International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 207-212, April-2024

https://doi.org/10.47001/IRJIET/2024.804029

Content HUB a Unified Content Aggregation Platform

[1]A Shruti Patil, [2]Vidhya Chavan, [3]Prof. Archana Chalwa

[1,2]Student, Department of Computer Engineering, Siddhant College of Engineering, Sudumbare, Pune, Maharashtra, India
[3]Professor, Department of Computer Engineering, Siddhant College of Engineering, Sudumbare, Pune, Maharashtra, India

***Abstract* - The project is a dynamic and versatile system designed to aggregate web content efficiently. This innovative framework leverages cutting-edge technologies to collect, organize, and present diverse online content in a unified and user-friendly manner. By enabling the aggregation of web data from various sources, including websites, social media, and news feeds, this project empowers users to access a comprehensive and curate stream of information. Whether for research, content duration, or staying informed, the framework simplifies the process of collecting and managing web content, enhancing the accessibility and utility of online information. The Carrier Content Aggregation and Preference Finding System is a comprehensive project designed to streamline and enhance the user's carrier content consumption experience. This system aggregates diverse carrier content from various sources, such as CVs, and Candidates, and employs advanced algorithms like Keyword Extraction & Text Mining to understand user preferences. Through User inter action sand feedback, it adapts and recommends personalized content tailored to individual tastes and interests. This project not only simplifies content discovery but also offers users a more engaging and relevant carrier experience in an increasingly digital world.**

Keywords: Content Acquisition, Web Content Aggregation, Data Mining, Keyword Extraction & Text Mining, etc.

I. INTRODUCTION

In an age where the internet serves as an abundant source of information, the challenge lies not in the availability of content but in its effective organization and accessibility. The project proposed system endeavors to address this challenge by introducing a dynamic and responsive system that simplifies the aggregation of web content. With the exponential growth of online data from diverse sources, including websites, social media, and news feeds, users are often over whelmed by the sheer volume of information.

This project seeks to streamline the process of collecting, crating, and presenting web content, making it more user friendly and coherent. It aspires to provide users with unified plat form that aggregates content from across the web, offering holistic view of relevant information. The frame work will offer valuable applications for a wide range of users, from researchers and content curators to those simply looking to stay informed.

The proposed system addresses the challenge of efficiently managing and accessing web content in the digital age. With the overwhelming volume of online information from various sources, this project aims to simplify content aggregation, offering users a unified plat form to access, curate, and interact with web content. This frame work has the potential to benefit researchers, content curators, and theses eking streamlined way to access and utilize online information.

II. LITERATURE SURVEY

"An Automatic Online Recruitment System based on Exploiting Multiple Semantic Resources and Concept-relatedness Measures". Authors: Aseel B. Mohammed Maree, Mohammed Belkhatir, Saadat M. Alhashmi [1]. They propose an automatic online recruitment system that employs multiple semantic resources to highlight the semantic contents of resumes and job posts. Additionally, it utilizes statistical concept-relatedness measures to further enrich the highlighted contents with relevant concepts that were not initially recognized by the used semantic resources. The proposed system has been instantiated and validated in a precision-recall based empirical framework. Recruitment is considered among the most challenging functions for job portals and human resource (HR) departments. This is because employers often receive a huge number of resumes some of which are uploaded as unstructured documents in different formats such as .pdf, .doc, and .rtf , while others are uploaded according to specific forms prepared by employers that are difficult to manually process and analyze. Recently, many companies have shifted to automatic online recruitment systems in an attempt to reduce the cost, time, and efforts required for screening out applicants and matching candidate resumes to their relevant job posts.

"Matching GitHub developer profiles to job advertisements." Authors: Claudia Hauff, Georgios Gousios [2]. GitHub is a social coding platform that enables developers to efficiently work on projects, connect with other developers, collaborate and generally be seen by the community. This visibility also extends to prospective employers and HR

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 8, Issue 4, pp 207-212, April-2024
https://doi.org/10.47001/IRJIET/2024.804029

personnel who may use GitHub to learn more about a developer's skills and interests. We propose a pipeline that automatizes this process and automatically suggests matching job advertisements to developers, based on signals extracting from their activities on GitHub. Over the years, job advertisements have asked for a larger number of skills from prospective employees. This has led to a situation where a developer matching half of the described requirements may actually be a very well qualified candidate for the advertised position.

"Matching Jobs and Resumes: a Deep Collaborative Filtering Task." Authors: Thomas Schmitt, Philippe Caillou, and Michele Sebag, [3] this paper tackles the automatic matching of job seekers and recruiters, based on the logs of a recruitment agency (CVs, job announcements and application clicks). Preliminary experiments reveal that good recommendation performances in collaborative filtering mode (emitting recommendations for a known recruiter using the click history) co-exist with poor perform a cold start mode (emitting recommendations based on the job announcement only). A tentative interpretation for these results is proposed, claiming that job seekers and recruiters whose mother tongue is French yet do not speak the same language. As first contribution, this paper shows that the information inferred from their interactions differs from the information contained in the CVs and job announcements. The second contribution is the hybrid system Majore (Matching Jobs and Resumes), where a deep neural net is trained to match the collaborative filtering representation properties. The experimental validation demonstrates Majore merits, with good matching performances in cold start mode.

"A Hybrid Approach to Conceptual Classification and Ranking of Resumes and Their Corresponding Job Posts." Authors: Abeer Zaroor, Mohammed Maree, and Muath Sabha, [4] Due to the constant growth in online recruitment, job portals are starting to receive thousands of resumes in diverse styles and formats from job seekers who have different fields of expertise and specialize in various domains. Accordingly, automatically extracting structured information from such resumes is needed not only to support the automatic matching between candidate resumes and their corresponding job offers, but also to efficiently route them to their appropriate occupational categories to minimize the effort required for managing and organizing them. As a result, instead of searching globally in the entire space of resumes and job posts, resumes that fall under a certain occupational category are only those that will be matched to their relevant job post. In this research work, we present a hybrid approach that employs conceptual based classification of resumes and job postings and automatically ranks candidate resumes (that fall under each category) to their corresponding job offers. In this context, we exploit an integrated knowledge base for carrying out the classification task and experimentally demonstrate - using a real-world recruitment dataset- achieving promising precision results compared to conventional machine learning based resume classification approaches.

"An Automatic Online Recruitment System based on Exploiting Multiple Semanticm Resources and Concept-relatedness Measures." Authors: Aseel B. Mohammed Maree. Mohammed Belkhatir Description. Saadat M. Alhashmi, [5] Due to the rapid development of job markets, traditional recruitment methods are becoming insufficient. This is because employers often receive an enormous number of applications (usually unstructured resumes) that are difficult to process and analyze manually. To address this issue, several automatic recruitment systems have been proposed. Although these systems have proved to be more effective in processing candidate resumes and matching them to their relevant job posts, they still suffer from low precision due to limitations of their underlying techniques. On the one hand, approaches based on key-word matching ignore the semantics of the job post and resume contents; and consequently a large portion of the matching results is irrelevant. On the other hand, the more recent semantics-based models are influenced by the limitations of the used semantic resources, namely the incompleteness of the knowledge captured by such resources and their limited domain coverage. In this paper, we propose an automatic online recruitment system that employs multiple semantic resources to highlight the semantic contents of resumes and job posts. Additionally, it utilizes statistical concept-relatedness measures to further enrich the highlighted contents with relevant concepts that were not initially recognized by the used semantic resources. The proposed system has been in satiated and validated in a precision-recall based empirical framework.

III. PROPOSED SYSTEM

In the proposed system represents a sophisticated and user-centric solution designed to alleviate the challenges associated with the exponential growth of web content. At its core, this framework aims to streamline the process of collecting, organizing, and accessing web content from a multitude of online sources, providing users with a unified and intuitive platform. One of the pivotal features of this proposed system is its ability to aggregate content from a diverse array of web sources, including websites, social media platforms, news feeds, blogs, forums, and more. By serving as a central hub for content aggregation, users can access information from various platforms without the need to visit each source individually. This unification significantly simplifies the process of content discovery and management. To enhance the user experience, the system allows for highly personalized

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 207-212, April-2024

https://doi.org/10.47001/IRJIET/2024.804029

content streams. Users can tailor their content feeds by specifying their areas of interest, selecting preferred sources, and defining keywords. This customization ensures that the content users receive is not only relevant but also aligned with their individual preferences and needs.

The interface of the system is designed to be intuitive and user-friendly, catering to users with varying levels of technical expertise. Navigation is seamless, allowing users to access their content streams, perform searches, and manage their content with ease. Content duration is a key aspect of this framework, and it provides users with a range of tools to make the process more efficient. Users can bookmark articles, tag content, and create collections, thereby simplifying the organization of their content. Additionally, robust search and discovery functionalities empower users to quickly locate content of interest through keyword searches, filters, and content categories. Custom notifications and alerts are yet another valuable feature. Users can set up notifications based on their selected topics, sources, or keywords, ensuring they stay up to date with the latest developments and don't miss critical updates.

For users seeking deeper insights, the system may offer data analytics tools. These tools can analyze trends, sentiment, and patterns within the aggregated content, providing valuable data-driven insights for researchers, businesses, and individuals. The system is designed with scalability in mind, capable of handling the ever-increasing volume of online content and adapting to the evolving needs and preferences of its users. Robust security measures are in place to protect user data and privacy, instilling trust in the safety of personal information and content interactions. It's important to note that while focusing on a single platform can offer these advantages, it may also limit the diversity of data and perspectives available to users. The choice between aggregating content from various platforms or focusing on a single one should be driven by the specific goals and requirements of the project.

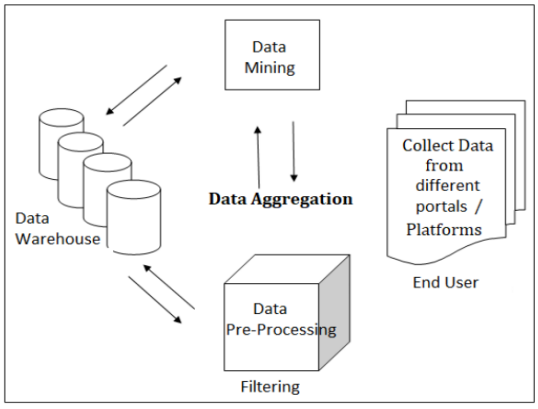

Figure 1: Proposed Methodology

System design for a content aggregation system focused on collecting content from one specific platform involves careful planning and considerations to ensure efficiency and user satisfaction. Here's a system design in paragraph format: The system design for our content aggregation platform, which concentrates on collecting content from a single platform, is centered on providing a seamless and efficient user experience while optimizing data collection, processing, and presentation. At its core, the system leverages the platform's API (Application Programming Interface) to establish a secure and autheorized connection for data retrieval. This API integration allows the system to interact with the platform's data resources while adhering to the platform's terms of use and ensuring legal compliance.

Data consistency is a fundamental aspect of the design, as the system will focus on maintaining uniformity in data structure, format, and presentation. This consistency not only enhances the user experience but also streamlines data processing and analysis. To improve data quality, accuracy, and relevance, the system will implement robust quality control measures, including data validation and verification mechanisms. Content relevance is a key priority, and the system will employ advanced algorithms and filters to extract the most pertinent and valuable information for users. The user interface and data presentation will be optimized to provide a user-friendly experience, tailored to the expectations and preferences of the platform's audience. The system design emphasizes efficient maintenance and adaptability. Regular updates to the platform's data structure can be accommodated seamlessly, ensuring that the system remains responsive and reliable. The simplicity of dealing with content from a single source also facilitates trustworthiness assessments, enabling the system to verify the accuracy and credibility of the content more effectively. Overall, the system design strikes a balance between a user-centric experience and effective content collection. By focusing on one specific platform, it aims to deliver a streamlined and reliable solution that optimizes data quality and relevance, offering users a valuable and consistent source of information.

IV. RESULTS AND DISCUSSIONS

Components Involved Figure 2 shows the various components involved in generating a user profile and the flow of data through these components. This entire system is built using Java as the programming language with Hibernate framework as the 'Object-Relational Mapping' technology and MySQL as the relational database system to store the data for the users. The system exposes a REST ful Web Service which provides the list of user profiles over the web in JSON format. Following are the components in detail:

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 207-212, April-2024

https://doi.org/10.47001/IRJIET/2024.804029

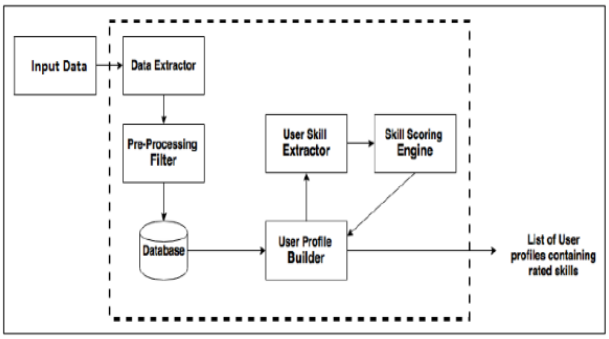

Figure 2: Methodology

A) Data Extractor The job of extracting the data about the user from the Google API or external webservices is done by the Data Extractor component. Since Git Hub provides REST ful API over HTTP to get the data, java's 'Http URL Connection' object is created using a URL, over which the data is available. After the setup of the connection, the JSON data is read as a stream using an 'Input Stream Reader' object. The string of JSON data is then mapped to a 'JSON Object' available in 'json' library which provides useful APIs to parse a JSON string. The output from this component is an object of type 'JSON Object' containing data pulled form Gi tHub.

B) Pre-Processing Filter the JSON data pulled from GitHub contained a lot of in-formation which is irrelevant to our problem statement like svn-url, merge or rebase permissions etc for a repository or avatar-url, gists-url, html-url etc for a user. The information we require is the general data of user like username, education, email id and number of followers, plus the push events (which are commits). modification done in the commits, issues events, issue comment events for the user and information about repositories created or starred by the user. So, we need a component that filter out the unnecessary information pulled from GitHub and provides only the required information. And that is the functionality of the Pre-Processing Filter component. The output from this component is a series of java objects which contain relevant information to the problem in hand.

C) Database Data base stores all the information required about the user to identify the skills an well as rate those skills based on the contributions done in repositories. As pointed out earlier, we use relational database management system, MySQL, as the database technology. Hibernate framework is used to map the object received from the Pre-Processing Filter to the tables present in the database.

D) User Profile Builder After having all the necessary data, the next step is to process the data. The User Profile Builder component pulls the data from the database and builds the Profile for the users. These Profiles contain basic information about the users, plus the contributions done by the users in the open source platform. To identify the skills possessed by a user and the proficiency in those skills, User Profile Builder sends the user Profile to the User Skill Extractor component.

E) User Skill Extractor User Skill Extractor receives Profile for all the users from the User Profile Builder. As the skills for the users are identified using a couple of steps: repository languages and file extensions. Languages used in repositories created or starred by the user are used to find user skills. But not all the languages are identified by the library used by web-services.

So we move to the next step of identifying skills, which is based on the file extensions of the Profile present in those repositories. The output from this component is all the user Profile containing list of skills for each user.

F) Skill Scoring Engine To score the proficiency of every skill identified for a user by User Skill Extractor component, Skill Scoring Engine is used. As three out of the six parameters are identified for rating the skills of the users: content of commits done, number of issues closed by the user and number of followers for the user.

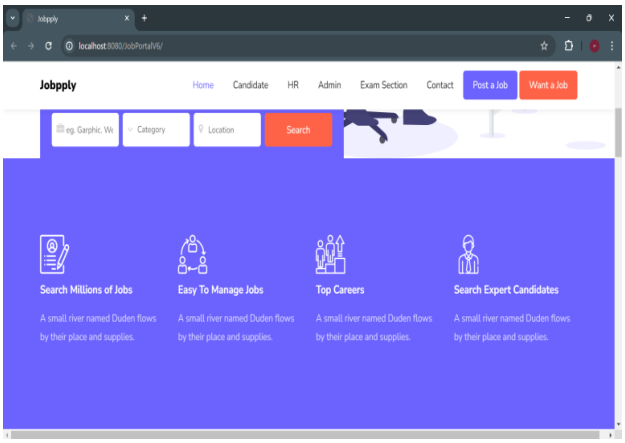

Figure 3: Home Page

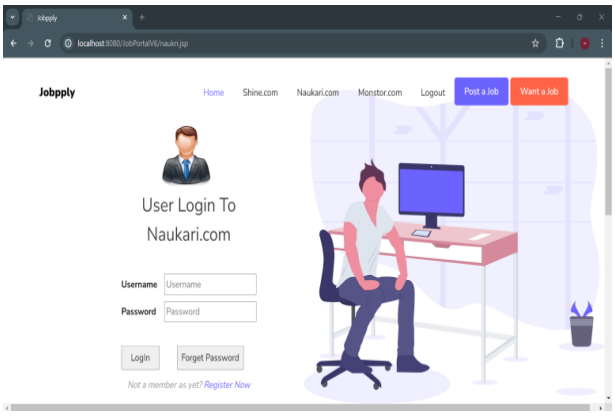

Figure 4: Login Page

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 207-212, April-2024

https://doi.org/10.47001/IRJIET/2024.804029

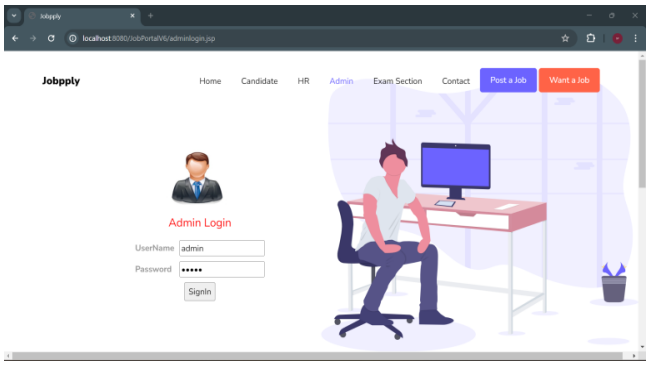

Figure 5: Admin

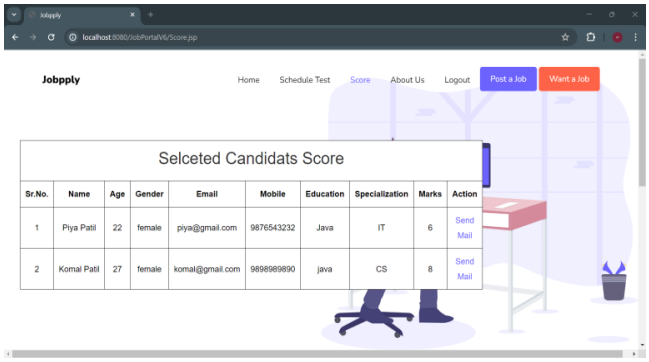

Figure 6: Candidates Score

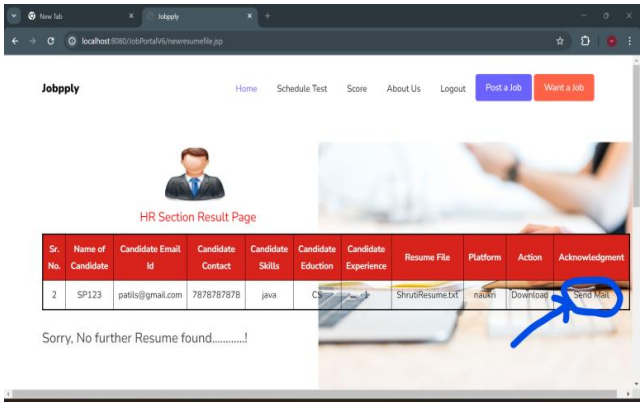

Figure 7: HR Mail

V. CONCLUSION

In this project, we have presented the concept of Web Content Acquisition in Web Content Aggregation concludes with the successful development and implementation of a robust system for efficiently gathering relevant web content from various online sources. Through meticulous design and rigorous testing, the system demonstrates its effectiveness in automating the acquisition process and crating a comprehensive dataset of web content. This dataset meets specific criteria or user preferences, ensuring the quality, relevance, and freshness of the acquired content. The project's success signifies a significant advancement in the field of web content aggregation, providing users with access to a rich and diverse collection of web content for various applications. Moving forward, further enhancements and optimizations can be explored to continually improve the system's performance and capabilities in meeting evolving user needs and preferences.

REFERENCES

[1] Abdulrahm an Aljuaid, Maysam Abbod, "Artificial Intelligence-Based E-Recruitments System", Proceedings of 2020 IEEE 10th International Conference on Intelligent Systems, 2020.

[2] Jongwoo Kim, Daniel X. Le, and George R. "Na¨ıve Bayes Classifier for Extracting Bibliographic Information from Biomedical Online Articles", National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA.

[3] Ajay S. Patil, B.V. Pawar, "Automated Classification of Web Sites using Naive Bayesian Algorithm", Proceedings of the International Multi-Conference of Engineers and Computer Scientists 2012 Vol I, IMECS 2012, March 14-16, 2012, Hong Kong.

[4] Md. Faisal Kabir "Enhanced Classification Accuracy on Naive Bayes Data Mining Models", International Journal of Computer Applications (0975 – 8887) Volume 28– No.3, August 2011.

[5] Mauricio A. Valle, Samuel Varas, Gonzalo A. Ruz "Job performance prediction in a call center using a naive Bayes classifier", Facultad de Ciencias Economicas y Administrative, ´ Universidad de Valpara´ıso, Santiago, Chile, 2011.

[6] Gla´ucia M. Bressan "Using Bayesian networks with rule extraction to infer the risk of weed infestation in a corn-crop Universidadede Sa˜o Paulo, Department of de Engenharia Ele´trica, 13566-590Sa˜o Carlos, SP, Brazil 2009.

[7] S.L. Ting, W.H. Ip, Albert H.C. Tsang. "Is Na¨ıve Bayes a Good Classifier for Document Classification?", International Journal of Software Engineering and Its Applications Vol. 5, No. 3, July, 2011.

[8] YasinUzun" Keyword Extraction Using Naive Bayes", Bilkent University, Department of Computer Science, Turkey.

[9] Binal A. Thakkar, Mosin I. Hasan, Mansi A. Desai "Health Care Decision Support System For Swine Flu Prediction Using Na¨ıve Bayes Classifier", International Conference on Advances in Recent Technologies in Communication and Computing, India, 2010.

[10] B. Zoltak. An Efficient Message Authentication Scheme for Stream Cipher. Cryptology ePrint Archive 2004.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 207-212, April-2024

https://doi.org/10.47001/IRJIET/2024.804029

AUTHORS BIOGRAPHY

A Shruti Patil, Student, Department of Computer Engineering, Siddhant College of Engineering, Sudumbare, Pune, Maharashtra, India.

Vidhya Chavan, Student, Department of Computer Engineering, Siddhant College of Engineering, Sudumbare, Pune, Maharashtra, India.

Citation of this Article:

A Shruti Patil, Vidhya Chavan, Prof. Archana Chalwa, "Content HUB a Unified Content Aggregation Platform", Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 8, Issue 4, pp 207-212, April 2024. Article DOI https://doi.org/10.47001/IRJIET/2024.804029

****

