IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 9 August 2023, accepted 29 August 2023, date of publication 1 September 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3311203

RESEARCH ARTICLE

Information Technology Job Profile Using Average-Linkage Hierarchical Clustering Analysis

PUJI CATUR SISWIPRAPTINI [1], HARCO LESLIE HENDRIC SPITS WARNARS [1], ARIEF RAMADHAN [2], (Senior Member, IEEE), AND WIDODO BUDIHARTO [3], (Senior Member, IEEE)

[1]Computer Science Department, BINUS Graduate Program-Doctor of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
[2]School of Computing, Telkom University, Bandung 40257, Indonesia
[3]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Corresponding author: Puji Catur Siswipraptini (puji.siswipraptini@binus.ac.id)

ABSTRACT The growth in Information Technology (IT) jobs is predicted to reach 15 percent between 2021 and 2031. The growth of IT jobs has resulted in a remarkable change in all infrastructure, such as information, skills, and domains covered in IT job profiles. Unfortunately, job roles and skills in this field remain undefined. The gap between the supply and demand needs in the IT workforce must be filled immediately with an appropriate strategy. To fulfill industry needs, an in-depth analysis of IT job profiles is important. Therefore, it is important for educational programs to identify the competencies needed by the industry to update their output. This study aims to identify the job profiles required for IT job specialists by analyzing real-world job posts published online to identify hidden meanings from a textual database. A systematic semantic methodology was proposed using an average-linkage hierarchical clustering analysis. It resembles a tree structure technique to discover relevant phrases, relationships, and hidden meanings through semantic analysis. Occurrences of the most frequent words and phrases were extracted to reveal the domain knowledge of each IT job cluster. The result is a systematic semantic analysis of the IT job profile comprising the programming language, specialized type, duty, database, tools, and frameworks. The justification for each job profile was validated by 10 IT professionals from various private and government companies in Indonesia through Focus Group Discussions (FGD).

INDEX TERMS Information technology job profile, skills, average-linkage hierarchical clustering analysis, most frequent word, most frequent phrase.

I. INTRODUCTION

Association of Computer Science Colleges-Indonesia (APTIKOM) has identified 550 types of jobs in the Information Technology (IT) Workforce and IT Enablers with various educational qualifications. The projected demand for IT workforce in Indonesia highly increase to 1.9 million workforces in 2022-2025 [1]. Meanwhile, a survey institute in the United States has identified 10 clusters of types of work in the IT sector. The IT jobs discussed in this study support computer applications, systems, and computer networks. U.S. Bureau of Labor Statistics states that the growth of IT jobs in the world is predicted up to 15 percent in line with the increase of 682,800 new jobs from 2021 to 2031. This prediction is much faster than the average for all occupations. The median annual wage for this group was $97,430 in May 2021, which is higher than the median annual wage for all occupations of $45,760 [2]. The workforce in the IT field must have specialized and professional skills in this digital era. They must have the ability to handle or apply successful digital transformation concepts, followed by different types of changes including different skills in each job [3], [4], [5], [6].

Job profile refers to duties and responsibilities as technical competencies that the IT workforce must fulfill. These competencies are described as hard and soft. Job advertisements describe the hard and soft skills required for every IT job. Hard skills in IT jobs, including programming languages,

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato.

databases, tools, and frameworks. This study limits the competencies analyzed to hard skills only, since we think that common soft skills, such as communication, problem-solving, and teamwork, are general skills that must be fulfilled for every job.

Recently, skills and job profiles have become a critical issue [7]. Unfortunately, skills and job profiles in the IT field are undefined [8]. Hundreds of variations of existing job names lead to confusion for job seekers as to what specific skills they need to achieve [7], [9]. However, the gap between the availability of supply and demand needs in the IT workforce is filled immediately with the right strategy [10]. Educational institutions need to customize their curricula to meet industrial needs [6]. Recent studies on job profile analysis have used various techniques. Data mining, text mining, web crawling, and Natural Language Processing (NLP) were used to analyze and compare job profiles from various job perspectives. A data mining technique was used to predict a job post area [11]. Extracting data structures with text mining has illustrated the relevant job skills in the medical informatics area and the skills required for computer vision and NLP specialists have been identified using the web crawling technique [12]. NLP methods are already used in the scientific context for the automatic extraction of relevant information from texts and are also efficient methods to summarize data [13], [14]. These studies have some limitations, such as the manual checking of labeling skills only by researchers and not validated by professionals in their fields, not explaining systematic semantic analysis, and in-depth analysis of job profiles in the IT field have not been found much.

The aim of this research is to identify the job profiles required for IT job specialists by analyzing real-world job posts published online using average linkage hierarchical clustering (ALHC) analysis. Hierarchical methods are clustering techniques that form a hierarchy based on levels that resemble tree structures. Thus, the grouping process is performed in a multi ranking or gradual manner. Different linkage methods are used in hierarchical clustering such as single, complete, and average methods. Average-linkage hierarchical clustering outperforms single and complete linkage methods for text mining [15], and can be improved by combining it with other methods to encourage accuracy [16].

This algorithm primarily aims to model a set of objects into two or more clusters based on the similarities between the various characteristics of each object. The objectives of this study are the contribution of two findings: first, the improvement of IT professionals' knowledge and skills can be evaluated; second, it can evaluate the course structure of educational programs as industry needs.

II. RELATED WORK

In recent years, research in skills and job profiles has become a critical topic for discussion. This condition is caused by every job being integrated with skills in the field of information technology, along with the development of the era, since everything is completely computerized. Some jobs that are integrated into the IT area include medical informatics, which requires expertise in database knowledge, computer networks, and software development that supports industry-standard applications to fulfill customer services. That study acquired only one job portal that could make the coverage less varied [7]. Access to information, database query language, and web technology are important knowledge that must be mastered by librarians [38], and this study does not address the large and specific profiles of IT jobs so that they cannot be used as a reference for the IT workforce.

Data mining techniques include classification, association, correlation, categorization, prediction, estimation, clustering, trend analysis, and visualization. The advantage of using data mining is that rules or knowledge recommendations can be updated based on new facts that may appear and affect the results [17], [18]. Information retrieval (IR) is the discovery of materials, such as structured and unstructured data, that meet information needs. Unstructured data are typically in the form of text documents, emails, videos, audio, websites, and data from social media [19], [20]. The representation and organization of information items should be such that users can easily access the desired information [21]. Recently, text mining and clustering techniques have been used to extract unstructured data such as job profiles [6], text summarization [22], text analysis [23], question answering [24], and topic interest [25]. Later, research on job selection and recommendations was inseparable from the Artificial Intelligence (AI) technology approach, as in clustering, classification, text mining, and NLP [26], [27], [28]. Research in this field includes job profiles that match those of job seekers [29], [30], [31], [32], and selection of appropriate courses [26], [33], student interest in choosing a study program that is in accordance with the career goals to be achieved [34], types of work and skills needed [27], [35], [36], [37], and learning paths with career goals [38].

Therefore, similar research on data mining techniques has been carried out by [26], which discusses the selection of courses that are tailored to job profiles and job seeker profiles using the k-NN algorithm to classify types of work. That study recommended jobs that would be available after the completion of each course but did not reveal detailed job profiles, so it cannot be used as a guide for job seekers [26]. K-means clustering and the Doc2Vec algorithm were used in [30] to provide recommendations that aim to help job seekers find suitable jobs. Jobs are grouped such that the model will recommend users who have suitable features to choose suitable jobs. Unfortunately, this study did not evaluate the accuracy of the model. Research by [29], [39] recommends relevant jobs for job-seekers by utilizing the concepts of recommender systems, information retrieval, and data mining. Their results prove that flooding jobseekers with thousands of irrelevant jobs is a frustrating and time-wasting process for finding the right job, and the model sometimes cannot provide the level of accuracy. Another technique in job profile research is text-mining-based, since the development of measuring the readiness of human resources to meet industrial needs [6], [7]. However, their work has some limitations related to the algorithm's distinct classes that bring simplicity and clarity to the proposed process, which can result in false job profiles that need to be refined.

The Average Linkage Hierarchical Clustering (ALHC) algorithm, as part of the agglomerative technique, seeks to build a hierarchy of objects based on its criterion that is used to merge two clusters of objects. ALHC provides a unique cluster at a higher level of hierarchy is provided by ALHC [40]. A systematic semantic analysis of the textual content of IT job advertisements was conducted using text mining and hierarchical clustering techniques has been revealed in this work. ALHC is using to reveal knowledge of IT job profiles. We chose ALHC because the resulting hierarchical cluster can be explored in both directions. This is a crucial method when considering data resumes, such as profiling jobs that can be described at different levels of granularity, from broader to more specific ones [41]. ALHC reveals a hierarchical cluster that can be explored in both directions. This is a crucial method when considering data resumes, such as profiling jobs that can be described at different levels of granularity, from broader to more specific ones. At various levels of granularity, this method proved splendid when dealing with new, hidden information, which can therefore be assigned to a profile based on the hierarchy. However, this algorithm requires less parameter optimization. Only two tunable parameters, the distance metric and linkage criterion, are needed to run the algorithm over the data. This allows for simpler adaptation to different environments and datasets. Recent studies have also proven that hierarchical clustering is a trend analytical technique for partitioning mechanisms and time-series data [42], [43]. This study uses two job portals, TechInAsia and Jobstreet Indonesia, which make the coverage of datasets richer and wider. This study was justified by a professional IT group, and good validation was obtained.

Therefore, the background of this study is addressed in two sub-sections: "IT Job Profiles" and "Average-Linkage Hierarchical Clustering".

A. INFORMATION TECHNOLOGY JOB PROFILE

Workforce skills are key to the success of a highly creative industry [44]. Job profiling based on summarization of text using hierarchical clustering algorithms has been widely researched. A semantic-relevant representation of summary text, including job profiles, can, in turn, allow for a better model. Various job profiles have been identified by [41] using sentence BERT to perform sentence embeddings. Unfortunately, research [41] only involved one job portal, so the dataset used was subjective, and the clustering results for each skill in the job profile had not been validated by professionals in their fields. In our study, we involve two job portals so that the resulting dataset is objective in its assessment, and the results of each skill on the job profile have been validated by IT professionals to obtain reliable results. A job profile and recommender model aimed at extracting meaningful data from job posts using text-clustering methods was developed by [30]. Job profiles are divided into job clusters based on their common features, and job offers are matched to jobseekers according to their interactions. That study did not reveal the skills of each job in detail. Our work complements previous research [30] by revealing the job profiles and skills required by the industry in a systematic and complete way.

Programming language is an essential skill in application development. In the 21st century, many programming languages have been developed to meet industrial needs in various fields [45], and they complement each other to achieve their purposes [5]. In this information explosion era, databases are needed to store, design, and manage information in many areas, such as health [46], the industrial sector [47], and finance [48]. Job profiles in the IT field include programming languages, databases, tools, and frameworks.

Tools are often combined with programming languages to develop more efficient software applications. Tools contain utilities, such as operating systems, platforms, and libraries. This framework is the basic conceptual structure that is used to solve complex problems. It is used to make it easier for the IT workforce to develop applications and reduce processing time and costs.

B. AVERAGE-LINKAGE HIERARCHICAL CLUSTERING

Data mining is a computer-assisted process of mining and analyzing datasets and extracting useful information and knowledge, such as trends and patterns, from unstructured or semi-structured databases to model data-driven processes including classification, clustering, and regression analysis [5]. In some cases, the clustering method is also used to extract the topic model from the text corpus to retrieve hidden information by analyzing and mining data [49]. Text mining using a hierarchical clustering algorithm as part of data mining analyzes unstructured or semi-structured text documents related to each other. It can be used to describe research trends by identifying relevant phrases and relationships to conclude research findings [50].

Clustering is one of the most popular techniques used in data analysis. In recent years, Hierarchical Clustering (HC) has been proven to provide higher quality clusters to reduce the sensitivity of clustering to problem types [40]. HC consists of two techniques: bottom-up, known as agglomerative and top-down, known as divisive [51]. In agglomerative techniques, such as Average-Linkage Hierarchical Clustering (ALHC), each instance is considered as a cluster, and then the clusters are merged to create larger clusters. This process continues until all clusters are merged into one large cluster that contains all instances. The hierarchical clustering method discovers standard features and job profiles in IT job advertisements, such as programming languages, enterprise software, databases, and other skills [5], [50]. Average-Linkage Hierarchical Clustering (ALHC), as one of a clustering technique that can form clusters of varying sizes and shapes [52] and is reasonably interpretable [17]. ALHC methods are easy to implement because they start with only one instance in each cluster and then gradually join the nearest clusters. Hence, this study proposes an ALHC-based clustering method.

C. CONTRIBUTIONS

This study presents an approach to model the average linkage hierarchical clustering to describe IT job profiles. The model works based on the most frequent words and phrases representing skills and knowledge as a new dataset in the IT field. An in-depth and systematic semantic analysis was conducted to explore job profiles. This study will also benefit the IT workforce by fulfilling its competencies based on industrial needs. The justification of each job's skills was validated by ten IT professionals from various private and government companies in Indonesia through Focus Group Discussions (FGD).

III. METHODOLOGY

This research methodology consists of four significant steps: data collection, text pre-processing, average-linkage hierarchical clustering, and visualization and job profile analysis, as presented in Figure 1.

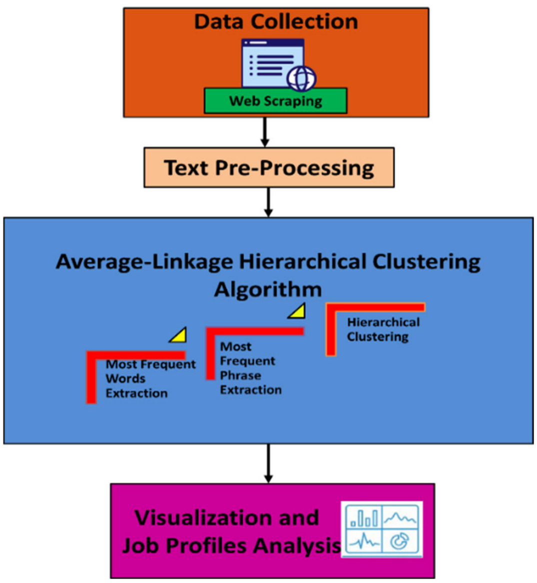

FIGURE 1. Research methodology of IT job profiles.

The following sections describe each step in the research methodology.

A. DATA COLLECTION

Recently, web-based real datasets have become popular in information retrieval research. IT job advertisements are the most interesting and important way to recruit employees. Traditionally, they have been published in newspapers, while in this digital era, they are mostly published online, either on social media or specialized websites. Many job advertisements such as LinkedIn [35], [50], Nukari.com [36], CareerBuilder.com [53], and Bebee [32] have been used in many studies. This study explored the Tech in Asia [54] and Jobstreet Indonesia [55] platforms for several reasons. First, they focus on IT job advertisements in Asia. Second, they presented a skill tag for every posted job advertisement. Finally, Tech in Asia has a tagline building and serves Asia's tech and startup communities. This aligns with the Indonesian government's program to develop 1.000 new digital start-up companies [56]. Jobstreet Indonesia has broad coverage of IT jobs. The technical skills justification of each IT job on the platforms is analyzed from computer and information technology occupations on the U.S. Bureau of Labor Statistics website [2].

B. TEXT PRE-PROCESSING

Text pre-processing should be performed on unstructured datasets to ensure precise data processing. Four steps of text pre-processing were conducted in this study, as presented in Figure 2.

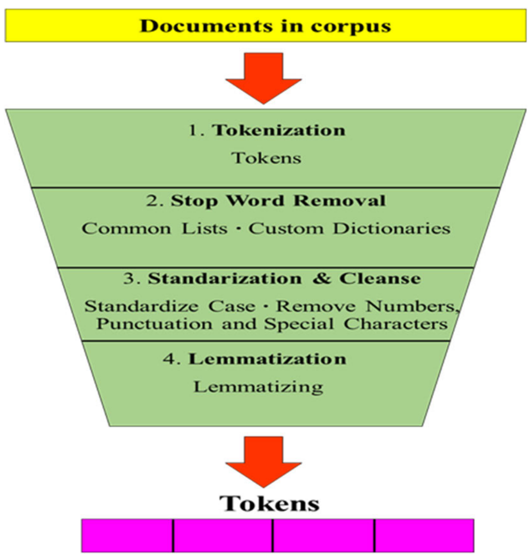

FIGURE 2. Text pre-processing stages of the average-linkage hierarchical clustering.

Documents in the corpus must be preprocessed to obtain normalized data. Tokenization is the process of separating text into tokens for subsequent analysis. Words, numbers, symbols, punctuation marks, and other essential entities can be considered as tokens. Stop word removal aims to form a common list and custom dictionaries by taking essential words from the document using a stop list algorithm (discarding less important words) such as "and," "or" "the" etc. Standardization and cleansing aim to standardize cases by removing numbers, punctuation marks, or special characters such as periods, commas (,) and question marks (?). etc.

The last step is lemmatization, which transforms a plural word into a single word or the past tense. In addition, other tense verbs were transformed into present-tense versions. The output of these steps is tokens to be processed in the most frequent words and phrases stage.

C. AVERAGE-LINKAGE HIERARCHICAL CLUSTERING ALGORITHM

The collected data were processed in three steps of ALHC process: most frequent word extraction, most frequent phrase extraction, and hierarchical clustering. Words were extracted to summarize IT job skills. The most frequent words, ranging from 50 to 230 job advertisements in each job cluster, were extracted. The most frequent phrases consisted of a maximum of three words reflecting job profiles such as duties and skill pairs in each job. This approach resulted in over 50 phrases for every cluster used to describe the job profiles. Both the most frequent words and phrases occur using the Term Frequency Inverse Document Frequency (TF∗IDF) to measure metric values to estimate the importance of a word or phrase in a textual corpus. Importance is based on the ratio of the frequency of words/phrases in a document to the total number of words/phrases in the textual corpus (TF). Along with the logarithm of the ratio of the total number of documents, the word/phrase that occurs in, and the total number of documents present in the collection (∗IDF), as shown in Equations (1) and (2).

$$W_{i,j} = tf_{i,j} x idf_{ij} \tag{1}$$

$$W_{i,j} = tf_{i,j} x log\left(D/df_{i,j}\right) \tag{2}$$

where the weight of the term (t) to the document ($W_{i,j}$) is calculated by the number of occurrences of the term in document ($tf_{i,j}$) multiplied by the number of documents containing the term (idf_{ij}). The rarer a term is between documents, the larger its (idf_{i_j}) value. Its value is the logarithm (log) of the reciprocal of the number of documents that have term t divided by the total number of documents (D). The more often a term appears in (D), the greater its value ($tf_{i,j}$).

Hierarchical clustering was conducted by measuring the similarity or distance between each observation from pairwise distances between clusters. The distances between clusters are based on the Jaccard similarity index, as shown in Equation 3, which identifies similarities between finite sample sets by comparing members for the two sets to determine which members are shared or different.

$$J(A, B) = |A \cap B| \Big/ |A \cup B| = |A \cap B| \Big/ |A| + |B| - |A \cup B| \tag{3}$$

The main function of the Jaccard Index is to calculate how similar or different two datasets are. Intersect is the intersection operation of first set (A) and second set (B) (A∩B), all members of A and includes members of B. Union is two sets A and B (AU B) is the set of all members of A or B or both. The average pairwise distances were then calculated. This value is a measure of the similarity between clusters, and the closest clusters were merged using Equation (4).

$$d_{12} = 1/kl \sum\nolimits_{i=1}^{k} \sum\nolimits_{j=1}^{l} d\left(X_i, Y_j\right) \tag{4}$$

where the distance (d_{12}) between the data in the first cluster (k) and the data in the second cluster (l) is formed by the number of items (X_i) in the first cluster (k) and the number of items (Y_j) in the second cluster (l). Furthermore, observations were made for k (X_1, X_2,..X_i) and l (Y_1, Y_2,..Y_j), data that are similar will have a close hierarchical relationship and form data clusters. Pairwise clusters continue to form until all data are connected in the hierarchical chart. This formula was calculated to create a hierarchical chart (dendrogram) to show the similarities between the data. Cluster analysis was used to extract job profiles, and the extracted phrases were embedded in an average-linkage hierarchical clustering algorithm to create a similarity matrix that highlighted phrases that were close to each other. The average linkage is where the distance between each pair of observations in each cluster is added and divided by the number of pairs to obtain the average intercluster distance. This is one of the most popular distance metrics used for hierarchical clustering.

D. VISUALIZATION AND JOB PROFILE ANALYSIS

The average-linkage hierarchical clustering algorithm is presented in some visualizations such as dendrograms, word/phrase frequency analysis, and link analysis, which describe words and phrases associated together. These visualization concepts were used to describe the job profile analysis in this study.

A dendrogram performs hierarchical clustering analysis, which allows computation of the similarity of documents based on words or phrases as keywords. Word/phrase frequency analysis allows a glance at the most common terms in the textual corpus. It is also helpful to identify keywords in the same clusters and the associated terms that can be used to retrieve information related to these clusters. This feature may also help identify co-occurring words, allowing the exploration of search terms for retrieving relevant text. A visual result can be presented using a table format and word cloud graphical display. The various sizes of each word are presented proportionally according to their relative frequency of use in the text. A network graph is presented as the connection between keywords, known as a Link Analysis. It offers a high level of interactivity and heterogeneous data types [57], which allows the exploration of relationships and the detection of underlying patterns and structures from the corpus. Each element is represented as a node and the edges describe the relationships between them. The thickness of this edge represents the strength of this relationship. The visualizations used to explore job profile analysis comprise programming languages, specialized types, duties, databases, tools, and frameworks.

Based on the experiment tested, this study has some advantages; for example, the number of clusters to be specified is flexible according to the job profile to be explored, computationally running fast, graphical representation in dendrogram simplifies the readers' understanding, and varying shapes and sizes of clusters can be adjusted to achieve the aims of the research. However, this experiment also found some disadvantages, such as high complexity has been determined, cutting off the dendrogram's level is not easy to decide, and the distance metric used determines the results of the clusters that are formed.

IV. RESULT AND ANALYSIS

Result and analysis describe stages in research methodology in IT job profiles. It consists of data collection, text pre-processing, average linkage hierarchical clustering algorithm, and visualization and job profiles analysis. To understand the IT job profile better, the results of the average linkage hierarchical clustering algorithm are presented in Table 5. Then, it was analyzed to describe the in-depth analysis using programming language, specialized type, duty, database, tools, and frameworks. The following section presents and discusses the results of the analysis. This has been visualized in several formats. Additionally, a Venn diagram was presented to explain the similarity of the functions of each programming language, tools of computer networks, tools of software, and databases among IT job fields.

A. DATA COLLECTION AND TEXT PRE-PROCESSING

Data collecting was conducted using a web scraper extension for data collection, Word Stat Provalis software version 8.0, and Excel software for text pre-processing, clustering, and visualization. The IT job advertisement dataset in English was scraped from the Tech in Asia and Jobstreet Indonesia websites [54], [55] from March to November 2022. In total number of 1065 IT jobs advertisements were obtained. It comprises ten clusters of computer and information technology occupations, as stated in [2]. These ten clusters are web developers, computer system analysts, software developer quality assurance and testers, information security analysts, computer programmers, computer support specialists, database administrators and architects, information research scientists, computer network architects, and network and computer system administrators. To extract relevant advertisements, this study focused on job advertisements that contained keywords strongly related to Information Technology jobs. A Boolean word combination was used when searching for relevant studies, as presented in Table 1. This dataset is limited to several conditions, including work location in Indonesia, the function of which must be related to IT, the limitation of work experience is one to four years, and the type of work is full-time. It eliminated all datasets except for the records in English. The dataset performs experiments on an AMD Ryzen™7 5800H Processor of 3.2 GHz and a minimum of 16 GB RAM.

TABLE 1. Keywords for information technology jobs advertisement.

Job Title	Keywords	Amount of data
Web Developers	"Web developer" or "UI/UX design"	230
Computer System Analysts	"Computer analyst" or "system analyst"	69
Software Developers, and Quality Assurance, and Testers	"Quality assurance and testing" or "software engineer"	180
Information Security Analysts	"Information security" or "security analysts"	81
Computer Programmers	"programming" or "object-oriented program"	128
Computer Support Specialists	"IT support"	114
Database Administrators and Architects	"Database administrators" or "data architects"	90
Computer and Information Research Scientists	"Data science," or "robotics" or "information research scientist"	50
Computer Network Architects	"Computer network" or "network architect" or "network engineer"	52
Network and Computer System Administrators	"Network administrator" or "system administrator"	71

Text mining is appropriate for IT job advertisements as an unstructured automated analysis of relevant information and should be conducted to extract helpful information [58], such as skill requirements in the industry [45]. IT experts includes a wide range of job titles and descriptions. The keywords options were utilized to acquire job ads that only covered ten clusters of IT jobs, as shown in Table 1. This is shows that Web Developers were the most available jobs, and Computer and Information Research Scientists were the least available jobs during the data collection period. One of the obstacles encountered in data collection was the difficulty in distinguishing between several types of jobs because they have similar descriptions and requirements. This forced us to analyze each job manually so that the job was in the right cluster.

Table 2 shows some examples of Tech in Asia and Jobstreet Indonesia data collection consisting of three main

TABLE 2. Samples of tech in asia data collection.

Job Title	Job Description	Required Skills
Web Developers	Web Developers are responsible for converting a software program design into a set of instructions that computers can follow. This requires proficiency in a variety of common programming languages. They are responsible for designing & implementing (i.e.: coding) web applications based on certain functional specifications. They will work closely with Web Designers & Product Owners. Proficient with PHP & JavaScript and have in-depth practical experience in one of these frameworks: Phalcon, Laravel, Yii, or CI. Having knowledge in another language such as Python, Ruby, ActionScript, Bash, etc is a plus. Must be familiar with Linux Operation, W3C Standard, website technology update and have knowledge in programming languages and environments including CSS, HTML, XML, and Json (understanding JavaScript, AJAX, Action Script and Ruby will be an advantage). Experience with Git or GitHub is mandatory. Excellent knowledge of MySQL or PostgreSQL databases.	HTML CSS MySQL Javascript PHP Web Development Linux
Computer System Analysts	Ensure the effectiveness of the system development process. Identify the user's need and available business process. Ensure the arrangement and system development align with the user's requirement. Monitor and enhance user's satisfaction towards various IS Application Development in Software Solution Department. Have experience in Business Process Management and SDLC Tools.	Operating Systems Business Analysis System Design

attributes. First, the job title explains various job names in the same area. Second, the job description explains the hard and soft skill requirements. Third, the required skills explain the hard skills that must be fulfilled for every job. Word Stat was applied with different iteration numbers and stabilized for 500 Gibbs sampling iterations. Another important parameter for the application of ALHC is the number of word frequencies, which is a user-specified parameter that adjusts the granularity level of discovered clusters. Different values were trained, and the desired modeling level was obtained at a rate of 10,000 words. When selecting the ideal number of clusters, the semantic consistency of the discovered clusters and the discovered clusters, and the distribution of the descriptive keywords in these clusters were considered.

TABLE 3. Most frequent words of web developers.

Words	Frequency	% Shown	No Cases	% Cases	TF*IDF
developer	224	11,88	169	73,80	29,6
javascript	131	6,95	131	57,21	31,8
css	101	5,36	100	43,67	36,3
html	101	5,36	100	43,67	36,3
frontend	99	5,25	76	33,19	47,4
react	95	5,04	86	37,55	40,4
php	78	4,14	72	31,44	39,2
backend	75	3,98	55	24,02	46,5
web	71	3,77	56	24,45	43,4
mysql	62	3,29	62	27,07	35,2
engineer	60	3,18	59	25,76	35,3
software	57	3,02	48	20,96	38,7
java	55	2,92	47	20,52	37,8
framework	44	2,33	42	18,34	32,4
node	40	2,12	36	15,72	32,1
sql	37	1,96	36	15,72	29,7
apis	30	1,59	30	13,10	26,5
python	27	1,43	25	10,92	26,0
laravel	26	1,38	24	10,48	25,5
application	22	1,17	22	9,61	22,4
design	22	1,17	22	9,61	22,4
rest	21	1,11	21	9,17	21,8
vue	21	1,11	21	9,17	21,8
server	18	0,95	18	7,86	19,9
golang	17	0,90	13	5,68	21,2
mobile	17	0,90	16	6,99	19,6
stack	17	0,90	16	6,99	19,6
postgresql	16	0,85	16	6,99	18,5
angular	14	0,74	12	5,24	17,9
angularjs	14	0,74	14	6,11	17,0
bootstrap	12	0,64	12	5,24	15,4
git	12	0,64	12	5,24	15,4
mongodb	12	0,64	12	5,24	15,4

B. MOST FREQUENT WORD AND PHRASE EXTRACTION

The IT job profiles yielded numerous results concerning the most frequent word and phrase extractions. For example, Tables 3 and 4 present the most frequent words and phrases reflecting the knowledge and technical skills related to Web Developers. The values of the metrics used to estimate the importance of a phrase in numerous documents are shown in the column Term Frequency Inverse Document Frequency (TF-IDF) in Tables 3 and 4.

The top ten most essential phrases according to their frequency are "html css", "front end", "javascript

TABLE 4. Most frequent phrases of web developers.

Phrase	Frequency	No Cases	% Cases	Length	TF*IDF
html css	89	89	38,86	2	36,5
front end	60	51	22,27	2	39,1
javascript framework	28	28	12,23	2	25,6
backend development	25	25	10,92	2	24,0
web development	24	24	10,48	2	23,5
front end development	23	23	10,04	3	23,0
backend developer	20	20	8,73	2	21,2
frontend developer	20	20	8,73	2	21,2
html css javascript	18	18	7,86	3	19,9
rest apis	18	18	7,86	2	19,9
sql server	18	18	7,86	2	19,9
full stack	17	16	6,99	2	19,6
software engineer	17	17	7,42	2	19,2
front end developer	16	16	6,99	3	18,5
html css javascript	18	18	7,86	3	19,9
rest apis	18	18	7,86	2	19,9

framework'', backend development'', ''web development'', ''front end development'', ''backend developer'', ''frontend developer'', ''html css javascript'', and ''rest apis''. The extracted phrases were divided into two groups: general and special. Special skills such as the html-css and JavaScript frameworks. Several types of specialized web developers, such as back-end and front-end developers, were also found in the list of phrases. Several phrases reflect general skills such as communication skills and Structured Query Language (SQL) servers, which are transferrable across different IT jobs. However, some phrases have overlapping meanings, such as front-end, front-end development, and front-end developer. To summarize the ten IT job clusters, the most frequent word and phrase extractions were performed to analyze each job profile.

The results of all job profiles consist of the programming language used in every job, the specialized type or other names that have the same area for each job, and duty as a job requirement and description. The database, tools, and framework were identified as support systems for each job. An in-depth analysis is provided in the next section.

C. AVERAGE LINKAGE HIERARCHICAL CLUSTERING

The clustering stage using the average linkage hierarchical clustering algorithm begins with calculating the similarity between data using the Jaccard similarity index. Jaccard similarity index as a measure of similarity between data that forms inter-cluster relationships. Clustering results can be seen through dendrogram visualization, as an example dendrogram of word reflecting knowledge in Computer Programmers' profiles as shown in Figure 5. Results present the group of words that occur together in job advertisements. For example, Computer programmers are expected to primary one or more types of programming languages, such as Java, Golang, PHP, ASP.Net, JavaScript, C++, Java, and Python. The databases that programmers often use are SQL Server, NoSQL, and PostgreSQL combined with additional APIs-Rest tools.

D. VISUALIZATION AND JOB PROFILES ANALYSIS

The results of the average linkage hierarchical clustering analysis consisted of six areas. First, the programming language is a set of syntax and semantic rules for defining computer programs, standard instructions, and instructing computers to perform certain functions. Second, the framework is a basic conceptual structure that is used to solve or deal with complex problems. Third, the specialized type includes other names that have the same area for each job. Fourth, duty is a job requirement and a description. Fifth, the database is an organized collection of data that is generally stored and accessed electronically from a computer system. Sixth, tools serve to perform commands in a particular application. This study provides each IT job profile based on Programming Language, Specialized Type or other names, Duty, Database, Tools, and Framework as follows:

1) WEB DEVELOPERS

The average-linkage hierarchical cluster analysis of words reflecting knowledge in Web Developers' profiles shows that Web Developer has some duties, such as designing and maintaining interface layout, features, and navigation for usability, known as user interface-user experience (UI-UX) design. Variants of programming language needed by a web developer, such as HTML-CSS, occur at the table most frequently because HTML is used to create webpages, and CSS is used to control the layout and style of webpages. The other programming languages in the table with the most frequent words are JavaScript, PHP, Java, and Python.

Databases and data warehousing, as integrated parts in IT fields, have process linkages, such as data retrieval, processing, and storage. Sometimes, a specialized web developer is called a front end/back end/full-stack developer. The results revealed that the SQL Server, MySQL, and MongoDB were the most demanded databases for web developers. Combination tools for web developer jobs include APIs-Rest, Git, Redux, and jQuery. APIs-Rest is an interface used by two computer systems to exchange information securely over the Internet [59]. Git is a distributed version of the control system used to coordinate work among programmers. Redux and jQuery are open-source JavaScript libraries used for managing and centralizing applications [60]. Many open-source frameworks are available for building websites, including Laravel, Angular, Angular JS, Vue JS, Bootstrap, and Node JS. Laravel and Vue, as front-end runtime environments, follow the model-view controller design pattern concept. Angular JS is typically used to build one-page client applications, and angular JS is typically used to develop dynamic websites with JavaScript. A bootstrap was designed to speed up the process of developing a responsive and mobile-first Web. Node JS is known as a cross-platform because it is an open-source server environment that can run on Windows, LINUX, MacOS, and so on [61]. The connection between entities of different types of nodes is mostly used to explain the relationship between them. Figure 3 shows the connections between the programming languages, tools, and frameworks using a force-based graph layout. Each node is represented by programming languages, tools, and frameworks, and their relationship is represented by an edge. The thickness of this line represents the strength of the relationship. As shown in Figure 3, CSS and HTML have a strong relationship because their appearances are always together. Their relationship is represented by a Jaccard similarity of 0.924. Laravel is an open-source PHP framework that contains many basic modules to optimize the PHP performance in web application development. Moreover, the PHP is a dynamic programming language, and Laravel can make web development faster, safer, and more straightforward.

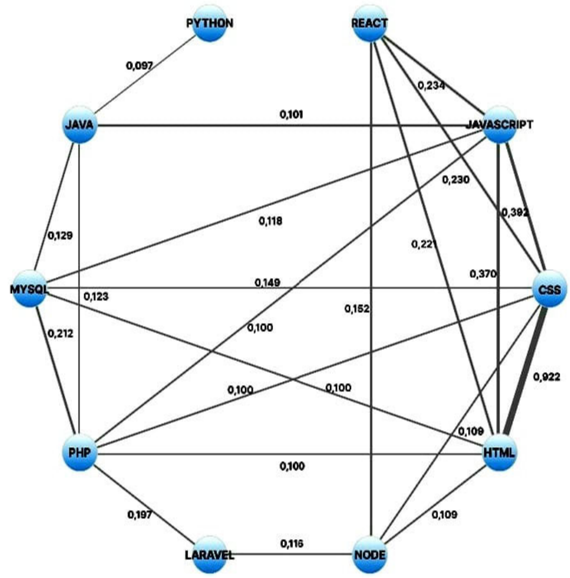

FIGURE 3. Link Analysis of word reflecting connections between language programming, tools, and framework as Web Developers.

FIGURE 4. Words cloud for Software Developer, Quality Assurance, and Testers advertisements with TF-IDF algorithm.

2) SOFTWARE DEVELOPER, QUALITY ASSURANCE, AND TESTERS (SD-QAT)

Software and senior engineers are also known as SD-QAT. This job includes duties such as developing software, designing applications, testing automation, and quality assurance, as shown in the word cloud in Figure 4. Software developers, quality assurance analysts, and testers are involved in creating a software program [62], [63]. To do so, they must know some language programming, such as HTML-CSS, JavaScript, Net, Golang, Python, PHP, and Java. The results revealed that MySQL, PostgreSQL, and MongoDB are the databases most in demand by SD-QAT.

Sustainable software development using an agile methodology is essential for SD-QAT. Some tools are required to fulfill the SD-QAT requirements. APIs-Rest supports the exchange of information as it follows secure, reliable, and efficient software communication standards. IOS and Android are part of the (OS) tools supporting software. The SD-QAT, such as Node JS, requires frameworks such as React JS and Spring JS. React JS as a front-end JavaScript library for building user interfaces. Spring is a java-based open-source framework that provides comprehensive infrastructure for developing Java applications.

3) COMPUTER SYSTEM ANALYSTS (CSA)

Computer System Analysts, known as system architects, design systems and analyze an organization's data, system, and business to increase efficiency. CSA usually uses SQL servers to manage databases. The SQL Server is client/server-based software because it has a client component that displays and manipulates data and a server component that

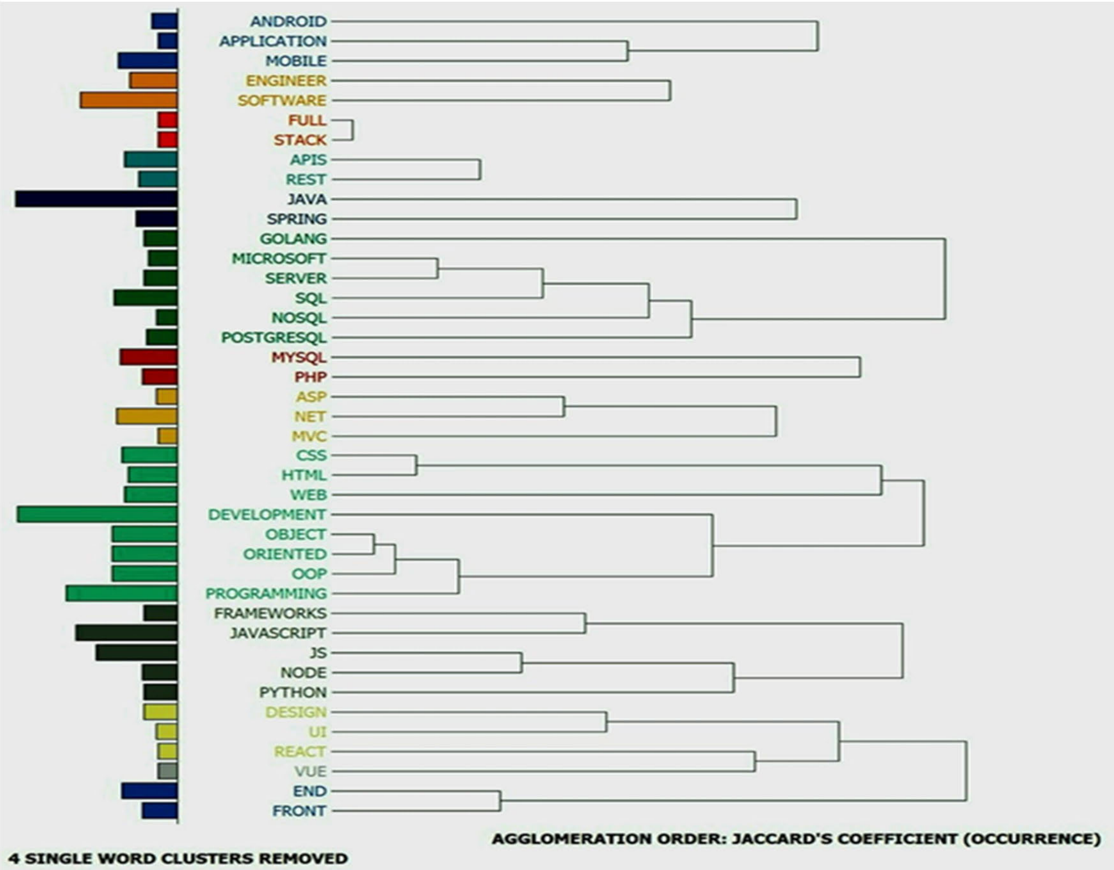

FIGURE 5. Average-Linkage Hierarchical Cluster Analysis of word reflecting knowledge in Computer Programmers' profiles.

stores, retrieves, and secures databases. To design computer systems, CSA uses various techniques such as data modeling, which allows analysts to view processes and data flows by implementing an agile approach, sometimes called project management. Nowadays, Tableau and SQL are tools widely used by companies to analyze and visualize data, including computer system analysts.

4) INFORMATION SECURITY ANALYSTS (ISA)

The information security analysts identified in this study were cyber-security analysts. It has duties such as controlling and monitoring risk infrastructure, data security, and developing standard security in an organization [64], [65]. Python is required for tool development, monitoring code reviews, and automation of certain ISAs. The development of cloud computing to improve information security in an organization requires the ISA to have technical capabilities to maintain data and information security in cloud data storage. The benefits of LINUX tools as an open-source framework that can reduce IT operational costs [66], authorize security, and can be run on all devices platforms [67] are the main attractions for business organizations to use them. To maintain the software, the ISA can use firewalls to protect sensitive information [68]. Several certifications required for the ISA profession are certified information security auditors (CISA), certified information security managers (CISM), and certified information systems security professionals (CISSP).

5) COMPUTER SUPPORT SPECIALISTS

Computer Support Specialists assist computer users and organizations by maintaining computer networks or by providing technical assistance directly to computer users.

Computer support specialists are also known as IT helpdesk technicians. Computer support specialists have duties such as testing and evaluating computer systems for clients/users and service management as provided in the Service Level Agreement (SLA). Service Level Agreement (SLA) is an agreement between a company or customer and its service provider that documents in detail the services that the provider must provide to fulfill customers' expectations. In the IT world, SLA must ensure that service providers meet customer expectations. Routine maintenance is carried out to ensure that the network operates appropriately, solving network problems in Local Area Networks (LAN), Wide Area Networks (WAN), and Internet systems, and guiding customers through problems step-by-step, such as re-establishing an Internet connection or troubleshooting Wi-Fi routers or other IT products. In other words, Computer Support Specialists' roles include IT Helpdesk with troubleshooting services, testing, and evaluating existing network systems. The tools commonly used by them are Linux and firewalls.

6) COMPUTER PROGRAMMER

Writing and testing codes or scripts to ensure that the function of the application or software runs properly is a duty of Computer Programmers. They are also responsible for functions that fit user requirements. If software fails, the programmer checks and modifies the code or scripts for errors. As presented in Figure 5, the average linkage hierarchical cluster analysis of words reflects knowledge in Computer Programmers' profiles. Programmers work closely with software developers, and their duties overlap in some businesses. Computer programmers are generally referred to as programmers or software engineers. Computer programmers are expected to primary one or more types of programming languages, such as Java, Golang, PHP, ASP.Net, JavaScript, C++, Java, and Python. The duties of computer programmers include implementing software design in programming languages, and developing existing programs. The databases that programmers often use are SQL Server, NoSQL, and PostgreSQL combined with additional APIs-Rest tools. This study identified the frameworks used by programmers, such as Node, Spring, React, and Vue. In addition, computer programmers can write or deploy software-as-a-service (SaaS) on online platforms. Sometimes, programmers must modify their programs to work on various system platforms implemented in SaaS.

7) DATABASE ADMINISTRATORS AND ARCHITECTS (DBA)

Database administrators and architects are responsible for creating or managing systems for storing, backing, maintaining, and securing various types of data, as well as ensuring that the data are available to authorized users or those whose access rights. New systems and applications require new databases to be designed and built using DBA. They also researched the organization's technical requirements, created and coded the data architecture, integrated the existing database, and checked for sources of inefficiency. Data can be an important asset for increasing company profits, especially in digital business [69]. In other words, this ensures that a company's database management system functions properly. As shown in Figure 10, some database versions such as structured query language (SQL) are used in this field to access, change, and manipulate data in a relational database. Oracle, NoSQL, MySQL, and PostgreSQL are familiar with these job advertisements; the differences lie in the functions and features that have their respective advantages. Tools used to support database administrator applications include SQL Server Integration Services (SSIS) and Online Analytical Processing (OLAP). One of the platforms for enterprise-level data integration and data transformation is SSIS. It solves business data problems by performing tasks, such as data mining and managing SQL Server objects and data. OLAP supports business intelligence using large databases. It consists of categories, and each category is organized and designed by the administrator to match the method of retrieving and analyzing data [70]. It is also provides aggregated data visualizations from different perspectives [71].

8) COMPUTER AND INFORMATION RESEARCH SCIENTISTS (CIRS)

Computer and Information Research Scientists have devised innovative uses for new information technologies, studying and solving complex problems in computing in business, science, and other fields. Theories and models must be developed based on data analysis using machine-learning techniques from a data visualization perspective. Python has become a popular programming language in the research and design industries, especially in relation to data science and artificial intelligence, such as machine learning and deep learning [72]. SQL is used to manage databases as a query language. In this digital transformation, the trend of Robotic Process Automation (RPA) research can be utilized to enhance digital workers [73]. The CIRS explores how a robot or machine can be developed and interact with the physical environment. The robot was designed to have functions, such as information processing and sensory feedback. Machine learning or deep learning frameworks, such as TensorFlow, Keras, and Caffe, make it easier for data scientists to develop machine-learning models [74].

9) COMPUTER NETWORK ARCHITECTS (CNA)

Computer network architects (CNA) design and build data communication networks, including local area networks (LANs), wide area networks (WANs), intranets, and cloud infrastructure. Updating hardware, such as routers or adapters, and software, such as network drivers, are also responsible. Computer network architects used. NET programming framework as an open-source platform to build desktop, web, and mobile applications that can run natively on any operating system. The .NET systems include tools, libraries, and languages that support modern, scalable, and high-performance software developments. Recently, cloud-based data warehouses have been used by CNAs for their usability and accessibility. This study found that the tools used by the CNA were CISCO, Firewall, LINUX, Azure, and Amazon Web Service (AWS). CISCO is the leading equipment widely used in wide-area networks (WAN) because information is forwarded to addresses far apart and on different computer networks. Azure is Microsoft's public cloud platform that offers a broad collection of services, spanning platform as a service (PaaS), infrastructure as a service (IaaS), and managed database service capabilities. Amazon Web Services (AWS) is a cloud service provider offering computing power, database storage space, content delivery networks, and other functionalities that help businesses grow and securely run applications [75].

TABLE 5. Result of average linkage hierarchical clustering algorithm for IT job profile.

Job Title	Programming Language	Specialized Type or Other Name	Duty	Database	Tools	Framework
Web Developers	HTML-CSS, JavaScript, Java, PHP, Python	Front-End Developer, Back-End Developer, Full-stack	Design UI-UX	SQL Server, MySQL, MongoDB	APIs-Rest, Git, Redux, jQuery	Laravel, Angular, Angular JS, Vue JS, Bootstrap, Node JS
Computer System Analysts		System Architects	Data analysis, business analysis, system analysis, design system	SQL Server	Agile Software, Microsoft Excel, SQL-Tableau	
Software Developers, and Quality Assurance, and Testers	HTML-CSS, JavaScript, Net, Golang, Python, PHP, Java,	Software Engineer, Senior Engineer	Develop software, design application, test automation, quality assurance	MySQL, PostgreSQL, MongoDB	Agile, APIs-Rest, IOS, Android	Node JS, React JS, Spring
Information Security Analysts	Python	Cyber security	Control and monitor risk infrastructure, and data security, develop standard security		LINUX base, firewall	
Computer Programmers	Java, Golang, PHP, ASP.Net, HTML-CSS, JavaScript, Python	Programmer, Developer, Software engineer	Design and develop a program	SQL Server, NoSQL, PostgreSQL,	APIs-Rest,	Node JS, Spring JS, React JS,Vue JS
Computer Support Specialists	.Net	IT Support	Test and evaluate computer systems for a client, Service Level Agreement, Helpdesk for problem service	Not available	LINUX, firewall,	
Database Administrators and Architects			Design and develop database, maintenance and backup database, control access rights	SQL, Oracle, NoSQL, MySQL, PostgreSQL	SSIS, OLAP	
Computer and Information Research Scientists	Python	Data Science	Data analysis, design, and experiment using Machine Learning technique	SQL		Machine Learning
Computer Network Architects	.Net	Network Engineer	Design infrastructure network, configure LAN and WAN		CISCO, Firewall, LINUX, Azure, AWS	
Network and Computer System Administrators		Network Engineer Staff	Maintain hardware and network infrastructure		CISCO, Firewall, LINUX, VMWare	

10) NETWORK AND COMPUTER SYSTEM ADMINISTRATORS (NCSA)

Network and computer system administrators (NCSA) are responsible for day-to-day network operations, such as setting up, installing, and supporting an organization's computer systems, including local area networks (LANs), wide area networks (WANs), network segments, intranets, and other data communication systems. NCSA is commonly referred to as network engineer staff. Duties at the NCSA determine organizational system requirements and install network hardware and software. It also performs network upgrades and repairs, and ensures that the system operates correctly. Maintain network and computer system security, and evaluate and optimize the network performance or system. Currently, the database used by NCSA is cloud server-based. VMWare offers virtualization and cloud computing to servers, computers, and other hardware devices. Other tools commonly used by NCSA are CISCO, Firewall, and LINUX.

A summary of all ten job profiles is presented in Table 5. This table is the result of a systematic semantic analysis using the ALHC algorithm and is expected to provide a clear and detailed description of the IT job profile. The validation of this IT job profile has also been justified by IT professionals from various private and governmental companies in Indonesia through Focus Group Discussions (FGD). Valid results were obtained using the FGD.

E. VENN DIAGRAM OF INFORMATION TECHNOLOGY JOBS

Venn diagrams, as an in-depth analysis of job profiles, explain the similarity of functions of each tool of computer networks, tools of software, frameworks, programming languages, and databases to each IT jobs fields. Each function of the Venn diagram shows the intersection of several jobs inthat have close proximity to the scope, duties, and responsibilities described in the job description.

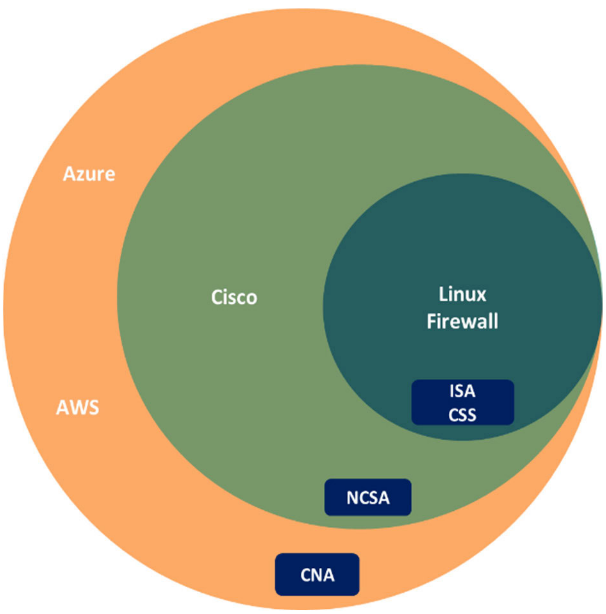

FIGURE 6. Venn Diagram of tools of computer network.

1) TOOLS OF COMPUTER NETWORK

Tools of a computer network in the IT Jobs field include a variety of operating systems and platforms used for managing computer network systems. Figure 6 shows that some of the tools of a computer network used by Information Security Analysts and Computer Support Systems, such as Linux and Firewall, are also used by computer network architects and network and computer system administrators because they have work related to organizing computer networks and supporting each other. Cisco acts as a tool for managing computer system networks that a CNA and NCSA need, because both jobs focus on the architecture and management of day-by-day computer system network operations. The CNA uses Azure and AWS as platforms for designing cloud-based computer network management.

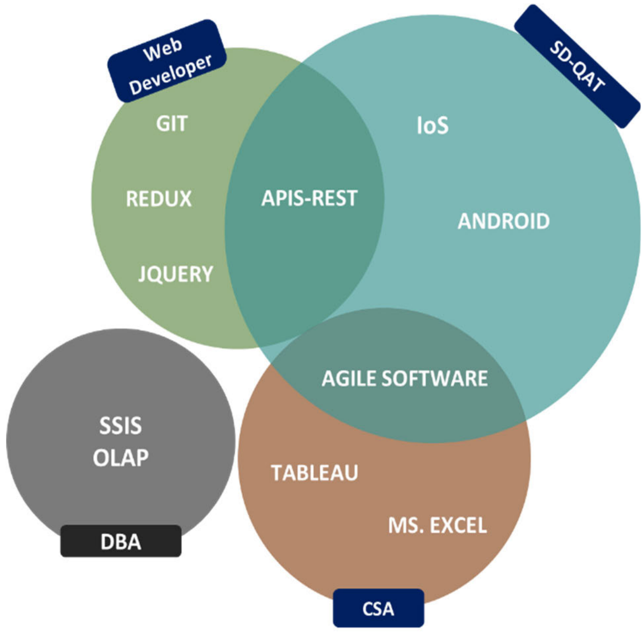

FIGURE 7. Venn diagram of tools of software.

2) TOOLS OF SOFTWARE

Tools of software in IT jobs are various libraries calling the functions, control, and operating systems needed in developing computer programs, as presented in Figure 7. Web developers and SD-QAT use APIS-Rest as an interface model for two or more computer systems to exchange information securely over the internet. CSA and SD-QAT use Agile Software to develop an effective effort estimation model to facilitate project planning and successful implementation.

Agile methods focus on customer satisfaction by ensuring that software products are faster and more economical, with enhanced quality [76] and minimizing risks after project completion [77]. iOS and Android are the most commonly used operating systems by software developers. Tableau and Microsoft Excel are tools used in data visualization across business intelligence areas to help computer system analysts understand complex data. Web developers use Git, Redux, and JQuery to manage and centralize applications. Nowadays, SSIS and OLAP have become real-time technologies used to organize large business databases and support business intelligence [70], chosen by Database Administrators and Architects.

3) FRAMEWORKS

Framework is a basic conceptual structure used to solve or deal with a complex problems. Frameworks make web developers, software developers, and computer programmers easier to develop websites, save processing time and application development costs, and help in writing lines of code more easily, quickly, and in a structured manner.

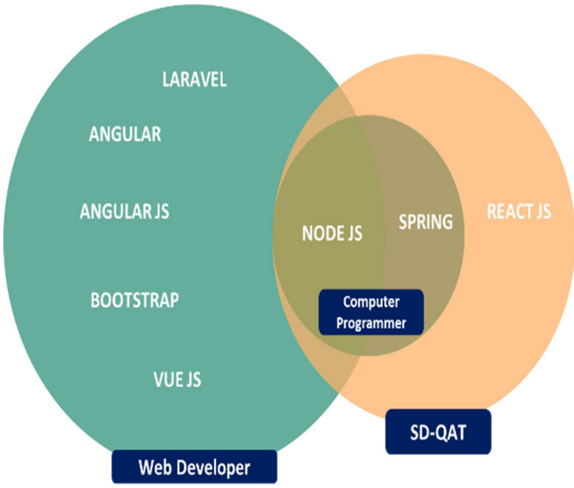

FIGURE 8. Venn Diagram of frameworks.

Figure 8 shows that the Node js framework is an open source cross-platform related to web developers, software developers, and computer programmers because node js are flexible for use in these jobs. Spring is an open-source framework that helps optimize application performance, particularly for software developers and computer programmers. Integrated frameworks used by web developers are generally Laravel, which can make websites more dynamic, and angular-typescript frameworks are used to create large-scale websites. Additionally, Angular JS is used to create single-page websites that are much more accessible. To speed up the work on creating responsive websites, web developers use bootstrap on it. In other cases, Vue JS is a single-page application that creates a User Interface (UI) attractively. React js is a unique framework that uses only SD-QAT for software development, as part of a single UI.

4) PROGRAMMING LANGUAGES

Programming language is a set of syntax and semantic rules for defining computer programs, standard instructions, and instructing computers to perform certain functions. Figure 9 shows that Python is the most popular programming language that can be used by the ISA, CIRC, Web Developer, SD-QAT, and computer programmers.

Computer programmers cover all the programming languages. Python is open-source programming and can be used by any programmer because it builds controls and bugs such as software tests, server-side websites, and data analysis for data science projects. Golang is a programming language used by SD-QAT to simplify the software development process and build complex process architectures with a high security impact. ASP.net represents active server-page network-enabled technologies, which are open sources developed by Microsoft to enable programmers to build dynamic websites, applications, and services. Computer programmers only use ASP.net.

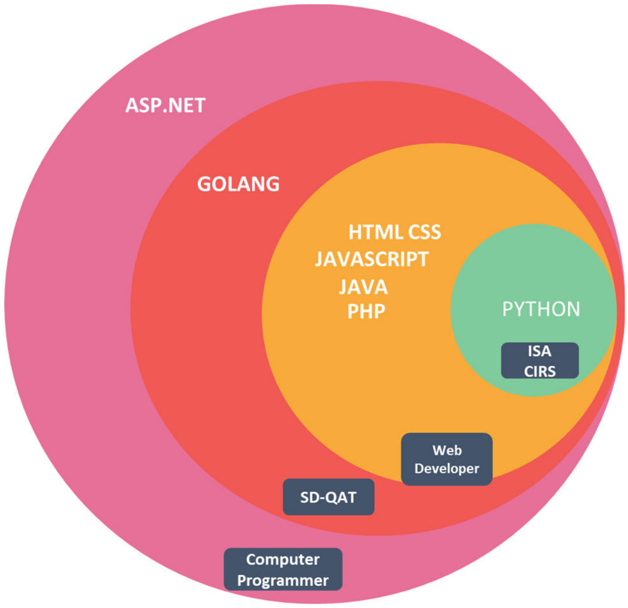

FIGURE 9. Venn diagram of programming language.

Web developers generally use the HTML-CSS programming language to create websites. JavaScript was used to make the website more interactive. Java is used to manage the website database section, and PHP exist on the server side to manage dynamic content, databases, session tracking, and even entire e-commerce sites.

5) DATABASES

Databases are organized data collections stored and accessed electronically from a computer system. As databases become more complex, they are being developed using formal design and modeling techniques. As shown in Figure 10, MySQL is used as a Database Management System (DBMS) for almost every IT job field. These are executed using SQL commands to create website-based or mobile applications that support large databases. SQL is a Structured Query Language that is helpful as a particular language in accessing and managing data in a relational database, focusing on table-based formats for storing data. NoSQL was created for a specific purpose for specific data models and has a flexible schema for creating modern applications that can work with unstructured data. It saves data in document format, JSON, key-value graphs, etc. PostgreSQL is an open-source RDMS used for more complex data processing. Database Administrator Architects

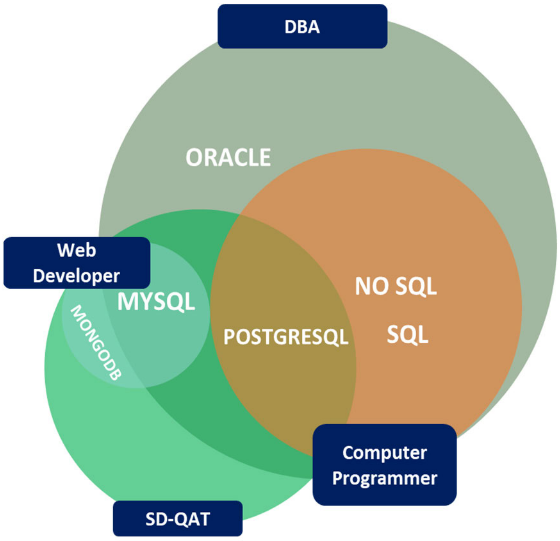

FIGURE 10. Venn diagram of databases.

and Computer Programmers have used Oracle to build large-scale systems. Usually, MongoDB is used by web developers and SD-QAT because it is suitable for storing unstructured data on social media, such as Twitter and Facebook.

V. COMPARISON

This study has comparisons with previous studies, namely:

1. It uses two job portals as datasets so that the data used are objective, more complex, and broader. Previous research used only one dataset from the job portal, so subjectivity could not be avoided.

2. The model was validated by ten IT professionals from various private and government companies, especially in Indonesia. Other studies have not used expert justification; therefore, the results still need further validation from the industrial world.

3. This study systematically analyzes job profiles specifically in the IT field so that they can be used as a guide for workers or students to prepare their technical skills when entering the world of work.

VI. CONCLUSION, LIMITATIONS, AND FUTURE WORKS

This study concludes that, to satisfy industrial requirements, the IT job profile in the hard skills field has been identified and can be considered in computer science or information technology. A practical experiment with 500 Gibbs sampling iterations was conducted using at a rate of per 10,000 words was performed. The proposed method can be improved by regularly updating the list of competencies based on the large and new requirements launched by industries with automatic analysis. A significant improvement in the communication level between industry and IT education programs hopefully can be achieved, and graduates of IT programs can gain the appropriate skills to fulfill the industry's requirements. The results of the job profile analysis in this study have been validated by professionals in their respective fields through focused discussion groups and declared valid according to practice in the field.

Some limitations of our model include:

1) It is difficult to distinguish each type of work in the IT field because of the similarities in terms of duties and responsibilities.

2) To solve this problem, we must manually analyze each type of job.

3) The number of clusters formed must also be adjusted to obtain proper accuracy.

As future research, some potential areas can be improved:

1) Integrate the clustering and classification algorithm to improve the accuracy of the model.

2) The big data of each IT job can be explored to develop a taxonomy of job competencies by implementing the Latent Dirichlet Allocation (LDA) algorithm using topic modeling structures.

REFERENCES

[1] A. of C. S. C.-I. (APTIKOM). (2019). *KKNI Curriculum Development Based on OBE in Informatics and Computer Science*. Jakarta, Indonesia. Accessed: Jan. 13, 2022. [Online]. Available: https://aptikom.org/

[2] U.S Bureau of Labor Statistics. *Occupational Outlook Handbook*. Accessed: Jan. 13, 2022. [Online]. Available: https://www.bls.gov/ooh/computer-and-information-technology/home.htm

[3] G. Karacay, "Talent development for Industry 4.0," in *Industry 4.0: Managing The Digital Transformation* (Springer Series in Advanced Manufacturing). Cham, Switzerland: Springer, 2018, pp. 123–136, doi: 10.1007/978-3-319-57870-5_7.

[4] F. Brunetti, D. T. Matt, A. Bonfanti, A. De Longhi, G. Pedrini, and G. Orzes, "Digital transformation challenges: Strategies emerging from a multi-stakeholder approach," *TQM J.*, vol. 32, no. 4, pp. 697–724, Jul. 2020, doi: 10.1108/TQM-12-2019-0309.

[5] K. Binici, "What are the information technology skills needed in information institutions? The case of 'code4li' job listings," *J. Academic Librarianship*, vol. 47, no. 3, May 2021, Art. no. 102360, doi: 10.1016/j.acalib.2021.102360.

[6] D. S. Wadhwa, "Text mining of job profiles: Finding relevant job skills in different industries," *J. Positive School Psychol.*, vol. 6, no. 2, pp. 5737–5745, 2022.

[7] S. Fareri, G. Fantoni, F. Chiarello, E. Coli, and A. Binda, "Estimating Industry 4.0 impact on job profiles and skills using text mining," *Comput. Ind.*, vol. 118, Jun. 2020, Art. no. 103222, doi: 10.1016/j.compind.2020.103222.

[8] A. De Mauro, M. Greco, M. Grimaldi, and P. Ritala, "Human resources for big data professions: A systematic classification of job roles and required skill sets," *Inf. Process. Manage.*, vol. 54, no. 5, pp. 807–817, Sep. 2018, doi: 10.1016/j.ipm.2017.05.004.

[9] J. Schedlbauer, G. Raptis, and B. Ludwig, "Medical informatics labor market analysis using web crawling, web scraping, and text mining," *Int. J. Med. Informat.*, vol. 150, Jun. 2021, Art. no. 104453, doi: 10.1016/j.ijmedinf.2021.104453.

[10] K. Nair, "Overcoming today's digital talent gap in organizations worldwide," *Develop. Learn. Org.*, vol. 33, no. 6, pp. 16–18, Nov. 2019, doi: 10.1108/DLO-02-2019-0044.

[11] S. U. Habiba, Md. K. Islam, and F. Tasnim, "A comparative study on fake job post prediction using different data mining techniques," in *Proc. 2nd Int. Conf. Robot., Electr. Signal Process. Techn. (ICREST)*, Jan. 2021, pp. 543–546, doi: 10.1109/ICREST51555.2021.9331230.

[12] H. Kortum, J. Rebstadt, and O. Thomas, "Dissection of AI job advertisements: A text mining-based analysis of employee skills in the disciplines computer vision and natural language processing," in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, Jan. 2022, pp. 5211–5220, doi: 10.24251/hicss.2022.635.

[13] M. Treviso et al., "Efficient methods for natural language processing: A survey," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 826–860, Jul. 2023, doi: 10.1162/tacl_a_00577.

[14] G. Varelas, D. Lagios, S. Ntouroukis, P. Zervas, K. Parsons, and G. Tzimas, "Employing natural language processing techniques for online job vacancies classification," in *Proc. FIP Int. Conf. Artif. Intell. Appl. Innov.*, 2022, pp. 333–344, doi: 10.1007/978-3-031-08341-9_27.

[15] F. Iqbal, R. Batool, B. C. M. Fung, S. Aleem, A. Abbasi, and A. R. Javed, "Toward tweet-mining framework for extracting terrorist attack-related information and reporting," *IEEE Access*, vol. 9, pp. 115535–115547, 2021, doi: 10.1109/ACCESS.2021.3102040.

[16] M. Bibi, W. Aziz, M. Almaraashi, I. H. Khan, M. S. A. Nadeem, and N. Habib, "A cooperative binary-clustering framework based on majority voting for Twitter sentiment analysis," *IEEE Access*, vol. 8, pp. 68580–68592, 2020, doi: 10.1109/ACCESS.2020.2983859.

[17] M. Mimis, M. El Hajji, Y. Es-Saady, A. O. Guejdi, H. Douzi, and D. Mammass, "A framework for smart academic guidance using educational data mining," *Educ. Inf. Technol.*, vol. 24, no. 2, pp. 1379–1393, Mar. 2019, doi: 10.1007/s10639-018-9838-8.

[18] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," *Educ. Inf. Technol.*, vol. 23, no. 1, pp. 537–553, Jan. 2018, doi: 10.1007/s10639-017-9616-z.

[19] E. Novak, L. Bizjak, D. Mladenić, and M. Grobelnik, "Why is a document relevant? Understanding the relevance scores in cross-lingual document retrieval," *Knowl.-Based Syst.*, vol. 244, May 2022, Art. no. 108545, doi: 10.1016/j.knosys.2022.108545.

[20] W. Budiharto, *Knowledge and Information Retrieval*. Yogyakarta, Indonesia: Deeppublish, 2016.

[21] T. Bai, Y. Ge, S. Guo, Z. Zhang, and L. Gong, "Enhanced natural language interface for web-based information retrieval," *IEEE Access*, vol. 9, pp. 4233–4241, 2021, doi: 10.1109/ACCESS.2020.3048164.

[22] P. Mahalakshmi and N. S. Fatima, "Summarization of text and image captioning in information retrieval using deep learning techniques," *IEEE Access*, vol. 10, pp. 18289–18297, 2022, doi: 10.1109/ACCESS.2022.3150414.

[23] J. Jiang and K. Srinivasan, "MoreThanSentiments: A text analysis package," *Softw. Impacts*, vol. 15, Mar. 2023, Art. no. 100456, doi: 10.1016/j.simpa.2022.100456.

[24] A. Otegi, I. S. Vicente, X. Saralegi, A. Peñas, B. Lozano, and E. Agirre, "Information retrieval and question answering: A case study on COVID-19 scientific literature," *Knowl.-Based Syst.*, vol. 240, Mar. 2022, Art. no. 108072, doi: 10.1016/j.knosys.2021.108072.

[25] F. Gurcan and N. E. Cagiltay, "Exploratory analysis of topic interests and their evolution in bioinformatics research using semantic text mining and probabilistic topic modeling," *IEEE Access*, vol. 10, pp. 31480–31493, 2022, doi: 10.1109/ACCESS.2022.3160795.

[26] M. E. Ibrahim, Y. Yang, D. L. Ndzi, G. Yang, and M. Al-Maliki, "Ontology-based personalized course recommendation framework," *IEEE Access*, vol. 7, pp. 5180–5199, 2019, doi: 10.1109/ACCESS.2018.2889635.

[27] M. C. Urdaneta-Ponte, A. Méndez-Zorrilla, and I. Oleagordia-Ruiz, "Lifelong learning courses recommendation system to improve professional skills using ontology and machine learning," *Appl. Sci.*, vol. 11, no. 9, p. 3839, Apr. 2021, doi: 10.3390/app11093839.

[28] I. Lauriola, A. Lavelli, and F. Aiolli, "An introduction to deep learning in natural language processing: Models, techniques, and tools," *Neurocomputing*, vol. 470, pp. 443–456, Jan. 2022, doi: 10.1016/j.neucom.2021.05.103.

[29] I. A. Heggo and N. Abdelbaki, "Hybrid information filtering engine for personalized job recommender system," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl.*, vol. 723, 2018, pp. 553–563, doi: 10.1007/978-3-319-74690-6_54.

[30] D. Mhamdi, R. Moulouki, M. Y. El Ghoumari, M. Azzouazi, and L. Moussaid, "Job recommendation based on job profile clustering and job seeker behavior," *Proc. Comput. Sci.*, vol. 175, pp. 695–699, 2020, doi: 10.1016/j.procs.2020.07.102.

[31] H. Jain and M. Kakkar, "Job recommendation system based on machine learning and data mining techniques using RESTful API and Android IDE," in *Proc. 9th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Jan. 2019, pp. 416–421.

[32] A. Rivas, P. Chamoso, A. González-Briones, R. Casado-Vara, and J. M. Corchado, "Hybrid job offer recommender system in a social network," *Expert Syst.*, vol. 36, no. 4, pp. 1–13, Aug. 2019, doi: 10.1111/exsy.12416.

[33] X. Ma and L. Ye, "Career goal-based e-learning recommendation using enhanced collaborative filtering and PrefixSpan," *Int. J. Mobile Blended Learn.*, vol. 10, no. 3, pp. 23–37, Jul. 2018, doi: 10.4018/IJMBL.2018070103.

[34] B. E. V. Comendador, W. F. C. Becbec, and J. R. P. D. Guzman, "Implementation of fuzzy logic technique in a decision support tool: Basis for choosing appropriate career path," *Int. J. Mach. Learn. Comput.*, vol. 10, no. 2, pp. 339–345, Feb. 2020, doi: 10.18178/ijmlc.2020.10.2.940.

[35] L. D. Kumalasari and A. Susanto, "Recommendation system of information technology jobs using collaborative filtering method based on LinkedIn skills endorsement," *Sisforma*, vol. 6, no. 2, pp. 63–72, Feb. 2020, doi: 10.24167/sisforma.v6i2.2240.

[36] R. Mishra and S. Rathi, "Enhanced DSSM (deep semantic structure modelling) technique for job recommendation," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7790–7802, Oct. 2022, doi: 10.1016/j.jksuci.2021.07.018.

[37] D. Shi, T. Wang, H. Xing, and H. Xu, "A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105618, doi: 10.1016/j.knosys.2020.105618.

[38] N. T. Son, J. Jaafar, I. A. Aziz, and B. N. Anh, "Meta-heuristic algorithms for learning path recommender at MOOC," *IEEE Access*, vol. 9, pp. 59093–59107, 2021, doi: 10.1109/ACCESS.2021.3072222.

[39] M. Tavakoli, A. Faraji, J. Vrolijk, M. Molavi, S. T. Mol, and G. Kismihók, "An AI-based open recommender system for personalized labor market driven education," *Adv. Eng. Informat.*, vol. 52, Apr. 2022, Art. no. 101508, doi: 10.1016/j.aei.2021.101508.

[40] T. Li, A. Rezaeipanah, and E. S. M. T. El Din, "An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3828–3842, 2022, doi: 10.1016/j.jksuci.2022.04.010.

[41] A. Bondielli and F. Marcelloni, "On the use of summarization and transformer architectures for profiling résumés," *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115521, doi: 10.1016/j.eswa.2021.115521.

[42] S. Pasupathi, V. Shanmuganathan, K. Madasamy, H. R. Yesudhas, and M. Kim, "Trend analysis using agglomerative hierarchical clustering approach for time series big data," *J. Supercomput.*, vol. 77, no. 7, pp. 6505–6524, Jul. 2021, doi: 10.1007/s11227-020-03580-9.

[43] P. R. Garikapati, K. Balamurugan, T. P. Latchoumi, and G. Shankar, "A quantitative study of small dataset machining by agglomerative hierarchical cluster and K-medoid," in *Emergent Converging Technologies and Biomedical Systems*. Singapore: Springer, 2022, doi: 10.1007/978-981-16-8774-7_59.

[44] A. Benešová and J. Tupa, "Requirements for education and qualification of people in Industry 4.0," *Proc. Manuf.*, vol. 11, pp. 2195–2202, 2017, doi: 10.1016/j.promfg.2017.07.366.

[45] F. Gurcan and N. E. Cagiltay, "Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling," *IEEE Access*, vol. 7, pp. 82541–82552, 2019, doi: 10.1109/ACCESS.2019.2924075.

[46] J. Yang, Y. Li, Q. Liu, L. Li, A. Feng, T. Wang, S. Zheng, A. Xu, and J. Lyu, "Brief introduction of medical database and data mining technology in big data era," *J. Evidence-Based Med.*, vol. 13, no. 1, pp. 57–69, Feb. 2020, doi: 10.1111/jebm.12373.

[47] G. Cornelli, J. Frost, L. Gambacorta, R. Rau, R. Wardrop, and T. Ziegler, "Fintech and big tech credit: A new database," Bank Int. Settlements, Basel, Switzerland, White Paper no. 887, 2020, pp. 1–35.

[48] R. Ray, K. P. Gallagher, W. Kring, J. Pitts, and B. A. Simmons, "Geolocated dataset of Chinese overseas development finance," *Sci. Data*, vol. 8, no. 1, pp. 1–12, Sep. 2021, doi: 10.1038/s41597-021-01021-7.

[49] A. Onan, "Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering," *IEEE Access*, vol. 7, pp. 145614–145633, 2019, doi: 10.1109/ACCESS.2019.2945911.

[50] M. Pejic-Bach, T. Bertoncel, M. Meško, and Ž. Krstić, "Text mining of Industry 4.0 job advertisements," *Int. J. Inf. Manage.*, vol. 50, pp. 416–431, Feb. 2020, doi: 10.1016/j.ijinfomgt.2019.07.014.

[51] D. Jiang, W. Lin, and N. Raghavan, "A Gaussian mixture model clustering ensemble regressor for semiconductor manufacturing final test yield prediction," *IEEE Access*, vol. 9, pp. 22253–22263, 2021, doi: 10.1109/ACCESS.2021.3055433.

[52] P. Govender and V. Sivakumar, "Application of *k*-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," *Turkish Nat. Committee Air Pollut. Res. Control*, vol. 11, no. 1, pp. 40–56, Jan. 2020.

[53] G. Zhu, N. A. Kopalle, Y. Wang, X. Liu, K. Jona, and K. Börner, "Community-based data integration of course and job data in support of personalized career-education recommendations," *Proc. Assoc. Inf. Sci. Technol.*, vol. 57, no. 1, p. e324, Oct. 2020, doi: 10.1002/pra2.324.

[54] Tech in Asia Pte Ltd. (2022). *Singapore. Tech in Asia Advertisement*. Accessed: Mar. 5, 2022. [Online]. Available: https://www.techinasia.com/jobs/search?country_name[]=Indonesia

[55] Jobstreet. (2022). *JobStreet Advertisement*. Accessed: Mar. 5, 2022. [Online]. Available: https://www.jobstreet.co.id/

[56] *Projection of Manpower Needs in Companies Based on Competency in the Information Technology & Communication Sector in 2022–2025*, Ministry of Manpower of the Republic of Indonesia, Jakarta, Indonesia, 2021.

[57] K. Abbas, M. K. Hasan, A. Abbasi, U. A. Mokhtar, A. Khan, S. N. H. S. Abdullah, S. Dong, S. Islam, D. Alboaneen, and F. R. A. Ahmed, "Predicting the future popularity of academic publications using deep learning by considering it as temporal citation networks," *IEEE Access*, vol. 11, pp. 83052–83068, 2023, doi: 10.1109/access.2023.3290906.

[58] L. Guo, C. J. Vargo, Z. Pan, W. Ding, and P. Ishwar, "Big social data analytics in journalism and mass communication," *J. Mass Commun. Quart.*, vol. 93, no. 2, pp. 332–359, Jun. 2016, doi: 10.1177/1077699016639231.

[59] N. Laranjeiro, J. Agnelo, and J. Bernardino, "A black box tool for robustness testing of REST services," *IEEE Access*, vol. 9, pp. 24738–24754, 2021, doi: 10.1109/ACCESS.2021.3056505.

[60] B. Bibeault, Y. Katz, and A. De Rosa, *jQuery in Action*, 3rd ed. Greenwich, CT, USA: Manning Publications, 2015.

[61] C. Northwood, *The Full Stack Developer*. Manchester, U.K.: Apress Media, 2018.

[62] M. S. Hossain, "Challenges of software quality assurance and testing," *Int. J. Softw. Eng. Comput. Syst.*, vol. 4, no. 1, pp. 133–144, Feb. 2018, doi: 10.15282/ijsecs.4.1.2018.11.0044.

[63] C. Laporte, A. April, *Software Quality Assurance*. Hoboken, NJ, USA: Wiley, 2018.

[64] M. M. Alhassan and A. Adjei-Quaye, "Information security in an organization," *Int. J. Comput.*, vol. 24, no. 1, pp. 100–116, 2017. [Online]. Available: http://ijcjournal.org/

[65] E. Karanja, "The role of the chief information security officer in the management of IT security," *Inf. Comput. Secur.*, vol. 25, no. 3, pp. 300–329, Jul. 2017, doi: 10.1108/ICS-02-2016-0013.

[66] H. T. Al-Rayes, "Studying main differences between Android & Linux operating systems," *Int. J. Electr. Comput. Sci.*, vol. 12, no. 5, pp. 46–49, 2012.

[67] H.-C. Kuo, D. Williams, R. Koller, and S. Mohan, "A Linux in unikernel clothing," in *Proc. 15th Eur. Conf. Comput. Syst.*, Apr. 2020, pp. 1–15, doi: 10.1145/3342195.3387526.

[68] H. H. Khairi, S. H. S. Ariffin, N. M. A. Latiff, K. M. Yusof, M. K. Hassan, and M. Rava, "The impact of firewall on TCP and UDP throughput in an Openflow software defined network," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 20, no. 1, pp. 256–263, 2020, doi: 10.11591/ijeecs.v20.i1.pp256-263.

[69] A. Ramadhan, "Data capital a systematic literature review," *DESIDOC J. Library Inf. Technol.*, vol. 42, no. 2, pp. 119–129, Feb. 2022, doi: 10.14429/djlit.42.2.17503.

[70] S. A. Shaikh and H. Kitagawa, "StreamingCube: Seamless integration of stream processing and OLAP analysis," *IEEE Access*, vol. 8, pp. 104632–104649, 2020, doi: 10.1109/ACCESS.2020.2999572.

[71] P. O. Queiroz-Sousa and A. C. Salgado, "A review on OLAP technologies applied to information networks," *ACM Trans. Knowl. Discovery Data*, vol. 14, no. 1, pp. 1–25, Feb. 2020, doi: 10.1145/3370912.

[72] J. Blank and K. Deb, "Pymoo: Multi-objective optimization in Python," *IEEE Access*, vol. 8, pp. 89497–89509, 2020, doi: 10.1109/ACCESS.2020.2990567.

[73] J. Ribeiro, R. Lima, T. Eckhardt, and S. Paiva, "Robotic process automation and artificial intelligence in industry 4.0—A literature review," *Proc. Comput. Sci.*, vol. 181, no. 2019, pp. 51–58, 2021, doi: 10.1016/j.procs.2021.01.104.

[74] W. G. Hatcher and W. Yu, "A survey of deep learning: Platforms, applications and emerging research trends," *IEEE Access*, vol. 6, pp. 24411–24432, 2018, doi: 10.1109/ACCESS.2018.2830661.

[75] S. Chenthara, K. Ahmed, H. Wang, and F. Whittaker, "Security and privacy-preserving challenges of e-health solutions in cloud computing," *IEEE Access*, vol. 7, pp. 74361–74382, 2019, doi: 10.1109/ACCESS.2019.2919982.

[76] M. Fernández-Diego, E. R. Méndez, F. González-Ladrón-De-Guevara, S. Abrahão, and E. Insfran, "An update on effort estimation in agile software development: A systematic literature review," *IEEE Access*, vol. 8, pp. 166768–166800, 2020, doi: 10.1109/ACCESS.2020.3021664.

[77] W. Aslam and F. Ijaz, "A quantitative framework for task allocation in distributed agile software development," *IEEE Access*, vol. 6, pp. 15380–15390, 2018, doi: 10.1109/ACCESS.2018.2803685.

PUJI CATUR SISWIPRAPTINI received the master's degree in information technology from Universitas Indonesia, in 2009. She is currently pursuing the Ph.D. degree with the Department of Computer Science, Bina Nusantara University. She is a Lecturer with Institut Teknologi PLN, Jakarta, Indonesia. Her research interests include intelligence and information systems, data mining, information retrieval, and artificial intelligence. She received the Best Presenter Award at the Asia Future Conference, Japan, in 2016, and the Best Researcher Award from the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia, in 2018.

HARCO LESLIE HENDRIC SPITS WARNARS received the Ph.D. degree in computer science from Manchester Metropolitan University, Manchester, U.K. His Ph.D. dissertation was on data mining from 2008 to 2012. His Ph.D. was funded by the Directorate General of Higher Education, Ministry of Education and Culture, Republic of Indonesia (DIKTI) Scholarship. He is the Head of Concentration of Information Systems at the Doctor of Computer Science (DCS), Bina Nusantara University, and a Supervisor of Ph.D. students in computer science. He has published more than 261 articles, which were cited by 1292 articles with Google Scholar H-index=21 and i10-index=51, as can be seen in his Google Scholar link. He has published SCOPUS, as can be seen in his SCOPUS link account with 157 documents, which were cited by 244 documents with an H-index of 12.

ARIEF RAMADHAN (Senior Member, IEEE) is an Assistant Professor with Telkom University, Bandung, Indonesia. His research interests include information systems, information technology, e-business, e-government, e-learning, e-tourism, business intelligence, enterprise architecture, gamification, data analytics, and metaverse. He currently supervises several Ph.D. students in his research fields. His Scopus ID is 53264599300.

WIDODO BUDIHARTO (Senior Member, IEEE) received the bachelor's degree in physics from the University of Indonesia, Jakarta, Indonesia, the master's degree in information technology from STT Benarif, Jakarta, and the Ph.D. degree in electrical engineering from the Institute of Technology Sepuluh Nopember, Surabaya, Indonesia. He took the Ph.D. Sandwich Program in robotics with Kumamoto University, Japan, and was a Postdoctoral Researcher in robotics and artificial intelligence with Hosei University, Japan. He was a Visiting Professor with the Erasmus Mundus French Indonesian Consortium (FICEM), France, Hosei University, in 2017 and 2016, and an Erasmus Mundus Scholar with the EU Universite de Bourgogne, France, in 2007. He is currently a Professor of artificial intelligence with the School of Computer Science, Bina Nusantara University, Jakarta. His research interests include intelligent systems, data science, robotic vision, and computational intelligence.