12-2019

Clustering models for topic analysis in graduate discussion forums

Mallika GOKARN NITIN
Singapore Management University, mallikang.2015@sis.smu.edu.sg

Swapna GOTTIPATI
Singapore Management University, SWAPNAG@smu.edu.sg

Venky SHANKARARAMAN
Singapore Management University, venks@smu.edu.sg

Citation

GOKARN NITIN, Mallika; GOTTIPATI, Swapna; and SHANKARARAMAN, Venky. Clustering models for topic analysis in graduate discussion forums. (2019). *Proceedings of the 27th International Conference on Computers in Education: Taiwan*. 1-10.
Available at: https://ink.library.smu.edu.sg/sis_research/4516

Shih, J. L. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Clustering Models for Topic Analysis in Graduate Discussion Forums

Gokarn Mallika NITIN, Swapna GOTTIPATI*, Venky SHANKARARAMAN
School of Information Systems, Singapore Management University, Singapore
mallikag.2015,*swapnag,venky@smu.edu.sg

Abstract: Discussion forums provide the base content for creating a knowledge repository. It contains discussion threads related to key course topics that are debated by the students. In order to better understand the student learning experience, the instructor needs to analyse these discussion threads. This paper proposes the use of clustering models and interactive visualizations to conduct a qualitative analysis of graduate discussion forums. Our goal is to identify the sub-topics and topic evolutions in the discussion forums by applying text mining techniques. Our approach generates insights into the topic analysis in the forums and discovers the students' cognitive understanding within and beyond the classroom learning settings. We developed the analysis model and conducted our experiments on a graduate course in Information Systems. The results show that the proposed techniques are useful in discovering knowledge from the forums and generating user-friendly visualizations. Such results can be used by the faculty to analyse the students' discussions and study the strengths and weaknesses of the students' cognitive knowledge on course topics.

Keywords: Topic analysis, Online Discussion Forums, Clustering Models, Topic evolutions

1. Introduction

Online discussion forums are advantageous as they provide an equitable space and flexibility for all students. They offer an alternative means for encouraging interactions between the students and the instructors. Discussion forums are classified into three types; standard forum for general use, single simple discussion forum and question and answering forum. General use forums are useful for any random discussions posted by any user with or without the goal of testing or learning more about the topics covered in the class. Simple discussion forum is focused on a single topic or a subject. Question and answer forums are focused on several diversified topics and aimed at supporting the learning process. Most of the student forums are usually designed as Q&A forums (Andy, 2013) where the instructor posts a questions and the students submit their answers.

To measure the students' learning, instructors always want to know what the students are discussing. For effective learning through discussions, it is important for the instructors to intervene in the threads (Simon et al. 2016). Effective intervention requires an understanding of the content and analysis of the forums (Chaturvedi, 2014). Use of analytics on discussion forums is focused on collecting, analysing, and displaying the "traces" that learners leave behind, with a purpose to improve learning (Atapattu, Falkner et al. 2016). For example, if the students are discussing less on certain topic, instructors can post additional supporting hints for the students to continue the discussions. Emergent topics or declining topics require more instructor intervention. Another example is that if the students are digressing from the topic, instructors can control the discussions or encourage further learning points in the digressed topics.

In this paper, we propose a solution for discovering the topical insights from the discussion forums based on a text analytics approach. In particular, understanding the topics and sub-topics that are emerging in the discussions provide useful insights into the student learning process. For example, if the students have discussed only the main topics that were covered in class, it indicates that the students are bounded to in-class learning and have not taken efforts to do further research on their own. If more sub-topics, that were not covered in the class, emerge from the main topic, it indicates the out of class learning process of the students. In this digital era, it is important for students to learn beyond the classroom, and further scaffold this learning, by instructors intervening to identify the links between the various sub-topics and providing a summary of the topical evolutions. However, a manual approach to this is very time-consuming, since the instructor has to read all the posts and generate topics, sub-topics and the evolutions. Using automated tools to help gain insights from the discussion forum posts holds great promise for providing adaptive support to individual students and collaborative groups.

We use data from the online discussion forum of a masters course, "Text Analytics and Applications" taught at the School of Information Systems, Singapore Management University. A study by Burge (1994) of Master of Education students enrolled in a web-based distance program identified challenges that related to peer interaction, difficulties associated with handling and managing large quantities of information and discussion fragmentation. Therefore, we design the discussion forum with the controlled and challenging threads, so that the students can appreciate and participate in the organized discussions. The knowledge generated from such posts can be applied to their project and exam preparations.

This paper is structured as follows. Section 2 will review the background of discussion forum analysis and topics analysis of posts. Section 3 provides a background of text analytics. Section 4 describes the research problem statement along with the context. Section 5 presents the solution models with details of the text analytics techniques that are used in the models. In section 6, we describe findings, analysis and answer our research questions, and we conclude in Section 7.

2. Related Work

2.1 Discussion Forum Analysis

Several researchers have studied interactions in the classroom since the 1960s to help quantify verbal behaviour. Applications of interaction analysis include improvement of teaching style and pupil achievement through reflection, using the classification of interaction type (Amidon, 1968). Additionally, an adapted form of Flanders' system of Interaction Analysis was used to understand and provide feedback on teaching behaviour in a foreign language classroom, to support future classroom planning and improve content delivery (Wragg, 1970). Lively online discussions can be facilitated by requiring participants to not only post their work but also comment and respond to each other's submissions. Classroom live discussions capturing and analysing is also an important research work for better learning process (Venky et al. 2018). As a result, the discussions become more than just an assignment; students learn from each other and become more engaged in the learning process.

Learning analytics is focused on collecting, analysing, and displaying the "traces" that learners leave behind, with a purpose to improve learning (Duval, 2011). The system developed by Leony et al. (2012) captures and visualizes the events of learning through the use of a dashboard which serves as a presentation layer to display important analytics insights. Lisa Lobry (2004) defined social, cognitive, and system responses that can be identified in the student postings. Scholars have argued that, theoretically, asynchronous discussion forums should be able to improve learning outcomes because of their unique technological affordances. As Allen et al. (2013) noted, because of the asynchronous nature of the technology, the very nature and method of discussions are different. "This means that students can think, edit, research, and post on a topic, even a couple of days after the original post".

2.2 Topic Analysis in Online Discussion Forums (ODFs)

Topics are latent and embedded in the textual data. There are multiple methods for topic discovery, including the Hierarchical Dirichlet Process that was used by Ma et al. (2016) to model topic evolution in online news articles published by Reuters. Latent Dirichlet Allocation model used by Feng Jian et al. (2018) to automatically extract topics from different time slices and thereby extract the evolutionary relationship among sub-topics in Microblogs. Ezen-Can et al. (2015) used clustering techniques to group discussion topics. They proposed the k-medoids algorithm and defined each forum post as a data point. To determine the number of clusters, they rely on the Bayesian Information Criterion. Atapattu, Falkner et al. (2016) proposed the ideas of using topic-wise classification of discussion threads on MOOC discussion forums. This work facilitates the instructors to locate and navigate the most influential topic clusters as well as the discussions that require intervention by connecting the topics with the corresponding weekly lectures. Li et al. (2016) proposed keyGraph (Ohsawa, 1998), an advanced mining technique, that helps to assess the students' knowledge discovery process and aids the instructors to create a new approach for transformative learning and teaching in education. Coffrin et al. (2015) proposed visualization techniques to understand the patterns of the students' engagement in MOOC discussions. Similar work by Vytasek et al. (2017) used topic modeling approach to discover topics and sub-topics from the entire discussion forum. Our goal is similar but the forum is designed with more organized Q&A forums format. Further, we also emphasized on the topics evolutions and visualizations in our work..

In our project, we adopt similar techniques from Ezen-Can et al. (2015). We make use of the k-means clustering algorithm to group documents or contributions during the discussions based on cosine similarity and TF-IDF vector space. Each document is assigned to a single cluster, while documents from the time slice (in our experiment - weeks), can represent multiple clusters. Further, we propose techniques for extracting the sub-topics and discovering the evolution of the topics which is a new contribution to the research area of analysing discussion forums.

3. Background

The goal of text analytics is to extract high-quality information from collections of documents. Text Mining and Natural Language Processing techniques are useful for data processing and discovering useful patterns. Text Analytics techniques comprise of multidisciplinary fields like information retrieval, extraction, natural language processing, and text mining. Some of the issues that should be considered during text mining are tokenization, stop word list, lemmatization, etc. A brief description of the key components and techniques used in our solution follows.

Tokenization: Tokenization is a common text data pre-processing step that deals with the splitting of data into smaller units. These units can be paragraphs, sentences, phrases of n-grams (n number of words), and single words. Every dataset might have a different delimiter used to make the distinction between these units. Some common delimiters include commas, semicolons, tabs, new line characters and space.

Stop Word Removal: When doing text mining, many of the frequently used words in English are useless and add "noise" to the document. Words such as "could", "and", "if", "the", etc. that are classified as pronouns, conjunctions and prepositions, do not add value or carry information of importance to the model. Therefore, these words are referred to as stop words and are removed at the pre-processing stage.

Stemming: Stemming is the process of retrieving the stem or root form of a word in a heuristic approach, in the hopes of achieving the common base form or root form of the word. For the purpose of our application, words are stemmed using the Porter Stemming algorithm proposed by Porter (1980).

Lemmatization: Lemmatization although similar to stemming in that the ultimate goal is to retrieve the base form of a word, it is more complex than stemming because it requires Parts of Speech categorization of words before lemmatization to retrieve the lemma or the canonical form of the word by removing the inflectional suffix or prefix only. The lemmatizer used is based on the WordNet Database. In our preliminary experiments, we observed that lemmatization is the best pre-processing step for our results. Therefore, our focus will be on using the dataset cleaned with the help of the WordNet Lemmatizer (Christiane).

TF-IDF Document Representation: For any statistical computations on text data such as similarity scoring, a vector space representation of the text data is required. This representation consists of each document being evaluated as a term-frequency (TF) and inverse document frequency (IDF) weighted vector. TF-IDF is a statistical weight that ensures the rarer the term, the higher the weight of the score, by checking the frequency of occurrence of the term in a document (TF) as compared to how significant that term is with respect to the whole corpus or collection(IDF). Both these measures, in our solution, aid in generating the aspects or topics from the discussions. One way to combine a word's term frequency and inverse document frequency into a single weight is a TF-IDF. Each document in the dataset is then represented as a document-term matrix. For a term *i* in document *j*, the TF-IDF representation can be written as:

$$w_{i,j} = tf_{i,j} \times \log \frac{N}{df_i}$$

where $tf_{i,j}$ is the number of occurrences of *i* in j, df_i is the number of documents containing i, and N is the total number of documents.

Document similarity score: The similarity score between two documents determines the co-occurrence of a primary topic in the two documents. We compute this score by computing the cosine angle between two documents which are modelled as vectors in a vector space suggested by Christopher Manning (2008).

K-means clustering: Clustering algorithms are exploratory data analysis tools that have proved to be essential for gaining valuable insights on various aspects and relationships of the underlying textual data. Clustering algorithms are used to find groups of similar objects in the data. The k-means clustering algorithm finds its clusters by initializing with k set of seeds to which each of the documents in the corpus are assigned (one to each) on the bases of closest similarity, in our case the cosine similarity. In the subsequent iterations, until convergence, the new seeds are defined by the cluster centroids of the current clusters derived (Aggarwal & Zhai, 2012).

Agglomerative Clustering: Agglomerative algorithms find the clusters by initially assigning each object to its own cluster and then repeatedly merging pairs of clusters until either the desired number of clusters has been obtained or all the objects have been merged into a single cluster leading to a complete agglomerative tree (Christopher Manning, 2008). The key step in these algorithms is the method, also referred to as clustering function, used to identify pairs of clusters to be merged iteratively.

4. Research Problem and Methodology

Research Problem

Recall that our motivation for this paper is to identify the sub-topics and the evolution of topics within the discussion forum. We study the two research questions.

RQ1: How the clustering technique performs in discovering sub-topics?

RQ2: Which visualizations are suitable for sub-topic and topic evolution representation?

Online Forum Settings

As part of the text analytics course for graduate students, the instructor designed a weekly question and answering forum. The information systems graduate courses are business-IT courses. We observed that the student were reluctant to use the discussion forum if it was a mere repetition of the course content. Therefore, we had to come up with a design, where questions were asked that promoted the student to do research and then participate in the forum. Moreover, in our experience, we also observed that the questions that related to topics beyond the class, were found to be more interesting and motivated the students to be active participants in the forum. Table 1 shows the discussion forum settings for the course. Note that the questions are a mix of business and technical aspects which align with the course objectives.

Table 1: Weekly topic and the related questions. Underlined phrases show the main topics.

Week	Discussion Forum Thread
0	General discussions General discussions including concepts, labs, class etc.
1	Text Mining Introduction What are applications of Text mining in education domain?
2	Text pre-processing and NLP How search engines (Bing or Google) use NLP? What are examples of applications of chatbots in different industries?
3	Document Similarity Explain the differences between the bag of words & vector space model.
4	Text Classification What are examples of text classification in the industry (Government, healthcare, banks, etc.)? What are various evaluation measures for text classification?
5	Text Clustering What are visuals for the displaying cluster results - Free draw and upload? Explain one clustering evaluation measure with an example.
6	Information Extraction What are applications of HMM models (Or any other Sequence Model)? What are examples of information Extraction in Industry (e.g. Finance, Retail, Travel, Healthcare, Media, Education etc.)

9	Sentiment analysis Discuss the technique to handle negation in opinions. Discuss technique to handle sarcasm in opinions. Discuss the technique to handle suggestions in opinions.

Participants

Out of the 55 students enrolled in the course, 37 students participated in the discussion forums. More than 50% of the students have past industry experience or were currently working in the industry.

5. Solution Design

The data from the discussion forums can be noisy and requires significant cleaning before we apply the mining techniques. To generate the topics, commonly used technique is LDA topic model. However, the limitation with this technique is the requirement of large datasets for better performance. Therefore, we took the clustering approach that also aids in the identifying the topics of discussions. With the several clustering techniques available from the machine learning research, we propose to study two popular models and evaluate them on their performance for discussion forum analysis. Finally, choosing visualizations is based on the previous research on the topic evolution and social network studies. Figure 1 depicts the solution overview showing the four stages; data processing, clustering, evaluations and visualizations.

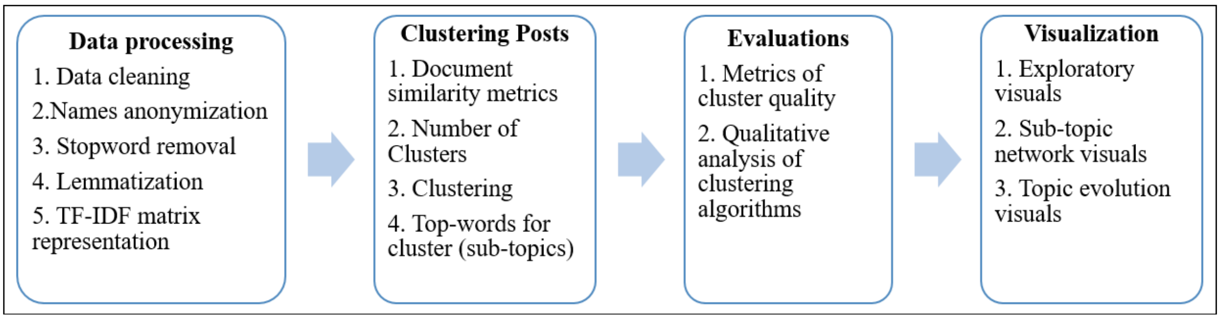

Data processing
1. Data cleaning
2.Names anonymization
3. Stopword removal
4. Lemmatization
5. TF-IDF matrix representation
Clustering Posts
1. Document similarity metrics
2. Number of Clusters
3. Clustering
4. Top-words for cluster (sub-topics)
Evaluations
1. Metrics of cluster quality
2. Qualitative analysis of clustering algorithms
Visualization
1. Exploratory visuals
2. Sub-topic network visuals
3. Topic evolution visuals

Figure 1: Solution model for topical analysis in discussion forums

Data processing: The first stage begins with the data preparation for representing each post as a matrix that becomes the input for the clustering algorithms. The posts consist of noisy data such as dates of posting, names, etc. We remove such data by regular expression method. Each discussion post is then represented as a TF-IDF matrix after the processing of stopword removal and lemmatization.

Clustering Posts: We provide the choice of two clustering algorithms for the users; k-means and agglomerative clustering. Document similarity scores for k-means algorithm from scikit learn is based on Euclidean distance (Gavin, 2017). The objective function is to minimize the within cluster sum of squares between the documents and centroids. In case of agglomerative clustering, the tool supports l1 norm (cityblock distance) and l2 norm (Euclidean distance). In our preliminary analysis, the cosine similarity performance has not performed well. Hence, even though the user interface is developed with the choice of different metrics, in this paper, we focus on Euclidean distance metric. The challenge of choosing the best number of clusters can be estimated using the elbow method. For each k value, initialise k-means and use the inertia attribute to identify the sum of squared distances of samples to the nearest cluster centre. As k increases, the sum of squared distance tends to zero. Plot sum of squared distances for k in the range specified. If the plot looks like an arm, then the elbow on the arm is optimal k. The final part is labelling the clusters. Since, clustering is unsupervised learning, to label each cluster, we use the high frequent words to label the cluster. If the top words are coherent, qualitatively, the cluster is of better quality, indicating the performance of the clustering algorithm.

Evaluations: The quantitative analysis for clusters can be based on true labels or the non-human clustering performance metrics such as Calinski and Harabaz score (Yanchi, 2010). We also perform the qualitative analysis by comparing both the clustering methods based on the top words representing the clusters. The clusters with coherence and non-repetitive representative words are considered to be of high-quality clusters.

Visualizations: For exploratory analysis to study the statistics of student's discussion, pie chart or bar chart are the best choice. To represent the topics and sub-topics, the network-based graphs are suitable (Aric et al. 2008). They are the advanced charts which are not only user-friendly but also suitable to represent the hierarchies and relationships. Finally, for the topic evolution visualization, we propose interactive line graphs, which are user-friendly charts with hovering feature.

6. Experiments and Findings

To examine our research questions from a more objective standpoint, we used exploratory analysis to study the general statistics on the student participations on various topics. For answering our RQ1, we conduct clustering evaluation as discussed in Section 5, and for answering RQ2, we develop visuals using python networkX and mathplot packages (Aric et al. 2008).

6.1 Exploratory analysis results

Figure 2 shows the distribution of discussions by topics posted by the instructor. The figure shows the proportions of posts over the main topics described in Table 1.

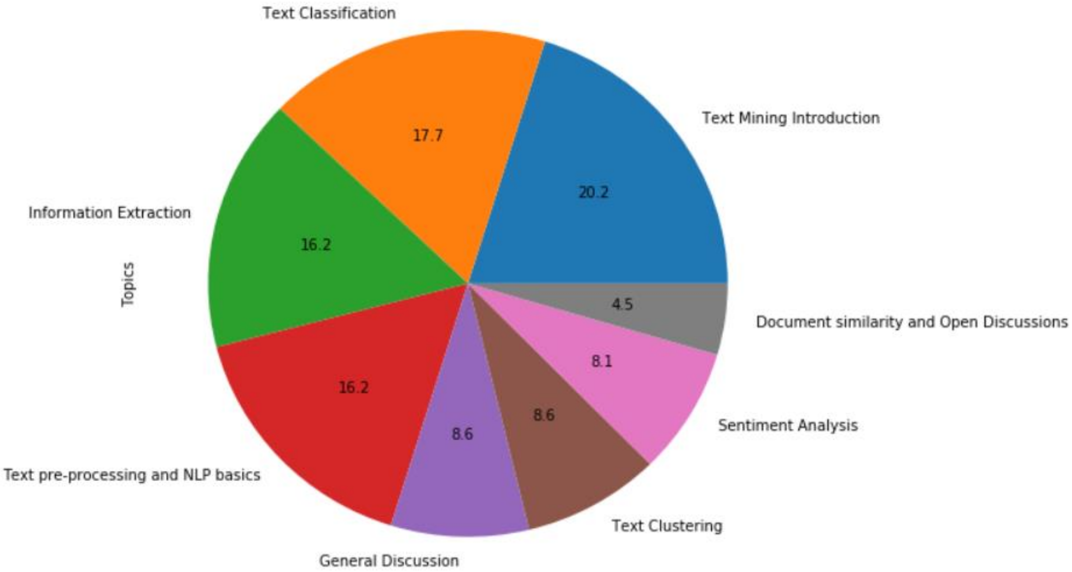


Figure 2: Overall participation in topics.

We observe that the first four topics have similar contributions more than 16% and the last four topics are around 8% except for the topic, "document similarity" which is lower than 5%. This is the most challenging technical topic during the early weeks of the course, and hence, the number of posts is lower. From the perspective of the type of questions, the different types of questions, namely understanding, analysis and discuss, received a similar number of posts from the students. The average number of words per participant is 676, which is quite high, thus indicating students' interest in proving detailed explanations.

6.2 Clustering Evaluations

As described in our solution, the k-means clustering and agglomerative clustering are used to cluster the posts. Major limitations with k-means the selection of the number of clusters and randomness in the clusters. On the bright side, it can be implemented easily and on well segregated data, the clusters can be very coherent. To define the number of clusters, we employed the elbow method with the k-means algorithm. The results show that the best number for clusters can be around 8 or 20. Since we are studying the sub-topics, we choose the bigger number, 20.

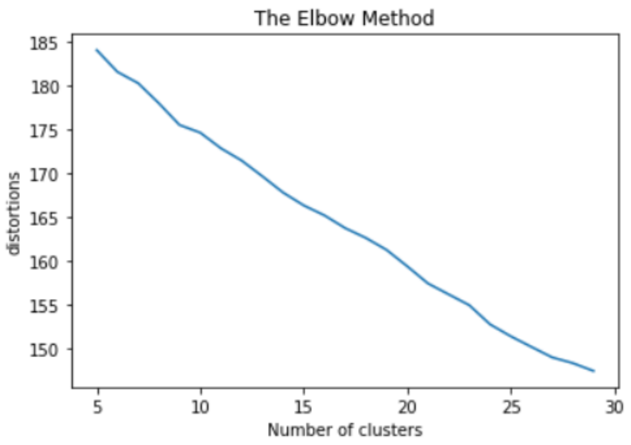


Figure 3: Elbow plot for k-means algorithm based on inertia.

Once the number of clusters is determined using the elbow method, we set the parameters for both clustering algorithms as below.

1. K-means settings: number of clusters is 20, initialization method is k-means++, number of time the k-means algorithm will be run with different centroid seeds is 20, and iterations is set to 500.

Agglomerative Cluster **Top words in each cluster**	***C*oherent/** ***R*epetitive**	**K-means** **Top words in each cluster**	***C*oherent/** ***R*epetitive**
['machine', 'tweet', 'comments', 'replying', 'google']	C	['machine', 'training', 'model', 'dataset', 'data']	C
['discuss', 'technique/s', 'opinions', 'handle', 'cluster']	C	['doubt', 'clarified', 'arun', 'comment', 'mislead']	R
['sentiment', 'negation', 'sarcasm', 'analysis', 'expressions']	C	['customer', 'chatbots', 'service', 'customers', 'chatbot']	C
['text', 'mining', 'classification', 'government', 'precision']	C	['doubt', 'clarified', 'arun', 'comment', 'mislead']	R
['in-class', 'question', 'thanks', 'students', 'compilation']	C	['sentiment', 'negation', 'analysis', 'scope', 'polarity']	C
['chatbots', 'customer', 'service', 'questions', 'banks']	C	['government', 'sites', 'classification', 'banks', 'examples']	R
['data', 'plagiarism', 'clinical', 'manually', 'website']	C	['patient', 'text', 'healthcare', 'doctors', 'analytics']	C
['doctors', 'symptoms', 'flu', 'doctor', 'analytics']	C	['words', 'slide', 'word', 'document', 'general']	C
['sequence', 'models', 'applications', 'state', 'model']	C	['doctor', 'medical', 'records', 'hospital', 'doctors']	C
['automated', 'model', 'learning', 'features', 'selection']	C	['patient', 'text', 'healthcare', 'doctors', 'analytics']	C
['doc', 'list', 'documents', 'docs', 'words']	C	['text', 'mining', 'students', 'education', 'analytics']	R
['words', 'training', 'dictionary', 'data', 'document']	C	['text', 'mining', 'students', 'education', 'analytics']	R
['extraction', 'finance', 'metadata', 'industry', 'used']	C	['text', 'mining', 'students', 'education', 'analytics']	R
['search', 'google', 'nlp', 'words', 'bing']	C	['google', 'search', 'nlp', 'engines', 'bing']	C
['news', 'readers', 'articles', 'organizations', 'travel']	C	['customer', 'chatbots', 'service', 'customers', 'chatbot']	R
['evaluation', 'clusters', 'measures', 'clustering', 'various']	C	['doc', 'documents', 'list', 'docs', 'w.lower']	C
['medical', 'records', 'problem', 'institutions', 'doctor']	C	['government', 'sites', 'classification', 'banks', 'examples']	R

['feedback', 'increase', 'customer', 'teacher', 'sort']	C	['text', 'mining', 'students', 'education', 'analytics']	R
['chat', 'robots', 'chatbot', 'technology', 'customer']	C	['named', 'article', 'entity', 'recognition', 'articles']	C
['events', 'recognition', 'drug', 'speech', 'gait']	C	['discuss', 'sarcasm', 'opinions', 'technique/s', 'handle']	C

1. Agglomerative Settings: number of clusters is 20, metric used to compute the linkage is “l2”, and linkage criterion for distance calculation is set to “complete”, the maximum distance between all observations.

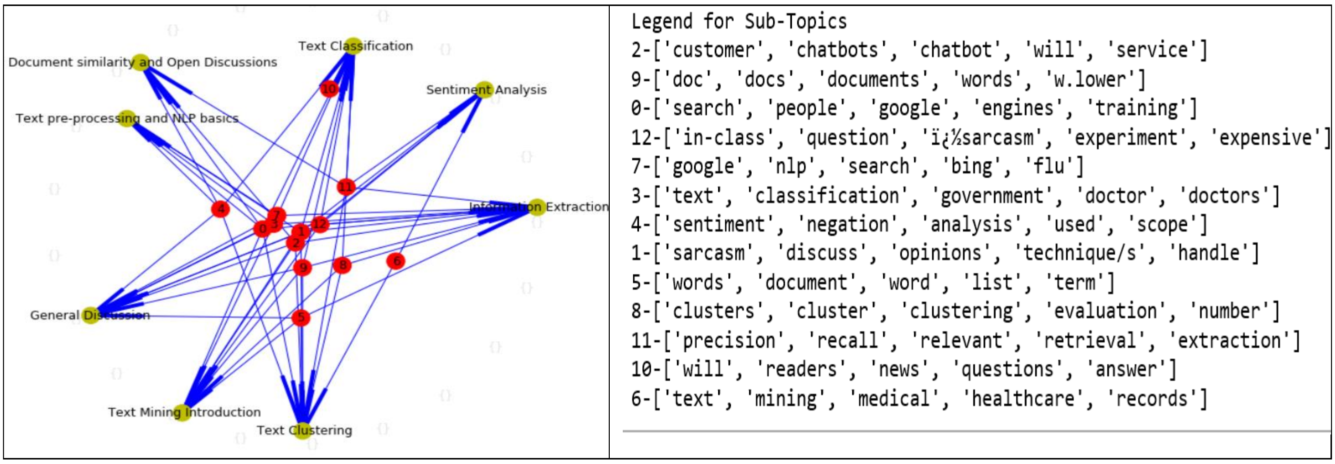
The representation for each cluster is the list of high-frequency words. To evaluate the clustering outcomes from k-means and agglomerative, we compare the clusters using qualitative analysis, as shown in Table 2. If the top words are coherent and non-repetitive, we label as ‘C’, and repetitive clusters are labeled as ‘R’.

Table 2: Qualitative analysis of agglomerative and k-means clustering. C represent coherently and R represents repetitive.

From the table, it is evident that agglomerative clustering has out performed k-means clustering. Hence, for final solution design evaluations, we use agglomerative clustering. This answers our first research question, RQ1.

6.3 Visualizations

Figure 4 shows the network graph for topics and sub-topics visualization. The topics are represented by yellow circles and the sub-topics are represented by red circle. The legend for the sub-topics is shown to the right. For the purpose of this paper, we are showing only 12 sub-topics instead of 20 for simplicity.



```
Legend for Sub-Topics
2-['customer', 'chatbots', 'chatbot', 'will', 'service']
9-['doc', 'docs', 'documents', 'words', 'w.lower']
0-['search', 'people', 'google', 'engines', 'training']
12-['in-class', 'question', 'ï¿½sarcasm', 'experiment', 'expensive']
7-['google', 'nlp', 'search', 'bing', 'flu']
3-['text', 'classification', 'government', 'doctor', 'doctors']
4-['sentiment', 'negation', 'analysis', 'used', 'scope']
1-['sarcasm', 'discuss', 'opinions', 'technique/s', 'handle']
5-['words', 'document', 'word', 'list', 'term']
8-['clusters', 'cluster', 'clustering', 'evaluation', 'number']
11-['precision', 'recall', 'relevant', 'retrieval', 'extraction']
10-['will', 'readers', 'news', 'questions', 'answer']
6-['text', 'mining', 'medical', 'healthcare', 'records']
```

Figure 4: Topics and Sub-topics network graph. The legend is displayed to show the sub-topics, red circles.

From Figure 4 we observe that sub-topics can be part of more than one main topic. For example topic, 6 - “medical and healthcare”, appears under “text classification”, “text mining introduction”, and “clustering”. This shows how students are connecting the sub-topics over various topics via the discussion posts. From such graphs, the instructor can identify the missing sub-topics and submit the posts under the main topic to lead the students in the learning process.

Figure 5 shows the sub-topic evolution over time. It depicts topic evolution over the weeks, given the percentage makeup of each week. A threshold value of minimum percentage makeup can be set using the slider below the graph. The interactive nature of the graph also enables the users to study each topic in detail and aid the instructor to decide on the need for intervention if the student misses the sub-topics. This chart is based on the results from the 12 clusters shown in Figure 4.

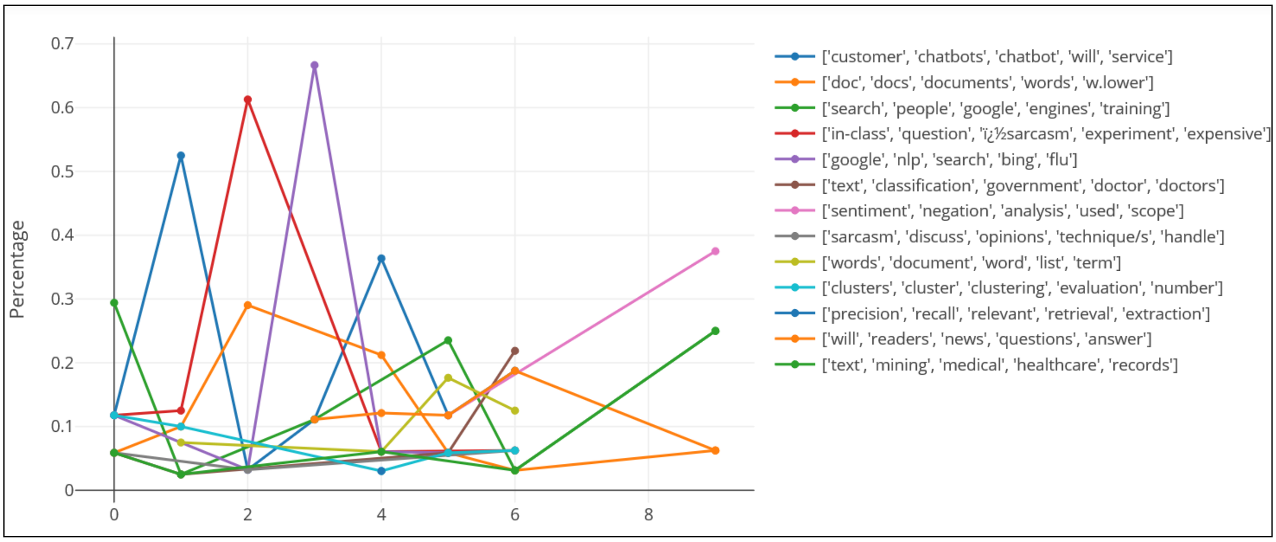


Figure 5: Sub-Topic Evolution over the weeks. Interactive graph with hovering features.

We observe some interesting patterns in this graph in terms of the short-lived vs repeated topics. Topics on “chatbot” and “healthcare” have occurred several times over the weeks. This is due to the examples that the students choose to apply the concepts in the given domains. The chart is also interactive that the selection of topics can be done and study the sub-topics comparison in details. This answers our RQ2 on types of visualizations for the topic of network analysis and evolution. We observed that standard non-interactive graphs can provide overwhelming information to the users where as the interative graphs enable the users to analyse the visuals and gain insights in effective manner.

6.4 Discussions

Our proposed solution worked well for the discussion forum within the chosen information systems graduate course. Our experiments show that agglomerative clustering model performs better than k-means clustering for IS technical courses. However, this may not be true for other courses. Therefore, our solution design provides the flexibility for the users to choose the algorithm, similarity techniques and a number of clusters. In this solution, we explored clustering techniques to discover the sub-topics. An interesting future work is to incorporate LDA models to extract the sub-topics by considering the tokenisation by sentences which can provide large data for LDA learning. It is a more suitable algorithm if the discussion posts tend to have multiple subtopics. Another interesting technique for clustering latent class analysis which can overcome some of the limitations of hierarchical clustering methods. The second interesting future work is the summarization based on the topics and sub-topics. We are currently working on the topic-based summarization models to generate automated summaries to the instructors that can be shared with the students.

7 Conclusion

In this paper, we propose a clustering based solution model for discovering sub-topics from the students’ discussions in an online asynchronous discussion forum. Our experiments show that agglomerative clustering model performs better than k-means clustering for IS technical courses. To generalize the solution model across other courses, we provide the flexibility to the users to choose the clustering settings. The second contribution is the discovery and visualization of topic-evolution over the time. The interactive model enables the users to study how students are connecting the previous topics over the period of the course and to what proportion of students post such topical connections every week of the course.

References

Allen, M., Omori, K., Burrell, N., Mabry, E., & Timmerman, E. (2013). Satisfaction with distance education. In M. G. Moore (Ed.), Handbook of distance education (3rd ed., pp. 143–154). New York, NY: Routledge.

Amidon E. J. (1968). Interaction Analysis. Theory Into Practice, Workshop in the Analysis of Teaching: 7:5, 159-167.

Andy, C. (2013), Reports Forum Graph, Moodle. https://moodle.org/plugins/report_forumgraph

Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart. (2008) "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

Atapattu, T., Falkner, K. and Tarmazdi, H. (2016). Topic-wise classification of MOOC discussions: A visual analytics approach. Proceedings of the 9th International Conference on Educational Data Mining (EDM), Raleigh, NC, USA.

Burge, E. (1994). Learning in computer conferenced contexts: The learners' perspective. Journal of Distance Education, 9(1), 19-43.

C. Coffrin, L. Corrin, P. de Barba, and G. Kennedy. (2015). Visualizing patterns of student engagement and performance in MOOCs. Pages 83–92, New York, New York, USA, 2014. ACM

Charu C. Aggarwal and Cheng Xiang Zhai. (2012). Mining Text Data. Springer Publishing Company, Incorporated. 2012

Chaturvedi, Snigdha and Goldwasser, Dan and Daumé III, Hal. (2014). Predicting Instructor's Intervention in MOOC forums Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014

Christiane Fellbaum (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. (2008). Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

De Bruyn, Lisa Lobry (2004) 'Monitoring online communication: can the development of convergence and social presence indicate an interactive learning environment?', Distance Education, 25: 1, 67 — 81

Duval, E. (2011). Attention please!: learning analytics for visualization and recommendation. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge (pp. 9-17). ACM.

Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modelling for understanding MOOC discussion forums: a learning analytics approach. In Proceedings of the Fifth International Conference on Learning Analytics and Knowledge (pp. 146-150).

Gavin Hackeling. (2017). Mastering Machine Learning with Scikit-Learn - Second Edition: Apply Effective Learning Algorithms to Real-World Problems Using Scikit-Learn (2nd ed.). Packt Publishing.

Jian, F., Yajiao, W., & Yuanyuan, D. (2018). Microblog topic evolution computing based on LDA algorithm. Open Physics, 16(1), 509-516.

Leony, D., Pardo, A., de la Fuente Valentín, L., de Castro, D. S., & Kloos, C. D. (2012). GLASS: a learning analytics visualization tool. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 162-163). ACM.

Li, S. & Wong, G. 2016, Educational Data Mining using Chance Discovery from Discussion Board. In Proceedings of GCCCE'16. The Hong Kong University of Education (pp. 712-715).

M.F. Porter. (1980) An algorithm for suffix stripping, Program, 14 (3):130-137, 1980

Ma, T., Qu, D., & Ma, R. (2016). Online topic evolution modelling based on hierarchical Dirichlet Process. IEEE International Conference on Data Science in Cyberspace, pp. 400–405.

Ohsawa, Y., Benson, N. E., & Yachida, M. (1998) KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on (pp. 12-18).

Simon, Y. K. Li, and Gary K. W. Wong. (2016) Visualizing the asynchronous discussion forum data with topic detection. In SIGGRAPH ASIA 2016 Symposium on Education: Talks (SA '16). ACM, New York, NY, USA, Article 17, 3 pages. DOI: https://doi.org/10.1145/2993363.2993367

Venky S, Swapna Gottipati, Ramaswami Seshan and Chirag Chhablani. (2018). Class Discussion Management and Analysis Application. In proceedings of 26th International Conference on Computers in Education (ICCE). 2018.

Vytasek, J. M., Wise, A. F., & Woloshen, S. (2017). Topic models to support instructors in MOOC forums. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference (pp. 610-611). ACM.

Wragg, E.C. (1970). Interaction Analysis in the Foreign Language Classroom. The Modern Language Journal: 54:2, 116-120.

Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. (2010) Understanding of Internal Clustering Validation Measures. In Proceedings of the 2010 IEEE International Conference on Data Mining (ICDM '10). IEEE Computer Society, Washington, DC, USA, 911-916. DOI=http://dx.doi.org/10.1109/ICDM.2010.35