Research on Information Demand of College Students in Emergency Environment based on Virtual Q&A Community - Take "Zhihu" as an Example

Yue Xiao

School of Economics and Management, China University of Petroleum (Beijing), Beijing, China
1902443892@qq.com

Abstract. *Due to the outbreak of COVID-19 school return was postponed, it's important for university administrators to understand the information demand of college students in time. Based on this we analyze the text of the virtual Question answering community platform to capture the information demand. First we collect questions that raised from January to May 2020 about the topic of college students on Zhihu platform as the initial data set. By using the method of text segmentation and stop words removing us come up with a list of words, then we calculate the TF-IDF value of each words to get the keywords. After then we combine these keywords with the question document to construct the vector space model matrix and using K-means to cluster the documents. After analysing we found that during COVID-19, the information demand of college students mainly include six aspects: study at home, graduating, online examination, employment environment, youth inspiring and postgraduate entrance examination. At the same time, we analyse the changing of specific discussion under each topic. And it is suggested that during the COVID-19 period, universities should improve the psychological adjustment of students, perfect the system of final evaluation of students, and give special care for some rural students.*

Keywords: *college students, text clustering, information demand, emergencies, Zhihu*

1. Introduction

Question answering websites provide a platform where users can post questions and receive answers. These systems take advantage of the collective intelligence of users to find information [1]. With the advent of Web2.0 era, the virtual Question answering community platform has entered a golden age of development, take Zhihu as an example, the number of registered users has reached 220 million by January

2019 [2]. At present, the number of active users has reached 26 million per day, and the number of users is still increasing.

Since the outbreak of COVID-19, colleges and universities have taken corresponding measures to deal with the “unconventional” life of college students who can not return to school, students also actively cooperate. In this case, knowing the information demand of college students during the COVID-19 period is helpful to deal with the problems that online work cannot obtain the timely and accurate information reflection caused, so as to better implement the corresponding policies. The aim of this paper is by using methods of text analysis to know the information demand of college student during the outbreak.

2. Literature review

Silverstein, Marc [3] analyze the information demand by designed and distributed questionnaire. The way they taken is suitable for a small sample, cosidering we want to know the information demand of the whole college students in China we decided to adopt other method.

At present, the research on online Question answering community mainly focuses on user behavior, user characteristics, personalized service and content quality evaluation. Jin J, Li Y [4] analyse why users continuously contribute knowledge to online social Q&A communities. Jing [5] present a system, VisQAC, which explores the patterns of Q&A sequence and user behavior.

Hamm[6] supported by Zhihu’s topic categorization algorithm, using social network analysis and critical discourse analysis, examine the 60 most popular question threads about drones on Zhihu.

Through the capture of comments on microblog platform, JiaWenjun [7] analyzes the experience of students in online education, and provides some advices for the future online education mode. They use distributed crawler technology, distributed database system, SnowNLP sentiment analysis model and K-Means algorithm. Huang Lucheng [8] focuse on gerontechnology, take Zhihu as the data source and adopted LDA model, analyse the public’s attention and attitude and constructe a framework of “attention-satisfaction”.

3. Research framework

3.1 TF-IDF

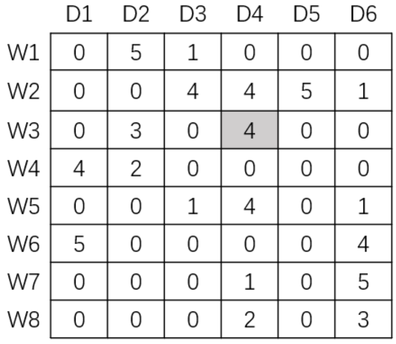
TF-IDF consists of two parts: term frequency (TF) refers to the frequency of a particular word appearing in the text, inverse document frequency (IDF) is the ratio of the amount of text to the number of times a particular word appears in a text set.

Suppose that the frequency of feature word “i” in text “d” is “$tf_i(d)$”, and “n_i”is the number of texts containing feature word “i”, then TF-IDF function is as follows:

$$\mathrm{TF\_\,IDF_i(d) = tf_i(d) * ln\left(\frac{N}{n_i + 1}\right)}$$

3.2 vector space Model (VSM)

Vector Space Model (VSM) maps documents to vector form and in the metrix, (D1, D2... Dn) represents the text sentences to be clustered, (W1, W2... Wn) represents the characteristic vocabulary in the document. The element in the matrix represent the number of words contained in a question text. For example, the element x[4][3]=4 represents Document4 contains four word3. Figure1 is a graphical representation of a document's eigenvector. So we combine one hunderd character words choosing from TF-IDF ranking list with the problem document to construct the vector space model matrix.



	D1	D2	D3	D4	D5	D6
W1	0	5	1	0	0	0
W2	0	0	4	4	5	1
W3	0	3	0	4	0	0
W4	4	2	0	0	0	0
W5	0	0	1	4	0	1
W6	5	0	0	0	0	4
W7	0	0	0	1	0	5
W8	0	0	0	2	0	3

Figure. 1 Document eigenvector

3.3 K-means clustering

After the vector matrix of text space is established we use K-means package for clustering. The vectors with smaller euclidean distance are grouped into the same cluster. After repeated iteration and transformation of the center points, the clustering results are as good as possible by adjusting the number of center points.

4. Model Application

Using the above algorithm and process to the collected data we obtained the corresponding clustering results.

4.1 Information topic discovery

4.1.1 Calculate cluster feature word list

Because the special model of Zhihu platform that users can add related labels of each qestions and delet the existing labels, it is executable for us to know exactly the information demand of college students by collecting the question about the topic of “cllege students”. First we use a spider program to collect question under “College Student” topic and the URL of each question, the latter is used for visiting the log of question raising. It’s important to notice that people proposing the question can chang it or add some statement, but most of time the aim of change is to make it easier to understand, so we chose the date that the question first raised.

Then we use Jieba, an open source Chinese word segmentation tool which can effectively extract words from sentences one by one, for text processing to get a a list of words. Yet the list consist many words that has no practical meaning, to deal with it we integrate stop list from Harbin Institute of Technology (HIT) and word list made by us according to the result of the first tentive test processing into a new stop list to delet the meanless words. After Chinese word segmentation and deleting stop words we cauculate TF-IDF value of each word, order the values from largest to samllest so that the more important one word is the higher it ranked.

And here is the top fifteen words: back to school, accommodation fee, COVID-19, notice, rural, home quarantine, rising waves, review, graduation ceremonies, graduates, employment environment, applacation, bachelor degree, COVID-19 situation, online. By analyzing the words appear in the list above we can understand the information demand of college students roughly. First, due to the postponement of the date to return school because of the COVID-19, there is a heated discussion on online learning, home quarantine, back to school, COVID-19 situation an so on. These are the discussions that appear for the first time under the COVID-19 situation. Meanwhile, there are some discussion that appear every year like graduation ceremony, employment environment, graduates, bachelor degree, but there exists several differents point of view compared with previous years, for example, under the topic of the employment environment, this year has focused more on the impact of the COVID-19 on employment.

4.1.2 Problem text clustering

First we combine one hunderd character words choosing from TF-IDF ranking list with the problem document to construct the vector space model matrix. Then we run the K-means clustering algorrithm which needs to provide the number of clusters in advance, to find the best parameters we adopt two methods described following.

In the first method we calculate the value of error variance within the cluster under different “K”, the result is shown in Figure2, and it’s obvious that when the cluster value is six, there is a large platform period, so we choose six as the parameter.

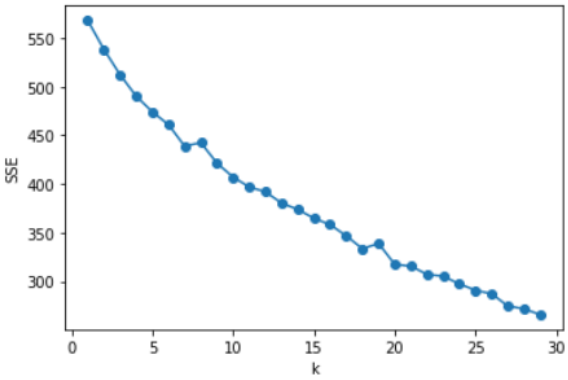


Figure. 2 Error value within cluster variance

In the second method we use Principal Component Analysis (PCA) method to reduct text vector matrix‘s dimension and produces visible results that each sentence of text is one point on the plane and each color represents all the problem texts under a cluster. Figure3 is the result when cluster number is six. We found out that when we choose “six” the distribution of points with different colors is clearer compared to other numbers. After the two methods we choose six as the cluster number.

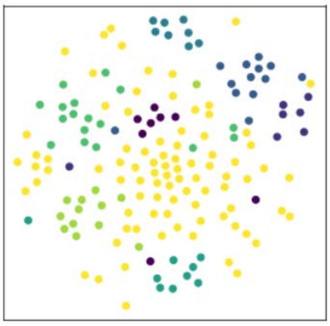


Figure. 3 clustering effect diagram

4.1.3 Cluster label generation

With the above process all the question texts are divided into six categories and each category is extracted as a separate question document. After Chinese word segmentation and deleting stop word in each category, we cauculate TF-IDF value of each word. The following table shows the words in the top of each category, and the tags of this kind of clustering are summarized through the observation of the results.

Table 1 cluster label

number	Clustering keywords	Cluster label
0	academic courses, classmates, stay at home, review, freshman, school, family, home, dormitory, disturb	Study at home
1	entrepreneurship, university, work, graduation, school, bachelor's degree, graduation ceremony, graduate reply, paper	Graduating
2	online, suspected, 2021, cheating, exam, experiment, COVID-19, final examination, back to school, formalism	Online examination
3	work, graduation, employment, ACCA, experience, 874, COVID-19 situation, anxiety, at home,	Employment environment
4	rising waves, bilibilibili, dedication, speech, evaluation, May 4th, China's youth day, short film	Youth inspiring
5	major, full-time, publicity, at home, online, review, junior, admission, enrollment, ministry of education	Postgraduate entrance examination

The word with high TF-IDF value are selected as the keywords of each cluster. And these keywords or phrases composed of key words are used as the labels of the cluster so as to master the main topics contained in the question text. For example, there are keywords such as at home, review, dormitory in cluster 0, which is inferred to be related to online learning at home adopted by universities during the COVID-19 period. Cluster 4 has keywords such as “rising waves" and "BiliBili", these words are related to the publicity video of "rising waves" which has caused much discussion, so this cluster is summarized as Youth inspiring.

4.2 Hot information demand discovery

Though we summerize the topic of each categories, it’s not enough for us to know exactly about information demand of college students. So based on the above six clusters, we analyze specific questions under each aspect and refer to the related resources and events.

4.2.1 Study at home aspect

COVID-19 began in winter vacation and is highly contagious. As college students' return to school is likely to bring more difficulties in COVID-19 prevention, colleges and universities across the country have postponed the start of school and adopted online teaching method to cope with the changes. The discussion on study at home included many aspects, mainly including three parts: at home, learning and school starting.

In the “home” part, the discussion focuses on how to avoid conflicts with family members when staying at home for a long time during the COVID-19, how to alleviate the anxiety bought on by not making very much progress because can’t return to school, how to schedule time at home, whether the college should refund the dormitory fee and how to refund it, etc.

Discussions on learning mainly include how to improve the efficiency of online learning and how to focuse on learning at home. The discussion about the beginning of the semester has always had a high degree of discussion, around this topic there are 210 pieces of relevant questions collected and the angle of questioning always changing. These questions mainly contain the school opening time forecast, discussion about return to school notices issued by individual universities and how to route students after returning to school.

4.2.2 Graduating aspect

The graduating aspect mainly involves the paper, graduation ceremony, job search and so on.

There are a lot of discussions about how to hold graduation ceremony and take graduation photo. And the choice some schools made that hold graduation ceremonies in virtual games has also been discussed by many people. Here are some spicific questions: “Is it sad to say that there is no graduation ceremony this year?”, “Changchun Institute of engineering has held a graduation ceremony in the virtual game “Minecraft”. What do you think of it?”

The way that online thesis defense has also caused a lot of discussion, mainly including how to make powerpoint of oral defense, the opinion of online defense, how to better perform in online defense and the impact of online defense on students.

4.2.3 Online examination aspect

The aspect of online examination mainly include the form of online examination, experience sharing of online examination, discussion of the fairness of online examination, etc.

The discussion of online examination mainly focuses on how to use the equipment combination for the examination and the method to use different examination systems. The experience sharing of online examination includes the influence of screen shifting in the examination under “Superstar” examination system, the problem that the network cannot be uploaded and how to deal with it. There is a discussion on the fairness of online examination, such as "how to treat some problems brought by online examination?"

4.2.4 Employment environment aspect

Aspect of employment environment involves the influence of COVID-19 on job searching, resume production, tips of job interview, psychological state transition of new job seekers, etc.

COVID-19 has had a great impact on many industries, and the employment problem of graduates has triggered a lot of discussion. For example, “In 2020, 8.74 million college students will graduate, how to find a job in such a severe employment situation?", "Under the background of the COVID-19 situation, the employment of college graduates in the past two years is too harsh, how to solve the pressure?", "Facing the most difficult graduation season in history, what do you want to say to this year's college graduates in Wuhan?”.

4.2.5 Youth inspiring aspect

After the release of the video "rising waves" from bilibili on May 4th Youth Day, a wave of discussion among netizens has been set off. The discussion on this issue has gradually evolved from the discussion of the video itself when the video was released to a deeper discussion and understanding of the concept “rising waves”.

After the release of the short film, the question focused on the assessment of the short film, such as: “how to view the video of rising waves". In the later period, there appeared opposition to the content expressed in the video of “rising waves” in Zhihu, such as: “Why has the speech "rising waves" in bilibili caused so many retort?”, “Why is there polarization in the evaluation of "rising waves" in Zhihu and bilibili?”

4.2.6 Postgraduate entrance examination aspect

The topic of postgraduate entrance examination mainly focuses on the interpretation of enrollment expansion policy, the discussion of initial examination scores, the interview experience of online re examination and the enrollment strategies of colleges and universities. And content of the discussion under this aspect changes with time according the process of graduate admission.

4.3 Change of information demand

Figure4 shows the proportion of different types of questions in each month from January to May. Because the COVID-19 broke out in early January, there were few discussions about the term begins, online examination, youth motivation and rising waves, the main topics in this month focused on home study, graduation, employment environment, postgraduate entrance examination and enrollment. Since February, there has been more and more discussion about unconventional activities under the COVID-19 situation. Based on the information learned from the diagram we can know the heat and emphsis of each aspect's change situation.

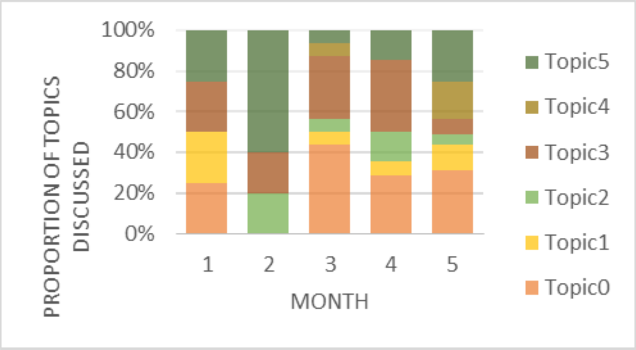


Figure. 4 proportion of time topic discussion

Most colleges and universities began to carry out online study at home in March. Due to the first large-scale application of this kind of teaching form, a large number of related problems have arisen. The number of questions raised in March account for the largest proportion, the questions mainly focus on the experience of online learning and how to operate the equipment. During April and May, the discussions mainly focused on how to improve the efficiency of online classes, how to adjust the anxiety of study at home and the worry about the final grades. As students gradually adapt to online teaching, the problems raised are more focused on learning itself.

The discussion on online examination began in February, mainly about the feasibility of online examination and how to ensure fairness. Before the examination week was coming, discussions about online examination requirements issued by some schools began to emerge. When many students have experienced the online examination, there were a large number of questions about how to ensure the fairness of online examination and how to treat dishonesty. After the end of the examination week, the number of such questions gradually decreased.

There has always been a large proportion of discussions on the issue of employment. Since most of the people who rasie employment questions are in different job seeking states, the change of emphises within the topic is smaller than that of other topics. It's mainly about resume, interview, career choice, suggestions for new graduates in the workplace, discussion on employment trend under the COVID-19 situation, etc.

The topic of postgraduate entrance examination occupies a high proportion of discussion from January to May. With the change of postgraduate entrance examination progress, the focus of discussion under this topic is also different. In early February, the results of the first postgraduate entrance examination in various provinces are announced one after another. Once after the score of each province is announced, relevant problems will appear. Then after the announcement of the national line in mid March, there were a lot of discussions on the experience and results of the online re examination which lasted until May. In addition, this topic also covered the sharing of experience in test, the situation of the 21st postgraduate entrance examination and the study scheme of each month.

5. Conclusion and suggestion

In 2020, the information demand of college students are mainly about study at home, graduation, online examination, employment environment, youth inspiring and postgraduate entrance examination. The information demand of study at home are mainly at home, online learning, and the school starting. The graduation aspect of this year involves the discussion of papers, graduation ceremony and work stuff. The aspect of online examination mainly include the form of online examination, experience sharing of it and discussion of the fairness of online examination, etc. The employment environmentaspect involves the influence of COVID-19 situation on work and the resume production. The discussion under the aspect of youth inspiring mainly caused by the release of "raising wave" vedio. The aspect of postgraduate entrance examination mainly focuses on the interpretation of enrollment expansion policy, the initial test scores and so on.

Combined with the above analysis, the school and relevant parties have taken a lot of methods to deal with the challenges in different periods, and on the whole, they have taken the most appropriate measures under specific circumstances. However, due to it is sudden of the epidemic, there still exist deficiencies in some aspects. Based on this, this paper puts forward some suggestions as reference for improvement.

First of all, in the analysis process, there are many words such as“anxiety”and “confusion” under several topics. Home isolation makes people divorced from normal life to a certain extent. The anxiety caused by the lack of social contact, entertainment and sports needs the school and teachers to understand and conduct psychological counseling in time. Secondly, the fairness of online examination has caused a lot of discussion, as a result, many students have an incorrect understanding of dishonesty. In this case schools and teachers should take corresponding "unconventional" measures to ensure maximum fairness, at the same time students should also adhere to their line in the sand. Finally, the discussion of rural students on many topics is very popular. The implementation of the method is doomed to be unable to take care of all the circumstances, but for the rural students' difficultis, such as hardware equipment, energy allocation in busy agricultural season and so on, schools and teachers need to give full understanding and flexible arrangements.

However, this article only selects the question from Zhihu platform as a source of text data, users have certain characteristics so that cannot reflect of the overall attitude well. The next step can combine with other platforms for research. Besides this paper only focuses on the question without considering the answer text. In the future, the satisfaction analysis can be combined with the answer text.

References

[1] Sneha C, Venkatesh G. Knowledge Sharing in the Online Social Network of Yahoo! Answers and Its Implications [C]// Acm International Conference on Information & Knowledge Management. ACM, 2012.

[2] Chen Guoquan. Ten years of Zhihu [J]. Chinese journalist, 2020, (6): 74-79 (in Chinese)

[3] Galnares-Cordero, L, Gutierrez-Ibarluzea. Information needs of health technology assessment units and agencies in Spain [J]. International Journal of Technology Assessment in Health Care, 2010.

[4] Jin J, Li Y, Zhong X, et al. Why users contribute knowledge to online communities: An empirical study of an online social Q&A community [J]. Information & Management, 2015, 52 (7): 840-849.

[5] Jing, Liang, Ruoyu. VisQAC: Visual Analytics for Online Q&A Communities [J]. Journal of Beijing Institute of Technology, 2019.

[6] Hamm, Andrea, Lin, Zihao. “Why Drones for Ordinary People?” Digital Representations, Topic Clusters, and Techno-Nationalization of Drones on Zhihu [J].Information, 2019, 10 (8)

[7] Research on online learning of college students by Guo Wenting, Gao Yujun [2020] (in chinese)

[8] Huang Lucheng, Jiang Linshan, Miao Hong, et al. Topic identification and analysis based on online Q & a community -- Taking Zhihu "elderly" topic as an example [J]. Library and information work, 2016, v.60; no.546 (05): 94-101(in Chinese)