计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200048

工业界需求工程关注点分析

贾经冬　张筱曼　郝　璐　谭火彬
北京航空航天大学软件学院　北京 100083

摘　要　为使理论有效指导实践以提高需求工程质量,了解工业界对需求工程的关注点是十分有必要的。为此,提出了基于数据挖掘的4步研究框架。首先筛选合适的工业界数据源,包括博客类和问答类网站,然后确定合适的关键词以爬取数据并进行清洗,随后根据不同的数据特点进行相似度分析和数据标注处理,最后完成数据统计分析。研究结果表明,国内外工业界对需求工程的关注点各有异同。国内外都关注敏捷需求;国内外都关注用户故事和用例的区别,其潜在反映了实践中传统和敏捷混合开发模式下的需求实践问题;国内外都关注实践中工具的应用,虽然国内使用工具种类多样,但自主开发的工具相对少;国内工业界还关注需求工程的概念和方法,以及需求工程师的职业发展,但国外基本不关注。此外,国内实践中关注需求分析多于需求变更,还关注与需求相关的测试和项目管理领域。该研究结果可有效指引需求工程相关理论在工业界的应用,以解决实践中的难点,并为学术界和工业界提供了可能的研究和发展方向。

关键词:需求工程;工业界;文本相似度分析;数据标注;数据分析
中图法分类号　TP391

Analysis of Focuses of Requirements Engineering in Industry

JIA Jing-dong,ZHANG Xiao-man,HAO Lu and TAN Huo-bin
School of Software,Beihang University,Beijing 100083,China

Abstract　In order to effectively guide theory into practice and further improve the quality of requirements engineering (RE),it is necessary to understand the focuses of RE in industry. To solve this problem,this paper proposes a research scheme with four steps based on data mining. Firstly,suitable data sources are selected,including blogs and Q&A websites. Secondly,suitable keywords are determined,and data related to RE,are crawled and cleaned. Then,according to the characters of different data,text similarity analysis and label data are conducted . Finally,data analysis are done. The research results show that the focuses of RE between domestic and foreign industry have similarities and differences. Both domestic and foreign industries focus on agile requirements,and both concern the difference between user story and use case,which potentially reflects the requirements issue of hybrid development combing traditional with agile in practice. The applications of RE tools are concerned by both,and,although the types of RE tools used in domestic practice are multiple,tools developed by domestic companies are relatively few. The concepts and methods of RE and the career development of requirements engineers are the focuses in domestic industry,but not in foreign industry. In addition,domestic industry pays more attention to requirements analysis than requirements change,and two fields (test and project management) related to RE are also focused on in domestic industry. The research results can effectively guide the application of related RE theory into focuses in industry,so as to solve the difficulties in RE practice,and provide possible research and development directions for academia and industry.

Keywords　Requirements engineering,Industry,Text similarity analysis,Data label,Data analysis

1　引言

需求工程(Requirements Engineering,RE)的概念最初于20世纪70年代被提出,经历了初始的模糊概念阶段,目前RE已逐渐成为一个有着完整的科学理论体系和方法论的软件工程子学科[1]。因为RE是软件工程生命周期的起点,是软件开发后续阶段的基础[2],所以学术界一直很重视对RE的研究,并且提出了一系列的理论和方法以提高RE的质量。从工业界来看,企业也认识到需求分析的好坏对软件开发项目的成败有极大影响,需求分析师的专业知识是软件行业所必需的知识技能[3],但是企业对需求阶段的重视并没有提高到对软件工程生命周期的其他阶段,如开发和测试阶段处于一样的地位。因此,一些RE领域的技术方法仅停留在理论层面,并未在工业界得到充分应用,这表明RE研究在学术界

到稿日期:2020-09-04　返修日期:2020-10-30　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(2018YFB1402600)
This work was supported by the National Key R&D Program of China(2018YFB1402600).
通信作者:贾经冬(jiajingdong@buaa.edu.cn)

和工业界上存在不匹配的现象[4-5]。

理论研究最终是要服务于实际应用的，工业界究竟关注RE的哪些方面是值得关注的，因为其代表着企业在RE领域中急需解决的问题。了解了工业界RE的关注点，才能把相关理论学术成果应用于这些方面，从而更有效地提高企业RE的质量。

虽然现在RE领域的期刊和会议中也有一些行业调研论文，但仅依赖这些行业调研论文来了解业界对RE的实际关注热点类似于管中窥豹，因为发表在学术期刊和会议集上的行业调研论文数量相对较少，而且往往是针对特定行业的。此外，由于学术和工业界的差异，很多需求工程师并不是通过学术论文来表达其对需求领域的关注点和疑惑，而是通过网络社交媒体来表达。

因此，为了更好地了解工业界RE关注热点，本文将筛选和爬取相关网站上与RE相关的数据，通过对数据的分析来挖掘工业界对RE的关注点。通过本文的研究，在了解工业界对RE领域的关注点后，可以指导RE的实践发展，并促进学术成果在这些热点上的应用，缩小学术和工业界的差距，提高RE的实践质量，并为学术和产业界提供可能的研究和发展方向。

2 总体研究方案

本文的研究重点是分析工业界对RE的关注热点，为解决此问题，本文总体研究方案如图1所示。

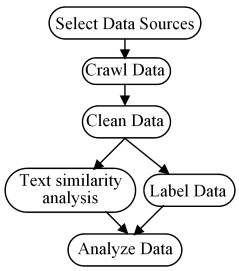


图1 总体研究方案

Fig. 1 Overall research scheme

从图1可以看出，本研究的第一步是选取合适的数据源。现在学术界的文献回顾是基于历来的文献总结研究学术趋势，而工业界的研究热点数据显然不是来自于期刊上的文献，而是来自于各种网络数据。如何从广泛的网络中选取合适的数据源是本研究的关键点，合理地选取数据才能保证后面数据分析的正确性。选定数据源后进行数据爬取，并完成必要的数据清洗，如去掉重复数据等。然后进行数据处理工作，根据不同的数据类型采用不同的数据分析方法，本文包括文本相似度分析和数据标注两种处理方法。最后，基于数据处理结果进行数据分析，以获取工业界RE的关注热点。下面对总体研究方案进行详细阐述。

3 数据源选取

对于工业界RE的相关数据，软件开发人员活跃的社交网站是重要的数据源。依据软件从业者常用网站类型和目前社交媒体的流行度，本文主要考虑两种类型的网站：博客类网站和问答类网站。前者上存在很多软件从业者对某个领域的观点、看法或对实际项目的描述，从后者上可以搜集到从业者对某领域的疑问。不论观点还是疑问都能反映从业者对某领域的关注点。

由于博客类和问答类的网站很多，如何选取数据源是关键。为了更好地全面评估业界情况，数据源选取包括中英文网站。对于中文类网站，我们依据经验选取了用户活跃度高且与软件工程相关的网站。由于本论文关注RE领域，因此确定了如人人都是产品经理(woshipm)这种与需求密切相关的网站作为候选数据源。英文问答类网站相对容易确定，Stack Overflow和Github issue等都是软件开发者常用的问答网站。而对于英文博客类网站，在进行选择之前，首先通过问答类网站及搜索引擎查询国外较好的技术博客网站，然后对其进行调研、分析并选择，最终确定了11个博客类和8个问答类候选网站，如表1所列。

本文设计了包含5个问题的质量评估列表(见表1)来评估候选网站是否可以作为数据源。

表1 数据源选取

Table 1 Data source selection

Blog Website	Q1	Q2	Q3	Q4	Q5	Choose
CSDN	√	√	√	√	√	√
cnblogs	√	√	√	√	√	√
woshipm		√			√	
SegmentFault		√			√	
jianshu	√		√	√		
CodeProject		√	√		√	
Medium		√	√	√		
BAtimes		√	√	√		
GeeksforGeeks				√	√	
GitHub	√		√	√		
Wechart-Official-Account	√	√	√	√		

Q&A Website	Q1	Q2	Q3	Q4	Q5	Choose
CSDN Q&A area			√		√	
q. cnblogs			√		√	
zhidao. baidu	√	√	√	√	√	√
zhihu			√	√	√	
Stack Overflow	√	√	√	√	√	√
Stack Exchange		√	√		√	
GitHub issue	√		√	√	√	
Quora			√	√	√	

Quality assessment checklist	
Q1	Does the website have rich and high-quality data?
Q2	Is the content in website relevant to our topic?
Q3	Is the data original instead of being reprinted?
Q4	Does the website have enough new data?
Q5	Is the data easy to crawl?

Q1关注网站数据是否丰富和高质量。根据搜索后的数据量是否至少有1000条、网站用户是否活跃、搜索结果中明显广告数据和重复数据是否较少来判断。Q2关注网站内容是否与研究主题相关。人工阅读网站前200条搜索数据，若其中RE内容涉及少，水帖或广告数据较多，则该网站被认为是与主题不相关的网站。Q3关注数据的原创性。根据每篇

帖子是否有转载字样或相关链接,或标题是否重复来判断是否为转载,从网站随机抽取 100 篇帖子,如果转载率小于 25%,则认为是原创率较高的网站。Q4 根据网站近期热帖数和回复数来判断网站是否有足够的新数据。Q5 主要关注网站数据是否容易爬取。待选网站反爬机制不能过于复杂,所有相关文章和评论等有效内容均应便于获取,且相关内容需以解析或描述为主,而非代码占主体。例如,微信公众号上 RE 数据丰富且用户活跃度高,但却难以爬取,因此不将其选为最终的数据源。表 1 列出了对 19 个候选数据源的评估选择。最终博客类网站只选择了 CSDN 和博客园(cnblogs)。虽然候选数据源中有国外博客网站,但是通过质量分析最终都舍弃了,因此博客类网站只分析中文网站。事实上 CSDN 的强大数据量是很多研究的数据源,如文献[6]。问答类数据源确定了中文的百度知道和英文的 Stack Overflow,二者是问答系统的典型代表,也有很多研究是基于其上的数据来完成的,如文献[7-8]。

4 数据获取和清洗

确定数据源后,我们使用网络爬虫技术获取相关数据。尽管 2 个博客网站和 Stack Overflow 都是软件开发技术类网站,但是其不只包含 RE 的内容,并且百度知道上涉及的问题领域很广,因此为了获取本论文关注的 RE 领域的数据,我们需要设置合适的搜索关键词在数据源上进行数据爬取。

在设计检索关键词时,为了获取尽可能多的数据,我们从 4 个维度设计搜索关键词。首先从 RE 的概念入手,选取了最相关的两个搜索关键词:需求工程和软件需求,且考虑到敏捷开发的特殊性,增加了敏捷需求关键词。其次,从 RE 阶段考虑,设置了需求分析、需求管理、需求调研、需求评审和需求变更 5 个关键词。然后从需求文档的角度选取了需求文档、需求规格说明书和用户故事 3 个关键词。此外,还设置了其他 7 个搜索关键词。因为本文也想关注企业对需求工具的使用情况,故直接把需求工具设为关键词;需求测试虽然常被用于描述测试问题,但也有不少网络发帖从测试角度提出软件需求的要求,因此也将其用作搜索关键词;根据前面的关键词搜索,所获文章标题中常包含另外 5 个关键词(需求研究、需求问题、需求方法、需求评估、需求设计),因此也选取这 5 个词作为搜索关键词,则共设计了 18 个搜索关键词。

接下来根据搜索关键词进行数据爬取。使用 node.js 语言,基于 Robots 协议进行数据抓取,然后分析网站 DOM 树爬取所需要的数据。通过 http 模块、https 模块和 superagent 模块对网页内容进行请求和获取,使用 cheerio 模块对获取的 html 代码进行解析,将数据存储到 mongoDB 数据库中。除此之外,为了简化代码编写,使用 async 模块将回调函数改写为按照顺序处理的函数。

最终在两个博客类网站上共得到 5910 条数据,每类关键词获取数据量的百分比如表 2 所列。当用 18 个关键词在问答类网站进行搜索时,我们发现不是所有关键词都是合适的,因为有些关键词的检索结果数量不多且质量不高,最终在对百度知道进行数据爬取时只选取了 7 个关键词,在对 Stack OverFlow 进行数据爬取时只选取了 5 个关键词,分别获得 2409 和 2262 条数据,每类关键词涉及数据比例见表 2。每条数据都是一个包括了被分析要素的对象。博客类数据包括文章标题、文章内容、来源网站、作者、日期;问答类数据包括问题标题、问题描述和回答。

表 2 数据爬取结果

Table 2 Results of data crawling

Website	Search Terms and Percentages in Search Results	Amount of Data
CSDN+cnblogs	"Requirements Engineering"−11%," Software Requirements"−11%," Agile Requirements"−10%," User Story"−6%," Requirements Analysis"−9%," Requirements Management"−8%, "Requirements Survey"−5%," Requirements Review"−3%," Requirements Change"−4%," Requirements Document"−6%," Requirements Specification"−4%," Requirements Tool"−2%," Requirements Research"−4%," Requirements Problem"−2%," Requirements Method"−5%," Requirements Evaluation"−3%," Requirements Design"−2%," Requirements testing"−5%	5910
zhidao, baidu	"Requirements Engineering"−22%," Software Requirements"−22%," Agile Requirements"−22%," User Story"−5%," Requirements Analysis"−15%," Requirements Management"−7%, "Requirements Document"−7%	2409
Stack Overflow	"Requirements Engineering"−21%," Agile Requirements"−14%, " User Story"−22%," Requirements Analysis"−21%," Requirements Management"−22%	2262

为获得有效的数据以进行后续分析,需要对已获取的原始数据进行初步清洗。数据清洗主要是删除重复内容和与 RE 主题无关的内容。博客类数据中,转载是重复内容出现的主要原因,因此将博客类数据按照标题进行排序,依据标题判断是否为重复文章或无关内容。问答类数据中重复数据的链接具有相同的问题号,则可通过正则表达式进行匹配、筛选和删除,之后再辅以人工检查来清洗数据[9-10]。最终的数据清洗结果如表 3 所列。

表 3 数据清洗结果

Table 3 Results of data cleaning

Website	Original Article	Irrelevant (Invalid) Article	Duplicate Article	Remaining Article
CSDN+cnblogs	5910	589 (9.97%)	762 (12.89%)	4559 (77.14%)
zhidao, baidu	2409	937 (38.90%)	390 (16.19%)	1082 (44.91%)
Stack Overflow	2262	1329 (58.75)	136 (6.01%)	797 (35.24%)

5 数据处理和分析

获取的博客类数据和问答类数据具有不同的特征，前者属于长文本，而后者属于短文本，因此我们采用不同的分析处理方法。

对于长文本博客，为了获取文章的关注点，采用对文章进行分类和加标签的数据标注方法，从而将非结构化的文章通过结构化的类别和标签进行表示，便于统计分析。问答类数据一般是问题标题文字较少但是回答较为详细，而问答数据的主要关注点是问题本身，因此对问答数据标题进行相似度分析，这是常见的问答类数据分析方法[11]，然后将问题进分类和合并以找到热点问题，这些热点问题代表了工业界 RE 的关注热点。按照这两种数据的分析方法，下面阐述每类数据的分析过程。

5.1 博客类数据分析

对博客类数据进行分析的关键在于合理的数据标注。首先每篇文章被标注为一个类别，如果有必要则进一步对文章进行标签标注。类别和标签都是自定义完成的，类别的确定是根据人工数据清洗过程中对文章概念的理解并随着每个文章的数据标注进展逐步扩展得到的，最终确定了正交的 10 个类别，如表 4 所列，表中给出了每类的含义，并按照每类数据量排序。

表 4 博客类数据类别

Table 4 Blog data classification

ID	Classification	Data Size(percent)	Meaning
1	Requirements engineering focuses	1617 (35.47%)	Strongly related to RE
2	Reading notes	645 (14.15%)	Reading notes on books, blogs, or papers related to RE
3	Areas related to requirements	635 (13.93%)	mainly talking about other areas but also involving requirements
4	Course content	399 (8.75%)	Project development, document writing, and exercise classes required for software engineering courses
5	Project requirements	346 (7.59%)	Practical project requirements
6	Document template	337 (7.39%)	Document template and how to write a RE document
7	Project document	330 (7.24%)	Practical project document
8	Requirements tool	113 (2.48%)	Software tools for requirements management
9	Recommended content	78 (1.71%)	Collection of tools, books, etc.
10	Practical question	59 (1.29%)	Practical work questions

表 4 中排首位的是与 RE 密切相关的，如讨论需求变更、需求分析等的文章，此类被命名为“需求工程焦点”。因其数据量大，之后通过对其添加标签的方法来细化分析。第 2 类是读书笔记，即发表的阅读 RE 相关书籍、博客或论文的笔记。从读书笔记的高占比反映出业界对 RE 学习是重视的。第 3 类命名为“需求相关领域”，此类文章虽然会涉及需求，但是实际关注的是与需求相关的其他领域。在第 10 类中我们也发现了这种现象，例如“没有需求文档的情况下应该如何进行测试”这一问题，虽然标题中有需求文档关键词，但是其实际关注的是测试领域。通过对第 3 类的分析，我们可以更好地了解业界关注 RE 的哪些相关领域。第 4 类是与软件工程课程相关的数据。其不能反映本文关注的工业界内容的主题，因此虽然数据量大(占比 8.75%)，也不再对其进行进一步分析。第 5 类项目需求是讨论实际项目的需求的文章，与本文主题密切相关。第 6 类和第 7 类都与文档相关，但第 6 类是指只讨论需求文档模板的文章，而第 7 类是指就实际项目讨论项目文档的文章。虽然第 8 类需求工具的数据量不大，但它反映了业界使用什么样的工具来管理需求，故也是本文的关注重点。第 9 类推荐内容包括对 RE 领域的工具、书籍等的推荐。虽然部分此类文章中也涉及工具，但是与第 8 类需求工具有区别，二者不重复。第 8 类需求工具偏实践，当文章主体内容是对某 RE 工具使用过程的讲解或记录时，将其归入第 8 类，表示作者使用此工具。当一篇文章不涉及或者较少涉及 RE 工具的使用，重点在于纯粹的工具介绍时，将其归类为推荐内容，表示作者推荐此工具，故推荐内容中涉及的需求工具偏理论。因为本文的研究主题强调工业界实践，故对第 9 类推荐内容也不做深入分析。第 10 类实际问题和本文主题密切相关，但是因为这类文章数量少，且以问题居多，如“产品需求文档应该由产品经理一个人编写还是大家编写”“如何避免需求频繁更改”等，这些都包含在问答类数据中，故本节不对此类数据进行进一步分析。下面对除第 4 类、第 9 类和第 10 类之外的 7 类数据进行重点分析。

5.1.1 需求工程焦点

此类数据量大，因此需进一步进行标签标注。所用标签一方面参考了 RE 书或文章上的相关概念，另一方面参考了相关数据源网站上本身的标签系统。一篇文章可以有多个标签，因此产生了很多标签。为了便于分析，把所有的标签进一步归类汇总，最终定义了 9 类标签，每类的含义和所含标签示例如表 5 所列。

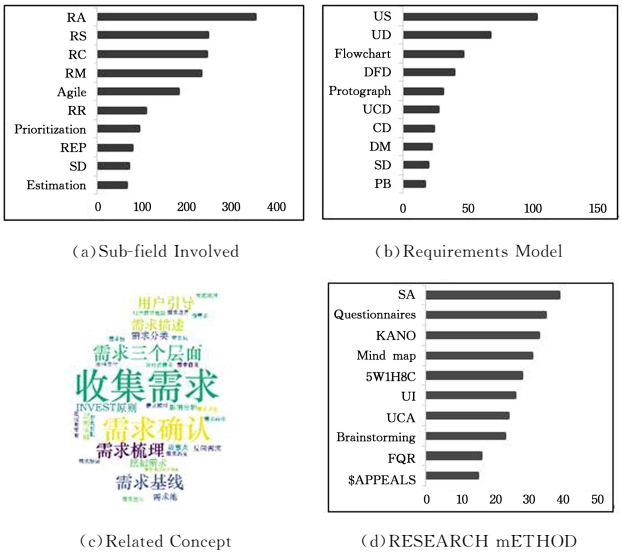
表 5 标签类别

Table 5 Tag categories

Category	Meaning	Tag Examples
Sub-field involved	Sub-fields of RE, mostly defined in books	Requirements analysis, Requirements development
Requirements Model	Models related to RE, mostly introduced in books	Use case diagram, Class diagram, Entity relationship diagram, User story
Related concept	Concepts related to RE, mostly used in industry	Gathering requirements, User description
Research method	Possible methods used in RE	KANO, Mind Map, Structured Method
Comparison	Comparison of two concepts	User story-Use case
Career	Possible occupations related to RE	Product manager, Requirements analyst
Document & plan	Documents or work plans related to RE	Requirements management plan
Meeting	Meeting related to RE	Requirements review meeting, Requirements determination meeting
Other	Some simple but important concepts	User, Function, Performance, Business

接下来依据标签类别对每类数据进行分析。

(1)涉及的子域。此标签主要包括在 RE 书中定义的很多子域概念，如需求分析、需求开发等与 RE 过程相关的概念，也包括软件工程的相关概念，如敏捷、软件设计等。经过统计，排名前 10 的子领域如图 2(a)所示。需求分析(Requirements Analysis，RA)、需求调研(Requirements Survey，RS)、需求变更(Requirements Change，RC)和需求管理(Requirements Management，RM)是整个 RE 过程中业界较关心的子领域，这与实际情况也是吻合的，因为需求变更是业界常见现象，如何做好需求管理则很重要。此外，敏捷也是常见的关注子领域，这是因为敏捷开发是现在的主流开发模式，所以讨论敏捷需求的较多。其他还包括需求评审(Requirement Review，RR)、优先级排序、需求工程过程(Requirement Engineering Process，REP)、软件设计(Software Design，SD)和估算。



(a)Sub-field Involved (b)Requirements Model

(c)Related Concept (d)RESEARCH mETHOD

图 2 涉及的子域、需求模型、相关概念和研究方法的数据分析

Fig. 2 Data analyses of sub-field involved, requirements model, related concept, and research methods

(2)需求模型。此类标签一般指 RE 理论中的需求描述模型，包括 UML 模型和敏捷需求建模概念，涉及的前 10 个标签如图 2(b)所示。可以看出，敏捷需求模型——用户故事(User Story，US)出现最频繁，排名第 10 的产品待办列表(Product Backlog，PB)也是敏捷需求中的相关概念，这再次体现了敏捷开发的流行性，因此敏捷 RE 备受关注；其次是用例图(Use Case Diagram，UD)、流程图、数据流图(Data Flow Diagram，DFD)，这些都是需求阶段常见的建模图形；原型图和用例说明(Use Case Description，UCD)也与 RE 密切相关。值得注意的是，类图(Class Diagram，CD)和时序图(Sequence Diagram，SD)并不是 RE 阶段的常用模型，但是也受到了关注，领域模型(Domain Model，DM)反映了现在业界对领域软件开发的重视。从这 10 类需求模型标签可以看出，传统结构化开发需求描述的数据流图仍然在实际中使用，但是提及率远低于面向对象开发的用例模型和敏捷中的用户故事。

(3)相关概念。此类别的概念虽然在 RE 理论中也会出现，但更多的是指 RE 实践过程中涉及的概念，尤其是业界定义的概念。有些概念在实际中叫法不统一，因此分析时对本质一样的概念进行了合并，例如将“收集需求”和“需求收集”合并为一个概念。最终涉及的相关标签词云如图 2(c)所示。显然，收集需求、需求确认、需求梳理、需求基线、需求三个层面、需求描述和用户引导为使用较多的概念。这里需求的收集、确认和梳理反映了业界关注的仍然是需求分析的过程。需求描述使用的范围较广，一般表示描述需求的方法，包括文档。

(4)研究方法。其主要涉及 RE 中使用的方法，包括理论和实践中的方法。前 10 个关注的热点方法如图 2(d)所示。可以看出，提及较多的研究方法是需求建模的结构化分析法(Structured Analysis，SA)，其次是需求获取的问卷调查法和用户访谈(User Interview，UI)。头脑风暴(Brainstorming)也是常见的获取用户需求的方法。这些方法在 RE 教材中很常见。但是，业界实践中也用到了很多 RE 书上不常提及的方法，如思维导图(Mind map)是业界常见的梳理用户需求的方法，四象限法则(Four Quadrant Rule，FQR)和 KANO 模型是业界确定需求优先级的方法，这两种方法受关注和涉及的子域包括优先级是统一的。用例分析(Use Case Analysis，UCA)方法说明需求描述中的用例描述仍然受到关注，这与需求模型中用例模型受到关注是一致的。5W1H8C 法则是受到广大业界认可的来自《面向对象葵花宝典》一书中提及的需求分析方法，5W 指 when，where，who，what，why，即需求在什么时候、什么地方使用，谁来使用，用户怎么做，对用户的价值是什么。1H 指 how，即采用用例方法来分析需求。8C 指需求的 8 个约束，包括性能、成本、时间、可靠性、安全性、合规性、技术性和兼容性。$APPEALS 是 IBM 公司总结出来的客户需求市场定位的分析方法，涉及此类标签的文章主要关注产品型软件的市场需求。

其他类别的标签涉及的数据总量较少，在此只列举涉及的各类标签的主要分析结果。

(5)对比。此类主要关注两个不同概念和方法的对比，最受关注的是用户故事和用例的对比，主要讨论二者对需求描述的不同。

(6)会议。此类关注 RE 中涉及什么样的会议，其中需求评审会最受重视。

(7)文档和计划。此类主要关注 RE 中有哪些文档类型，主要关注需求规格说明书和需求管理计划。

(8)职业。此类关注与 RE 相关的职业，产品经理是最受关注的职业，需求分析师、开发人员和项目负责人职业数量相当，处于第二梯队。

需求工程焦点包含大量的标签，将所有的标签绘制词云图，可得到图 3。从中可以看出，业界主要关注焦点集中在 3 个方面。首先是对 RE 过程的关注，尤其是对需求开发阶段的关注，体现为高频词：需求分析、需求调研和收集需求；需求管理阶段也受到关注，但此阶段主要的高频词是需求变更，需求状态变化和需求跟踪很少被提及，这可能是因为业界习惯

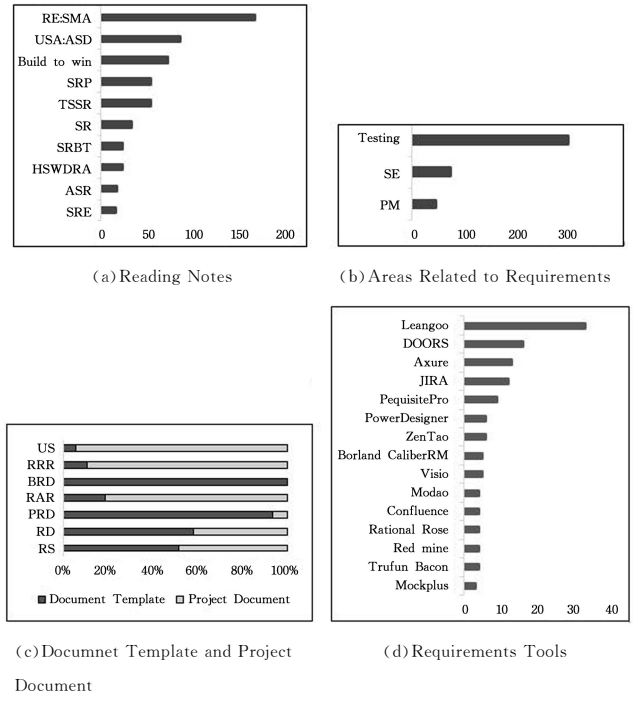
用需求管理一词涵盖此阶段的所有 RE 活动，也就是说，实践中也许不需要像学术界那样细化需求管理活动。其次，虽然敏捷的词频并不高，但是用户的词频很高，这可能是实践中用户故事的反映，而且其也很好地表明了现在的 RE 内容以用户作为核心的特点。第三，可以观察到产品经理、需求分析师、项目负责人等高频词，这说明业界对 RE 中角色和职业的关注。此外，管理、沟通、敏捷、文档这些标签关键词也表明业界 RE 正在向敏捷化、管理规范化、文档规范化以及更加以人为本的方向发展。

图 3 需求工程焦点词云图

Fig. 3 Word cloud chart of requirements engineering focuses

5.1.2 读书笔记

读书笔记指阅读 RE 相关书籍、博客或论文的笔记，这表明很多从业者通过读书的方式学习 RE 的相关内容。对于此类文章，结合第 9 类推荐内容中给出的书籍推荐，我们得到前十大热门读物，如图 4(a)所示。按照涉及文章数量 9 本书依次为：《需求工程：软件建模与分析》(Requirements Engineering: Software Modeling and Analysis, RE: SMA)、《用户故事与敏捷方法》(User Stories Applied: for Agile Software Development, USA: ASD)、《构建之法》(Build to win)、《软件需求模式》(Software Requirements Patterns, SRP)、《软件需求十步走》(Ten Steps of Software Requirements, TSSR)、《软件需求》(Software Requirements, SR)、《软件需求最佳实践：SERU 过程框架原理与应用》(Software Requirements Best Practices-Principles and Applications of SERU Process Framework, SRBT)、《敏捷软件需求：团队、项目群与企业级的精益需求》(Agile Software Requirements-Lean Requirements Practices for Team, Program, and the Enterprise, ASR)和《软件需求工程》(Software Requirement Engineering, SRE)。值得注意的是，《构建之法》并不是专门讲需求的书籍，但阅读量却很高，通过对涉及此书的文章分析了解到该书被一些学校的软件工程课作为教材，学生需要阅读并做笔记，故相关数据量高。对于这 9 本高阅读量的书籍，仅有 2 本与敏捷需求相关且都是外国作者的作品。此外，《我们应当怎样做需求分析》(How should we do requirements analysis, HSWDRA)是阅读量和引用量都非常高的一篇博客文章，经过分析发现，此博客也是一些课程要求的课后读物。此外，第 1 类需求工程焦点中研究方法提及的 5W1H8C 法则的出处《面向对象葵花宝典》并不在读书笔记这一类中，这主要是因为该书包含内容较多，RE 内容只占其中一章。



(a) Reading Notes (b) Areas Related to Requirements

(c) Documnet Template and Project Document (d) Requirements Tools

图 4 读书笔记、需求相关领域、文档和项目模板、需求工具的数据分析

Fig. 4 Data analyses of reading notes, areas related to requirements, document template and project document and requirements tool

5.1.3 需求相关领域

与需求相关的前 3 个领域如图 4(b)所示。可以看出最相关领域是测试，有 200 多篇文章关注需求测试，其次是讨论软件工程(Software Engineering, SE)领域，RE 作为其一部分而被讨论，此外讨论项目管理(Project Management, PM)时涉及 RE，这与第 1 类中相关子领域的估算子领域是吻合的，需求估算属于项目管理的范畴。这个类别的结果表明，业界做需求分析时不能忽视其与测试和项目管理的关系。

5.1.4 项目需求

项目需求是对实际项目的描述，占比 7.59%，采用同第 1 类需求工程焦点相同的标签对此类进行标注，且对此类标签进行词云分析，如图 5 所示。

图 5 项目需求类标签词云

Fig. 5 Word cloud chart of project requirement tags

项目需求类最能代表业界在具体项目中关注的 RE 内容，从图 5 可以看出，需求分析是此类关注频率最高的标签，与图 3 中第 1 类需求工程焦点词云的分析结果一致。但从图 5 看出，项目需求类中需求模型类的标签占比明显较大，如图中高频的用例图、原型设计、流程图、用户故事等，而需求工程焦点类中占比较大的是所涉及的子域标签类别，如图 3 中的高频词需求调研、需求变更等。此差异反映了 RE 的理论和实践差异：尽管现实中需求变更和需求管理很重要，理论上也非常强调这两个方面，但实践中可能得到了有效应对，因此不如理论上那么强调，相反实践中更关注具体的需求描述模型。对比图 3 和图 5 还可以看出，高频词“功能描述”可以看成实际项目中最常用的需求方法之一，但完全没有出现在需求工程焦点的热点标签中，而需求工程焦点中的诸多需求模型用例图和用户故事是在项目中主要使用的模型。

5.1.5 文档模板和项目文档

因为文档模板和项目文档两类分别可代表理论研究和实际应用中的需求文档，且二者涉及的文章数量也基本相同（分别为 337 和 330），所以将二者做对比分析，结果如图 4(c)所示。可以看出，理论角度关注了 7 种文档模板，但实际项目中并不存在商业需求文档（Business Requirement Document，BRD）。对于实际项目中涉及的 6 种文档，需求规格说明书（Requirements Specification，RS）和需求文档（Requirements Document，RD）在理论和实践中的受重视程度差不多。事实上，这两种文档都是重文档的传统开发模型下的典型文档，企业如果采用传统开发模式，那么一定会重视需求规格说明书的撰写。而产品需求文档（Product Requirements Document，PRD）虽然作为模板给出，但实际中真正使用此文档的很少。与之相反的是，用户故事（US）、需求调研报告（Requirements Research Report，RRR）和需求分析报告（Requirements Analysis Reports，RAR）则是实践中应用远多于理论研究的文档。用户故事文档在实践中的广泛应用再次说明了敏捷开发在实践中非常流行。

5.1.6 需求工具

需求工具类数据量虽然只占博客类数据的 2.48%，但它是我们关注的一个重要方面。其数据量少，一方面是因为大部分需求工具的使用需要结合具体项目来描述，另一方面只有全篇都在描述需求工具使用的文章才被归为此类。结合第 9 类推荐内容中的推荐工具，最终涉及数据量排名前 15 的需求工具如图 4(d)所示。排名前 3 的需求工具分别是 Leangoo，DOORS 和 Axure。这 3 种工具分别代表了 RE 实践中应用的 3 种类型工具。第 1 种是项目管理工具，如 Leangoo、JIRA、禅道（ZenTao）和 Redmine，这些工具中包含需求管理功能，并且前两个工具主要应用于敏捷项目。值得关注的是，这类软件中的 3 个（Leangoo，ZenTao 和 Redmine）是国内公司开发的。第 2 种是专门的需求管理工具，如 IBM 的 DOORS，RequisitePro，Borland CaliberRM 和 Trufun Bacon，仅最后一个软件是国内公司开发的。第 3 种是需求的原型设计工具，如 Axure、Mockplus 和墨刀，后两个软件都是国内软件公司开发的。此外，涉及的 RE 工具还包括需求文档编写时的辅助工具，如 Visio 和 Rational Rose 可以实现用例图绘制。Confluence 实际是知识管理和协同软件，可以实现团队成员之间的信息共享和文档协作，在 RE 中也有所使用。PowerDesigner 在 RE 中更多地是用于制作数据流程图。由此看出，RE 实践中工具类别较丰富，但使用较广的并不多。

综合以上分析，博客类网站 RE 关注点具有以下特点：各种 RE 概念、方法和领域都有所涉及，但敏捷开发的需求工程在各个方面都有体现，占主体；从 RE 过程看，对需求开发的关注高于需求管理；注重从测试和项目管理角度讨论需求；对 RE 的相关职业也有较多探讨；此外，RE 实践中关注了多种类型的工具。

5.2 问答类数据分析

问答类数据来自中文的百度知道和英文的 Stack Overflow。百度知道是一个包含领域广泛的问答网站，根据问题标题进行相似度分析。Stack Overflow 是软件开发领域的技术问答网站，其标题中包括了问题所属领域，故不需要进行相似度分析就可直接进行领域分析。本文对两个网站的数据分别进行了分析，以对比中英文工业界热点。

5.2.1 百度知道数据分析

百度知道数据经过清洗后有 1 082 条，把所有数据的标题从数据库读取，使用 Python 的第三方工具包 gensim 对问题标题建立数据字典，以 0.5 为阈值进行相似度分析，此步处理后数据共 844 类。然后对初步分类结果进行人工分析，把相似结果进一步合并以了解问答类数据实际关注的热点问题。问答类每条数据中出现 2 个或 2 个以上问题的情况较少，但如果遇到一条数据包含多个问题时，则根据含义将其分割为多个问题然后分别进行归类。人工分析中，凡是与 RE 产业实践无关的问题则直接删除，例如“软件工程的目标、性质、内容是什么？作业题求高手回答”，这种显然是学校作业的问题，直接删除。最终得到的百度知道的十大热点问题如表 6 所列，表中频率表示与此问题相关的原始数据量。

进一步，把经过相似度分析后的所有问题总结分类，将所有问题分为 5 类：第 1 类是需求分析问题，如表 6 中问题 1，6，9，即实践中大部分从业者将 RE 等同于需求分析，提到需求工程时最关心的是需求分析，这与图 3 和图 5 中将需求分析作为高频词是一致的；第 2 类是 RE 和 SE 的概念问题，如表 6 中的问题 5 和 8，表明有些从业者对 SE 和 RE 的相关概念有困惑；第 3 类是需求文档的编写和询问需求文档的相关内容，如表 6 中的问题 2 和 4，表明很多从业者不会编写 RE 文档；第 4 类是需求管理问题，如表 6 的问题 10，说明需求管理是 RE 领域除了需求分析之外的第 2 个热门领域，但问题数量远低于需求分析相关问题；第 5 类是需求工程师的相关问题，如表 6 中的问题 3 和 7，说明从业者关心 RE 的职业发展。从百度知道的数据分析结果可以看出，大部分国内从业者对 RE 的关注问题仍停留在基本概念层面，现有理论可以指导如何撰写文档和实施需求分析等基本问题；实践中对具体 RE 问题的探讨则很少，关注点主要集中在需求分析和需求管理两个子领域。其次，大家对 RE 的职业发展也非常关注，对这个职业的关注也说明业界现在对于 RE 的重视。

表 6 百度知道的十大热点问题

Table 6 Top 10 hotspots of zhidao.baidu

Number	Title	Frequency	Classification
1	How to do requirements analysis?	45	Requirements Analysis
2	How to write the requirements document (or requirements specification)?	41	Requirements Documents
3	What are career development, job content and job requirements of requirements engineer/Analyst?	34	Requirements Engineers
4	Ask for requirements document (or requirements specification)	23	Requirements Documents
5	What is certain definition and classification of software requirements?	22	Concept of RE & SE
6	What is included in requirements analysis and what is the final output?	21	Requirements Analysis
7	What are the benefits and advantages of requirements engineers (or requirements analysts)?	19	Requirements Engineers
8	What are the stages involved in software development?	19	Concept of RE & SE
9	How to write requirements analysis?	15	Requirements Analysis
10	How to manage requirements and how to do well?	13	Requirements Management

5.2.2 Stack Overflow 数据分析

Stack Overflow 的问题标题中包含了问题所属领域信息，例如，“architecture-How to get correct user stories in BDD?”表示此问题与架构相关，因此充分利用标题领域对清洗后的 797 条数据进行人工分析。首先对问题标题进行排序，并依据标题领域，直接选择明显多次出现的与 RE 相关的领域，然后从全部标题中筛选与前面被选领域同领域的内容加入进去，同时注意是否有新的领域出现。因为数据清洗时已多次观察问题标题，故此处分析虽然是人工处理，但结果相对准确、无遗漏。通过多次筛选与分类，最终得到十大热点问题如表 7 所列，且所有问题可以归纳为表中的 4 大类。

表 7 Stack Overflow 热点问题

Table 7 Stack Overflow hotspots

Number	Question	Frequency	Type
1	How to create, modify, export, retrieve user stories by Rally?	51	Rally
2	How to create, modify, export, retrieve user stories by TFS?	46	TFS
3	Other questions about operations and use of TFS software	44	TFS
4	Questions about user story in agile	42	Agile and Scrum
5	Questions about organizational structure, team and other practical issues in agile development	42	Agile and Scrum
6	Conceptual issues such as the definition of agile and the differences between agile with other development models	39	Agile and Scrum
7	Questions about use of Rally and its API interfaces with other software	36	Rally
8	Questions encountered during the use of Rally	13	Rally
9	Practical questions about UML design and modeling	12	UML
10	Theoretical questions such as UML concepts, differences, changes, etc.	12	UML

从表 7 可以看出，国外 RE 实践中关注的问题首先是 RE 工具的使用问题，尤其是两个项目管理软件 Rally 和 TFS 的使用问题。值得说明的是，博客类需求工具中提及的 Jira 和 DOORS 两个工具的使用在 Stack Overflow 中也有提及，但是数据量相对较少。这说明了国外从业者关注如何用工具来进行需求管理。其次，国外从业者对敏捷相关的内容关注较多，如表 7 中问题 4,5,6 所反映的，事实上，表 7 中的前两个问题是关于软件中用户故事的问题，本质上也是敏捷 RE 的问题。这些都反映了敏捷开发模式的流行以及业界关注如何做好敏捷开发中的需求。最后一个方面是业界对 UML 的关注，关注其概念和建模实践，这与博客类数据分析中用例分析的关注是一致的，也意味着很多从业者用 UML 进行 RE 的建模。

由于 Stack Overflow 上问题不仅有标题，还有详细的问题描述，为了使分析更加完整，本文进一步通过 Python 中的 NLTK 库对 Stack Overflow 上问题的标题和问题描述进行分词和词形还原，然后根据词性进行名词的词频统计，结果如图 6 所示。从图中可以看出提及最多的是 story，其可能是问题描述中 stories 的词性还原，主要来源仍然是 user story，而且图中 user story 也是高频词，因此可以推断，在 Stack Overflow 上 RE 更多的是探讨敏捷中的用户故事的问题。这个结果与表 7 是一致的。

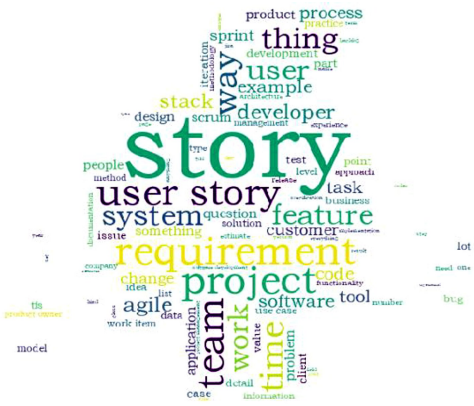


图 6 Stack Overflow 数据词云

Fig. 6 Data word cloud of Stack Overflow

综合表 7 和图 6，Stack Overflow 上关于 RE 基础概念的问题很少，但是关于 RE 工具的问题较多，一定程度上反映了国外 RE 实践中工具使用的广泛性。此外，国外工业界 RE 更关注敏捷需求。

5.2.3 国内外问答类数据对比分析

问答类问题更代表了业界对 RE 的关注点,对比百度知道和 Stack Overflow 上的数据,可以发现国内外问答类数据既有相同点也有不同点。

一个显著的共同点是它们都关注用户故事和用例的关系,两个问答类网站出现了相同的问题,中文的“user story 和 use case 有什么区别”和英文的 “Difference between use case and user stories?”这说明国内外从业者都关注两种需求描述形式的异同。但国内外问答数据也有很多不同点,主要包括以下 4 个方面:

(1)百度知道上几乎没有需求管理软件的操作问题,而 Stack Overflow 上有很多这类问题,这可能表明国内 RE 实践中工具应用的普遍性相对较弱。

(2)中文网站上有许多与需求分析师职业相关的提问,但英文网站上几乎没有。这可能反映了国外此职业发展得更成熟,故国外从业者关于此职业发展的疑惑较少。其也可能反映了国内软件行业和 RE 的快速发展。软件开发项目规模变大和团队规模变大,使得团队中角色更加细分,更多从业者担任需求分析师,但此角色不如程序员、测试员的职业发展定位明确,故受到关注。

(3)中文问答网站上关于敏捷开发中的 RE 问题远少于英文问答网站。这可能说明国外软件企业中敏捷方法的普及性更高。

(4)中文问答网站上有很多基础性的 RE 概念性问题,英文问答网站上则很少。这可能表明国内 RE 职业的新生力量远大于国外,因为新从事 RE 职业的从业者可能会有这种基本概念问题;其次也可能表明国内 RE 发展的成熟度不够,因此从业者会有一些基础性的问题。

5.3 综合分析

综合博客类和问答类数据分析,可以看出产业界对 RE 的关注热点,以及对学术界和产业界的启示。

(1)国内外都关注 RE 实践中工具的使用

国外实践中主要使用两个敏捷项目管理工具。国内实践中使用的工具多样,包括 3 类功能的工具(专门的 RE 软件、项目管理软件和原型设计软件)。工具功能的多样性固然可以满足 RE 实践的不同方面,但是也会弱化每种功能工具的使用市场。因为 RE 只是软件工程的一部分,而且正如前文所分析,RE 与测试及项目管理都有关联,所以把多种功能集成在一种工具上,不仅能满足 RE 实践需要,而且能够支撑软件开发全过程,这可能是工具开发的一个思考方向。其次,尽管国内 RE 实践中涉及的工具很多,但是由国内公司开发的 RE 实践工具数量相对较少(6/15),虽然 3 类工具都有国内开发的,但却没有占主导地位的工具。在当今时代背景下,需要加强国内自主的 RE 工具开发和集成,使其做强做大,以更好地满足国内以及国外市场的需要。事实上,据了解,国内大型软件公司有能够很好地满足软件开发全过程的工具,但是如何使这些工具从满足内部需要到大量普及甚至国外应用,值得相关业界思考。

(2)国内外都关注敏捷需求

数据分析中用户故事不仅作为项目文档使用频率高,而且用户故事建模和敏捷子领域都受到业界关注。这和敏捷开发日益成为主流开发模式有密切关系。但这并不意味着传统开发模式就完全不被关注,UML 模型,尤其是用例图仍然很受业界关注。

用户故事和用例为敏捷和传统两种开发模式下的需求描述方式,原则上从业者对二者的区别没有疑惑,因为在一种开发模式下只需要关注对应的需求表示方法即可。但是,博客类对比标签和国内外问答类数据都反映了业界很关注用例和用户故事的区别,这可能反映了两个实践中的问题。第一,当企业从传统开发模式转换为敏捷开发时,企业中的开发者可能会面临如何从用例描述转化为用户故事的问题,理解二者的区别有助于开发者尽快适应这种转变。第二,实践中存在混合开发模式,即把需求的用例建模集成到敏捷开发上。文献[12-13]也存在对这种混合模式的方法和实践的描述,而且根据最近的一篇敏捷 RE 的综述文献[14]可知,虽然用户故事在敏捷 RE 中仍然占主体,但是用例也在敏捷 RE 中被使用,并提出了敏捷用例的概念。这种混合模式可能会使得从业者关注用例和用户故事的区别。理清二者的区别并实现两种需求描述方式的转化可能有助于业界在这种混合模式下的实践。从学术角度来看,混合开发模式下的各种问题可能值得研究。

此外,尽管敏捷 RE 受到关注,但是根据读书笔记数据分析,业界关注的读物中只有两本与敏捷需求相关,且都是国外著作,国内著作仍然集中于传统 RE。这个结果从侧面反映出学术界应该积累出版敏捷 RE 的相关读物以适应当今产业界的发展,并为实践做出指导。

(3)从 RE 过程来看,数据分析结果表明国内产业界更关注需求分析

需求分析领域有众多的学术研究,相对来说学术界认为需求变更是软件项目的一个重大问题。但是,产业界对需求变更的关注度并不如需求分析高。这可能因为盛行的敏捷开发拥抱需求变化或企业内部流程有效解决了变更,使企业面对需求变更时不再谈虎色变,所以实践中的需求变更可能并不如学术界渲染的那么严重。既然产业界更关注需求分析,从学术角度可以更充分实证调研业界需求分析的痛点,把更多的理论上可行的需求分析方法应用到业界,或者提出新的方法来解决业界的需求实践困扰。

(4)国内企业并不仅仅是孤立地关注需求,还关注需求的关联领域

测试和项目管理是与需求工程关联密切的两个受关注的子域。这对 RE 从业者的启示包括两个方面:首先,RE 实践中要尽量考虑全面,不能遗漏关联的问题分析;其次,如果想扩展自己的职业知识和技能,从这两个关联的领域入手则可能比找一个全新子域入手更容易。

(5)国内业界很多人讨论 RE 的基本概念、定义和方法,关注需求工程师的职业发展,这从侧面反映出相关教育中对 RE 的关注和普及不够,应尽早填补此缺陷,使得 RE 发展得更成熟。

结束语 本文旨在研究工业界对 RE 的关注热点，为解决此问题，首先筛选确定了 4 个网站作为数据源，然后选取合适的关键词从数据源上爬取 RE 相关数据，经过数据清洗后，根据不同数据源的特点采用了不同的数据处理方法，包括相似度分析和加标签的数据标注，最后对处理后的数据进行统计分析。研究发现，国内外工业界对 RE 的关注点既有相同点也有不同之处。

国内外都很关注敏捷开发中的 RE 问题，尤其是用户故事，而且都对传统和敏捷开发中需求描述方法的异同点很关注，即关注用户故事和用例的区别。这可能是混合模式开发下或模式转变开发下软件从业者的一个困惑，其反映了 RE 发展中不同方法的衔接问题。在敏捷盛行的情况下，这个关注热点为学术界和产业界提供了一定的启示和研究方向，如何帮助企业实现不同需求描述的转化是需要解决的问题。此外，国内外都关注需求工具的使用，但是国内外 RE 实践工具的应用情况有差别，如何做大做强国内自主开发的 RE 工具是国内产业界需要考虑的。最后，国内对 RE 的概念、定义和方法的关注远高于国外，这潜在说明了未来应重视国内 RE 的教育和行业发展。

参 考 文 献

[1] GREGORY S C, TERZAKIS J. Viewpoint: effectiveness of focused mentoring to improve requirements engineering industrial practice[J]. Requirements Engineering, 2017, 22(3): 1-5.

[2] WU HG, WU G Q, CHEN S, et al. A software behavior oriented requirements model and properties verification[J]. Journal of Computer Research and Development, 2011, 48(5): 869-876.

[3] HERTZ K, SPOLETINI P. Are requirements engineering courses covering what industry needs? a preliminary analysis of the united states situation[C]// IEEE International Workshop on Requirements Engineering Education & Training. Banff, AB: IEEE, 2018: 20-23.

[4] SIKORA E, TENBERGEN B, POHL K. Industry needs and research directions in requirements engineering for embedded systems[J]. Requirements Engineering, 2012, 17(1): 57-78.

[5] AMBREEN T, IKRAM N, USMAN M, et al. Empirical research in requirements engineering: trends and opportunities[J]. Requirements Engineering, 2018, 23(1): 63-95.

[6] GAN T, LIN F, CHEN C, et al. User behaviors analysis in website identification registration[J]. China Communications, 2013, 10(3): 76-81.

[7] ABDALKAREEM R, SHIHAB E, RILLING J. On code reuse from StackOverflow: An exploratory study on Android apps[J]. Information and Software Technology, 2017, 88: 148-158.

[8] ZHU Z X, ZOU Y Z, HUA C Y, et al. Mining and organizing software functional features based on stackoverflow data[J]. Journal of Software, 2018, 29(8): 2210-2225.

[9] YOU F C, GONG H C, GUAN X X, et al. Design of data mining of wechat public platform based on Python[J]. Journal of Physics: Conference Series, 2018, 1069: 012017.

[10] DENG L. Innovative application of python in data crawling-chinese version of movie recommendation platform[J]. Journal of Physics Conference Series, 2019, 1168(3): 032083.

[11] HALKIDI M, SPINELLIS D, TSATSARONIS G, et al. Data mining in software engineering[J]. Intelligent Data Analysis. 2011, 15(3): 413-441.

[12] YAO J W. Automated sentiment analysis of text data with NLTK[J]. Journal of Physics: Conference Series, 2019, 1187(5): 052020.

[13] XU C, LIU D. Chinese text summarization algorithm based on Word2vec[J]. Journal of Physics Conference Series, 2018, 976(1): 012006.

[14] ZHANG W, LIU T, YANG Y, et al. A topic clustering approach to finding similar questions from large question and answer archives[J]. PLoS One, 2014, 9(3): e71511.

JIA Jing-dong, born in 1975, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include software engineering and machine learning.