ISSN 2395-1621

A MACHINE LEARNING APPROACH FOR FINDING HIDDEN JOBS USING WEB MINING

[#1]Vidya Patil, [#2]Sayali Gund, [#3]Pranjal Hande, [#4]Pooja Kale,
[#5]Prof. Shubhangi R. Khade

[1]vidyapatil2397@gmail.com
[2]sayaligund143@gmail.com
[3]handepranju1610@gmail.com
[4]prkale96@gmail.com

[#12345]Department of Computer
Modern Education Society's College of Engineering Pune-411011

ABSTRACT

In the job classification field, precise classification of jobs to profession categories is important for harmonizing job seekers with appropriate jobs. An example of such a job title classification system is an automatic text job post classification system that utilizes machine learning. Machine learning based job type classification techniques for text and related entities have been well researched and successfully applied in many industrial settings. Digital recruitment is a popular online method that has been widely used for attracting individuals who are seeking for career opportunities. This is approach for machine learning-based semi-supervised job title classification system. Our method contains varied collection of classification and techniques to full the challenges of designing a scalable classification system for a large taxonomy of job categories. It encompasses these techniques in cascade classification architecture. We first present the architecture of our system, which consists of a two-stage Capture with filtration and fine level classification algorithm. This approach is presenting Experimental results on real world live data which is twitter feeds.

ARTICLE INFO

Article History

Received: 25th November 2017

Received in revised form :

25th November 2017

Accepted: 27th November 2017

Published online :

4th December 2017

Keywords: Cloud, Twitter, Web Mining, Mongo DB, Machine learning.

I. INTRODUCTION

The World Wide Web (WWW) is a popular and interactive medium with tremendous growth of amount of data or information available today. The World Wide Web is the collection of documents, text files, images, and other forms of data in structured, semi structured and unstructured form. It is also huge, diverse, and dynamic, hence raises the scalability. The primary aim of web mining is to extract useful information and knowledge from web. The web data store becomes the important source of information for many users in various domains. The web mining becomes the challenging task due to the heterogeneity and lack of structure in web resources. Because of these situations, the web users currently drowning in information and facing information overload [8]. Most of the web users could encounter the following problems, while interaction with the web;

Finding Appropriate Information related to job:

When a user wants to find specific information in the web, they input a simple keyword query. The query response will be the list of pages ranked depends on their similarity to the query. Though, today's search tools have some problems such as Low precision (due to the irrelevance of search results) and Low recall (inability to index all the information available).

Goal and Objective:

Goal of propose system is too efficiently and e effectively process the Job data for extracting the

© 2017, IERJ All Rights Reserved

Knowledge or Pattern from it and we can make job easily available for job seekers.

The social networking site twitter is used to achieve the goal for Pre-processing of raw job data from Social Networking Sites. The application of NLP API's is used for acquisition and literation. The Naive Bayes algorithm is used for Data Text Classification and Extracting the Knowledge from this processed Data. This approach will Extract the Job patterns related to Technology and Science.

II. LITERATURE SURVEY

Grant Williams and Ana's Mahmoud, "Mining Twitter Data for a More Responsive Software Engineering Process", in this paper Twitter has created an unprecedented opportunity for software developers to monitor the opinions of large populations of end-users of their software. How-ever, automatically classifying useful tweets is not a trivial task. Challenges stem from scale of the data available, its unique format, diverse nature, and high percentage of spam. To overcome these challenges, this extended abstract introduces a three-fold procedure that is aimed at leveraging Twitter as a main source of technical feedback that software developers can benefit from. The main objective is to enable a more responsive, interactive, and adaptive software engineering process. Our analysis is conducted using a dataset of tweets collected from the Twitter feeds of three software systems. Our results provide an initial proof of the technical value of software-relevant tweets and uncover several challenges to be pursued in our future work.

Bholane Savita Dattu, Prof. Dipali V. Gore, "A Survey on Sentiment Analysis on Twitter Data Using Different Techniques", this paper explain Sentiment analysis has many applications in various domains like political domain, sociology and real time event detection like earthquakes. Previously research was carried out to model and track public sentiments. But with the advancement in research, today we can use it for interpreting the reasons of the sentiment change in public opinion, mining and summarizing products reviews, to solve the polarity shift problem by performing dual sentiment analysis. Here we use different algorithms/models to perform the above tasks like LDA approach, DSA model, Nave Bayes classifier, Support Vector Machine algorithm and so on.

Simon Fong, Raymond Wong, and Athanasios V. Vasilakos, "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data", Big Data though it is a hype up-springing many technical challenges that con-front both academic research communities and commercial IT deployment, the root sources of Big Data are founded on data streams and the curse of dimensionality. It is generally known that data which are sourced from data streams accumulate continuously making traditional batch-based model induction algorithms infeasible for real-time data mining. Feature selection has been popularly used to lighten the processing load in inducing a data mining model. However, when it comes to mining over high dimensional data the search space from which an optimal feature subset is derived grows exponentially in size, leading to an intractable demand in computation. In order to tackle this problem which is mainly based on the high-dimensionality and streaming format of data feeds in Big Data, a novel lightweight feature se-lection is proposed. The feature selection is designed particularly for mining streaming data on the y, by using accelerated particle swarm optimization (APSO) type of swarm search that achieves enhanced analytical accuracy within reasonable processing time. In this paper, a collection of Big Data with exceptionally large degree of dimensionality are put under test of our new feature selection algorithm for performance evaluation.

Johan Bollen, Huina, Mao, Xiao-Jun Zeng, "Twitter mood predicts the stock market", Due to the sheer volume of text generated by a micro blog site like Twitter, it is often difficult to fully understand what is being said about various topics. In an attempt to understand micro blogs better, this paper compares algorithms for extractive summarization of micro blog posts. We present two algorithms that produce summaries by selecting several posts from a given set. We evaluate the generated summaries by comparing them to both manually produced summaries and summaries produced by several leading traditional summarization systems. In order to shed light on the special nature of Twitter posts, we include extensive analysis of our results, some of which are unexpected.

David Inouye and Jugal K. Kalita, "Comparing Twitter Summarization Algorithms for Multiple Post Summaries", summarizing micro blogs can be viewed as an instance of the more general problem of automated text summarization, which is the problem of automatically generating a condensed version of the most important content from one or more documents. A number of algorithms have been developed for various aspects of document summarization during recent years. Notable algorithms include SumBasic and the centroid algorithm. SumBasics underlying premise is that words that occur more frequently across documents have a higher probability of being selected for human created multidocument sum-maries

© 2017, IERJ All Rights Reserved

than words that occur less frequently. The centroid algorithm takes into consideration a centrality measure of a sentence in relation to the over-all topic of the document cluster or in relation to a document in the case of single document summarization. The Lex Rank algorithm. for computing the relative importance of sentences or other textual units in a document creates an adjacency matrix among the textual units and then computes the stationary distribution considering it to be a Markov chain.

III. PROPOSED SYSTEM

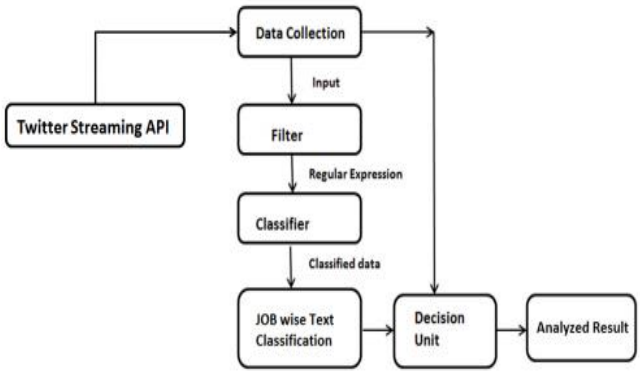

Fig 1. System architecture

Data Collection:

The data collection is the discovery of hidden information and usage pattern trends, which could aid the Web managers for improving the management, performance and controlling of the Web servers.

Data Pre-processing:

The selection of useful data is an important task in the data pre-processing stage. The data's were selected in each data type to generate the cluster models for finding web user access and server usage patterns. The removal of irrelevant and noisy data is an initial step in this task. The most recently accessed data were indexed with higher value of 'time index' while the least recently accessed data were placed at the bottom with lowest value. This becomes the critical step to obtain more precise analysis result due to time dependence characteristics of Web usage data.

Data Clustering:

The method of clustering is broadly used in different projects by researchers for finding the usage patterns or user profiles. The clustering algorithms become the most mining method in websites and the cluster objects include user groups (to describe user actions) and web pages.

IV. ALGORITHM USED

Algorithm I
Filtration Algorithm
Input: Live Data Feed
Steps:
1. Filter related data (likewise, URL, Special Characters, Emotions, and Re-tweets). All other unnecessary data will be removed.
2. Divide the Data into Appropriate Key Value Pair.
Output: Filtered data.

Algorithm II
Analysis and Classification Algorithm
Input: Filtered Data.
1. Gather the filtered data from data store.
2. Apply NLP using Machine Learning APIs for Individual Data Item from Data Store.
3. Persist the final summary into data store. Output: Analyzed and Classified Data.

Algorithm III
Job Trend Summarization
Input: Analyzed and Classified Data
Steps:
1. For each job event data or for the Technology data, Technology wise Categorical Data is extracted.
2. Summarize the data for all the live feed.
3. Persist the data into data store. Output: Trend Summarization for each job Category.

V. CONCLUSION

The unstructured data is analyzed, categorized and classified to job search entity based upon specific keywords, experience and area of interest. Job Trend Analysis can be done for various Job categories on Social Media Data feed. Various deep machine learning API techniques helps to improve the performance in job classification domain.

REFERENCES

1. T. Joachims, Transductive inference for text classication using support vector machines, In Proceedings of ICML 1999, pp. 200-209, 2016.

2. K. Crammer et al., Automatic code assignment to medical text, In BioNLP07, Association for Computational Linguistics, pp. 129-136, 2016

3. M. E. Ruize and P. Srinivasan, Hierarchical text categorization using neural networks, Information Retrieval, vol. 5(1), pp. 87-118,2016.

© 2017, IERJ All Rights Reserved

4. R. Bekkerman, M. Bilenko and J. Langford, Scaling up machine learn-ing: parallel and distributed approaches. Cambridge University, Press, 2015.

5. Guo, H. Wang, D. Bell, Y. Bi and K. Greer, Using kNN model for automatic text categorization, Soft Computing, vol. 10(5), pp. 423-430, 2015.

6. M. Zviran and W. J. Haga, User authentication by cognitive passwords: an empirical assessment, in Information Technology, 1990.Next Decade in Information Technology, Proceedings of the 5th Jerusalem Confer-ence on (Cat. No.90TH0326-9). IEEE, 1990, pp. 137144.

7. N. Roy, H. Wang, and R. R. Choudhury, I am a smartphone and I can tell my user's walking direction, in Proc. ACM MobiSys, 2014, pp.329342.

8. S. Fong, J. Liang, R. Wong, and M. Ghanavati, A novel feature selection by clustering coefficients of variations, in Proc. 9th Int. Conf. Digital Inf. Manag., Sep. 29, 2014, pp. 205213

9. W. Fan and A. Bifet, Mining big data: Current status, and forecast to the future, SIGKDD Explorations, vol. 14, no. 2, pp. 15, Dec. 2014.

10 W. Fan and A. Bifet, Mining big data: Current status, and forecast to the future, SIGKDD Explorations, vol. 14, no. 2, pp. 15, Dec. 2013.

11. A. Murdopo, Distributed decision tree learning for mining big data streams, Masters of Science thesis, European Master Distrib. Comput. Jul. 2013

12. S.Fong, X. S. Yang, and S. Deb, Swarm search for feature selection in classification, in Proc. 2nd Int. Conf. Big Data Sci. Eng., Dec. 2013, pp. 902909.

© 2017, IERJ All Rights Reserved