International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

A Study of Job Aggregation Systems using Web Scraping, APIs, TF-IDF, and Cosine Similarity

Jagruti Boraste
Department of Information Technology
Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra, India

Dhanashri Nanekar
Department of Information Technology
Bharati Vidyapeeth's College of Engineering for Women Pune, Maharashtra, India

Sanika Deokar
Department of Information Technology
Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra, India

Vaishanvi Patil
Department of Information Technology
Bharati Vidyapeeth's College of Engineering for Women Pune, Maharashtra, India

Prof. Dr. N. A. Mulla
Department of Information Technology
Bharati Vidyapeeth's College of Engineering for Women
Pune, Maharashtra, India

Abstract **- In the current scenario of a rapidly changing job market, it is essential to remain updated about the latest job opportunities, especially for engineering graduates looking to join premier IT companies. However, the presence of many job portals on the internet often leads to fragmented job postings, repeated postings, and unrelated search results, thereby making the job search process inefficient and time-consuming. This paper discusses a detailed analysis of the current online job recruitment systems, job aggregation systems, and data extraction methods such as web scraping and API extraction. The drawbacks of conventional job**

search systems, such as the absence of common access and inefficient relevance ranking, are critically analysed. Based on the identified research gaps, this study highlights the need for a hybrid job aggregation framework integrating automated data acquisition and relevance-based ranking mechanisms.

The study discusses how a hybrid job aggregation approach integrating automated data extraction and TF-IDF-based relevance ranking can provide centralized and relevance-oriented job discovery. The paper clearly illustrates that conventional information retrieval methods are still efficient for relevance-ranked job matching and can provide transparency and computational simplicity without the need for complex machine learning algorithms.

Index Terms **- Job Aggregation, Information Retrieval, Web Scraping, API-Based Retrieval, TF-IDF, Cosine Similarity, Job Search Systems**

Although these platforms have their benefits, the presence of many job platforms has also created some issues. These platforms work independently and sometimes display disjointed, repetitive, or inconsistent information about job opportunities. Organizations do not usually display job opportunities on all platforms at the same time, and job seekers have to visit all platforms to monitor the available job opportunities.

To overcome these issues, job aggregation systems have been developed [3]. A job aggregator is a software program or web-based service that automatically gathers job listings from various online sources and provides them in a single interface. Aggregation systems are very common in other areas as well, such as review aggregation, search aggregation, video aggregation, and news aggregation. For example, news aggregators are software programs that automatically gather news articles from various online sources and provide them in a single interface. Similarly, job aggregators automatically gather job postings from various job portals and provide them in a single interface. However, most existing aggregation systems do not have an efficient relevance ranking system, which results in non-personalized job recommendations.

This paper presents a study on job aggregation systems and information retrieval systems used in job search software [4], [5]. Based on the results of the literature review, the study examines the applicability of Term Frequency–Inverse Document Frequency (TF-IDF) and cosine similarity for relevance-based job aggregation. TF-IDF is a mathematical method used in information retrieval and natural language processing to calculate the importance of a word in a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

document compared to a set of documents. It converts text data into numerical vector representations, which can be computationally analysed.

Cosine similarity is then used to calculate the relevance of these vectors [5]. It is used to find the cosine of the angle between two non-zero vectors to find their similarity, especially in high-dimensional sparse data. In the context of job aggregation, this technique allows the comparison of job descriptions with candidate skill sets, thus making job recommendations more accurate and relevant.

Apart from relevance ranking, job aggregation also needs automated data acquisition systems to ensure continuous and updated job aggregation. Since job postings are often updated, automated data acquisition systems are essential in ensuring the accuracy of the system [6]. Web scraping tools are often used to gather structured data from various web sources [7], [8]. Similarly, Application Programming Interfaces (APIs) are used to ensure standardized and accurate access to job data from various platforms [9]. Hybrid systems that combine web crawling, web scraping, APIs, and automated schedulers have been found to enhance scalability, coverage, and system efficiency in aggregation systems [10], [11].

In general, although the existence of various job portals has improved job accessibility, it has also led to dispersed information and a lack of relevance in search results. Job aggregation systems can help alleviate these problems, but ranking remains a challenge. This paper explores existing job aggregation systems and popular information retrieval methods, such as TF-IDF and Cosine Similarity, to analyse how structured aggregation and relevance-based ranking mechanisms can improve job search effectiveness.

The conceptual difference between fragmented job search and aggregated job search is illustrated in Fig. 1.

II. LITERATURE REVIEW

A. Traditional Job Portals

Online job portals are web-based services that allow employers to post vacancies and job seekers to search and apply for jobs [12]. Traditional job portals such as Naukri.com, LinkedIn Jobs, and Indeed India provide extensive databases of job openings with filtering options by role, location, and experience level. The evolution of these online job portals into major employment sources for both freshers and experienced candidates is owing to the capabilities offered by these websites in terms of searching, resume-uploading, and more. These websites are designed to work in a standalone manner, which means users have to deal with all of these websites individually, thereby making searching for a job more cumbersome and time-consuming. Due to the presence of multiple online independent job portals, job vacancies are distributed sporadically across these websites.

B. Job Aggregation and Information Retrieval

Job aggregation systems have been developed as a solution to the problem of job fragmentation under the traditional job portal system by collecting different job sources into a unified system [3]. The "Job Aggregator – A Final Year Project Report (2017)" thoroughly explains the process of an online job aggregator that retrieves job listings from various online portals based on job title and company names. This project shows that the concept of job aggregation facilitates the job search process by avoiding the need to browse several portals individually; instead, the job search system provides a unified source of job listings for the user to search through. Although the system allows for a unified source of job listings, the limitations of not being able to update job listings instantly were noted by the system itself.

There is already research on aggregation systems in other fields, such as federated search. This provides foundational information on how to combine search results from disparate sources and organize the results into one interface, which would be useful in job aggregators [13]. Federated search involves broadcasting the search query to numerous sources, combining the results of the sources, and then displaying the results with minimal duplication – concepts that can help build more sophisticated job aggregation systems.

Aggregators gather job postings. However, merging these lists of texts does not achieve good quality matching. The literature highlights that feature representation and document ranking techniques drawn from IR studies are crucial [14].

Content-based approaches have their job postings and queries modelled as document objects. By transforming these document objects into numerical vectors, the classic IR algorithms apply to traditional text matching and ranking.

C. Data Collection Techniques in Job Aggregation Systems

Automated data acquisition is an essential component of job aggregation systems, as jobs are posted on a variety of online platforms and are often updated. To efficiently acquire job data, various automated methods such as web scraping, API-based job retrieval, web crawling, and hybrid methods have been investigated.

2.3.1 Web Scraping

Web scraping is the process of automatically extracting data from web pages, converting unstructured HTML data into structured data that can be stored and analysed. A detailed survey by Ferrara et al. [7] discusses the classification of web data extraction methods and their significance in large-scale information retrieval systems. Also, a recent survey on web scraping methods by Lotfi et al. [15] discusses the challenges of dynamic content processing and changes in web structures.

With respect to the automated extraction of academic and employment-related data, Naing et al. [8] have shown the effectiveness of structured web scraping tools for the systematic extraction and organization of online data. Also, the implementation of automated tools for the extraction of employment-related data has been discussed by Tewari et al. [16].

2.3.2 API-Based Data Retrieval

datasets. Moreover, intelligent job recommendation assistance systems that leverage scraping and analysis have been conceptualized by Nayana et al. [18], which illustrates

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

the role of automated data acquisition in enabling personalized job recommendations.

2.3.5 Ethical and Operational Issues

Online portals offer Application Programming Interfaces (APIs) for structured data access of job postings. APIs deliver standardized data formats like JSON or XML, which are less complex to parse compared to HTML. Kumar et al. [17] introduce a job suggestion system that combines API-based data retrieval with crawling techniques to improve data acquisition and system robustness.

APIs enable structured and robust data access but can limit scalability by rate limiting, authentication, or restricted query access [17].

2.3.3 Web Crawling

Web crawling is the process of navigating through a network of web pages to extract job postings. Ferrara et al. [7] highlight the importance of crawling techniques for continuous data acquisition and automated dataset extension. Web crawlers can keep job databases up to date by periodically revisiting web pages to identify new postings.

2.3.4 Hybrid Data Acquisition Strategies

Hybrid data acquisition techniques leverage the use of APIs, web scraping, and crawling to achieve the best possible results. Kumar et al. [17] show that the combination of APIs and crawling is effective in enhancing the completeness of job recommendation

Automated data acquisition needs to be carried out in accordance with platform policies and regulations. Research on web data extraction stresses the need to adhere to terms of service, robots.txt policies, and privacy laws for sustainable implementation [7], [15].

TABLE I. Comparison of Data Acquisition Techniques in Job Aggregation Systems

Technique	Key Characteristics	Advantages	Limitations	References
Web Scraping	Extracts structured data from HTML web pages	Flexible, works without API support, suitable for large-scale extraction	Sensitive to website structure changes, requires maintenance	[7], [8], [15], [16]
API-Based Retrieval	Uses official platform APIs for structured data access	Standardized formats (JSON/XML), reliable, cleaner data	Rate limits, authentication restrictions, limited access scope	[9], [17]
Web Crawling	Automatically navigates web pages to discover new job listings	Continuous discovery of new data, scalable with scheduling	High resource consumption, duplication issues	[7], [10]
Hybrid Approach	Combines scraping, APIs, and crawling	Improved coverage, higher completeness, better robustness	Increased architectural complexity	[11], [17], [18]

Based on the literature, a hybrid data acquisition strategy is preferred because it balances coverage and reliability. Web scraping ensures coverage where APIs are unavailable, while APIs provide structured and stable data access.

D. Information Retrieval Techniques

The Vector Space Model represents text documents as high-dimensional vectors, with each dimension representing a term, although not necessarily relevant to the text document. With this model, it becomes easy to perform any quantified comparison between two vectors that share similarities. For instance, VSM with TF-IDF can be useful in document ranking and job text matching [19].

This is done in a shared vector space, allowing for the computation of the distance or similarity score.

E. TF-IDF in Document Ranking

The most widely used term-weighting scheme in IR is Term Frequency–Inverse Document Frequency, TF-IDF [20]. It calculates the importance of a term for a document in relation to a larger collection of documents.

TF-IDF increases the weight of terms that are frequent in a certain document but rare across other documents in the corpus. It allows discriminating terms to be better distinguished. This weighting is crucial when comparing job descriptions and queries because it raises key requirements while down-weighting common words. Most research proves that TF-IDF combined with vector space representations yields reliable relevance scores in the task of text similarity, hence suitable for content-based job matching and ranking.

F. Cosine Similarity for Text Matching

Cosine similarity measures the cosine of the angle between two vectors in a high-dimensional space, which is useful for document representation using TF-IDF. It is defined as the dot product of two normalized vectors divided by their magnitude, enabling similarity comparison for different lengths of documents. In information retrieval and job matching contexts, a cosine similarity measure is used to calculate the similarity between a user query and job descriptions. More specifically, a higher cosine similarity implies a better match [20].

Various empirical evaluations carried out in different scholarly works affirm that cosine similarity using TF-IDF has yielded promising results with short and medium-length documents, usually forming the basis of the ranking mechanism in content-based recommender systems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

TABLE II. Comparison of Information Retrieval Techniques for Job Matching

Technique	Method Type	Strengths	Limitations	References
Vector Space Model (VSM)	Algebraic document representation	Enables quantitative similarity computation	Does not assign term importance by default	[19]
TF-IDF	Term weighting scheme	Highlights discriminative terms, simple implementation, low computational cost	Does not capture semantic meaning	[20], [21]
Cosine Similarity	Vector similarity measure	Length normalization, effective for sparse text data	Depends on quality of vector representation	[5], [20]
Deep Learning-Based Models	Neural embedding models	Captures contextual meaning, higher personalization	High computational cost, low interpretability	[4]

G. Applications in Recruitment and Job Matching

Although pure job aggregator research using TF-IDF and cosine similarity does not exist, many resume ranking and job matching studies still adopt these IR techniques.

The conference paper Resume Ranking with TF-IDF, Cosine Similarity and Named Entity Recognition by IEEE ICDCC 2024 applies TF-IDF and cosine similarity for ranking resumes against job descriptions. This directly reflects the efficiency of these IR measures in recruitment contexts parallel to the relevance ranking of job aggregators.

Other studies related to resume screening and ranking also confirm that TF-IDF and cosine similarity provide better performance to increase the accuracy of matching and reduce the manual screening effort [21].

III. PROBLEM IDENTIFICATION & RESEARCH GAP

The current job portals work independently, leading to scattered job information.

Existing job aggregation systems use periodic or limited data collection mechanisms, which can cause the job information to be outdated or incomplete.

Existing job aggregation systems focus on data collection but do not use efficient relevance-based ranking systems.

Existing job recommendation systems use advanced machine learning algorithms; however, they increase the complexity of computation and decrease interpretability.

Because of the above issues, there is still a research gap in finding a balanced approach that combines efficient data acquisition strategies with transparent and simple information retrieval systems for relevance-based job searching.

IV. CONCLUSION

This research paper has discussed the online recruitment platforms, job aggregation systems, and information retrieval methods used in job search applications. The paper has emphasized the problems associated with job information fragmentation, redundancy, and the absence of relevance-based ranking in job search applications.

The study suggests that integrating hybrid data acquisition strategies with traditional information retrieval techniques can enhance job aggregation systems. The study indicates that hybrid data acquisition strategies combined with traditional information retrieval methods can improve job aggregation effectiveness. Traditional information retrieval methods offer transparency and computational efficiency, making them suitable for relevance-based job matching.

This research paper clearly shows that traditional information retrieval methods are still effective for relevance-based job aggregation, and future improvements can be done using advanced NLP methods for personalization.

REFERENCES

[1] IU Jharkhand, "A Study on Online Recruitment," IU Journal. [Online]. Available: https://journal.iujharkhand.edu.in/A-Study-On-Online-Recruitment.pdf

[2] P. Shrivastav, N. Shrivastav, and P. Ranjan, "A Study of Role of Online Platforms in Modern Recruitment Process," Journal of Informatics Education and Research, 2025.

[3] "Job Aggregator – A Final Year Project Report," ResearchGate, 2017.

[4] IEEE, "Deep Learning Based Job Recommendation Analysis with NLP," IEEE Xplore, 2024.

[5] IEEE, "Text Similarity and Information Retrieval Using Cosine Similarity," IEEE Xplore, 2016.

[6] S. Kumar et al., "Web Scraping Techniques and Applications," SCRS Publications.

[7] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web Data Extraction, Applications and Techniques: A Survey," Knowledge-Based Syst., 2014.

[8] I. Naing, S. T. Aung, K. H. Wai, and N. Funabiki, "A Reference Paper Collection System Using Web Scraping," Electronics, vol. 13, no. 14, 2024.

[9] ResearchGate, "APIs and Data Collection," 2024.

[10] O. Etzioni et al., "Scheduling Algorithms for Web Crawling," Computer Networks.

[11] Elsevier, "Automated Data Collection and Intelligent Systems," Information Sciences, 2025.

[12] "Online Job Portal Systems and Their Impact," JETIR Journal, 2024.

[13] Wikipedia Contributors, "Federated search," Wikipedia.

[14] Springer, "Aggregation and Text Matching in Search Systems," Journal of Big Data, 2025.

[15] C. Lotfi et al., "Web Scraping Techniques and Applications: A Literature Review," 2023.

[16] P. Tewari et al., "From Web to Insights: Automating and Optimizing Job Data Collection with Selenium," World Acad. J. of Eng. Sci., vol. 11, no. 4, 2024.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[17] N. Kumar et al., "Technical Job Recommendation System Using APIs and Web Crawling," Med. & Biol. Eng. & Comput., 2022.

[18] N. R. et al., "Smart Job-Seeking Assistant and Web Scraping," Int. J. Comput. Eng. Res. Trends, vol. 12, no. 11, 2025.

[19] Wikipedia Contributors, "Vector Space Model," Wikipedia.

[20] M. Chahal, "Term Frequency–Inverse Document Frequency and Similarity Measures," 2018; Sitikhu et al., "A Comparison of Semantic Similarity Methods for Maximum Human Interpretability," 2019.

[21] S. Singh and A. Garg, "Resume Ranking with TF-IDF, Cosine Similarity and Named Entity Recognition," Proc. ICDCC, 2024.