Review on Project Management using Data Mining

[1]Abhijeet Jagtap, [2]Balwant Chauhan, [3]Pratik Sonawane, [4]Vidyadhar Suryawanshi , [5]Prof. D.V.Bhate
Department of Computer Engineering Zeal College of Engineering & Research, University of Pune.

Abstract- *This project is incredibly helpful for those organization UN agency work on project management like school organization. as a result of it's entirely effective of all project that is finished by students and project guide. the method from cluster formation to project submission is all done by this method. All the phases like project approvals, FTR (Formal Technical Review)'s, project submission and analysis are going to be attainable. it'll facilitate to stay the track of all the project work for locating potency of project. It will method abstract and also the result like settle for or reject are often determined. conjointly we have a tendency to develop an efficient automation tool which can notice a site from abstract victimization keyword techniques and allot guide to student mechanically in step with their connected domain. In these we have a tendency to live overall performance of project like poor, average and best.*

KEYWORDS- Abstract Selection, Project Evaluation, Guide Allocation, Data Mining, Data filtering ,Text Mining

1. INTRODUCTION

The project is devoted to automation of all project work that is finished by students and project coordinators manually. These are helpful from cluster formation to project submission. All the phases like cluster formation, project approvals, FTR (Formal Technical Review)'s, project submission and analysis are potential to done through this internet app. it'll facilitate to stay the track of all the project work and within the analysis method of comes.

The web app provides the facility to students to search information about groups formed, faculty and their domains. Also information on past projects is available. The plan of action to be followed is also provided. This web app provides a platform for guide student communication. All the deadline notifications will be sent to students regularly. Also the remarks given by faculty member can be known anytime students wish. The project will be given ratings by the faculty at all levels. At the time of final evaluation the staff can refer the graphs representing the overall performance. Thus the web app helps in proper communication and systematic evaluation.

The objective of connectedness feature discovery (RFD) is to find the helpful options accessible in text documents, as well as each relevant and irrelevant ones, for describing text mining results. This is often a very difficult task in trendy data analysis, from each associate empirical and theoretical perspective. This drawback is additionally of central interest in several internet customized applications, and has received attention from researchers in data processing, Machine Learning, data Retrieval and internet Intelligence communities. There square measure 2 difficult problems in victimization pattern mining techniques for finding connectedness options in each relevant and irrelevant document. The first is that the low-support drawback. Given a subject, long patterns square measure typically additional specific for the subject, however they sometimes seem in documents with low support or frequency. If the minimum support is minimized, plenty of noisy patterns may be discovered. The second issue is that the misunderstanding drawback, which implies the measures (e.g., "support" and "confidence") utilized in pattern mining end up to be not appropriate in victimization patterns for determination issues. For instance, a extremely frequent pattern (normally a brief pattern) is also a general pattern since it may be ofttimes utilized in each relevant and irrelevant documents. Hence, the difficult drawback is a way to use discovered patterns to accurately weight helpful options. There square measure many existing strategies for determination the 2 difficult problems in text mining. Pattern taxonomy mining (PTM) models are projected, in which, mining closed consecutive patterns in text paragraphs and deploying them over a term area to weight helpful options. Concept-based model (CBM) has conjointly been projected to find ideas by victimization tongue process (NLP) techniques.

3. PROPOSED SYSTEM

In this system the project abstract get accepted and then certain analysis perform on that abstract for finding that the project is done in previously. Techniques can be applied on abstract like TF-IDFs a which is used to count a particular count for matching and finding domain. In this technique only the new topic are accepted and repeatable topic get rejected. In this firstly student send abstract towards organization. Then the work get started organization assess the abstract. Information retrieval techniques get applied on particular abstract and the result is calculated that it is newer project or it is older one. Then result is calculated for abstract get accepted or rejected. Then according to domain of project the guide are get allocated for particular project. And finally the result will calculated on FTR base that is project is best, poor or

average. And finally the feedback will send to student and student will able to see their performance.

The project is dedicated to effective automation of all project work which is done by students and project coordinators manually. These will be useful from group formation to project submission. All the phases such as group formation, project approvals, FTR (Formal Technical Review)'s, project submission and evaluation will be possible to done through this web application. It will help to keep the track of all the project work and in the evaluation process of projects. In this we develop an effective automation tool which finding a domain from abstract using keyword techniques and allocate guide to student automatically according to their related domain. In these we measure overall performance of project such as poor, average and best.

TECHNIQUE

TF-IDF:

We will currently examine the structure and implementation of TF-IDF for a collection of documents. we'll initial introduce the mathematical background of the algorithmic program and examine its behavior relative to every variable. we tend to then gift the algorithmic program as we tend to enforced it. we'll provides a fast informal rationalization of TF-IDF before continuing. primarily, TF-IDF works by decisive the ratio of words in a very specific document compared to the inverse proportion of that word over the complete document corpus. Intuitively, this calculation determines however relevant a given word is in a very specific document. Words that ar common in a very single or alittle cluster of documents tend to possess higher TFIDF numbers than common words like articles and prepositions.

TF = (Word Count/ Total number of Words)

IDF = (Total no of document/Words in Actual Document)

Term frequency

In the case of the term frequency tf(t,d), the best selection is to use the raw frequency of a term during a document, i.e. the quantity of times that term t happens in document d. If we have a tendency to denote the raw frequency of t by foot,d, then the straightforward tf theme is tf(t,d) = ft,d. different potentialities include

Boolean "frequencies": tf(t,d) = one if t happens in d and zero otherwise;

logarithmically scaled frequency: tf(t,d) = one + log foot,d, or zero if foot,d is zero;

augmented frequency, to stop a bias towards longer documents, e.g. raw frequency divided by the utmost raw frequency of any term within the document:

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 \times f_{t,d}}{\max\{f_{t,d} : t \in d\}}$$

Inverse document frequency

The inverse document frequency could be a live of what proportion info the word provides, that is, whether or not the term is common or rare across all documents. it's the logarithmically scaled fraction of the documents that contain the word, obtained by dividing the full range of documents by the quantity of documents containing the term, then taking the index of that quotient.

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

With N : total range of documents within the corpus $N = |D|$

$$|\{d \in D : t \in d\}|$$

Range of documents wherever the term seems (i.e., $\mathrm{tf}(t,d) \neq 0$.) If the term isn't within the corpus, this can cause a division-by-zero. it's so common to regulate the divisor to .

$$1 + |\{d \in D : t \in d\}|$$

Mathematically the bottom of the log perform doesn't matter and constitutes a continuing increasing issue towards the result.

Term frequency–Inverse document frequency

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$

- Then tf–idf is calculated as

4. IMPLEMENTATION

Effective Project Management System Architecture based on client server architecture. It consist of four modules like student app, staff app, admin app and college sever. The client can be student, staff and admin and sever were all resources and database is stored. Student will send abstract, apply text mining check repeatability of topic if found cancel topic and get domain form abstract and allocate a faculty according to domain. Staff will receive the abstract according to domain and able to give the FTR Rating according to performance of student. Staff can list performance of student like best, poor, average. At the server side we place the database, admin. Admin will manage the student and staff.

4.1 Student Module

In Student Module New student can signup, fill the details of own, they have facility to login in system and then send the abstract. We applying a text mining on the abstract and check repeatability of topic if it found then it directly reject that topic. If no repetition is found then accept the topic and get domain from abstract and allot the faculty according to domain, and student will be able to view of performance of project and FTR Rating details. Here we are applying SHR algorithm for generating 16bit hash character password automatically which takes 5 character randomly. Here we are using Information Retrieval techniques TF-IDF for analyzing domain and repetition of project

Fig 1: Student Send Abstract

4.2 Staff Module

A staff can have facility to login the system staff will see the allocated student according to domain and giving the FTR Rating based on there performance . The staff will able to see graph of student which shows overall performance of student. They have facility to see rank of student, the highest rank of student will be in top list. He can analyze the best, poor, average performance of student based on the FTR Rating.

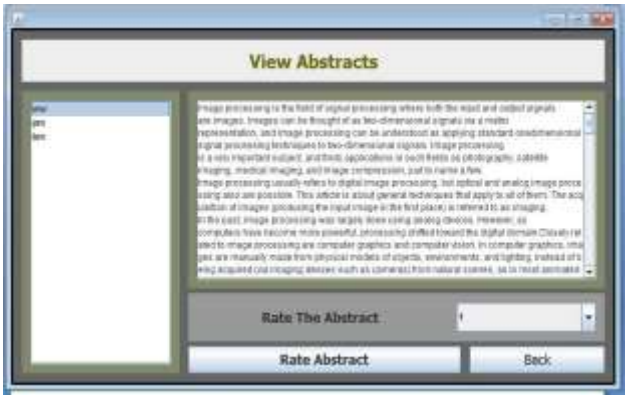

Fig 2 : Staff View Abstract

4.3 Admin Module

Admin which is a project coordinator which have facility to add and manage the staff and it can also manage the registered user. And add the reference paper for finding of domain from from the Abstract

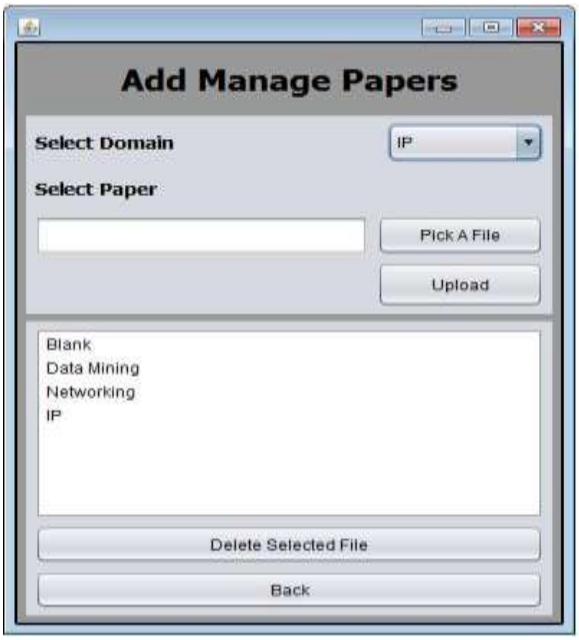

Fig 3 : Coordinator Add Abstract

4.4 Server

The databases are stored on the server side and web services are written which are used for importing the data from the client.

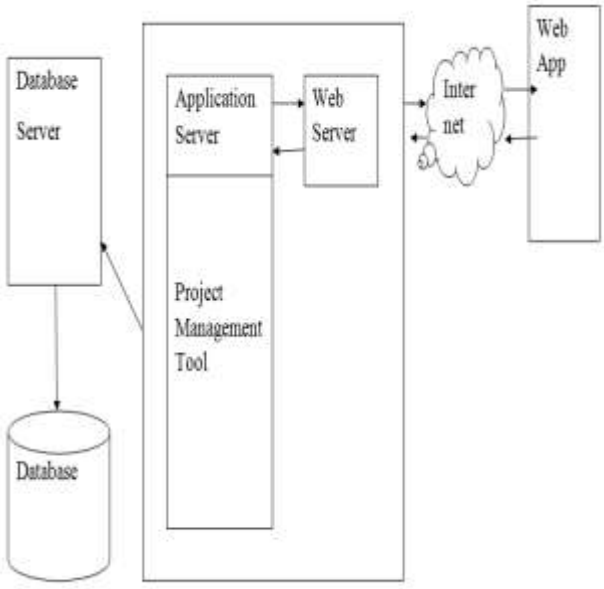

Fig 4 : Block Diagram

5. SYSTEM ARCHITECTURE

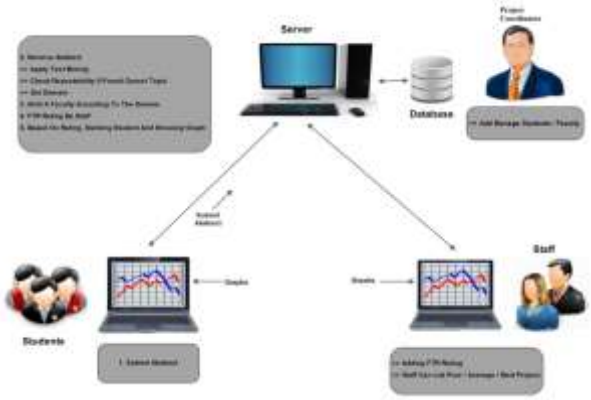

Fig 5: System Architecture

6. CONCLISION

Thus we have completed the analysis on project management tool successfully where the main motto of the project is to provide a single tool for all project data set handling and proper evaluation of the quality of projects. The tool is designed in order to maintain the efficiency and accuracy in quality evaluation and working

REFERENSES

[1] Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana "Relevance Feature Discovery for Text Mining" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 6, JUNE 2015.

[2] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu. "Effective Pattern Discovery for Text Mining" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012.

[3] Shaidah Jusoh and Hejab M. Alfawareh "Techniques, Applications and Challenging Issue in Text Mining" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012.

[4] M. Aghdam, N. Ghasem-Aghaee, and M. Basiri, "Text feature selection using ant colony optimization," in Expert Syst. Appl. vol. 36, pp. 6843–6853, 2009.

[5] A. Algarni and Y. Li, "Mining specific features for acquiring user information needs," in Proc. Pacific Asia Knowl. Discovery Data Mining, 2013, pp. 532–543.

[6] A. Algarni, Y. Li, and Y. Xu, "Selected new training documents to update user profile," in Proc. Int. Conf. Inf. Knowl. Manage., 2010,pp. 799–808.

[7] N. Azam and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," Expert Syst. Appl., vol. 39, no. 5, pp. 4760–4768,2012.

[8] R. Bekkerman and M. Gavish, "High-precision phrase-based document classification on a modern scale," in Proc. 11th ACM SIGKDD Knowl. Discovery Data Mining, 2011, pp. 231–239.

[9] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artif. Intell., vol. 97, nos. 1/2, pp. 245– 271, 1997.

[10] C. Buckley, G. Salton, and J. Allan, "The effect of adding relevance information in a relevance feedback environment," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1994, pp. 292–300.

[11] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, "Selecting good expansion terms for pseudo-relevance feedback," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 243–250.

[12] G. Chandrashekar and F. Sahin, "Asurvey on feature selection methods," in Comput. Electr. Eng., vol. 40, pp. 16–28, 2014.

[13] B. Croft, D. Metzler, and T. Strohman, Search Engines: Information Retrieval in Practice. Reading, MA, USA: Addison-Wesley, 2009.

[14] F. Debole and F. Sebastiani, "An analysis of the relative hardness of Reuters-21578 subsets," J. Amer. Soc. Inf. Sci. Technol., vol. 56, no. 6, pp. 584–596, 2005.

[15] J. Eisenstein, A. Ahmed, and E. P. Xing, "Sparse additive generative models of text," in Proc. Annu. Int. Conf. Mach. Learn., 2011, pp. 274–281.

[16] G. Forman, "An extensive empirical study of feature selection metrics for text classification," in J. Mach. Learn. Res., vol. 3, pp. 1289–1305, 2003.

[17] Y. Gao, Y. Xu, and Y. Li, "Topical pattern based document modeling and relevance ranking," in Proc. 15th Int. Conf. Web Inf. Syst. Eng., 2014, pp. 186–201.

[18] X. Geng, T.-Y. Liu, T. Qin, A. Arnold, H. Li, and H.-Y. Shum, "Query dependent ranking using k-nearest neighbor," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 115–122.

[19] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization," Technometrics, vol. 49, no. 3, pp. 291–304, 2007.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," in J. Mach. Learn. Res., vol. 3, no. 1, pp. 1157–1182, 2003.

[21] J. Ha n, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000, pp. 1–12.

[22] Y.-F. Huang and S.-Y. Lin, "Mining sequential patterns using graph search techniques," in Proc. Annu. Int. Conf. Comput. Softw. Appl., 2003, pp. 4–9.

[23] G. Ifrim, G. Bakir, and G. Weikum, "Fast logistic regression for text categorization with variable-length n-grams," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2008, pp. 354–362.