International Journal of Data Science and Artificial Intelligence (IJDSAI)
Volume 02, Issue 03, May – June (2024)

RESEARCH ARTICLE

SEMANTIC FEATURE ENABLED AGGLOMERATIVE CLUSTERING FOR INFORMATION TECHNOLOGY JOB PROFILE ANALYSIS

B. Jaison [1, *], R. Gladis Kiruba [2] and G Belshia Jebamalar [3]

[1] Department of Computer Science and Engineering, R.M.K. Engineering College, Kavaraipettai-601206 India.
[2] Department of Electronics and Communication Engineering and AIML, Bangalore College of Engineering and Technology, Chandapura, Bengaluru, India.
[3] Department of Computer Science Engineering, S.A Engineering College, Thiruverkadu, Tamil Nadu 600077 India.

Corresponding e-mail: bjn.cse@rmkec.ac.in

Abstract – The maintenance and implementation of computer systems are the core activities of information technology. Database administration and network architecture are also included in information technology. Professionals have access to a working environment that facilitates the setup of internal networks and the development of computer systems. There is an immediate need for a suitable approach to close the gap between supply and demand for IT workers. Extensive research into IT job profiles is crucial to meeting industry demands. Educational programs must identify the abilities that the industry requires to modernize its manufacturing. Semantic Feature-Enabled Agglomerative Clustering for Information Technology Job Profiling (SEA-IT) has been proposed to overcome these challenges. Semantic analysis is performed using a tree-like strategy. The most frequently used phrases and words from each cluster of IT professions were collected to demonstrate specific knowledge. Initially, the data from the online job posting sources will be collected and pre-processed using techniques such as stemming, normalization, text correction, removing stop words, and tokenization. Secondly, the pre-processed data can extract features using a bag of words. After feature extraction, the cluster is generated using an agglomerative algorithm to form an IT job analysis result, so that the knowledge and capabilities of IT professionals can be upgraded. The simulation findings, based on evaluation criteria and other statistical tests, demonstrated the suggested algorithm. Experiments demonstrated that SEA-IT functions well with a variety of descriptive methodologies and is independent of the dataset's dimensions.

Keywords *– Information Technology, Preprocessing, bag of words, agglomerative algorithm.*

1. INTRODUCTION

Informatics engineering is a subject that teaches students the principles of computer science and mathematical analysis in order to create, build, test, and evaluate software [1]. Database administrators, data scientists, UI designers, IT project managers, network engineers, UX designers, and other professions in the field of informatics technology are all open to graduates of the Department of Informatics Engineering with a computer science degree [2]. It differs from other majors in that a student with a doctoral degree in education will work as a doctor, a student with a nursing degree will work as a nurse, a student with a pharmacy degree will work as a pharmacist, a student with an education degree will work as a teacher, and so on [3].

However, students studying informatics engineering won't be sufficient if they merely have a good education without complementing talents [4]. Students majoring in informatics engineering must broaden their skill sets in order to meet the requirements for expert workers in certain IT domains [5]. Career seekers must be aware of the specialized knowledge and skills required for each IT career field from the wide variety of IT job fields [6]. The scarcity of graduates prepared for the workplace in informatics technology areas is caused by the fact that informatics engineering students frequently do not know the benchmark of their capacity for work skill requirements in IT industries [7].

Application system suggestion careers in IT sectors based on IT knowledge and abilities are therefore essential [8]. Even though each student who completes the Informatics Engineering program is equal, they all have unique knowledge and talents [9]. The terms and phrases that are most frequently used in the information technology sector to represent skills and knowledge as a new dataset serve as the foundation for the model's functionality. Ten IT professionals from various commercial and government enterprises in Indonesia confirmed the skills needed for each position through focus group discussions (FGD) [10]. The main contribution of the suggested method is as follows:

- Initially, the data from the online job posting sources will be collected and pre-processed using techniques such as stemming, normalization, text correction, removing stop words, and tokenization.

- Secondly, the pre-processed data can extract features using a bag of words. After feature extraction, the cluster is generated using an agglomerative algorithm to form an IT job analysis result. so that the knowledge and capabilities of IT professionals can be upgraded;
- The simulation findings, based on evaluation criteria and other statistical tests, demonstrated the suggested algorithm.
- Experiments demonstrated that SEA-IT functions well with a variety of descriptive methodologies and is independent of the dataset's dimensions.

The remaining sections of the paper are arranged as follows. Section 2 offers a review of related work. The model design and the main suggested algorithms are then presented in Sections 3 and 4, respectively. Additionally, Section 5 provides analysis and evaluation outcomes. The profession is concluded in Section 6 lastly.

2. LITERATURE SURVEY

In 2022 E. Novak et al. [11] proposed to identify the job profiles that IT specialists need. A systematic semantic technique was suggested using a hierarchical clustering analysis based on average linkage. Semantic analysis, which is akin to a tree structure technique, is used to uncover pertinent phrases, connections, and hidden meanings. The end result is a methodical semantic examination of the programming language, specialized kind, task, database, tools, and frameworks included in the IT job profile. Ten IT experts from various government and commercial enterprises in Indonesia participated in focus groups (FGD) to confirm the rationale behind each job profile.

In 2020 T. Bai, et al. [12] proposed the users should be able to retrieve needed information with ease thanks to the depiction and structuring of information elements. Thus, through a thorough literature review and the administration of two surveys, one for IT employers and the other for graduates, this paper seeks to determine the key variables influencing IT graduates' employability and capacity to compete in local, regional, and global labor markets. Then, using the Statistical Package for Social Sciences (SPSS) 28.0, data were gathered and examined in order to construct our suggested framework. This framework would incorporate all relevant variables and stakeholders in order to improve graduates' employability and align with market expectations.

In 2021 K. Binici, et al. [13] suggested to determine the information technology competencies needed in information institutions. Information science and library technology developers frequently use the code4lib platform, which is where the study's data was gathered. Among the outcomes is a list of information technology competencies required in information institutions, especially for technologists, instructors, and aspiring information workers.

In 2022 Mehirig, A., et al. [14] suggested a brand-new, analytical approach that is totally reproducible, semi-automated, and built on a blend of expert judgment and machine learning algorithms. In this approach to creates a comprehensible classification of job responsibilities and skill sets by utilizing a sizable volume of online job ads that were gathered through web scraping. The findings can help HR managers and corporate executives create clear plans for acquiring and developing the skills necessary to fully utilize big data.

3. PROPOSED METHODOLOGY

3.1 Information Technology jobs

In this study a Novel Semantic Feature Enabled Agglomerative Clustering for Information Technology job profiling (SEA-IT) has been proposed.

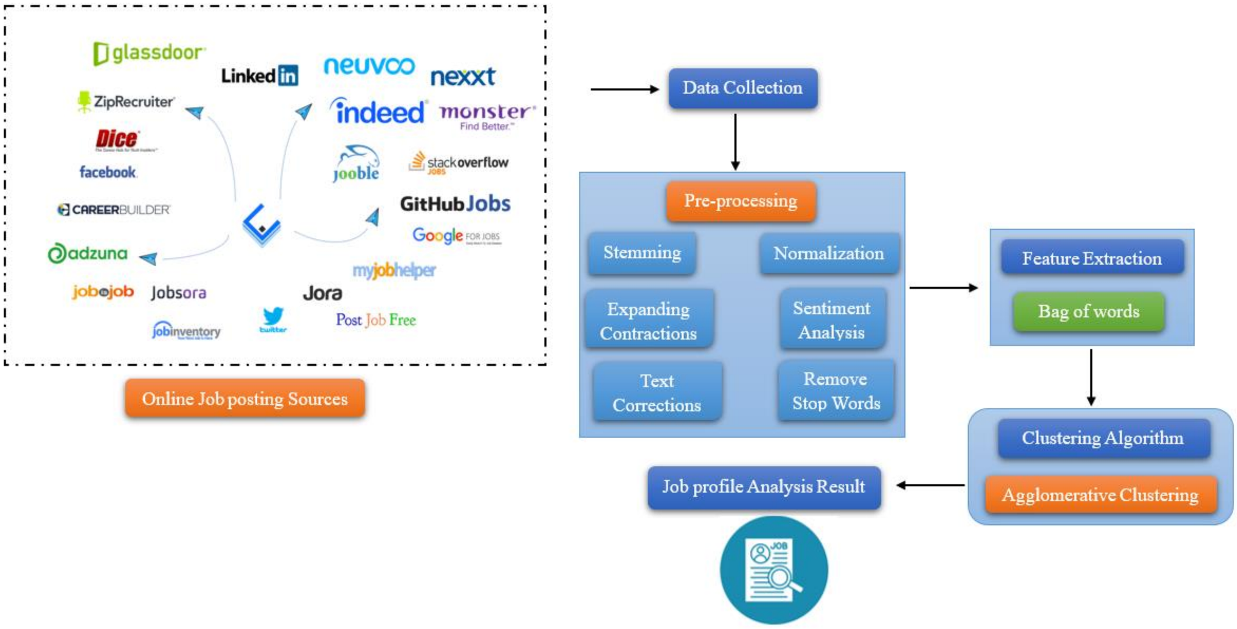

Figure 1. Block Diagram of Proposed methodology

Initially, the data from the online job posting sources will be collected and pre-processed using techniques such as stemming, normalization, text correction, removing stop words, and tokenization. Secondly, the pre-processed data can be utilized to extract features using a bag of words. After feature extraction, the cluster is generated using an agglomerative algorithm to form an IT job analysis result, so that the knowledge and capabilities of IT professionals can be upgraded. The overall planned workflow is shown in Figure 1.

3.2 Data Collection

Web-based actual datasets have gained popularity lately in information retrieval studies. IT job postings are the most significant and engaging source of new hires. They have typically been distributed through newspapers, but in the current digital era, they are primarily published online, on specialist websites or social media. Numerous research has made use of job adverts from sites including Bebee, Naukri.com, CareerBuilder.com, and LinkedIn. For a number of reasons, this study investigated the Job Street Indonesia and Tech in Asia platforms. They start by concentrating on IT job postings in Asia. Secondly, they provided a talent tag for each job posting that was advertised. Lastly, Tech in Asia supports the tech and entrepreneurial communities in Asia and has a catchy tagline.

3.3 Preprocessing

To guarantee accurate data processing, pre-processing must to be carried out on unstructured databases. This study involved six pre-processing techniques, as shown in below.

- **Stemming:** This process, often referred to as text standardization, involves stemming or reducing words to their most basic or root form.
- **Normalization:** Textual data may contain several words with comparable meanings. Normalization is the act of reducing the various forms of words with comparable meanings to a single standard form in order to standardize literature. By reducing the amount of randomness, this procedure seeks to enhance text quality, lower vocabulary size, boost text processing effectiveness, and optimize NLP model performance.
- **Text Correction:** Text correction is the process of fixing spelling and grammatical problems in order to ensure consistency with terms in a text and avoid different vector representations for misspelled words.
- **Remove Stop Words:** Words that add nothing to the meaning of a sentence are known as stop words. Consequently, their removal would not alter the sentence's meaning.
- **Tokenization:** Tokenization, in natural language processing, is the act of splitting texts into smaller word or sentence portions known as tokens. Tokenization divides text into discrete pieces so that models can identify patterns within them. This helps the model comprehend and represent the content more accurately.

3.4 Feature Extraction Using Bag of Words Algorithm

Dense SIFT features that are obtained from every SDP image using Euclidean distance are used for visual word matching. A numerical vector with a K-dimensional representation of the fault sample is created by counting the frequency of a specific type of word in the image. A hierarchical pyramid structure is formed by breaking down all different types of frequency vectors in picture words into blocks using spatial pyramid matching. This structure may be thought of as a bag of words with spatial-scale features.

3.5 Agglomerative Clustering Algorithm

Among the most well-liked and successful techniques for grouping data are agglomerative methods. However, a comprehensive comparison of these approaches concerning the crucial problem of inaccurate results during cluster search has not been conducted. A cluster model with a higher density centre encircled by a transition and outliers is used to measure the significance of the discovered clusters and deal with the false positive issue.

Algorithm 1: Pseudocode of the agglomerative hierarchical clustering

```
Input: Dataset M with S instances as $M = \{X_1, X_2 \dots\dots X_S\}$ and cluster distance function L$(d_i, d_j)$
Output: Black partition dendrogram Z each 1$\leq Z \leq S$
    $d_i$ = $\{X_i\}$,$\forall_i$=1, 2…S
    For Z=S down to 1 do
        $Dendrogram_Z$=$\{d_1, d_2, \dots d_S\}$
        L (i, j) = L $(L_i, L_j)$ ,$\forall_i$= 1,2, …. S
        Let (z, f) = $argmin_{(i,j)}\{L(d_i, d_j): 1 \leq i \leq j \leq Z + 1\}$
        CZ=Join $(C_Z, C_f)$
        Remove $(C_f)$
    End
```

3.6 Job profile Analysis

The median-LinkedIn certain visualizations, such as word/phrase frequency analysis and link analysis, which describe words and phrases connected together, use the hierarchical clustering technique. This study's employment profile analysis was explained using these visualization concepts. The most frequently used terms in the textual corpus can be quickly viewed through word or phrase frequency analysis. Finding keywords that belong to the same clusters and the related terms that may be utilized to get information about these clusters is also beneficial. This function might also aid in the identification of co-occurring words, enabling the investigation of search phrases to find pertinent text. A word cloud graphical display and table format can be used to present a visual result. Each word's different sizes are displayed proportionately to how frequently it appears in the text. A link analysis, which shows the relationship between keywords, is displayed as a network graph. Based on the experiment conducted, this study has several benefits. For instance, the number of clusters that can be specified can be changed based on the job profile that needs to be investigated; it can run quickly computationally; the graphical representation makes it easier for readers to understand; and different cluster sizes and shapes can be chosen to meet the objectives of the study.

4. RESULT AND DISCUSSION

The performance of the Novel Semantic Feature Enabled Agglomerative Clustering for Information Technology Job Profiling (SEA-IT) has been discussed in this section. Matlab is one program that can be used to simulate the blockchain process. This can be used to spread the blockchain and mine blocks with incorrect hashes for testing, as multiple nodes can carry out the activity in the simulation. With Matlab, one gigabyte of RAM at the minimum and sixty gigabytes of disk space at most are available for each worker. mimic procedures and perform Matlab algorithms on historical blockchain data. one gigabyte of RAM at a minimum and sixty gigabytes of disk space at most for each worker.

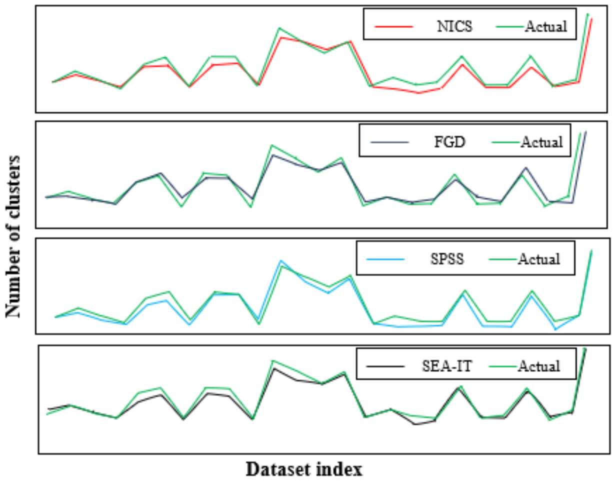

Figure 2. Comparison of SEA-IT with Existing Algorithms in the number of optimal clusters

Figure. 2 displays the comparison findings between the ensemble clustering and SEA-IT algorithms. For every dataset, two counts are provided: the number of clusters found by the clustering algorithms and the number of clusters in reality. The results unequivocally demonstrate that the SEA-IT curve and the curve corresponding to the number of real clusters diverge less, with the SPSS algorithm producing superior results. In SEA-IT, the average discrepancy between the number of clusters and the actual number of clusters is 15.5. The variations are documented for the algorithms, which are 16.9, 15.4, and 15.0, in that order.

Figure 3 presents the comparison findings for several algorithms. The number of dataset instances that are accurately allocated to the appropriate class determines the clustering accuracy. As a result, the ratio of correctly grouped instances to total instances is known as clustering accuracy. As demonstrated, SEA-IT provides more accuracy than other algorithms in most regards. In terms of data clustering, SEA-IT outperforms, SPSS, FGD, and NICS by an average of 5.2%, 2.3%, and 1.0%, respectively.

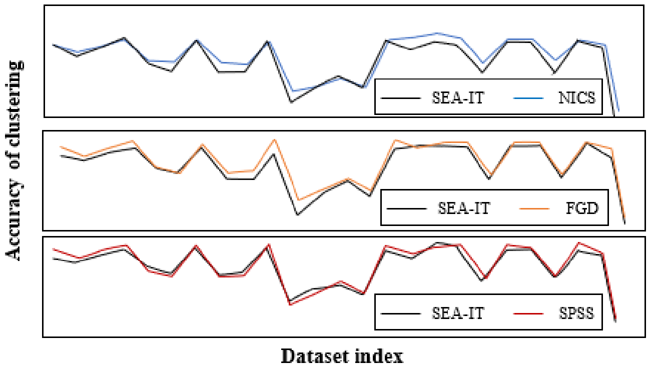

Figure 3. Comparison of SEA-IT with existing Algorithms in clustering Accuracy

5. CONCLUSION

In this research, a novel Semantic Feature Enabled Agglomerative Clustering for Information Technology job profiling (SEA-IT) has been proposed. Semantic analysis is performed using a tree-like strategy. The most frequently used phrases and words from each cluster of IT professions were collected to demonstrate specific knowledge. Initially, the data from the online job posting sources will be collected and pre-processed using techniques such as stemming, normalization, text correction, removing stop words, and tokenization. Secondly, the pre-processed data can extract features using a bag of words. After feature extraction, the cluster is generated using an agglomerative algorithm to form an IT job analysis result, so that the knowledge and capabilities of IT professionals can be upgraded. The simulation findings, based on evaluation criteria and other statistical tests, demonstrated the suggested algorithm. Experiments demonstrated the SEA-IT functions well with a variety of descriptive methodologies and is independent of the dataset's dimensions.

CONFLICTS OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING STATEMENT

Not applicable.

ACKNOWLEDGEMENTS

The author would like to express his heartfelt gratitude to the supervisor for his guidance and unwavering support during this research for his guidance and support.

REFERENCES

[1] O. Semenikhina, V. Proshkin, and O. Naboka, "Application of Computer Mathematical Tools in University Training of Computer Science and Mathematics Pre-service Teachers", *International Journal of Research in E-learning*, vol. 6, no. 2, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] 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, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] T.S. Prøitz, and L. Wittek, "New directions in doctoral programmes: bridging tensions between theory and practice?", *Teaching in Higher Education*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] J. Qadir, and A. Al-Fuqaha, "A Student Primer on How to Thrive in Engineering Education during and beyond COVID-19", *Education Sciences,* vol. 10, no. 9, pp. 236, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] L. Vyas, and N. Butakhieo, "The impact of working from home during COVID-19 on work and life domains: an exploratory study on Hong Kong", *Policy design and practice*, vol. 4, no. 1, pp. 59-76, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] S. More, and T. Rosenbloom, "Job-field underemployment and its impact on the demand for higher education at the Israeli labor market", *Israel Affairs*, vol. 28, no. 2, pp. 316-334, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang, "Industry 4.0 and Industry 5.0—Inception, conception and perception", *Journal of manufacturing systems*, vol. 61, pp. 530-535, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] M.E. Armstrong, K.S. Jones, A.S. Namin, and D.C. Newton, "Knowledge, skills, and abilities for specialized curricula in cyber defense: Results from interviews with cyber professionals", *ACM Transactions on Computing Education (TOCE)*, vol. 20, no. 4, pp.1-25, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] M. Torres, N. Flores, and R. Torres, "Fostering soft and hard skills for innovation among informatics engineering students: An emancipatory approach", *Journal of Innovation Management,* vol. 8, no. 1, pp. 20-38, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] J. Hammerschmidt, S. Durst, S. Kraus, and K. Puumalainen, "Professional football clubs and empirical evidence from the COVID-19 crisis: Time for sport entrepreneurship?", *Technological Forecasting and Social Change*, vol. 165, pp. 120572, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] E. Novak, L. Bizjak, D. Mladenić, and M. Grobelnik, "Why is a document relevant? Understanding the relevance scores in cross-lingual document retrieval", *Knowledge-Based Syst.*, vol. 244, pp. 108545. [CrossRef] [Google Scholar] [Publisher Link]

[12] T. Bai, Y. Ge, S. Guo, Z. Zhang, and L. Gong, "Enhanced Natural Language Interface for Web-Based Information Retrieval," *IEEE Access*, vol. IX, pp. 4233–4241, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] K. Binici, "What are the information technology skills needed in information institutions? The case of 'code4lib' job listings", *J. Acad. Librariansh.*, vol. 47, no. 3, pp. 102360, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] A. Mehirig, "Skills required in Big Data professions", *Journal of Advanced Economic Research//V,* vol. 7, no. 01, 2022. [CrossRef] [Google Scholar] [Publisher Link]

AUTHORS

B. Jaison is currently working as Professor in the Department of Computer Science and Engineering, RMK Engineering College, Kavarapettai, Chennai, India. He is having more than 24 years of teaching Experience in different Institutions in various cadres. He completed his M.E degree in Computer Science & Engineering from G.K.M College of Engineering and Technology, Affiliated to Anna University, Chennai in the year 2007 and Ph.D in Information and Communication Engineering from Anna University, Chennai in the Year 2015. He has published more than 50 Research Articles in International Journals and attended many International Conferences. He is the recognized Research Supervisor of Anna University Chennai and produced Five Research Scholars. His areas of interest include Data mining, Image Processing and Cloud Computing. He is a life member in IAENG, IACSIT and ISTE.

R. Gladis Kiruba received her UG degree in B. E (ECE) in PSN college of Engineering and Technology during the year 2004-2008 and received her master degree MTech (VLSI design) in karunya university during the year 2008-2010.She is currently working as Assistant professor in Bangalore college of Engineering and technology, chandapura, Bangaluru in the year April (2024).

G Belshia Jebamalar is currently working as Assistant professor in SA Engineering college in Department of computer science in the year 2024.

Arrived: 18.05.2024

Accepted: 25.06.2024