Project Management using Data Mining

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

Abstract- *It could be a huge challenge to ensure the standard of discovered connection options in text documents for describing user preferences due to giant scale terms and information patterns. Most existing in style text mining and classification strategies has adopted term-based approaches. However, they need all suffered from the issues of ambiguity and synonymousness. Over the years, there has been usually command the hypothesis that pattern-based strategies ought to perform higher than term-based ones in describing user preferences; nonetheless, a way to effectively use giant scale patterns remains a tough downside in text mining. to create a breakthrough during this difficult issue, this paper presents associate innovative model for connection feature discovery. It discovers each positive and negative patterns in text documents as higher level options and deploys them over low-level options (terms). It conjointly classifies terms into classes and updates term weights supported their specificity and their distributions in patterns. Substantial experiments victimization this model on RCV1, TREC topics and Reuters-21578 show that the projected model significantly outperforms each the progressive term-based strategies and therefore the pattern primarily based strategies.*

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

1. INTRODUCTION

The aim is to develop an internet app wherever students and guides are going to be able to perform project connected activities. These portals are going to be an excellent relief to the scholars and therefore the various guides. Tasks like cluster formation, project approval and guide allocation, communication on varied project aspects are going to be done on-line additional efficiently. Additionally with this efficient internet app we are able to have systematic follow of all the activities and correct analysis of identical.

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. It projected verb-argument structures to find ideas in sentences. These pattern (or concepts) based mostly approaches have shown a vital improvement within the effectiveness. However, fewer significant enhancements square measure created compared with the most effective term-based technique as a result of a way to effectively integrate patterns in each relevant associated irrelevant documents continues to be an open drawback. Over the years, folks have developed several mature term-based techniques for ranking documents, data filtering and text classification. Recently, many hybrid approaches were projected for text classification. to be told term options inside solely relevant documents and unlabeled documents, paper used 2 term-based models. Within the first stage, it utilised a Rocchio classifier to extract a collection of reliable irrelevant documents from the unlabelled set. within the second stage, it engineered a SVM classifier to classify text documents. A two-stage model was conjointly projected in , that verified that the combination of the rough analysis (a term-based model) and pattern taxonomy mining is that the best thanks to style a two-stage model for data filtering systems. for several years, we've ascertained that several terms with larger weights square measure additional general as a result of they're seemingly to be ofttimes utilized in each relevant and irrelevant documents . for instance, word "LIB" is also additional ofttimes used than word "JDK"; however "JDK" is additional specific than "LIB" for describing "Java Programming Languages"; and "LIB" is additional general than "JDK" as a result of "LIB" is

additionally ofttimes utilized in different programming languages like C or C++. Therefore, we have a tendency to suggest the thought of each term's distributions and specificities for connectedness feature discovery.

the goal of connexion feature discovery in text documents is to find a collection of helpful options, as well as patterns, terms and their weights, in an exceedingly coaching set D, that consists of a collection of relevant documents, Dþ, and a collection of impertinent documents, D. during this paper, we tend to assume that every one text documents, d, square measure split into paragraphs, during this section, we tend to introduce the fundamental definitions regarding patterns and therefore the deploying technique.

2. RELATED WORK

Feature choice could be a technique that selects a set of options from information for modeling Systems (see http://en.wikipedia.org/wiki/Feature_selection). Over the years, a range of feature choice strategies (e.g., Filter, Wrapper, Embedded and Hybrid approaches, and unattended or semi-supervised methods) are projected in numerous fields . Feature choice is additionally one among necessary steps for text classification and data filtering that is that the task of distribution documents to predefined categories. To date, several classifiers, like Naive Bayes, Rocchio, kNN, SVM and Lasso regression are developed, additionally several believe that SVM is additionally a promising classifier . The categoryification issues embrace the only class and multi-class downside. the foremost common answer to the multi-class downside is to decompose it into some independence binary classifiers, wherever a binary one is assigned to 1 of 2 predefined categories (e.g., relevant class or extraneous category). Most ancient text feature choice strategies used the bag of words to pick out a group of options for the multi-class downside . There square measure some feature choice criteria for text categorization, together with document frequency (DF), the world military force, data gain, mutual data (MI), Chi-Square (x2) and term strength. during this paper we tend to specialise in relevant feature choice in text documents. connection could be a massive analysis issue for internet search, that discusses a documents connection to a user or a question . However, the standard feature choice strategies aren't effective for choosing text options for determination connection issue as a result of connection could be a single category downside . The efficient approach of feature choice for connection is predicated on a feature coefficient operate. A feature coefficient operate indicates the degree of knowledge painted by the feature occurrences in an exceedingly document and reflects the connection of the feature. the favored term-based ranking models embrace tf*idf based mostly techniques, Rocchio algorithmic rule, Probabilistic models and Okapia johnstoni BM25. Recently, one among the necessary problems for multimedia system information is that the identification of the best feature set with none redundancy but, the difficult issue for text feature choice in text documents is that the identification of that format or wherever the relevant options square measure in an exceedingly text document attributable to the massive quantity of buzzing data within the document . Text options may be straightforward structures (words), complex linguistic structures or applied mathematics structures. we tend to in the main discuss 3 advanced structures below for choosing relevant features: ngrams, concept sand patterns. In summary, the prevailing strategies for finding connection options may be classified into 3 approaches. The first approach tries to diminish weights of terms that seem in each relevant documents and extraneous documents (e.g., Rocchio-based models). This heuristic is clear if we tend to assume that terms square measure isolated atoms. The other is predicated on however usually options seem or don't seem in relevant and extraneous documents (e.g., probabilistic based mostly models or BM25). The third one is predicated on finding options through positive patterns . The projected model more develops the third approach by grouping options into 3 categories: "positive specific features", "general features", and "negative specific features".

In previous system solely the project get accepted directly. there's no analysis such the way that it are often found that the bound project is completed antecedently or not .The allocation of guide to student don't seem to be get mechanically in keeping with their connected domain. the general performance of project like poor, average and best don't seem to be live in exiting system.

3. PROPOSED SYSTEM

In this system the project abstract get accepted and so bound analysis perform thereon abstract for finding that the project is completed in antecedently. Techniques are often applied on abstract like TF-IDFs that is employed to count a specific count for matching and finding domain. during this technique solely the new topic ar accepted and repeatable topic get rejected. during this firstly student send abstract towards organization. Then the work start organization assesses the abstract. info retrieval techniques get applied on specific abstract and also the result's calculated that it's newer project or it's older one. Then result's calculated for abstract get accepted or rejected. Then in keeping with domain of project the guide ar get allotted for specific project. And finally the result can calculated on FTR base that's project is best, poor or average. And finally the feedback can send to student and student can ready to see their performance.

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. DISCUSSION

The planned model has 3 major steps: feature discovery and deploying, term classification, and term advisement. Wrongdoer choice plays a very important role for mistreatment feedback within the method of feature discovery and deploying. during this section, we have a tendency to first discuss the difficulty of wrongdoer choice. we have a tendency to additionally discuss alternative problems for the planned model like term classification and specificities. We believe that the feedback is a lot of constructive than the feedback since the target of relevancy feature discovery is to find relevant information. However, we have a tendency to believe that feedback contains some helpful data that may facilitate to spot the boundary between relevant and extraneous data for up the effectiveness of relevancy feature discovery. the plain downside for mistreatment extraneous documents is that the majority of the extraneous documents aren't closed to the given topic thanks to the terribly great amount of negative data. Therefore, it's needed to decide on some helpful extraneous documents (offenders) to determine the teams of terms for the 3 classes.

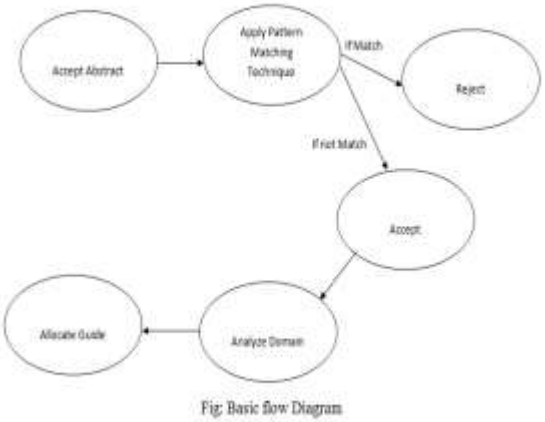

Fig 1: Basic Flow Diagram

5. SYSTEM ARCHITECTURE

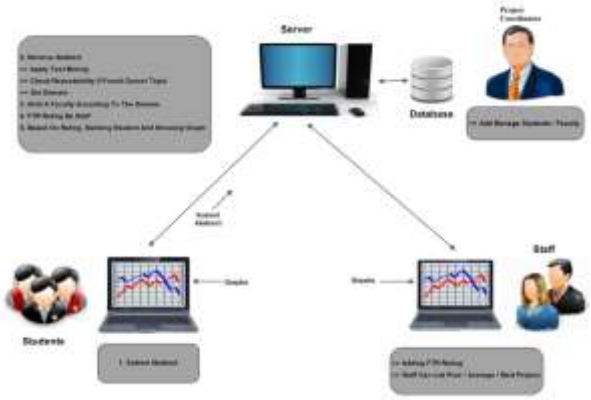

Figure 2: System Architecture

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.