Automatic Resume Parsing using Greywolf Algorithm Integrated with Strategically Constructed Semantic Skill Ontologies

Ayush Kumar A [1], Gerard Deepak [2*], and Sheeba Priyadarshini J[3]

[1] *National Institute of Technology, Tiruchirappalli, India*
[2] *Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India*
[3] *CHRIST (Deemed to be University), Bangalore, India*

Abstract

The quest for finding the right candidate for their post has made the recruiters employ several methods since the beginning of corporate job recruitment. Apart from the skills and the quality of interview, a thing that matters the most and forms the basis of selection is the candidate's resume. Recruiters and companies have a tough time dealing with the several thousands resumes of the candidates which apply, as manually scanning them and finding the right selection can be tough most of the time. In this paper, Natural Language Processing(NLP) methods have been integrated with ontologies to improve the pace and quality of the recruitment process by proposing an automatic resume parser model. The resume of a candidate, along with his LinkedIn and GitHub profiles are weighted and using the Greywolf algorithm, the global maxima of the most deserving and qualified candidate are found and are recommended with a high accuracy of 96.13%.

Keywords

Resume matching; Natural Language Processing; Greywolf Algorithm; Ontology.

1. Introduction

Modern world demands optimization for almost all the tasks which are manually tiring and take a lot of time to complete. Resume matching and parsing is one of the tasks which, if automated can save a lot of time and human labor. In this highly competitive and dynamic world, the recruiters must select only a handful of candidates among the several thousand applicants which flood the company for a vacant post. The candidates, along with their technical skills and interview performance, are judged based on their resumes, which is the description of the skills of the candidate. The conventional error prone method of manually skimming all the individual resumes and shortlisting the candidates without a formal weighting algorithm is usually error prone and may lead to many discrepancies as it is also prone to corruption and favoritism. Automatic Resume Parsing (ARP) based on NLP and Machine Learning methods can be very useful for the transparent judging process and for faster results.

The proposed system aims to automate the entire selection process by integrating the social media presence of the candidate along with his/her GitHub and LinkedIn profiles which are meant to be professional. Some recruiters do not consider the social media profiles of the candidates, but it may form an important part as LinkedIn may be needed after joining the company and a candidate with an already proficient presence may perform better than others. Hence, many such aspects are considered and weighted in this system making it fool-proof and robust which considers all-round qualities of the candidate.

Motivation: The motivation behind this work is to reduce and eliminate the time taken for resume parsing and matching along with the malpractices that may happen in the case of manual resume parsing

International Workshop on Semantic IoT (IWSIoT-2022) and Multilingual Semantic Web (IWMSW-2022) and Deep Learning for Question Answering (IWDLQ-2022)
∗Corresponding author
EMAIL: gerard.deepak.christuni@gmail.com (G. Deepak)
ORCID: 0000-0003-0466-2143 (G. Deepak)

if left unregulated. Human energy and time are valuable and should be saved in order to maximize the efficiency of the organization. ARP can also result in elimination of the unskilled candidates who appear for the interview process despite human checking which may be erroneous at times. Hence, this system and model will form an essential part of the recruitment process and will result in an overall efficient and transparent recruitment system.

Contribution: The information stored in the resume is reasonably structured in most of the cases but not always. Also, the level of structuring is not enough for direct processing and so, the information is retrieved and is stored in a structured form of an ontology. Segmentation, Standardization and Structuring of the information is done and Greywolf optimization algorithm is used to maximize the objective function which is the weighted relevance of the candidate with the vacant job position. Hybrid resume matching and ranking is performed, and the identification and recommendation values are found to be in a best-in-class range. An overall accuracy of 96.13% is obtained for this ARP system with a dataset of Indian resumes.

Organization: The remainder of the paper is organized as follows: Section 2 consists of the relevant Literature Review. In Section 3, the System Architecture is described. Ontology modeling and conceptualization is dealt in Section 4. Section 5 consists of the Implementation, Results, and Discussions. The paper is finally concluded in Section 6.

2. Literature Review

Bhatia et al., [1] have parsed the LinkedIn resumes of several candidates and have achieved an accuracy of 73% for applicant recommendation. They have also used BERT sentence pair arrangement to perform positioning dependent on their appropriateness to the set of working responsibilities. Then, they have presented a start to finish answer for positioning up-and-comers dependent on their appropriateness to a set of working responsibilities. Amin et al., [2] have analyzed resumes presented by the applicants are then contrasted and the activity profile prerequisite posted by the organization spotter by utilizing strategies like AI and Natural Language Processing (NLP). A web application has been proposed by them which is supposed to support the activity candidate applicants just as the enrollment specialists to utilize it for requests for employment and screening of resumes. Yan et al., [3] have suggested the advantages to get familiar with the inclination of the two scouts and employment searchers from past meeting chronicles and expect such inclination is useful to improve work continue coordinating. Along these lines, they concluded that selection representatives and abilities to be sure have inclination and such inclination can improve work continue coordinating hands on showcase. Le et al., [4] have considered the decent variety of occupation necessities and the multifaceted nature of employment searchers' capacities which set forward higher prerequisites for the precision and interpretability of Person-Job Fit framework. They have proposed an Interpretable Person-Job Fit use profound intelligent portrayal figuring out how to consequently become familiar with the reliance between a resume and occupation necessities without depending on an away from of employment searcher's capacities and sends the advancing issue as a figuring out how to rank issue. Das et al., [5] have addressed the problem of Named Entity Extraction based on big data tools for parsing resume. Deepak et al., [6] have proposed a Resume Parser System using the Firefly algorithm which accomplished Named Entity Recognition effectively. Mirjalili et al., [7] have proposed the grey Wolf optimizer which is highly suitable for finding the global maxima or the minima of a function. This method has outperformed several of the existing methods and is one of the most efficient optimization algorithms. Tobing et al., [8] have proposed resume parsing system based on special headers in Indonesian language specifically for parsing resumes in Indonesia. Papers [9-18] discuss the various uses of Ontologies in various fields.

3. Proposed System Architecture

The ARP system consists of several preprocessing steps for the efficient computation of the rank of the resume, as the tool must be robust enough to handle all the exceptions that occur. As Resumes of different candidates are in different format or style, a segmentation step must be performed where the

text extracted from the PDF resume is split into several subcategories which carry meaningful information. All the attributes like Education. Soft Skills, Technical Skills, Certifications are extracted here. these attributes are also present in the Ontology which is used for the efficient comparison of the text and weighting of the skills. The clustering is done such as all the information, which is unnecessary to consider, like Name, Contact, Email etc., are placed in a single cluster, and other valuable information is placed in another. The text is then standardized to match all the input resumes for efficient computation. After structuring and integrating the weights of the skills with the relevant GitHub and LinkedIn profiles, the profile is the compared with the skill ontology. Using the insights from the Ideal Resume from the company's perspective and the skill ontology, the resumes are ranked using a hybrid form of Greywolf Algorithm and the results are displayed.

The system architecture of the AMD system is shown in Fig. 1. Each resume from the database is taken through a rigorous preprocessing step which converts all the structured and unstructured data present in the resume to a properly structured form ready for usage in the algorithm. The data from the PDF is extracted. After the preprocessing stage, the qualification of the candidate is weighed using a novel weighing metric which evaluate the candidate's profile and assigns him a score. These are then integrated with the GitHub and LinkedIn profile of the candidate, and an overall score describing the social media presence of the candidate is also added to the net score which always is an advantage to the recruiter. These profiles are then taken through the Greywolf algorithm which uses an innovative methodology to rank these resumes and returns the best person possible. The weighing is done using the Ontology of the skills and the jobs, and these features are linked with the profile strategically. Overall, the process of Named Entity Recognition is carried on with the novel algorithm, and the list of the top candidates is obtained.

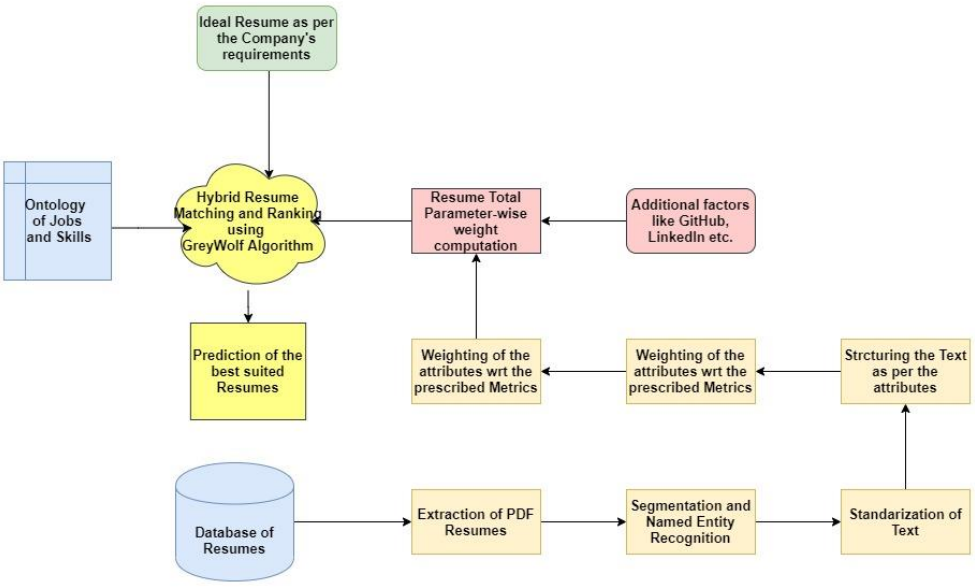

Figure 1: System Architecture

4. Ontology Modeling

A dynamic Ontology is created with the complete data out the job and the skills required for it with complete details and all the possible attributes. The classes are formed by the jobs and the postings, and the instances of the class are the skills required which come under a separate class housing all the skills required and their strategic categorization. For example, an ontology is built for an FMCG company and all the posts in the domain of Administration, Marketing, Engineering, and several others are spanned in the ontology which makes this ARP system the best and a comprehensive solution for all the job requirements for the company. All the posts are covered, and the respective skills required for them are matched with those classes which can then be integrated using a suitable measure and the process of resume parsing can be carried on effectively. The overall efficiency of the ARP system can

be improved, as compared to other methods, the usage of ontology can be far more efficient and accurate, as the skills can be better explained using RDF schema. A hierarchical classification is done for the job and the skills are aptly arranged so that while weighting the skills, there is no error as candidates often have extraordinary skills which are not expected, and using data mining, an efficient ontology can be created which then houses all the skills and relevant skills can get their deserved recognition. The description of skills is shown in Table 1.

Table 1
Definition of Skills

```
<AnnotationAssertion>
    <AnnotationProperty abbreviatedIRI="rdfs:label"/>  <Literal>Soft Skills</Literal>
</AnnotationAssertion>
<AnnotationAssertion>

<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
    <Literal>Soft skills are a cluster of productive personality traits that characterize one's relationships in a
milieu. These skills can include social graces, communication abilities, language skills, personal habits,
cognitive or emotional empathy, time management, teamwork and leadership traits.</Literal>
</AnnotationAssertion>

<AnnotationAssertion>
    <AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/><Literal>Adaptability</Literal>
    <AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/> <Literal>Communication</Literal>
    <AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/><Literal>Enthusiasm</Literal>
    <AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/><Literal>Reliability</Literal>
    <AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/><Literal>Teamwork</Literal>
</AnnotationAssertion>
```

Entity graphs for the classes of the ontology can be created for easy representation, and hence, they are an effective method of representation of the class. The relationships between the objects and the classes are represented in a graphical format. A semantic network is formed by means of ontologies which are very efficient in terms of representing and retrieving the knowledge in a structured matter. Several properties are represented as attributes of the classes. The ontology is created in XML format, and then a method of transfiguration from XML to OWL is proposed to enhance the efficiency of the ontology. Fig. 2 shows the entity graph of a related skillset.

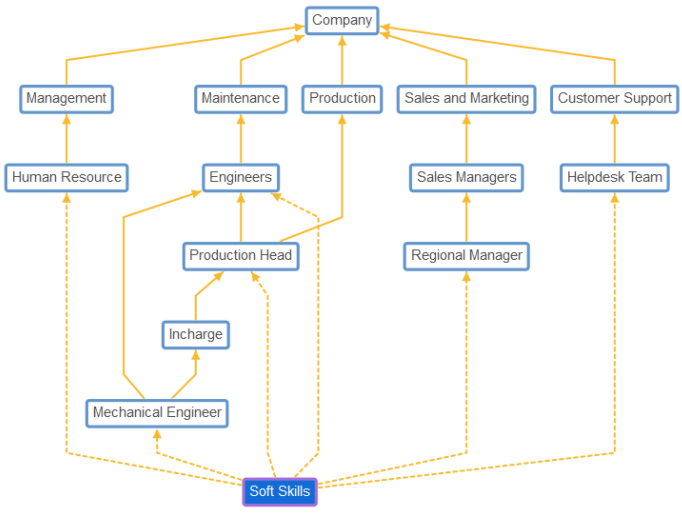

Figure 2: Entity Graph of Soft Skills

Ontologies can be used to use the ranking attributes which can be detrimental for the job post like the Current pay, expected pay, experience etc. These are the metrics needs to be considered while deciding the salary of the candidate, and hence, they also should be considered for resume parsing. Location, Work Gap, Education, and Experience should be considered, and hence, these details must be collected from the candidate. If not available, average values should be assigned to them, and the overall score of the candidate should be reduced for not mentioning these essential details as this indicates the lack of professionalism. The skill ontology is visualized using an online ontology visualization tool called WebVOWL. This is done by creating the .owl file of the ontology first, and then uploading it to the portal. The result of visualization is shown in Figure 3.

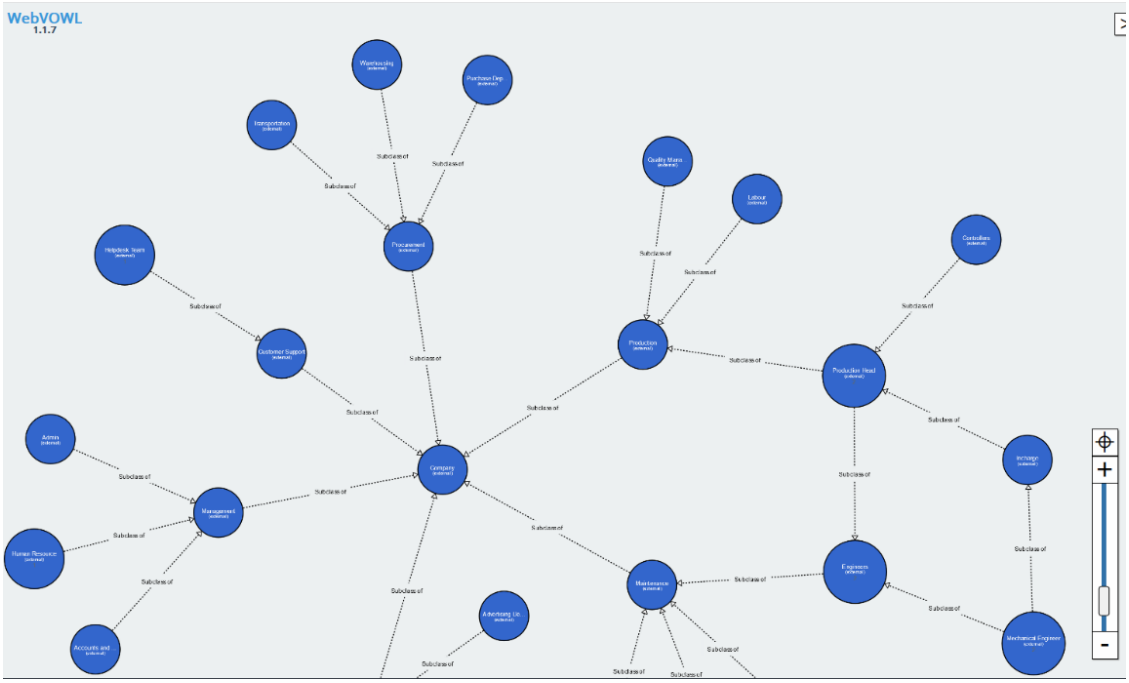

Figure 3: Visualization of Ontology

5. Implementation

The algorithm for Automatic Resume Parsing has been designed using the Greywolf Algorithm and implemented in Windows 10 operating system. The implementation was done on a computer with Intel i5 processor and includes 8GB of RAM. The dataset used for the implementation purpose is the Resume Corpus Collection. The pool of resumes is collected and used for implementation of this novel ARP system. Tokenizing the data obtained from the Data Extraction step of the Resume PDF is the first step of implementation. Then, the Segmentation and Structuring of the Data occurs which makes the data ready for processing. The final result of the algorithm is the filtering of the resumes and remove the resumes which are not suitable for the job post and which do not satisfy the minimum eligibility criteria for the job. The resumes are weighed, and the relative scores are used for the first stage processing, as there are several other factors which need to be considered while offering the job to the deserving candidate. The main aim is to select the set of highly qualified resumes which could be interviewed. The algorithm involves the novel Greywolf algorithm with many changes to suit the process of resume parsing. Table 2 shows the proposed algorithm.

Table 2
Algorithm for Automatic Resume Parsing

begin

Step 1: From the database containing applicant's resumes, each resume is denoted as a dictionary with the unique ID assigned by the recruiter as the variable name. The dictionary has the parameters as keys and the values as the Weight computed by the entry in the user resume.

Step 2: Let $f(x) = (x_1, x_2,...,x_n)$ be the objective function

Step 3: A list of resumes list_resume = $[R_1, R_2, R_3 ... R_n]$ is formed

Step 4: The consolidated weights of the resumes are calculated using the weights of the criterion and the value computed earlier.

Step 5: A constant λ is defined as the resume ranking coefficient and Candidate_List is defined as the list of selected candidates.

Step 6: The values for the coefficient vectors are initialized

Step 7:The top 3 resumes using their consolidated values are initialized as X_α, X_β, X_δ

Step 7:iter = 0, saturation = number of candidates required for the job

```
for candidate in range(0,saturation):
    while (iter < max_iter)
        for i in range(1,n):
            Update the position by X(iter+1) = (X1 + X2 + X3) / 3
        end for iter
        Update the values for the coefficient vectors
        Update Xα, Xβ, Xδ
        iter = iter + 1
    end while
    Candidate_List.append(Xα)
    list_resume.remove(Xα)
end for
```

Step 8: Post processing the results and visualization

end

The algorithm starts with the extraction of data from the database pool. A unique ID is allotted, and the resume is taken in the form of a dictionary. The objective function is defined, and it is the value of the qualifications and the relevance of the candidate with respect to the job. The list of resumes is taken and a constant λ is defined. The Greywolf algorithm is used for this implementation. The variable saturation is the total number of candidates to be selected. The top 3 resumes using their consolidated values are initialized as alpha resume, beta resume, and the delta resume. For all the resumes in this selection, the position of the resume is updated according to the position of the alpha resume, beta resume, and the delta resumes. The coefficient is updated, and the loop continues till the value of the maximum (highest similarity with the ideal resume) of the function is computed. The postprocessing step is then carried out, and the results are visualized.

5.1. Results and Evaluation

A confusion matrix is formed for this system, as this is basically a classification task. So, there may be some resumes which may be misclassified. The four metrics which are considered are Precision, Recall, Accuracy, and the False Positive Rate.

Precision is defined as a proportion of the recommended resumes that are truly highly qualified. Precision is measured by the formula given in Equation 1.

$$Precision = \frac{TP}{TP + FP} \quad - \quad (1)$$

Recall can be defined as the proportion of the recommended resumes predicted w.r.t total number of highly qualified candidates available in the application pool. is measured by the formula given in Equation 2.

$$Recall = \frac{TP}{TP + FN} \quad - (2)$$

Accuracy can be described as in Equation 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} - (3)$$

The False Positive Rate (FPR) is the number candidates who do not have the relevant skills, but are identified as highly-qualified by the system (all FPs), divided by the total number of rejected candidates. False Positive Rate (FPR) is defined as follows in Equation 4.

$$FPR = \frac{FP}{TN + FP} \quad - (4)$$

where the parameters are defined in Table 3.

Table 3
Parameter definition

TP - True Positive	Number of candidates who are qualified and are considered qualified by the system.
TN - True Negative	Number of candidates who are not qualified and are rightfully rejected by the system.
FP - False Positive	Number of qualified candidates which are rejected by the system.
FN - False Negative	Number of undeserving candidates which are identified wrongly as qualified by the system.

Table 4
Performance Measures of the proposed ARP system

Number of Resumes	Precision%	Recall%	Accuracy %	FPR%
20	98.36	96.62	97.49	2.52
40	98.21	96.31	97.26	2.75
60	97.59	95.87	96.73	3.28
80	96.42	93.88	95.15	4.87
100	95.23	92.89	94.06	5.96
Average	97.16	95.11	96.13	3.88

Table 4 shows the results obtained for the Performance Measures of the proposed algorithm. The average accuracy is 96.13% and the average FPR of the ARP algorithm is 3.88%. The parameters increase with the increasing number of applications taken into consideration as expected by a neural network system. The graph of Precision vs Number of Resumes, Recall vs Number of Resumes, and Accuracy vs Number of Resumes are depicted in Fig. 4.

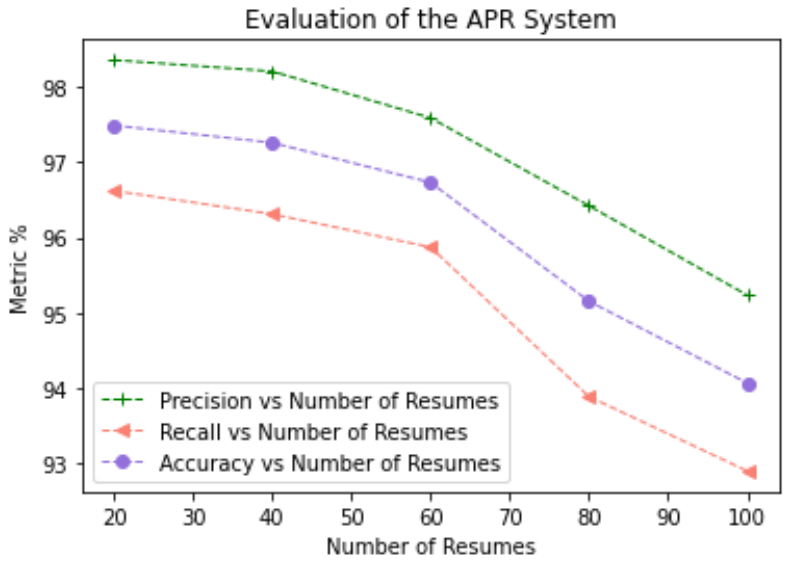

Figure 4: Evaluation of the ARP System

The proposed ARP system using Greywolf Algorithm gives an overall average accuracy of 96.13%. The proposed approach is better than many other existing approaches for Resume Parsing because the proposed system incorporates the data and its attributes in the form of an Ontology which is efficient. Also, the Greywolf algorithm is better than other optimization algorithms. Hence, the maxima of the function predicting the qualifications of the candidate is computed in an efficient manner, and the ARP system is efficient enough to yield such a high accuracy.

6. Conclusions

A robust and efficient APR system is the need of the hour because of the high number of applicants applying for a job position and the smaller number of posts. The recruiting process needs to be optimized and automated because this can prevent the wasting of resources such as time and human effort. The experimental results have indicated that the APR system proposed can be dynamically integrated with the current recruitment process and the screening of resumes can be done with high accuracy. The application of ontology can be improved by using a universal ontology for all the job purposes regardless of the industry so that the accuracy can be improved. Also, the data segmentation and structuring can be improved which can result in an even higher accuracy and more precision. Overall, the Greywolf algorithm is an efficient methodology for optimization purposes and integrated with ontology, it can be used for several other purposes. An overall average accuracy of 96.13% is obtained which shows that the results obtained are best-in-class and can be incorporated in real-time recruitment process.

References

[1] Bhatia, V., Rawat, P., Kumar, A., & Shah, R. R. (2019). End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT. arXiv preprint arXiv:1910.03089.

[2] Amin, S., Jayakar, N., Sunny, S., Babu, P., Kiruthika, M., & Gurjar, A. Web Application for Screening Resume. In 2019 International Conference on Nascent Technologies in Engineering (ICNTE) IEEE. 2019, pp. 1-7.

[3] Yan, R., Le, R., Song, Y., Zhang, T., Zhang, X., & Zhao, D. Interview Choice Reveals Your Preference on the Market: To Improve Job-Resume Matching through Profiling Memories. In

Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 914-922.
[4] Le, R., Hu, W., Song, Y., Zhang, T., Zhao, D., & Yan, R. Towards Effective and Interpretable Person-Job Fitting. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 1883-1892.
[5] Das, P., Pandey, M., & Rautaray, S.S. A CV Parser Model using Entity Extraction Process and Big Data Tools. International Journal of Information Technology and Computer Science, 10, 2018, pp. 21-31.
[6] Deepak, G., Teja, V., & Santhanavijayan, A. A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm. Journal of Discrete Mathematical Sciences and Cryptography, 23(1), 2020, pp. 157-165.
[7] Mirjalili, S., Mirjalili, S. M., & Lewis, A. Grey wolf optimizer. Advances in engineering software, 69, 2014, pp. 46-61.
[8] Tobing, B.C.L., Suhendra, I.R. and Halim, C. Catapa Resume Parser: End to End Indonesian Resume Extraction. In Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval (NLPIR 2019). Association for Computing Machinery, New York, NY, USA, 2019, pp. 68–74
[9] Deepak, Gerard, and J. Sheeba Priyadarshini. "Personalized and Enhanced Hybridized Semantic Algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis." Computers & Electrical Engineering 72, 2018: pp. 14-25.
[10] Shreyas, K., Deepak, G., & Santhanavijayan, A. GenMOnto: A strategic domain ontology modelling approach for conceptualisation and evaluation of collective knowledge for mapping genomes. Journal of Statistics and Management Systems, 23(2), 2020, pp. 445-452.
[11] Kaushik, I. S., Deepak, G., & Santhanavijayan, A. QuantQueryEXP: A novel strategic approach for query expansion based on quantum computing principles. Journal of Discrete Mathematical Sciences and Cryptography, 23(2), 2020, pp. 573-584.
[12] Pushpa, C. N., Deepak, G., Kumar, A., Thriveni, J., & Venugopal, K. R. (2020, July). OntoDisco: Improving Web Service Discovery by Hybridization of Ontology Focused Concept Clustering and Interface Semantics. In 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India. IEEE, 2020, pp. 1-5.
[13] Deepak, Gerard, Ansaf Ahmed, and B. Skanda. "An intelligent inventive system for personalised webpage recommendation based on ontology semantics." International Journal of Intelligent Systems Technologies and Applications 18, no. 1-2, 2019, pp. 115-132.
[14] Deepak, G., Shwetha, B.N., Pushpa, C.N., Thriveni, J. and Venugopal, K.R. A hybridized semantic trust-based framework for personalized web page recommendation. International Journal of Computers and Applications, 2020, pp.1-11.
[15] Ortiz-Rodriguez, F., Tiwari, S., Panchal, R., Medina-Quintero, J. M., & Barrera, R. MEXIN: multidialectal ontology supporting NLP approach to improve government electronic communication with the Mexican Ethnic Groups. In *DG. O* 2022: The 23rd Annual International Conference on Digital Government Research, 2022, pp. 461-463.
[16] Mehta, S., Tiwari, S., Siarry, P., & Jabbar, M. A. (Eds.). Tools, Languages, Methodologies for Representing Semantics on the Web of Things. John Wiley & Sons, 2022.
[17] Panchal, R., Swaminarayan, P., Tiwari, S., & Ortiz-Rodriguez, F. AISHE-Onto: a semantic model for public higher education universities. In *DG. O2021: The 22nd Annual International Conference on Digital Government Research, 2021*, pp. 545-547.
[18] Ortiz-Rodriguez, F., Medina-Quintero, J. M., Tiwari, S., & Villanueva, V. (2022). EGODO ontology: sharing, retrieving, and exchanging legal documentation across e-government. In *Futuristic Trends for Sustainable Development and Sustainable Ecosystems* (pp. 261-276). IGI Global.