

# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 11, Issue 5, May 2024



INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA

**Impact Factor: 7.802**



+91 99405 72462



+9163819 07438



ijmrsetm@gmail.com



www.ijmrsetm.com

# Web Content Acquisition in Web Content Aggregation

Akshada Khamat<sup>1</sup>, Alim Tamboli<sup>2</sup>, Yasmin Shaikh<sup>3</sup>, Prof. Sapike N.S.<sup>4</sup>

<sup>1,2,3</sup> Students, Department of Computer Engineering, Vishwabharti Academy's College of Engineering, Ahmednagar, Maharashtra, India

<sup>4</sup> Asst. Professor, Department of Computer Engineering, Vishwabharti Academy's College of Engineering, Ahmednagar, Maharashtra, India

**ABSTRACT:** In 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 curated stream of information. Whether for research, content curation, 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 interactions and 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.

**INDEX TERMS:** Content Acquisition, Web Content Aggregation, Data Mining, Keyword Extraction & Text Mining, etc

---

## I.INTRODUCTION

The main motive of this project becomes to construct trust evaluation Mechanism for consumer Recruitment in community Crowd-Sensing. So as to be built on Google's Cloud. Big enterprises and head-hunters get hold of numerous lots of resumes from activity candidates each day. HRs and bosses go through loads of resumes manually. Resumes or Profiles are unstructured files and feature generally ranges of various formats. As a result, manually reviewing more than one profile is a very time consuming techniques. How to ensure you have the correct Candidate in the right jobs on the proper time. That is a considerable hassle confronted by way of massive agencies nowadays in the market. Now a day's many task portals are to be had however the fundamental trouble in available device are it required manual efforts for both candidates and Employers. Candidate has to offer entire records in given text filed and organization also desires to use many filters to select the candidate. Even though employer has implemented many filters he might get heaps of resume even going via it and choosing applicants is very inefficient and time ingesting mission. Some costly extraction systems are available within the marketplace that still do the quest on key-word foundation and has many extraction boundaries like forcing applicants to fill templates and preserve updating the templates as in keeping with job profiles. Not an unmarried smart device available within the market which has advantages of information mining in addition to in an effort to take consideration of information found in social networking.

## II.RELATED WORK

An classification paradigm is a data mining framework containing all the concepts extracted from the training dataset to differentiate one class from other classes existed in data. The primary goal of the classification frameworks is to provide a better result in terms of accuracy. However, in most of the cases we cannot get better accuracy particularly for huge dataset and dataset with several groups of data. When a classification framework considers whole dataset for training then the algorithm may become un useable because dataset consists of several group of data. The alternative way of making classification useable is to identify a similar group of data from the whole training data set and then training each group of similar data. In our paper, we first split the training data using k-means clustering and then train each group with Naive Bayes Classification algorithm. In addition, we saved each model to classify sample or

unknown or test data. For unknown data, we classify with the best match group/model and attain higher accuracy rate than the conventional Naive Bayes classifier.

### III. PROBLEM STATEMENT

When it comes to recruiting and hiring, resumes are still the coin of the realm. While the Internet has lived up to its promise of opening access to new sources of talent, it has also made it much easier for job seekers to apply for jobs. The result has been a profusion of resumes – it's not uncommon for employers to receive hundreds or even thousands every time they post a job. Recruitment professionals need robust tools to help them take control of the resume flow, capture relevant information automatically, and upload candidate data directly to their database or applicant tracking system as efficiently as possible.

### IV. PROPOSED SYSTEM

HRs and Managers go through a hundreds of resumes manually. Resumes or Profiles are unstructured documents and have typically number of different formats (eg: .doc, .txt). As a result, manually reviewing multiple profiles is a very time consuming processes. How to ensure you have the Appropriate Candidate in the right jobs at the right time. This is a significant problem faced by large companies today in the market. Automated Resume Extraction and Candidate Selection System is a product which can be best suited for any organization's recruitment process. The system will be robust enough which will automatically extract the resume content and store it in a structure form within the Data Base. Classification algorithm (Naïve Bayes) will be run on the profiles to identify profile Categories or classes. Also the employer can specify his criteria and also decide the importance level. As the internet grows, amount of electronic text increases rapidly. This brings the advantage of reaching the information sources in a cheap and quick way. Keywords are useful tools as they give the shortest summary of the document. But they are rarely included in the texts. There are proposed methods for automated keyword extraction. This paper also introduces such a method, which identifies the keywords with their frequencies and positions in the training set. It uses Naïve Bayesian Classifier with supervised learning.

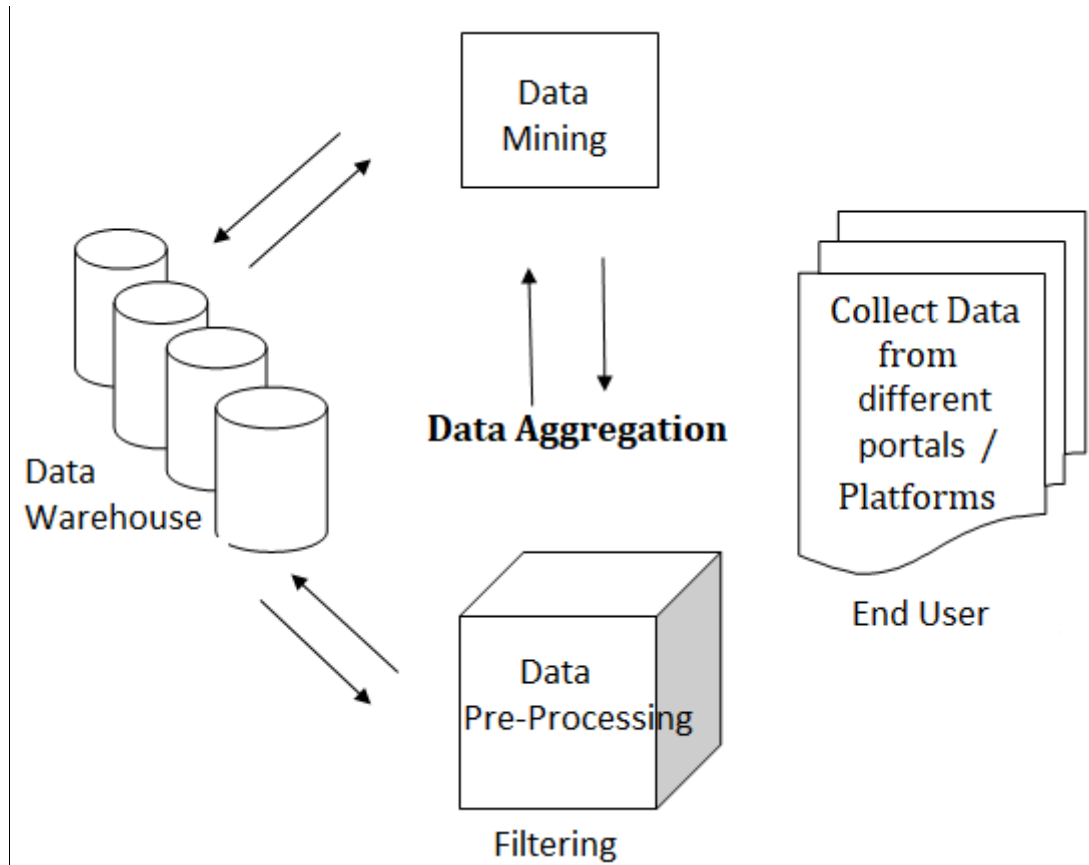


Fig.1: Proposed System Architecture

## V. RESEARCH METHODOLOGY

The above fig.1 illustrates the system architecture; it carried the overall classification process on sentiment analysis using movie review data set. Initially pre-processing analyzes the opinions from syntactical point of view and original syntax of sentence is not disturbed. In this phase, the several techniques like POS tagging, Stemming and Stop word removal are applied to data set for noise reduction and facilitating feature extraction.

### 1. Data Pre-processing

In the data pre-processing phase, we first process the data which is extracted from training as well as testing documents. Various methods have been used for data pre-processing these are describe in below section

### 2. Stop Word Removal:

Stop words are common and high frequency words like “a”, “the”, “of”, “and”, “an”. Different methods available for stop-word elimination; ultimately enhance performance of feature extraction algorithm.

### 3. Stemming

Stemming and Lemmatization are two essential morphological processes of pre- processing module during feature extraction. The stemming process converts all the inflected words present in the text into a root form called a stem. For example, ‘automatic,’ ‘automate,’ and ‘automation’ are each converted into the stem ‘automat.’ Stemming gives faster performance in applications where accuracy is not major issue.

### 4. Lemmatization (lemmas):

The lemma of a word includes its base form plus inflected forms. For example, the words “plays”, “played and “playing” have “play” as their lemma. Lemmatization groups together various inflected forms of word into a single one. Stemming also removes word inflections only whereas; Lemmatization replaces words with their base form. For example, the words “caring” and “cars” are reduced to “car” in a stemming process whereas lemmatization reduces it to “care” and “car” respectively, hence lemmatization is considered to be more accurate.

### 5. Part of speech (POS) tagging:

Parts of speech or POS tagging is a linguistic technique used which is used many existing researchers, for product feature extraction as product aspects are generally nouns or noun phrases. POS tagging assigns a tag to each word in a text and classifies a word to a specific morphological category such as noun, verb, adjective, etc. POS taggers are efficient for explicit feature extraction in terms of accuracy they achieved, however problem arises when review contains implicit features.

### 6. Features Extraction

In this phase system extract various feature set using machine learning methods for sentiment classification. We extract four basic features from preprocessed data like unigram features, Bi-tagged features, dependency rule base features etc. all these feature extraction techniques have illustrated in below section.

### 7. Unigram features

A Unigram feature is nothing but a bag of words extort by separating text by spaces as well as noise characters. The unigram model is additionally called the bag of words model.

### 8. Keyword Extraction

Hybrid method has used for feature selection from full extracted features. Basically three types of features have been extracted from given data. The purpose of select the best feature which increase the accuracy of classification. Many irrelevant features appear during the feature extraction; it needs to eliminate when we select the features. We used TF-IDF, Maximum Relevance and co-relation base hybrid method has used to select the features. The benefit of this method provides respective features selection for individual features set. The TF-IDF cosine similarity, TF-IDF Co-occurrence matrix and MRMR method has used for keyword extraction.

### 9. Classification:

After we get the training model, we can feed the testing data into it and get the prediction of classification. The testing stage includes preprocessing of testing text, vectorization and classification of the testing text.

## VI. CONCLUSION

In this paper, we have proposed a Resume Extractor and Candidate Recruitment System based on content aggregation an integrated skills knowledge base and an automatic matching procedure between candidate resumes and their corresponding job postings. The proposed System aspires to simplify and enhance the aggregation and management of web content, ensuring that users can easily access, organize, and interact with information from diverse online sources. Its broad applications across various real-life scenarios, ranging from research and content creation to staying informed and making data-driven decisions, position it as an invaluable tool for users seeking to navigate and harness the vast online landscape efficiently. Ultimately, this framework aims to elevate the accessibility and utility of online information for users from all walks of life. In the future work, we plan to utilize the extracted information from applicants’ resumes to dynamically generate user profiles to be further used for recommending jobs to job seekers.

## VII.ACKNOWLEDGMENT

We would prefer to give thanks the researchers likewise publishers for creating their resources available. We are conjointly grateful to guide, reviewer for their valuable suggestions and also thank the college authorities for providing the required infrastructure and support.

## REFERENCES

- [1] Jongwoo Kim, Daniel X. Le, and George R. “Naïve Bayes Classifier for Extracting Bibliographic Information from Biomedical Online Articles”, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA.
- [2] 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
- [3] 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.
- [4] Mauricio A. Valle, Samuel Varas, Gonzalo A. Ruz “Job performance prediction in a call centre using a naive Bayes classifier”, Facultad de Ciencias Económicas y Administrativas, Universidad de Valparaíso, Santiago, Chile, 2011.
- [5] Glaucia M. Bressan “Using Bayesian networks with rule extraction to infer the risk of weed infestation in a corn-crop Universidade de São Paulo, Department Engenharia Elétrica, 13566-590 São Carlos, SP, Brazil 2009.
- [6] S.L. Ting, W.H. Ip, Albert H.C. Tsang. “Is Naïve Bayes a Good Classifier for Document Classification?”, International Journal of Software Engineering and Its Applications Vol. 5, No. 3, July, 2011.
- [7] Yasin Uzun “Keyword Extraction Using Naive Bayes”, Bilkent University, Department of Computer Science, Turkey.
- [8] Binal A. Thakkar, Mosin I. Hasan, Mansi A. Desai “Health Care Decision Support System for Swine Flu Prediction Using Naïve Bayes Classifier”, International Conference on Advances in Recent Technologies in Communication and Computing, India, 2010
- [9] E. Ferrara and R. Baumgartner, “Design of Automatically Adaptable Web Wrappers,” Proc. The 3rd International Conference on Agents and Artificial Intelligence, 2011.
- [10] R. Baumgartner, G. Gottlob, and M. Herzog, “Scalable web data extraction for online market intelligence,” Proc. The VLDB Endowment, vol. 2(2), pp. 1512–1523, 2009.
- [11] E. Ferrara, and R. Baumgartner, “Automatic Wrapper Adaptation by Tree Edit Distance Matching,” Proc. The 22th IEEE International Conference on Tools with Artificial Intelligence, 2010



# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT



+91 99405 72462



+91 63819 07438



ijmrsetm@gmail.com

[www.ijmrsetm.com](http://www.ijmrsetm.com)