CLASSIFICATION OF COURSE WITHDRAWAL REASONS AMONG COLLEGE OF TECHNOLOGY STUDENTS USING RULE BASED TEXT MINING ALGORITHM AND MODIFIED MAP REDUCE ALGORITHM ON TEXT FILES– A COMPARISON

SIVAKUMAR

Abstract— Mining raw text data to comprehend the sentiment and classify the students interest is the central objective of the paper. The students in the college of Technology, Oman use an online registration system for majority of student transactions. In this research we attempt to conduct a study on the raw text the students entered as 'reasons' for dropping or withdrawing a registered course. Since the database records the raw text data it is important that we discover the interest of the students and try to classify according to the dictionary of popular reasons of withdrawals.

Index Terms— classification, CIMS, data mining, Hadoop, HDFS, map-reduce, Rule based mining.

I. INTRODUCTION

The users of internet today is on the rise year by year and by april 2017 it was close to 4 billion and proportionally there is also the availability of massive amounts of text data along with data of other forms has been always an interesting platform for datamining research, particularly text-mining. Text Mining is a branch of data mining where the raw text or unstructured data is subjected to an analysis to discover or classify the data based on an interestingness measure. The explosive growth of data has given rise to a new branch of study called Bigdata and there are many analytical techniques and algorithms such as hadoop in practice to analyse and experiment the massive data sets.

In this paper … argue that the general data mining methods are also suitable to text mining tasks and in particular we provide a frame work for text classification using rule based mining and map-reduce algorithm on windows file system.(not on HDFS). The attempt of applying two different methodologies is to evaluate the efficiency of the proposed framework, by comparing the results produced by the language implementation.

Related Work

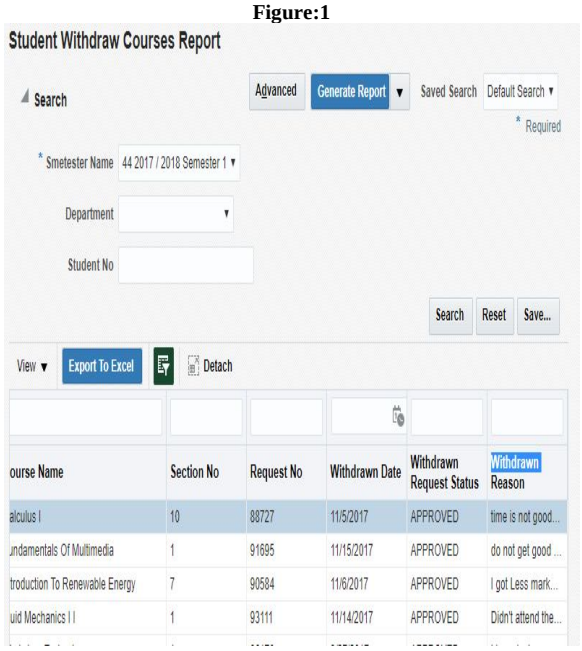
Text mining refers to the art of discovering subjective information from structured or unstructured original data using text-analytics, computational linguistics and natural language processing. Two methods are generally popular where one uses dictionaries and the other uses semantic dictionary either manually or semi automatically built. Pak and paroubek were able to build a classifier that determines positive, negative and neutral sentiments successfully for a document.

Using Hadoop eco system, many big data projects are also being done to discover interesting patterns and for classifications. It is relatively faster than the traditional databases as it has vertical and horizontal scalabilities. No SQL

The (CIMS) system that records reasons for withdrawals in the current system accepts text data. The verbatim extracted is not a quantified data and hence it is important that we use an algorithmic approach to discover any interesting information or summarized view or the hidden knowledge in-terms of the major factors that influences the decision to withdraw a course.

Following screen shots are actual display screens in the system where the withdrawal reports are present and can be downloaded as pdf or as excel sheets

Figure:1



Colleges of technologies in Oman use Colleges Information Management System (CIMS) application developed using Oracle fusion technology, to allow the staff and students to handle

student / staff transactions. The application has a big bunch of features to accomplish a variety of tasks related to Academic activities, Academic advising, Registration, timetabling, degree plans, program plans, attendance, many statistical data, finance related information, OJT, Analysis of results etc. From among this long list of different activities that is performed, students can use the system to drop / withdraw a course from the set of registered courses for the semester. Every student is entitled to drop / withdraw a course in a semester. The students can use their account in the CIMS and raise a request to their advisor to execute the withdrawal instruction in the system. The advisor may / may not approve the request after analyzing the consequences of dropping the course. The advisor can refuse to accept the request due to several reasons which is a reflection of the policies and decisions that the department / college observes on withdrawals.

The interesting fact in the request is that the students are supposed to record reasons for their withdrawal, as a text data, which is rather a large object than a choice from a limited set of reasons, students shall enter (type) anything that they think is the reason for withdrawals. This verbatim data can be seen and printed by all users of level advisors, teachers or above. This data is not quantified and it can be of various sizes and it is just the sentiment/state of mind of the student that expresses the reason for withdrawal. Since we have the data from all the students withdrawing, it would be interesting to analyze the reasons and classify

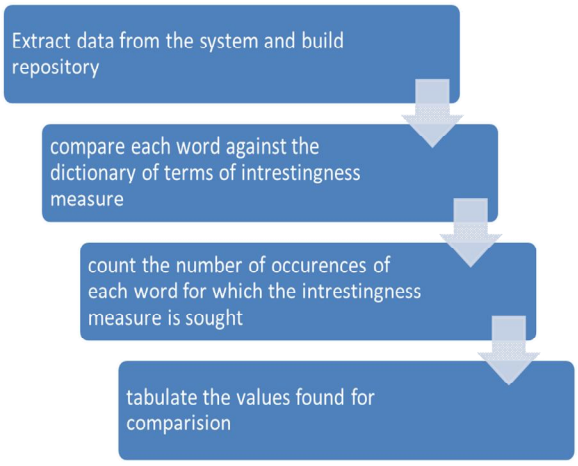
The study has been done on 659 student records from various levels and departments. Manual interpretation was performed initially to classify the reason under five popular elements that triggers withdrawals such as, Mid-Exam marks, Teaching, Timing, Transportation and for-Improvement. Students' inputs were preprocessed to check for spelling and autocorrected. But for experimental reasons the raw data was actually tried in the language compiler that implemented the algorithm and we used the 'stemming' approach in a simple way to find occurrences of data of classified interestingness.

Once the values are computed in each of the three methods namely manual, rule-based mining and map-reduce method, the accuracy is calculated according to the following formula

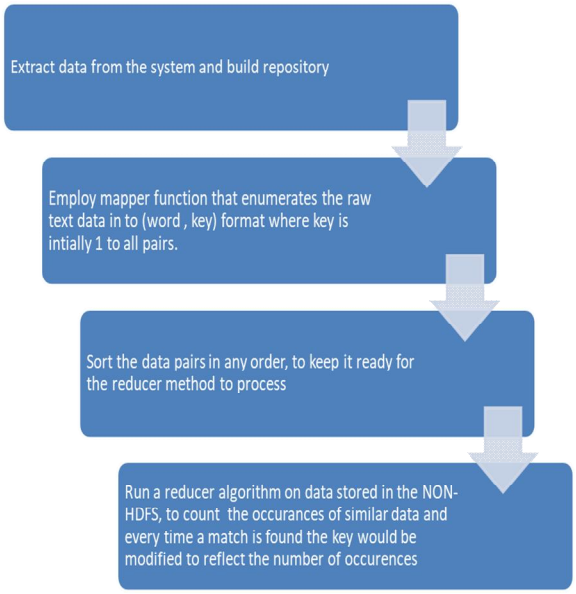
$$\textit{The accuracy index} = \frac{\text{The occurance of target factor in the proposed method}}{\text{The actual occurance of the factor}}$$

II. PROPOSED METHODOLOGY

Rule based Text mining Algorithm



Modified Map-reduce Algorithm for non-HDFS platforms



III. IMPLEMENTATION STEPS FOR RULE BASED TEXT MINING ALGORITHM USING A LANGUAGE COMPILER (FOR THIS RESEARCH PHP HAS BEEN USED FOR IT IS SIMPLE AND IT IS AN OPEN SOURCE)

Step1:
Extract the raw data from the CIMS as a CSV file.
Step2:
Build dictionary of terms of interestingness for classification
Step 3:
Repeat step 4 until End of File
Step 4: Read the CSV file
Split the contents word by word
Using rule based comparison increment counter when match found
Step5: Aggregate the results to display in a table
Figure:2

	Mid Exam	Teaching	Time	Transport	Improvement
Sum	556	16	75	2	8

Implementation steps modified Map-reduce Algorithm for non-HDFS platforms using a language compiler (for this research php has been used for it is simple and it is an open source)

Step-1:

Extract the students raw data from the CIMS as a csv file

Step 2:

Build a map that produces a 'name–value' pair from the csv file where name is the word and the value is an integer 1 for all words.

Step 3:

Sort this name-value pairs in any order

Step 4:

Make a copy of the list of 'name–value' pairs without duplicates for the use of the reducer procedure

Step4:

Count the frequently repeated pattern of identical name-value pairs and reduce them to one 'name–value' pair where name is unchanged and the value shows the frequency of the pattern as an integer

Step 5:

When step 4 is completed the reducer procedure outputs a list of 'name–value' pairs, examining which against the patterns of interesting data we filter the result to get the needed output.

Figure:3

(mid,1)	158
(Mid,1)	158
(MID,1)	158
(mark,1)	141
(MARK,1)	141

Figure: 4

Mid Exam	Teaching	Time	Transport	Improvement
496	15	85	1	1

The results of the two algorithmic approaches are calculated using the php program implemented. The actual results are displayed in the above figures namely figure-2 for the rule based text classification, figure-4 for the map-reduce approach for text classification the implementation results were not exactly same. The figure :3 above gives an idea of the reducer cycle of the map reduce approach, where the occurrences are displayed in a separate column for easy understating.

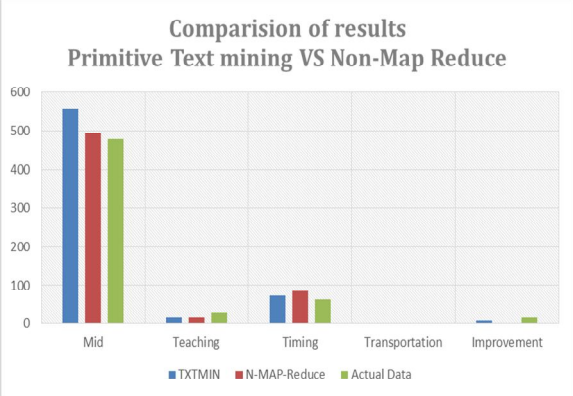
IV. EVALUATION

After the results are produced using the two different algorithmic approaches, to evaluate the model proposed, the accuracy of the implementation results were calculated and the results are given below in figure: 5 and table:1

Table: 1

				Accuracy in %	
	TXTMIN	N-MAP-Reduce	Actual Data	TXTMIN	N-MAP-Reduce
Mid	557	496	479	86	96.57
Teaching	16	15	28	57.14	53.57
Timing	74	85	64	86.49	75.29
Transportation	2	1	1	50	100
Improvement	8	1	15	53.33	6.667
	Average of accuracy			66.59	66.42

Figure : 5



CONCLUSION

It is discovered that the two methods proposed are almost same in terms of accuracy and produced reliable results. There are few problems in the raw data as there were lots of noise, since students entered data with lots of spelling mistakes and to overcome this the reducer procedure finds the match even if the initial few letters are matching. Another way to solve this issue is to cleanse the data using an auto spell check.

ACKNOWLEDGMENT

Sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page, not here.

REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] Abbott, R.D., Wulff, D.H., Nyquist, J.D., Ropp, V.A. & Hess, C.W. (1990). Satisfaction with processes of collecting student

opinions about instruction: The student perspective. Journal of Educational Psychology, 82, 201–206.

[3] Ah-Hwee Tan, 1999. Text mining: The state of the art and the challenges. In Proceedings PAKDD'99 Workshopon Knowledge Discovery from Advanced Databases (KDAD'99), pages 71–76, 1999.

[4] Biggs, J. (2003). Teaching for quality learning at university: What the student does. Buckingham, UK: SRHE and Open University Press.

[5] Bowden, J., & Marton, F. (1998). The university of learning: Beyond quality and competence in higher education. London: Kogan Page.

[6] Chen, Y., & Hoshower, L.B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. Assessment & Evaluation in Higher Education, 28(1), 71–88.

[7] Cohen KB, Hunter L (2008) Getting started in text mining. PLoS Comput Biol 4(1): e20. doi:10. 1371/journal.pcbi.0040020

[8] Coussement K., Vandenpoel ,D. (2008) Improving Customer Complaint Management by

[9] Automatic Email Classification Using Linguistic Style Features as Predictors, Decision Support Systems, 44(4), 870-882

[10] Delgado, M., M.J. Martín-Bautista, D. Sánchez, M.A. Vila, "Mining Text Data: Special Features and Patterns". In Proc. of EPS Exploratory Workshop on Pattern Detection and Discovery in Data Mining, London, September 2002. Lecture Notes in Computer Science 2447, D. Hand et.al., Eds., pp. 140–153. Springer-Verlag.

[11] Himmel W., Reincke, U., Michelmann H. W. (2009) Text mining and Natural language Processing Approaches for automatic categorization of lay requests to web-based expert forums. Journal of Medical Internet Research, 11(3):e25.

[12] Hounsell, D. (2003). The evaluation of teaching. In H. Fry, S. Ketteridge & S. Marshall (Eds.), A handbook for teaching and learning in higher education: Enhancing academic practice (pp. 200–212). London: Kogan Page

[13] Jackson P, Moulinier I (2007) Natural language processing for online applications: Text retrieval, extraction, and classification. 2nd edition. Herndon (Virginia): John Benjamins Publishing Company.

[14] Kember, D., Leung, D. Y. P., & Kwan, K. P. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? Assessment and Evaluation in Higher Education, 27(5), 411–425.

[15] Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues and directions for future research. International Journal of Educational Research, 11(2), 253–388.

[16] McKeachie, W.J. (1997). Student ratings the validity of use. American Psychologist, 52(11),1218–1225.

[17] Ramsden, P. (2003). Learning to teach in higher education (2nd ed.). London: Routledge Falmer, p. 223.

★★★