)基本持平,这表明,无论技术热点如何变化,通用技术的讨论始终如一,表明通用技术对于每一个开发者来说都有学习、掌握的价值.

最具讨论热度的开发工具:Spring 框架.

在主题分析的 20 个结果中,Spring 主题显得格格不入.这是由于 Spring 是其中唯一一个技术工具,在上文的分析中,本文分析了 Spring 作为唯一的技术工具单独成为主题的两个原因:(1) 由于 Spring 相对于其他热门开发工具更新,用户对 Spring 的特性更加陌生,会更多专门地针对 Spring 本身的用法、功能进行讨论.(2) 由于 Spring 相对于其他传统的开发工具,具有更多特性,这些特性会成为 Spring 的特征词出现在文本中,使 Spring 相关的问题更容易凝聚为一类.在热度趋势分析中,Spring 主题的热度在近年来快速增长.

结合以上两点,认为开发者和学习者应该重视 Spring 的系统学习,这是因为 Spring 本身具有大量用户并不熟悉的特性,学习这些特性能够帮助开发者和学习者在 Spring 方面更具竞争力.对于程序员培训的相关从业者或是网站管理者,建议更多开设 Spring 的系统培训.

4.2 启 示

本节说明本文得出的实验结果可对哪些对象起到帮助作用.

对项目开发者与学习者.

对于开发者而言,本文的分析内容可以帮助他们更好地了解技术动向,有助于他们在开发中对技术、平台、工具进行选择,也有助于他们在职业生涯中保持竞争力.开发者可以通过研究问题 4 中关键技术热度趋势的分析,了解到更受欢迎的技术工具或解决方案,为自己项目中技术的选择起到参考作用.例如,对于需要架设服务器的开发者来说,可以通过对服务器主题关键词的分析,了解 nginx、apache 等不同服务器的技术讨论热度趋势,得知 nginx 服务器因其小巧、便捷成为更多开发者的选择,从而为自己服务器的选择提供参考.

对于软件开发技术学习者来说,通过研究问题 1 中主题和其内容的分析,可以了解众多用户讨论的技术主题有哪些以及这些主题的类别,可以形成对软件开发领域知识内容的宏观认识.例如,后端开发类别下有项目部署、服务器配置、后端数据、分布式系统和系统设计这 5 个主题,这些主题是后端开发类别下用户主要讨论的内容.对后端技术感兴趣的学习者,可以根据这些主题来安排自己要学习的内容.通过研究问题 2 中分析技术内容的热度趋势变化,也能帮助这些学习者更好地决定要投入精力学习的技术内容.

在上文给出的重要结论总结中,同样从前端开发、后端开发、通用技术和 Spring 这 4 个方面给出了对于项目开发者和学习者的建议或启示.

对社区管理者.

对于社区管理者来说,本文的研究方法可以帮助管理者更好地获取网站中用户的讨论内容和热度趋势.通过对问答社区主题内容的展示与分析,可以加强社区中的知识传播,吸引更多的用户参与社区讨论.从而使问答社区更加健康,用户体验更好.

对图书出版商与培训机构.

对于图书出版商和培训机构来说,可以结合本文研究问题 4 中每一个具体技术的讨论热度趋势,更多地发行讨论热点的相关工具书或培训班.

也可以结合本文研究问题 3 中的回答情况分析,重点对热度较高但是回答难度较高的技术主题进行系统介绍.例如,在回答难度分析中,数据展示主题的难度最大,同时拥有不低的浏览量,如果出版商与培训机构推出关于数据展示的教学,那么开发者和学习者会由于在网上难以找到合适的答案而购买相关书籍或教程.

4.3 有效性威胁

为了保证实验结果具有有效性、普遍性,本文在这一环节讨论本次课题方法与结论的内部有效性和外部有效性.

4.3.1 内部有效性

内部有效性一般指的是结论的可靠程度.在本次研究中,通过主题模型的方法从实验数据问题中提取出主题,并对主题进行分析.实验过程中影响内部有效性的因素有:原始数据的获取与预处理是否保证正确以及核心算法的正确性.

首先,在数据获取和预处理阶段,本文获取了开源中国问答社区研究的时间范围内的全部有效数据.从而能够保证在数据选择上不存在选择偏见.在预处理阶段,本文使用成熟的中文分词工具:jieba 分词,从而显著降低了数据预处理阶段的有效性威胁.

在核心算法方面,本文采用 LDA 算法,其算法本身的有效性已被众多实验所证明[8,18].同时,这种方法经常被使用于国内外主题提取、主题分析的研究中[1−5],并且得到了众多有价值的成果.因此,本文通过采用 LDA 算法进行主题分析,有效降低了对分析结果的威胁.

4.3.2 外部有效性

外部有效性是指本文研究方法与结论的普适性,是否具有代表性、是否可推广.首先,本文研究数据来源于开源中国问答社区,是中国最大的软件问答社区之一,在中文结果上具有一定的代表性.但在数据量上仍有不足.未来会考虑研究一些其他中文问答社区,以增强结论的普适性和可信度.

5 总 结

随着计算机相关从业者和关注者的增多,越来越多的用户使用软件问答社区来进行技术交流,这使得众多软件问答社区上用户讨论的内容不断积累,这些社区成为了软件开发、软件工程专业知识的仓库.近年来,对于英文软件问答社区的内容分析逐年增多,研究者们通过对这些网站问答内容的分析,了解技术动向以及社区用户的关注点变化.但是,对于中文社区的分析仍十分有限.本文选用开源中国软件技术问答社区,能够通过研究分析中文软件问答社区的数据,来了解国内用户和开发者的关注热点,发现国内计算机行业的技术变化趋势.

本文针对研究对象和研究目标,提出了 4 个研究问题和有针对性的研究方案,分别是:(1) 用户在开源中国问答社区主要讨论的主题有哪些?(2) 这些讨论主题的热度随时间变化的趋势是什么样的?(3) 这些讨论主题的回答情况如何?(4) 用户对特定技术的兴趣是如何随时间发生变化的?

本文首先从开源中国问答社区获取 92 383 个问题,之后通过分词工具和正则表达式对原始数据进行了去噪和预处理.本文采用 LDA 算法从预处理后的数据提取出了 20 个主题.之后对这 20 个主题进行了详细分析.分析出用户在开源中国问答社区讨论的主题可以分为前端开发、后端开发、数据库、操作系统、通用技术和其他 6 个类别.其中,前端开发类别在社区中的讨论占比最大.之后通过热度趋势分析发现,前端开发类别下的主题在 2015 年之后都呈下降趋势,而后端开发下的主题中用户的关注重点从传统的项目部署、服务器配置转移到较新的分布式系统等主题.在回答情况分析中,本文通过对主题进行回答情况的统计,发现数据展示主题回答难度相对较大,数据类型主题回答难度较小.同时,从类别上看,前端开发类别下的主题都相对较难.最后,本文在关键技术分析中分析了一些关键技术的热度趋势.

本文的研究结论对于开发者来说,可以通过对比具体技术的热度变化,让他们更好地选择具体技术以实现

自己的应用.同时也可以通过用户讨论主题的热度趋势变化来了解到最新的技术热点动向,以免错过机会.网站管理者可以通过本文的研究成果更好地分析并展示网站内容,从而加快社区中知识的传播,进而使网站本身对于用户更有价值.

References:

[1] Barua A, Thomas S, Hassan A. What are developers talking about? An analysis of topics and trends in stack overflow. Empirical Software Engineering, 2014,19(3):619–654

[2] Yang X L, Lo D, Xia X, *et al*. What security questions do developers ask? A large-scale study of stack overflow posts. Journal of Computer Science and Technology, 2016,31(5):910–924.

[3] Wang S, David LO, Jiang L. An empirical study on developer interactions in stack overflow.In: Proc. of the 28th Annual ACM Symp. on Applied Computing. Coimbra, 2013. 1019–1024

[4] Rosen C, Shihab E. What are mobile developers asking about? A large scale study using stack overflow. Empirical Software Engineering, 2016,21(3):1192–1223.

[5] Syed A, Mehdi B. What do concurrency developers ask about? a large-scale study using stack overflowIn: Proc. of the 12th ACM/IEEE Int'l Symp. on Empirical Software Engineering and Measurement. Oulu, 2018. NO.30.

[6] Papadimitriou CH, Raghavan P, Tamaki H, Vempala S. Latent semantic indexing: A probabilistic analysis. Journal of Computer and System Sciences, 2000,61(2):217–235.

[7] Hofmann T. Probabilistic latent semantic analysis. In: Proc. of the 15th Conf. on Uncertainty in artificial intelligence. Stockholm, 1999. 289–296.

[8] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3(Jan):993–1022.

[9] Bicego M, Lovato P, Perina A, *et al*. Investigating topic models' capabilities in expression microarray data classification. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2012,9(6):1831–1836.

[10] Li WB, Sun L, Zhang DK. Text classification based on Labeled-LDA model. Chinese Journal of Computers, 2008,31(4):620–627 (in Chinese with English abstract).

[11] Binkley D, Heinz D, Lawrie D, *et al*. Understanding LDA in source code analysis. In: Proc. of the 22nd Int'l Conf. on Program Comprehension. Cyderabad, 2014. 26–36.

[12] Huang B, Yang Y, Mahmood A, *et al*. Microblog topic detection based on LDA model and single-pass clustering. In: Proc. of the Int'l Conf. on Rough Sets and Current Trends in Computing. Berlin, Heidelberg: Springer-Verlag, 2012. 166–171.

[13] Mehrotra R, Sanner S, Buntine W, *et al*. Improving LDA topic models for microblogs via Tweet pooling and automatic labelin. In: Proc. of the 36th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Dublin, 2013. 889–892.

[14] Krestel R, Fankhauser P, Nejdl W. Latent dirichlet allocation for tag recommendation. In: Proc. of the 3rd ACM Conf. on Recommender Systems. ACM, 2009. 61–68.

[15] Zhang L, Pu MY, Liu YJ, Tian JH, Yue T, Jiang J. Investigation of empirical researches in software engineering. Ruan Jian Xue Bao/Journal of Software, 2018,29(5):1422−1450 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5520.htm [doi: 10.13328/j.cnki.jos.005520]

[16] Christoph T, Ohad B, Margaret-Anne DS. How do programmers ask and answer questions on the Web? (NIER Track). In: Proc. of the 33rd Int'l Conf. on Software Engineering. Honolulu, 2011. 804–807.

[17] Venkatesh PK, Wang S, Zhang F, *et al*. What do client developers concern when using Web APIs? An empirical study on developer forums and stack overflow. In: Proc. of the 2016 IEEE Int'l Conf. on Web Services. San Francisco, 2016. 131–138.

[18] Bagherzadeh M, Khatchadourian R. Going big: A large-scale study on what big data developers ask. In: Proc. of the 27th ACM Joint Meeting on European Software Engineering Conf. and Symp. on the Foundations of Software Engineering. Tallinn, 2019. 432–442.

[19] Bangash AA, Sahar H, Chowdhury S, *et al*. What do developers know about machine learning: A study of ML discussions on StackOverflow. In: Proc. of the 16th Int'l Conf. on Mining Software Repositories. IEEE Press, 2019. 260–264.

[20] Griffiths TL, Steyvers M, Tenenbaum JB. Topics in semantic representation. Psychological Review, 2007,114(2):211–244.

[21] Blei DM. Probabilistic topic models. Communications of the ACM, 2012,55(4):77–84.

[22] Teh YW, Jordan MI, Beal MJ, *et al*. Sharing clusters among related groups: Hierarchical Dirichlet processes. In: Advances in Neural Information Processing Systems. 2005. 1385–1392.

[23] Thomas SW. Mining software repositories using topic models. In: Proc. of the 33rd Int'l Conf. on Software Engineering. Honolulu, 2011. 1138–1139.

[24] Cai L, Wang H, Xu B, *et al*. AnswerBot: an answer summary generation tool based on stack overflow. In: Proc. of the 27th ACM Joint Meeting on European Software Engineering Conf. and Symp. on the Foundations of Software Engineering. ACM, 2019. 1134–1138.

[25] Xinogalos S, Pitner T, Ivanović M, *et al*. Students' perspective on the first programming language: C-like or Pascal-like languages. Education and Information Technologies, 2018,23(1):287–302.

附中文参考文献:

[10] 李文波,孙乐,张大鲲.基于 Labeled-LDA 模型的文本分类新算法.计算机学报,2008,31(4):620–627

[15] 张莉,蒲梦媛,刘奕君,田家豪,岳涛,蒋竞.对软件工程中经验研究的调查.软件学报,2018,29(5):1422–1450. http://www.jos.org.cn/1000-9825/5520.htm [doi: 10.13328/j.cnki.jos.005520]

蒋竞(1985－),女,重庆人,博士,助理教授,CCF 专业会员,主要研究领域为经验软件工程,开源软件,基于数据的分析与推荐.

吕江枫(1997－),男,硕士生,主要研究领域为自然语言处理.

张莉(1968－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为软件建模与分析,需求工程,经验研究工程,软件体系结构.