◆特邀栏目◆

基于大数据技术的AI岗位需求分析研究*

徐正丽[1],文博奚[2],谢梅英[3],蔡　翔[1**]

(1.桂林电子科技大学,广西桂林　541004;2.广西建设职业技术学院,广西南宁　530007;3.南京信息工程大学,江苏南京　210044)

摘要:近年来,我国人才市场出现供需失配的结构性矛盾,尤其是在人工智能(AI)领域。准确感知并描述劳动力市场的需求是解决该问题的重要手段。本研究首先使用网络爬虫抓取智联招聘网站发布的AI岗位相关招聘信息,通过中文分词、K-means等大数据分析方法对招聘岗位名称进行聚类处理,识别出软件工程师、算法工程师、产品经理及产品架构师等4个岗位簇;然后利用概率主题模型(Latent Dirichlet Allocation,LDA)对招聘岗位要求继续进行聚类处理,得到数据库、机器学习、模式识别、大数据、程序设计等5个技能集;最后利用LDA求得岗位簇对其技能集的需求矩阵,以分析各岗位簇对其岗位技能的需求程度。结果表明:程序设计能力对AI软件工程师最重要,模式识别的理论与技术对算法工程师最重要;产品经理岗位对数据库、机器学习和大数据技术等均有较强的技能需求;机器学习的理论与技术对产品架构师最重要。本研究成果可为高校、企业常态化或实时准确感知并描述AI劳动力市场需求提供技术支持。

关键词:数据分析　人工智能　网络爬虫　岗位角色　岗位技能　岗位词典

中图分类号:TP18　　**文献标识码**:A　　**文章编号**:1005-9164(2021)03-0321-09

DOI:10.13656/j.cnki.gxkx.20210830.003

0　引言

近年来,我国人才市场出现供需失配的结构性矛盾,尤其是在人工智能领域。准确感知并描述劳动力市场的需求是解决该问题的重要手段。人工智能(AI)技术已成为全球新一轮科技革命和产业变革的着力点,对于推动产业转型升级至关重要,越来越多的公司把AI视为竞争力的关键要素[1]。根据2017年Gartner的统计显示,到2021年,AI预计将创造230万以上相关岗位,但人才缺口却非常严重[2]。由于AI是应用领域非常广泛和快速发展的新技术[3],人力资源管理部门对AI领域的专业认知更新却比较缓慢,对AI岗位职责及所需技能的认知往往是模糊、主观和过于简化的理解[4],甚至会将“AI”与“大

收稿日期:2021-04-02

*国家自然科学基金项目(71463010),教育部人文社会科学研究专项(17JDGC022)和江苏省研究生教育教学改革课题重点课题(JGZZ19_021)资助。

【作者简介】

徐正丽(1982-),女,博士研究生,讲师,主要从事数据统计、商业管理研究。

【**通信作者】

蔡　翔(1968-),男,博士,教授,博士生导师,主要从事大数据、商业管理研究,E-mail:ross8866@126.com。

【引用本文】

徐正丽,文博奚,谢梅英,等.基于大数据技术的AI岗位需求分析研究[J].广西科学,2021,28(3):321-329.

XU Z L,WEN B X,XIE M Y,et al.Research on AI Job Demand Analysis Based on Big Data Technology [J].Guangxi Sciences,2021,28(3):321-329.

数据”“机器学习”“深度学习”等概念混为一谈[5]。AI 岗位内容的广泛性及所需工作技能的复杂多样性[6,7]给准确把握 AI 岗位的需求带来很大的挑战。

为准确感知并描述劳动力市场对 AI 的需求,本研究采用大数据分析手段,对 AI 岗位簇的工作角色及所需技能进行类型学研究,为基于大数据分析 AI 岗位簇的角色及其所需技能需求提供了一个结构化框架,可有效提升人力资源管理部门的科学决策水平,同时促进高校提高 AI 人才培养的针对性。

1 算法框架

本算法主要包括 4 个部分:第一步,使用网络爬虫技术从招聘网站爬取 AI 相关岗位的招聘信息,然后实施数据清洗;第二步,利用 K-means 聚类与专家判断相结合的方法,分析 AI 的岗位簇;第三步,利用概率主题模型(Latent Dirichlet Allocation,LDA)与专家判断相结合的方法,分析 AI 相关领域的技能集;第四步,通过构建岗位簇与各技能集之间的需求矩阵,评估工作技能集对工作岗位簇的重要性,从而更准确地把握工作 AI 各岗位簇对工作技能的需求程度(图 1)。

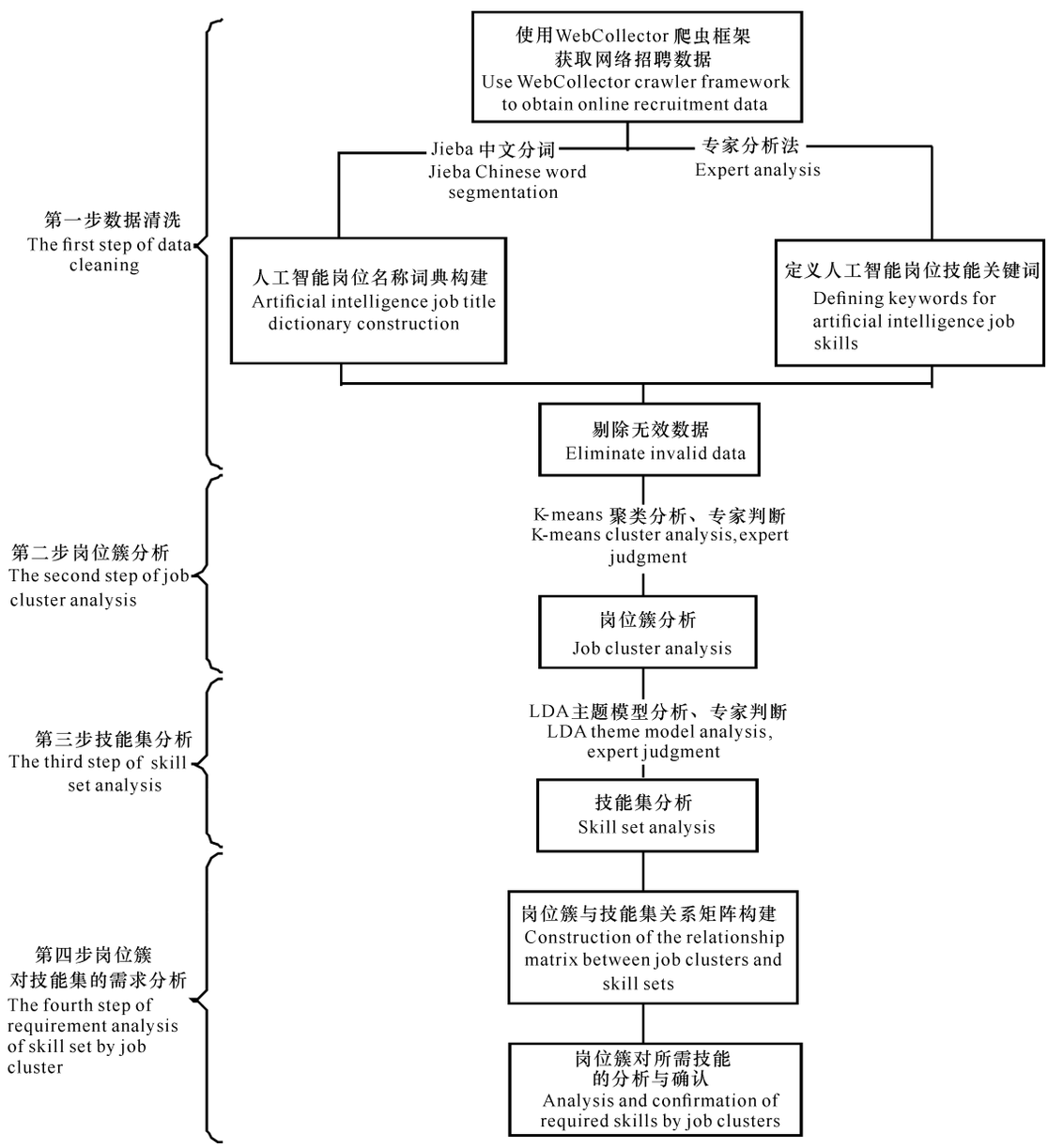

图 1 算法步骤

Fig.1 Arithmetic processes

2 数据来源及清洗

2.1 数据来源

选择智联招聘作为数据来源。相比其他招聘网站，智联招聘的招聘岗位页面 HTML 结构的标准化程度高，数据可获取性较好，Web 抓取可行性更高[8]。在 2019 年 3 月 - 2019 年 5 月期间，采用 WebCollector 爬虫框架对智联招聘网站在 2018 年全年的招聘岗位标题、岗位描述或岗位要求中包含关键词“AI”的岗位信息进行抓取，最终获得 10 656 条与 AI 相关的招聘信息。获取的招聘信息包括招聘信息 ID、公司名称、招聘岗位名称、岗位要求、薪酬、工作地点、工作年限要求、学历要求、信息公布时间等内容。

从需求时间看，2018 年 AI 岗位人才需求旺盛，呈现爆发式增长态势，尽管 7 月份达到最高峰（正值我国应届毕业生的毕业时间），但是下半年对 AI 的需求是上半年的 5.29 倍（图 2）。从需求地域看，2018 年 AI 专业人才需求主要集中在一线城市（北京、上海、广州、深圳）以及 15 个新一线城市（成都、杭州、武汉、南京、长沙、天津等）。这些经济发达城市 AI 产业发展迅速（图 3）。从学历要求看，2018 年 AI 领域对本科学历的需求最大，一定程度上表明了企业对 AI 应用开发的需求旺盛，而对 AI 研发人才的需求要小（图 4）。

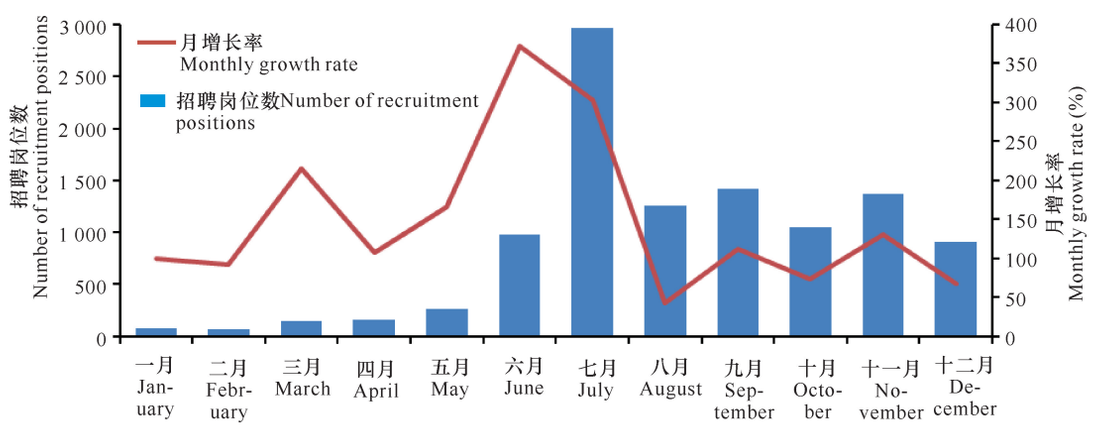

图 2 2018 年智联招聘发布的 AI 岗位招聘数

Fig. 2 Number of AI job recruitments released by Zhaopin. com in 2018

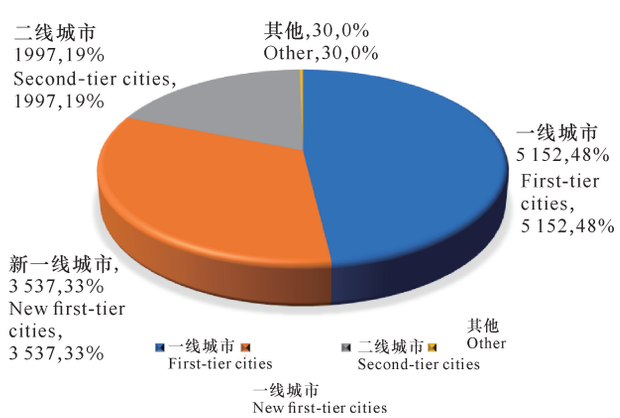

图 3 2018 年 AI 岗位工作地点分布

Fig. 3 Distribution of working place for AI in 2018

2.2 数据清洗

数据清洗按以下步骤进行：第一，使用网络爬虫获取的10 656条招聘信息中，有小部分为同一企业在不同时间点发布的对同一岗位的招聘信息，因此需要去掉这部分重复信息。第二，一些企业在互联网上发布招聘信息并不规范，例如招聘岗位名称中填写“博士”一词。这类招聘岗位名称属于无效值，不能作为岗位名称进行分析，需要视为无效数据予以剔除。如果某个岗位的招聘岗位名称中的技能词与 AI 岗位无关，那么这条招聘信息也属于无效数据而予以剔除。第三，鉴于中文的书写方式与英文不同，词汇之间缺少明显间隔，需要对中文文本采取“jieba 中文分词”处理，使计算机能准确地识别中英文词汇，分词之后需要对去除分词结果中的停用词和无效词（如“和”“或”“与”等），以消除停用词和无效词对数据分析的不利影响。然后，利用这些词构建岗位名称词典。岗位名称词典的构建还可以采取机器学习的方法[9]，考虑到算法的成熟度，本文采用“jieba 中文分词”工具。

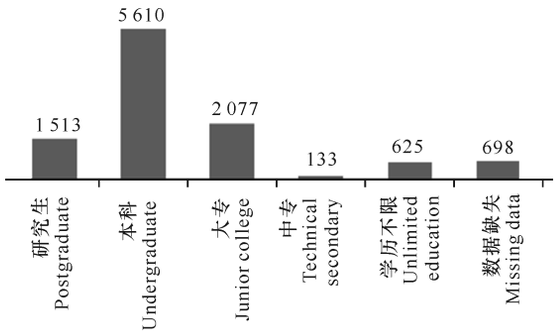

图 4　2018 年 AI 岗位的学历要求分布

Fig. 4　Distribution of education requirement for AI in 2018

在对招聘岗位名称进行分词和去停用词处理后，进一步选取在结果中出现次数超过 5 次的 194 个名词构成岗位名称词典，将招聘岗位名称中不包含岗位名称词典中词汇的招聘信息标记为无效数据予以剔除。图 5 展示了出现次数最多的前 50 个岗位名称名词的可视化词云图。每个名词的字体大小与每个名词出现的次数成正比。

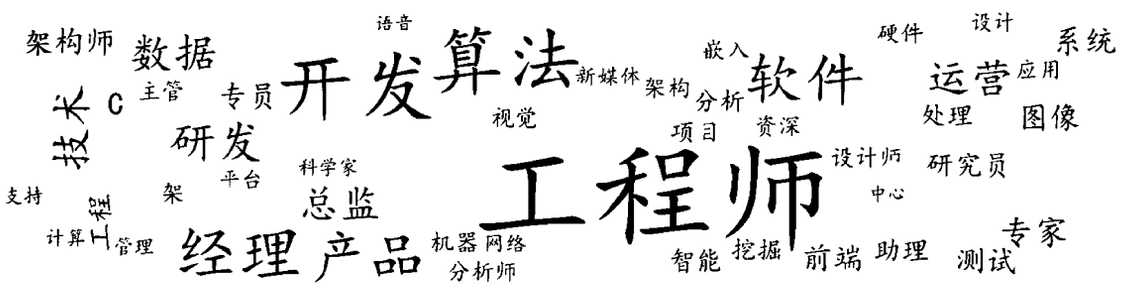

图 5　AI 岗位名称中重复出现的前 50 个单词

Fig. 5　The first 50 words that appear repeatedly in AI job names

对岗位要求进行预处理时，参考 IT 职业技能图谱，预先选择了与 AI 领域相关的 232 个技能词，将各个招聘岗位的岗位要求描述转化成技能词的集合。在前面处理的基础上，将岗位要求中不包含 AI 领域技能词的招聘信息标记为无效数据予以剔除。表 1 显示了岗位要求中词频最高的前 50 个技能词。

通过去重和两次清洗剔除重复数据和无效数据，最终保留 6 705 条数据作为有效样本数据。据此，可以对岗位名称进行 K-means 聚类分析获取岗位簇，并对岗位要求进行 LDA 分析获取技能集。

表 1　频率前 50 的岗位要求技能词

Table 1　Skills required for the top 50 positions

技能词 Skill word	频次 Frequency	技能词 Skill word	频次 Frequency	技能词 Skill word	频次 Frequency	技能词 Skill word	频次 Frequency	技能词 Skill word	频次 Frequency
数据挖掘 Data mining	1 054	Caffe	750	NLP	449	CSS	386	SVM	292
TensorFlow	1 050	云计算语言 Cloud computing language	729	聚类 Clustering	446	多线程 Multithreading	351	C#	291
分布式存储 Distributed storage	889	分类 Classification	709	shell	438	Android	321	XNET	290
建模 Modeling	867	计算机视觉 Computer vision	705	UI	437	知识图谱 Knowledge graph	320	MXNet	284
自然语言处理 NLP	838	模式识别 Pattern recognition	701	人脸识别 Face recognition	435	http	317	脚本语言 Scripting language	279
MySQL	830	Go 语言 Go language	659	OpenCV	420	RNN	313	git	271

续表 1

Continued table 1

技能词 Skill word	频次 Frequency	技能词 Skill word	频次 Frequency	技能词 Skill word	频次 Frequency	技能词 Skill word	频次 Frequency	技能词 Skill word	频次 Frequency
图像处理 Image processing	830	MATLAB	527	嵌入式 Embedded	409	HBase	302	.net	270
SPARK	792	语音识别 Speech recognition	514	CNN	405	hive	301	Scala	264
Hadoop	782	图像识别 Image identification	507	Redis	400	面向对象 Object-oriented	295	开源框架 Open source framework	255
数据结构 Structure of data	753	oracle	479	回归 Regression	394	信号处理 Signal processing	294	C语言 C language	243

3 数据分析

3.1 岗位簇识别

目前尚未有明确的AI岗位类别划分。因此,本研究使用AI招聘岗位名称作为输入,通过K-means聚类算法将获取的岗位名称进行聚类,从而识别出AI岗位簇[10]。为实现岗位簇的提取,需要将所有的岗位名称向量化,通过词袋模型,利用数据预处理时得到的岗位名称词典,将各个岗位名称分别转化为一个194维的0-1向量(岗位名称中出现词典中的单词记为1,未出现记为0)。将岗位名称向量化之后,再使用K-means聚类算法对所有岗位名称进行聚类。

K-means聚类需事前确定聚类数量,因此本研究利用肘部法则(图6)确定聚类数量为4。然后统计各簇中词对的出现频次。表2展示了各簇中出现频次最高的15项。这里需要特别指出的是,由于某些岗位名称书写不规范,致使通过分词和去停用词后该名称只剩一个名词。通过专家分析,将4类AI岗位簇分别命名为产品架构师、算法工程师、产品经理和软件工程师。

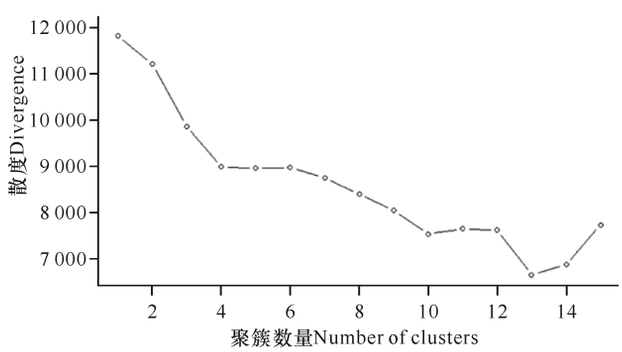

图6 K-means聚类肘部法则分析图

Fig. 6 Elbow rule analysis diagram of K-means clustering

表2 K-means聚类分析得出的4个岗位簇

Table 2 Four job clusters derived from K-means cluster analysis

岗位簇 Job clusters	词对 Word pair
软件工程师 Software engineer	开发工程师、软件开发、前端开发、CSS开发、C开发、开发工程、系统开发、android开发、web前端、后端开发、应用开发、测试开发、嵌入软件、PHP开发、工程师助理 Development engineer, software development, front-end development, CSS development, C development, development engineering, system development, android development, web frontend, back-end development, application development, test development, embedded software, PHP development, engineer assistant
算法工程师 Algorithm engineer	算法工程师、数据工程师、研发工程师、软件工程师、测试工程师、图像算法、数据挖掘、处理工程师、挖掘工程师、硬件工程师、分析工程师、前端工程师、视觉算法、系统工程师、C工程师 Algorithm engineer, data engineer, R & D engineer, software engineer, test engineer, image algorithm, data mining, processing engineer, mining engineer, hardware engineer, analysis engineer, front-end engineer, visual algorithm, system engineer, C engineer
产品经理 Product manager	产品经理、项目经理、技术经理、运营经理、研发经理、开发经理、经理总监、经理主管、硬件产品、智能产品、数据产品、方案经理、互联网产品、软件产品、平台运营 Product manager, project manager, technical manager, operations manager, R & D manager, development manager, director of manager, manager supervisor, hardware products, smart products, data product, program manager, internet products, software products, platform operation

续表 2

Continued table 2

岗位簇 Job clusters	词对 Word pair
产品架构师 Product architect	架构师、技术总监、新媒体运营、运营专员、架构、工程师助理、研发总监、运营、产品总监、算法专家、研究员、技术专家、数据分析师、运维工程师、程序员 Architect, technical director, new media operations, operations specialist, architecture, engineer assistant, R & D director, operation, product director, algorithm expert, researcher, technical experts, data analyst, operation and maintenance engineer, programmer

3.2 技能集识别

按照“能岗匹配”和“胜任力”理论,同一类型岗位所需的技能也应该是相似的[11]。反过来,相似的技能更有可能出现在同一份岗位说明书中。为分析岗位簇所对应的技能集,继续使用聚类分析方法对岗位簇所需的技能词进行聚类。为了识别工作岗位中的技能集,采用 LDA 进行聚类[12-14]。

LDA 的输入是招聘信息中的招聘岗位要求和需要识别的主题数量。为得到合适的主题数量,首先计算了主题数量 k 分别为 2-10 时的多个结果,然后组织专家对这些结果进行评估,最终得出主题数量 k 为 5 最合适,因此将技能集划分为 5 类最合理。表 3 显示了通过 LDA 分析出来的 5 个技能集,以及每个技能集中出现频次最高的 15 个技能词。组织专家对技能词所涉及的工作内容进行综合研判,确定将这 5 个技能词集合分别命名为数据库、机器学习、模式识别、大数据和程序设计。

表 3 基于 LDA 的技能集分析

Table 3 LDA-based skill set analysis

技能集名称 Name of skill set	技能词 Skill word
数据库 Database	MySQL、Redis、GO、分布式存储、CSS、Oracle、jQuery、MyBatis、MongoDB、云计算语言、Tomcat、J2ee、Html5、开源框架、多线程 MySQL, Redis, GO, distributed storage, CSS, Oracle, jQuery, MyBatis, MongoDB, cloud computing language, Tomcat, J2ee, Html5, open source framework, multithreading
机器学习 Machine learning	TensorFlow、Caffe、CNN、自然语言处理、RNN、XNET、MXNet、分类、计算机视觉、Theano、图像处理、NLP、PyTorch、数据挖掘、模式识别 TensorFlow, Caffe, CNN, NLP, RNN, XNET, MXNet, classification, computer vision, Theano, image processing, NLP, PyTorch, data mining, pattern recognition
模式识别 Pattern recognition	图像处理、计算机视觉、模式识别、语音识别、OpenCV、图像识别、人脸识别、MATLAB、信号处理、嵌入式、数据结构、目标检测、数据挖掘、自然语言处理、TensorFlow Image processing, computer vision, pattern recognition, speech recognition, OpenCV, image identification, face recognition, MATLAB, signal processing, embedded, structure of data, target detection, data mining, NLP, TensorFlow
大数据 Big data	数据挖掘、Spark、Hadoop、爬虫、分布式存储、分类、自然语言处理、聚类、hive、回归、数据结构、Scala、NLP、HBase、Storm Data mining, Spark, Hadoop, crawler, distributed storage, classification, NLP, clustering, hive, regression, structure of data, Scala, NLP, HBase, Storm
程序设计 Programming	云计算语言、UI、.net、Android、C 语言、嵌入式、C#、Http、数据结构、GO、多线程、软件测试、面向对象、ARM、Shell Cloud computing language, UI, .net, Android, C language, embedded, C#, Http, structure of data, GO, multithreading, software test, object-oriented, ARM, Shell

3.3 需求矩阵设计

在使用 LDA 分析技能集时,会输出每个岗位任职要求属于每个主题(技能集)的概率。每一项岗位任职要求代表一个工作岗位,因此该结果可理解为每个岗位对于每个主题(技能集)的需求程度。

为了得到各岗位簇对每个技能集的需求情况,首先选取位于同一个岗位簇中所有岗位对每一个技能集需求程度的平均值,将其作为该岗位簇对每一个技能集的需求程度,从而得到 4 个岗位簇对于 5 个技能集的需求矩阵 C。然后,将需求矩阵 C 的每一列除以其平均值来归一化矩阵 C,得到矩阵 T(表 4)。由于分析的工作岗位都是 AI 相关,同时岗位要求分析中用到的词都是和 AI 相关的词汇,因此不同岗位簇对技能集的需求程度区别不大。其中,元素 $T_{i,j}$ 表示岗位簇 i 对特定技能集 j 的需求程度。为了更清楚地描述岗位簇对各个技能集需求的重要程度,采用以

下方法予以简化处理,得到表5。

—T_(i,j)≥1.00:技能集 j 对岗位簇 i 特别重要;

—T_(i,j)<1.00:技能集 j 对岗位簇 i 不是特别重要。

表4 AI岗位簇对所需技能集的需求矩阵($T_{i,j}$)

Table 4 Requirement matrix of AI post cluster for required skill set ($T_{i,j}$)

	模式识别 Pattern recognition	程序设计 Programming	数据库 Database	大数据 Big data	机器学习 Machine learning
软件工程师 Software engineer	1.04	1.06	1.05	0.93	0.92
算法工程师 Algorithm engineer	1.04	1.01	0.95	1.02	0.97
产品经理 Product manager	0.99	0.92	1.10	1.02	1.03
产品架构师 Product architect	0.94	1.01	0.90	1.02	1.08

表5 岗位簇对所需技能集的需求评估

Table 5 Need assessment of job clusters for required skill sets

	模式识别 Pattern recognition	程序设计 Programming	数据库 Database	大数据 Big data	机器学习 Machine learning
软件工程师 Software engineer	* *	* * *	* *		
算法工程师 Algorithm engineer	* *	*		*	
产品经理 Product manager			* * *	*	* *
产品架构师 Product architect		*		*	* * *

注:每个单元格中的点数 * 表示AI岗位簇对于技能集的需求程度

Note: The number of stars in each cell indicates the importance of AI post cluster to skill set

4 结果可视化与分析

根据上述方法,可画出岗位簇映射技能集的冲击图,如图7所示。在图7中,对每一个AI岗位簇设置了识别标签,对岗位簇与所需技能集的映射关系进行了可视化处理,更为直观地描述了岗位簇对技能集的需求程度。其中,左侧是4类岗位簇,右侧是5类技能集,中间连接线的宽度表示各岗位簇对每个技能集的需求程度或相关度。

4.1 软件工程师

软件工程师的主要角色是从事AI软件开发相关工作。具体来说,AI软件工程师主要负责AI产品软件设计与构架、编写项目的核心代码、解决在产品的研发过程中遇到的技术难点、协调项目组成员之间的合作并参与代码开发规范编制。为此,AI软件工程师既要熟练掌握程序设计,又要了解模式识别[15]。根据图7可发现,程序设计对于AI软件工程师最为重要,其次是数据库和模式识别。该岗位簇的招聘信息中也多次提到对于程序设计(精通C#或Java语言,精通面向对象分析和设计技术,有足够的.net或Java开发经验)、模式识别(熟悉深度学习、AI、机器学习、神经网络等技术在图像处理领域的应用)以及数据库(熟练掌握MySQL、Oracle等数据库,有SQL性能调优经验优先)等技能要求。

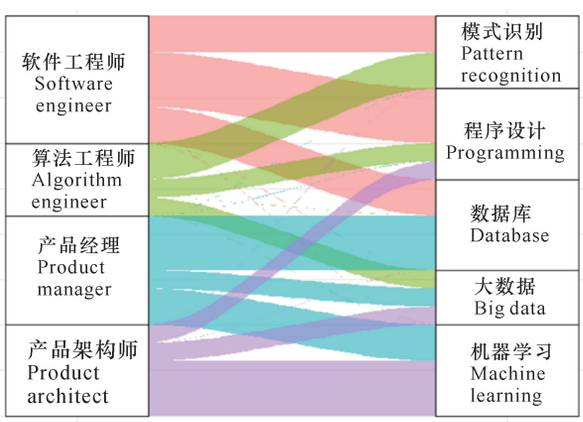

图7 岗位簇映射技能集的冲击图

Fig. 7 Impact map of job cluster mapping skill set

4.2 算法工程师

算法工程师是AI领域的稀缺核心岗位,其主要角色是通过模式识别等算法来完成不同的逻辑运算和优化业务。算法工程师的工作职责主要包括利用模式识别相关的手段分析大数据,然后将算法用伪代码描述出来,交由软件工程师实现[16]。根据图7可发现,模式识别对算法工程师最重要,其次是程序设计和大数据。该岗位簇的招聘信息中多次提到对模式识别(有图像处理、模式识别等项目经验优先)、程序设计(熟悉UI、.net和云计算、android和C#/C++等编程语言)和大数据(熟悉数据挖掘、spark、Hadoop和分布式存储)等技能要求。

4.3 产品经理

产品经理是需要将AI技术和行业知识相结合,并通过AI产品和项目的落地,最终实现企业商业目标的复合型岗位,需对AI产品进行规划设计、提炼使用场景、推动用户交互使用体验、推进产品上线。为此,AI产品经理既要掌握AI技术,同时又要熟悉

商业分析和产品开发管理,在工作中需要与产品构架师、算法工程师和软件工程师等充分沟通协作,保证产品功能落地[17]。根据图 7 可发现,除了行业市场知识、项目管理技能外,产品经理岗位对数据库、机器学习和大数据技术等有较强的技能需求。该岗位簇的招聘信息中多次提到对数据库(熟悉 MySQL、Oracle 等数据库)、机器学习(对 TensorFlow、Caffe 等算法有初步了解)和大数据(熟悉 Hadoop 底层文件系统,对大规模数据并行计算传输处理等有丰富的经验)这些领域的技能需求。

4.4 产品架构师

产品架构师是将 AI 落地解决问题的执行者、不同业务场景下的技术统筹人,主要着眼于 AI 系统的技术实现,需对产品全局掌控并能够及时洞悉局部技术瓶颈,并依据具体的 AI 业务场景给出解决方案。其主要职责是负责 AI 系统架构设计和技术架构选型,主导功能模块设计、数据结构设计、对外接口设计,针对行业客户设计场景化的解决方案,承担系统核心功能的研发工作和系统优化,负责制定 AI 业务规划等。为此,产品架构师必须能够熟练地与软件工程师、算法工程师以及 AI 产品经理沟通,充分了解 AI 的前沿理论与技术动态[18]。根据图 7 可发现,深度学习的理论与技术对产品架构师最重要,其次是大数据和程序设计能力。该岗位簇的招聘信息中多次提到对机器学习(深度学习、计算机视觉等领域工作经验,熟悉 TensorFlow/Caffe 框架)、大数据(丰富的 Hadoop 实战经验,熟悉 Hadoop 底层文件系统及分布式计算框架)和程序设计(熟悉.net、WCF、WPF 等相关技术开发优先)等技术领域有要求。

5 结论

与发展迅猛的 AI 技术领域比较,AI 领域的人力资源实践和研究均明显落后太多,人力资源管理实务界和学术界均迫切需要对 AI 岗位及所需具体技能有一个清晰的完整性理解。本研究基于 WebCollector 爬虫框架抓取了 10 656 条 AI 岗位的网络招聘数据,采用文本挖掘、K-means 聚类分析、主题模型构建、专家判断的半自动分析模型等方法,对 AI 岗位的岗位簇和技能集进行了类型学分析,得出如下结论:①AI 岗位可分为软件工程师、算法工程师、产品架构师和产品经理等 4 个岗位簇,以及数据库、机器学习、模式识别、大数据和程序设计等 5 个所需的技能集。②基于岗位簇对每个技能集的需求矩阵和基于冲击图的映射关系可视化结果显示,程序设计对于 AI 软件工程师最为重要,其次是数据库和模式识别;模式识别对算法工程师最重要,其次是程序设计和大数据;产品经理岗位对数据库、机器学习和大数据技术等有较强的技能需求;机器学习对产品架构师最重要,其次是大数据和程序设计能力。

本研究结果为精准感知劳动力市场对 AI 人才的需求提供了可能,对 AI 岗位词典编撰有一定贡献,有助于人力资源管理学术界和实务界对 AI 岗位及所需具体技能有一个清晰的完整性理解;从实践指导上可以帮助人力资源管理部门制定更精准的岗位管理、招聘遴选、培训开发方案,完善绩效管理等流程;高等学校也可根据本研究结果完善 AI 专业培养方案和课程体系建设,培养符合企业 AI 岗位所需专业人才,缓和 AI 领域的人才供需失配的问题。

由于本研究仅对智联招聘网站上的 AI 招聘岗位数据进行爬取,且未能考虑到欧美和日本、韩国等 AI 产业发展较好的其他地区和国家的情况,如何进一步高效拓展数据的爬取范围,将是下一步的工作重点。

参考文献

[1] 封帅.人工智能时代的国际关系:走向变革且不平等的世界[J].外交评论,2018,35(1):128-156.
[2] 陈劲,吕文晶.人工智能与新工科人才培养:重大转向[J].高等工程教育研究,2017,35(6):18-23.
[3] 张鑫,王明辉.中国人工智能发展态势及其促进策略[J].改革,2019,32(9):31-44.
[4] 于飞.我国企业人力资源管理信息系统建设分析[J].情报科学,2010,28(7):1117-1120.
[5] 马世龙,乌尼日其其格,李小平.大数据与深度学习综述[J].智能系统学报,2016,11(6):728-742.
[6] 蔡跃洲,陈楠.新技术革命下人工智能与高质量增长、高质量就业[J].数量经济技术经济研究,2019,36(5):3-22.
[7] 周文斌.机器人应用对人力资源管理的影响研究[J].南京大学学报(哲学·人文科学·社会科学),2017(6):23-34.
[8] 汤洋,汤敏倩.网络招聘信息中职业类型与专业领域的情报分析[J].情报杂志,2017,36(6):72-77.
[9] 文益民,杨鹏,文博奚,等.基于深度学习的中文网络招聘文本中的技能词抽取方法[J].桂林电子科技大学学报,2020,40(4):338-348.
[10] 周爱武,于亚飞.K-Means 聚类算法的研究[J].计算机技术与发展,2011,21(2):62-65.

[11] 魏新,杨俊.基于能岗匹配原理的人员招聘分析[J].中国人力资源开发,2010,27(10):70-72.
[12] ANDREA D M,MARCO G,MICHELE G,et al. Human resources for big data professions:A systematic classification of job roles and required skill sets [J]. Information Processing and Management,2018,54(5):807-817.
[13] GURCAN F,CAGILTAY N E. Big data software engineering:Analysis of knowledge domains and skill sets using LDA-based topic modeling [J]. IEEE Access,2019,7:82541-82552. DOI: 10.1109/ACCESS.2019.2924075.
[14] 李锋刚,梁钰等,GAO X Z,等.基于LDA-wSVM模型的文本分类研究[J].计算机应用研究,2015,32(1):21-25.
[15] 苑俊英,陈海山,杨智.校企合作培养卓越软件工程师模式的探索与实践[J].武汉大学学报:理学版,2012,58(S2):239-243.
[16] PACKEL E W. The algorithm designer versus nature:A game-theoretic approach to information-based complexity [J]. Journal of Complexity,1987,3:244-257.
[17] 任传宏.产品经理的定位与职责[J].企业管理,2011,32(11):91-92.
[18] 周爱民.做人、做事,做架构师——架构师能力模型解析[J].程序员,2008(4):70-73.

Research on AI Job Demand Analysis Based on Big Data Technology

XU Zhengli[1],WEN Boxi[2],XIE Meiying[3],CAI Xiang[1]
(1. Guilin University of Electronic Technology,Guilin,Guangxi,541004,China; 2. Guangxi Polytechnic of Construction,Nanning,Guangxi,530007,China; 3. Nanjing University of Information Science & Technology,Nanjing,Jiangsu,210044,China)

Abstract:In recent years,there has been a structural contradiction of mismatch between supply and demand in China's talent market,especially in the field of artificial intelligence (AI). Accurately perceiving and describing the demand of the labor market is an important method to solve this problem. This study first takes a web crawler to capture AI job-related recruitment information published by the recruitment website of Zhilian. Through Chinese word segmentation,K-means and other big data analysis methods,the recruitment job names are clustered to identify four job clusters:Software engineers,algorithm engineers,product managers and product architects. Then,the model of Latent Dirichlet Allocation (LDA) is used to cluster the job requirements,and five skill sets including database,machine learning,pattern recognition,big data and program design are obtained. Finally,LDA is used to obtain the demand matrix of job clusters for their skill sets to analyze the demand degree of each job cluster to their job skills. The results show that the programming ability is the most important for AI software engineers,and the theory and technology of pattern recognition are the most important for algorithm engineers. Product manager positions have strong skills requirements for databases,machine learning and big data technology. The theory and technology of machine learning are the most important to product architects. The results of this research can provide technical support for universities and enterprises to accurately perceive and describe the needs of the AI labor market in real time.
Key words:big data,artificial intelligence,web crawler,job roles,job skills,job dictionary

责任编辑:陆 雁