Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Advising Projects to Students using Data Mining and Natural Language Processing

Pratik Solim[1], Ankit Sagwekar[1], Rishikesh Patil[1], Swapnali Kurhade[2]
[1]BE student, Department of IT, Sardar Patel Institute of Technology, Mumbai, India
[2]Asistant Professor, Department of IT, Sardar Patel Institute of Technology, Mumbai, India

Abstract— Selecting project topics gives a tough time to students, as they have to go through a pool of past projects & research papers to shortlist doable projects. These projects may depend upon their field of interest. A lot of time is spent while topic selection, which eventually gets changed in later phases of development. By the use of proposed system, which is based on Natural Language Processing and data mining, they can select their domain to get a list of related projects & then can select a topic from that list

Keywords— Natural Language Processing (NLP), Machine learning (ML), Support Vector Machine (SVM), Natural Language Tool Kit (NLTK), Search Results Clustering (SRC)

I. INTRODUCTION

Most of the students have to do various projects in their curriculum. For number of subjects and fields, they need to select many project topics. These projects topics are required to have research papers associated with it. These papers give some idea to students for design, analysis, implementation and testing. Sometimes, they just have idea about domain in which they want to work. Due to large number of dataset on research paper's website, it is very time consuming task to find research papers which have work on similar problem with different technique to solve that problem. With 'Project Advisor', students can select field of their domain of interest and list of problems in that domain with list of related research papers will be shown as an output to students. This system is split into four phases- scrapping, classification, clustering, natural language processing. In the first phase of our system crawl through the web pages of research papers website to scrap title, abstract and various types of keywords. In the classification phase, it classifies papers into preselected four domains. The clusters of research papers with similar topics are found. In the final phase, a problem statement is generated for each cluster based on title and keywords. The input data is called training data and has a known label or result such as health-care or security at a time. A model is prepared through a training process where it is required to make predictions and is corrected when those predictions are wrong. Thus the system is trained till the model attains a required accuracy on the training data.

II. LITERATURE SURVEY

In this section, the previous work is reviewed. Also some of their drawbacks are discussed. The current system on the website of research papers does not classify the research papers based on domains. Specific domain related keywords are needed to be added in order to get required papers. Some general patterns in past project were observed but natural language processing could be used to further enhance it [1]. The algorithm scatters data into a small number of groups, and presents a short summary of that cluster to the user [2]. However, scatter-gather system had limitations because of its defects while using a traditional heuristic clustering algorithm. However, some well-known disadvantages are the noise induced by the snippets and the complexity. Here the SRC problems such as synonymy, polysemy, flat clustering were analysed and discussed [3]. Another well-known SRC algorithm named Lingo known by the-description-comes-first approach, the main idea of approach is that they proceed by generating meaningful labels to the clusters and assign each snippet to the right cluster. However, Lingo is unable to generate a structure for clustering and it is time-consuming for large dataset [4].

III. PROPOSED WORK

Tools and Technologies used to develop this system is discussed below:

A. *PostgreSQL:* PostgreSQL is an open source relational database management system used for managing system database which stores paper number, title, different types of keywords, abstract data. It runs on numerous platforms including Linux, most flavors of UNIX , Mac OS X and Windows. It also supports text, images, sounds, video, and includes programming interfaces for C/C++, Python. In this way it can easily achieve platform independence.

B. *Beautiful Soup:* Beautiful Soup is a Python library for actuation knowledge out of markup language and XML files. It works together with program to produce formulation ways that of navigating, searching, and modifying the take apart tree.

C. Support Vector Machine: Support Vector Machine (SVM) is a discriminative classifier formally outlined by a separating hyperplane. In alternative words, given labeled coaching knowledge (supervised learning), the algorithmic program outputs associate optimum hyperplane that categorizes new examples.

D. NLTK is a platform for building Python programs to figure with human language knowledge. It provides interfaces to over fifty corpora and lexical resources like WordNet, in conjunction with a collection of text process libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

IV. PROCESS OVERVIEW

The system crawls each web page to get title, abstract, and keywords of each research paper. In this way, training dataset is obtained. Using machine learning, the system is trained to classify the data into four classes. In each of the classes, we cluster various papers in order to get papers with similar problem statement. By using natural language processing, problem statement is formed for each cluster. User can select particular domain from available domains. User will be shown the list of problem statements in that domain. After selection of problem statement, various research papers in a selected problem statement are shown. The system model is as depicted in fig 1.

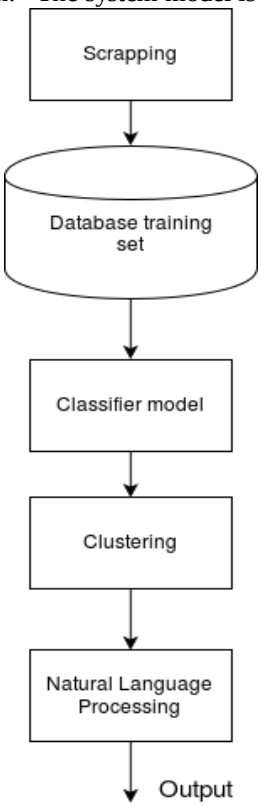

Fig 1. The System Model

V. IMPLEMENTATION

A. Data collection: Using beautiful soap and Selenium Webdriver, we crawl and collect title, abstract and keywords of recent research papers from HTML pages. Collected data is stored in PostgreSQL database.

B. Machine Learning and Classification: The keywords from PostgreSQL database are obtained which are then manually tagged as per domain and those keywords are shuffled for proper training. Stop-words which affect accuracy are removed. Keywords from all the domains are added into single file. Common keywords from each research paper are removed in order to improve accuracy. Each keyword is replaced by unique number. The system compare keywords of each research paper with training data set in order to classify the paper into respective domains as shown in fig 2.

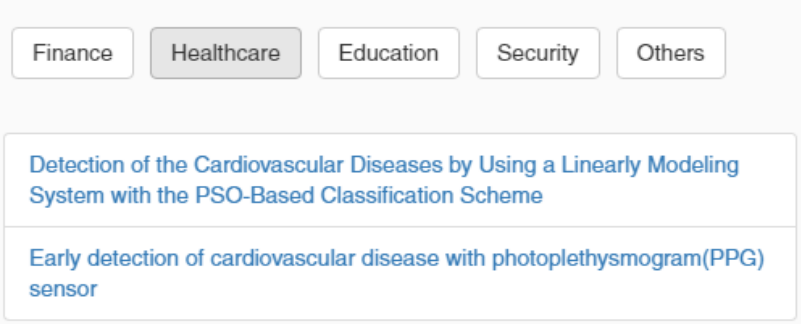

Fig 2. Research Papers listed within various domains

C. Clustering: Two-dimensional array of keywords is created with each row having all the keywords related to given paper of particular domain. Single dimensional array which contains all the keywords in the particular domain is created. Two dimensional array with rows as papers and all the keywords as columns is created. Cosine similarity is found between the papers with using euclidean formula. For finding clusters, k-means clustering can be used. Repeated k-means clustering can be used in order to reduce size of clusters.

D. Stemming: In this phase, words are transformed to its shortened version. Thus, all words with different grammatical forms can be represented by one basic form (e.g. "compute", "computing" and "computation" would all have the same form, which is "compute").

E. Natural Language Processing: A method for generating sentence from keywords is used. This method consists of two main parts: construction and evaluation. The construction part generates text sentences in the form of dependency trees. The evaluation part forms a model to generate an appropriate text when given keywords [5].

VI. RESULT

The classifier model was tested for 210 research papers with training data of 550 papers and accuracy of 92 percent was found. This shows that the system could classify the research papers into the predefined domains with higher level of accuracy.

VII. CONCLUSION

First and foremost we want to help students to decide their projects based on their area of interest. Based on our survey most of the students face problems while deciding their project topics. Our application will help every such user by giving list of research papers and technologies that can be used to solve the problem. Natural Language Processing is used in order to suggest the students project topics along with available research papers.

VIII. FUTURE WORK

The accuracy of classification and clustering can be improved further. In this system domains are predefined. Further the system features can be enhanced so that it can define the domains by its own.

REFERENCES

[1] Andrew Poon, "What Project Should I Choose"

[2] Aslı Çaliş, Ahmet Boyaci, Kasım Baynal, "Data mining application in banking sector with clustering and classification methods", 2015 International Conference on Industrial Engineering and Operations Management Dubai, United Arab Emirates

[3] Hansaem Park, Kyunglag Kwon, Abdel-ilah Zakaria Khiati, Jeungmin Lee, and In-Jeong Chung, "Agglomerative Hierarchical Clustering for Information Retrieval Using Latent Semantic Index", 2015 IEEE

[4] S. Osiński, J. Stefanowski, and D. Weiss, "Lingo: Search results clustering algorithm based on singular value decomposition" in Intelligent information processing and web mining, Springer, 2005

[5] Kiyotaka Uchimoto, Satoshi Sekine, Hitoshi Isahara, "Text Generation from Keywords".

[6] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary, "Twitter trending topic classification".

