International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 5, Issue 05, May-2019

College Recommendation with TF-IDF

Vaishnavi Nagose[1], Anjali Raut[2]

[1]Computer Science And Engineering, SGBAU,
[2]Computer Science And Engineering, SGBAU,

Abstract— ***Education is the most important factor for growth and development of any nation for that getting appropriate college is most important. Nowadays colleges are rapidly increasing so to attract students college can make fake publicity. To guide candidates to choose an appropriate college TF-IDF technique is used. The term TF-IDF refers to the term frequency-inverse document frequency which is used to calculate the score of the text taken in the form of comments. The main term comment classification is used with classification algorithm. The comment classification leads to the new term sentiment analysis and opinion mining. The opinions are taken in the form of comments.***

Keywords— Opinion Mining, Sentiment Analysis, Comment Classification, Text Mining, Recommendation System, Data Mining

I. INRODUCTION

Today's generation is getting fast due to the wide use of social networks which provides information very fast. The individuals and organization utilize these platforms to spread information. This leads to social networking as a major form of communication used daily. In the present situation, there are many college recommendation apps and websites which are based on different techniques and parameters. This college recommendation system is somewhat different from other systems. The systems which are previously developed are mainly focused on cut off, ratings, admission intake and publicity. In previous systems the tools and technology used for recommendation are different and as it also based on different parameters having their own efficiency and pitfalls. TF-IDF is the technique used in the previous system which calculates the TF-IDF score on the basis of comments and data is taken from social networking sites like Facebook. But when comes to college recommendation, it's a tedious task and candidate searching for college should get proper college according to interest. The system targets those candidates who need admission in college accordingly with proper guidance.

II. LITERATURE REVIEW

Text analysis focuses on comments analysis as a part of the user-generated text. The researcher has studied comments posted on pictures, blogs, announcements, and comments on a comment which require comment reference. The most important point in comment classification is sentiment analysis also known as opinion mining. The sentiment analysis is used to represent comment classification problems.

[1] The authors *F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh* proposed this recommendation system in which searching for information in a large amount of dynamically generated data which may cause the problem of information overload is solved to provide personalized content and services to the user. There are different roles and characteristics of different prediction techniques in a recommendation system which provides scope for research and practice in the recommendation system.

[2] *J. Gemmell, T. Schimoler, B. Mobasher, and R. Burke* proposed a Decision tree approach which is used for selection of students in any course program. Data mining is used with the optimally utilizing resources available to improve business intelligence process including education system to enhance efficiency.

[3] *ParneetKaur, Manpreet Singh, Gurpreet Singh Josan. Lakshmi M. Gadhikar, Deepa Vincent, Lavanya Mohan, and Megha V. Chaudhari*, discusses data mining techniques to process a dataset and identify the relevance of classification test data. It shows the process of WEKA analysis of file converts and the selection of attributes to be mined and comparison with Knowledge Extraction of Evolutionary Learning. It shows each Decision tree achieve a high rate of accuracy. It classifies the information into the properly and incorrectly instance.

[4] *Inbal Yahav, Onn Shehory, and David Schwartz* discuss the mining of comments in social networking. The technique TF-IDF is used with adjustment for the bias. Where the illustration is taken from various Facebook fan pages includes different domains. But using only this technique cannot give more accuracy in education domain.

[5] The field of text analysis focuses on the analysis of comments as an important part of the understanding user-generated online comment. Comments posted on pictures and YouTube videos, comments on blogs, comments on press releases, comments on corporate communication, comments on public service announcements, and even comments on comments are studied as a reference data.

[6],[8] Classification of comments studied vary from sarcasm and nastiness to attitude toward the comments misinformation can also spread such as rumors, information related to the document posted, previous comments, the uniqueness of comments.

[7] Comment classification, often denoted message classification in the statistical or rule-based natural language from a computational approach. Natural Language Processing (NLP) literature is the analysis of comments within a comment-sphere.

[10] The most common question in comment classification is sentiment analysis, also known as opinion mining. Sentiment analysis and opinion mining represent a large problem space, often defined slightly differently, covering, for example, opinion extraction, subjectivity study and emotion analysis. The aim of using sentiment analysis is to find out what people think or feel toward products, services, individuals, events, news articles, and topics. Practically the comment classification is achieved using content extraction and various classification techniques. In IS research, the term "sentiment analysis" is used to represent all the comment classification problems.

[12] In machine learning, data-driven analysis of user-generated content requires text pre-processing. This step commonly involves removal of stop words, word stemming or word lemmatization, and controlling for word frequency. In many kinds of literature on comment, classification is the analysis of comments as independent observations. The structure of the comments and their proximity are largely recognized as having a great impact on users opinion and sentiment and often used for comment classification, comments are still treated independently for the purpose of word frequency control.

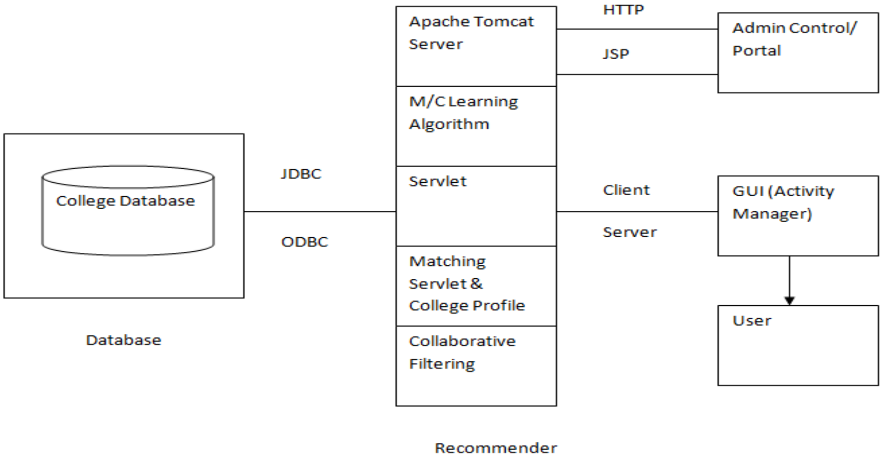

Fig. 1 Recommender System Layer Architecture

A. *Text Pre-Processing With TF-IDF*

The TF-IDF weighting stands for term frequency (TF) × inverse document frequency (IDF). TF-IDF weighting is commonly used in text mining and information retrieval to evaluate the important term in a studied corpus. Term importance (weight) increases with the term's frequency in the text. Given a collection of terms $t \in T$ that appear in a set of N documents $d \in D$, each of length n_d, TF-IDF weighting is computed as follows:

$tf_{t,d} = f_{t,d} / n_d$
$idf_t = \log(N/df_t)$
$W_{t,d} = tf_{t,d} \times idf_t$,
Where,
$f_{t,d}$: The frequency of term t in document d.
df_t: The document frequency of term t, that is, the number of documents in which term t appears.

For the task of comment classification, ‘document’ is replaced by ‘class’, e.g. sentiment class which is further divided as positive and negative in sentiment analysis. Term frequency (tf) is then computed per class. Inverse to document (IDF) becomes "inverse to comments", means that N is the size of the set of comments, and the document frequency for a term (df_t) is computed on that set.

Birmingham and *Smeaton* define this method as sentiment TF-IDF. Comments are then classified into classes using probabilistic (e.g., Naive Bayes) or discriminative models (e.g., SVMs). A common variant of the classic tf-idf is delta idf weighting, in which idf is calculated for each class separately, and then the difference between the values is used for sentiment classification. This variant is proved to be efficient for classification at the sentence level.

B. Comments Dependency Structures

Comments dependency can be classified in the following ways.

1) *Comment-to-document*

Comments which often discuss the document's content. Comments are given on document by different commenter's separately. The same sets of terms used by several commenters, commenting on the same article are observed.

2) *Comment-to-comment*

Comments are dependent on previous comments because they reply on them or influenced by them. This dependency is time-dependent. That is, once a term appears, it had more possibility to reappear in succeeding comments. This dependency possibly belongs to the thread of comments. It may have a smaller impact on comments which may thread further. As compared to flat discussion threaded comments exhibit higher correlation.

C. Comparison of Existing Techniques

Naive mathematician rule is predicated on chance and J48 rule is predicated on call tree. The KNN classify an object based on the majority class amongst its k nearest neighbours. The experiments results shown are about classification accuracy. The results in the paper on this dataset also show that the efficiency and accuracy of KNN are good. KNN achieves a high rate of accuracy. It classifies the data into the distance.

III. PROPOSED WORK

The proposed work includes a technique along with a classification algorithm to classify submitted comments into a particular category. In that the polarity detection of the comment which will guide fresher which college is a correct choice for them. As the classification algorithm will use in the proposed system the system will also work in case of ambiguous keywords. Along with classification, the sentiment analysis will also be used for DSS generation.

The system comprises of following modules:

1) Admin
2) College Admin
3) Students
4) Comments Classification
5) DSS Report Generation

IV. CONCLUSION

TF-IDF is a widely used technique for text pre-processing which also includes information retrieval and text mining. The text mining deals with the sentiments also known as an opinion. There are various techniques used for classification and analysis of comments with TF-IDF. The problems with existing systems can be solved by including the classification algorithm which is having better performance. The proposed system consists of five modules which show recommendation at different levels.

REFERENCES

[1] *F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh,* Recommendation systems: Principles, methods, and evaluation, Received 13 March 2015; revised 8 June 2015; accepted 30 June 2015.

[2] *J. Gemmell, T. Schimoler, B. Mobasher, and R. Burke*, “Recommendation by Example in Social Annotation Systems,” pp. 209–220, 2011.

[3] *Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan. Lakshmi M. Gadhikar, Deepa Vincent, Lavanya Mohan, and Megha V. Chaudhari* Classification and prediction primarily based data processing algorithms to predict slow learners in education sector rd International Conference on Recent Trends in Computing 2015(ICRTC-2015)

[4] Comments Mining With TF-IDF: The Inherent Bias and Its Removal *Inbal Yahav, Onn Shehory, and David Schwartz*. VOL. 14, NO. 8, AUGUST 2015

[5]*M.Potthast, B.Stein, F.Loose, and S.Becker,* "Information retrieval in the comment-sphere," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 4, p. 68, 2012.

[6] *D. Davidov, O. Tsur, and A. Rappoport,* "Semi-supervised recognition of sardonic sentences in twitter and Amazon," in Proceedings of the fourteenth conference on computational natural language learning. Association for Computational Linguistics, 2010, pp. 107–116.

[7] *D. Feng, J. Kim, E. Shaw, and E. Hovy*, "Towards modeling threaded discussions using induced ontology knowledge, "inProc. 21st National Conf. on Artificial Intelligence, vol. 21, no. 2. AAAI Press, 2006, pp. 1289–1294.

[8] *R. Justo, T. Corcoran, S. M. Lukin, M. Walker, and M. I. Torres*, "Extracting relevant information for the detection of witticism and nastiness within the social net," Knowledge-Based Systems, vol. 69, pp. 124–133, 2014.

[9] *A. Hassan, V. Qazvinian, and D. Radev*, "What's with the attitude?: distinguishing sentences with perspective in online discussions," in Proceedings of the 2010 Conference on Empirical strategies in tongue process. Association for Computational Linguistics, 2010, pp. 1245–1255.

[10] *T. Nasukawa and J. Yi*, "Sentiment analysis: Capturing favorability mistreatment tongue process," in Proceedings of the 2nd international conference on Knowledge capture. ACM, 2003, pp. 70– 77.

[11] *N. FitzGerald, G. Carenini, G. Murray, and S. Joty,* "Exploiting informal options to find high-quality diary comments," Advances in Artificial Intelligence, pp. 122–127, 2011.

[12] *M. Kantrowitz, B. Mohit, and V. Mittal*, "Stemming and its effects on tfidf ranking (poster session)," in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2000, pp. 357–359.

[13] *M. Toman, R. Tesar, and K. Jezek*, "Influence of word social control on text classification," Proc. InSciT, vol. 4, pp. 354–358, 2006.

[14] *M. Gamon,* "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," in Proc. 20th intl. conf. on Computational Linguistics. Association for Computational Linguistics, 2004, p. 841.

[15*] A. Kennedy and D. Inkpen*, "Sentiment classification of movie reviews using contextual valence shifters, "Computational intelligence, vol. 22, no. 2, pp. 110–125, 2006.

[16] *G. Salton and C.-S. Yang,* "On the specification of term values in automatic indexing," Journal of Documentation, vol. 29, no. 4, pp. 351–372, 1973.

[17] *C. Manning, P. Raghavan, and H. Sch¨utze, "*Language models for information retrieval," Introduction to Information Retrieval, pp. 237–252, 2008.

[18] *G. Paltoglou and M. Thelwall*, "A study of information retrieval weighting schemes for sentiment analysis," in Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010, pp. 1386–1395.

[19] *S. Robertson*, "Understanding inverse document frequency: on theoretical arguments for idf," Journal of Documentation, vol. 60, no. 5, pp. 503–520, 2004.

[20] *G. Salton and C. Buckley*, "Term-weighting approaches in automatic text retrieval," Information processing & management, vol. 24, no. 5, pp. 513–523, 1988.

[21] *A. Bermingham and A. F. Smeaton,* "On using Twitter to monitor political sentiment and predict election results, "SentimentAnalys is where AI meets Psychology (SAAIP), pp. 2–10, 2011.

[22] *J. Martineau and T. Finin*, "Delta tfidf: An improved feature space for sentiment analysis." Icwsm, vol. 9, p. 106, 2009.

[23] *Anshul Goyal and Rajni Mehta,* Performance Comparison of Naive Bayes and J48 Classification Algorithms, International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012).