Research on Data Analysis and Application for College Teacher- Student Communication

Yangming Wang
School of Economics and Management
Beijing Jiaotong University
Beijing, China
yangmingw@bjtu.edu.cn

Juanqiong Gou
School of Economics and Management
Beijing Jiaotong University
Beijing, China
jqgou@bjtu.edu.cn

Wenxin Mu
School of Economics and Management
Beijing Jiaotong University
Beijing, China
wxmu@bjtu.edu.cn

***Abstract*—The teacher-student communication is an important link of college education, in which the data generated are valuable, and havent been fully collected or applied, especially communication records. However, few systematic methods or frameworks are available to guide the teachers to communicate with students efficiently based on these data. Therefore, this paper proposes a framework for analyzing and applying teacher-student communication data. In analysis level, the approach is based on scene pattern acquisition and topic extraction. The application mainly serves before-intercourse and after-intercourse processes. Furthermore, a real case of teacher-student communication scenario is given to illustrate the proposed framework.**

***Keywords*—teacher-student communication, big data in education, scene data, data analysis, data application**

I. Introduction

Teacher-student communication is an indispensable but easily overlooked link in college education. Lack of support for educational media and applications is one of the main causes of communication barriers between teachers and students [1]. The emergence of big data in education related technologies has brought new ideas to the solution of above problems, which refers to the collection of data generated during the entire educational process. These data have the characteristics of real-time, coherence, comprehensiveness and naturalness [2], and most of them (including teacher-student intercourse data) have not been well collected and used. If these data are fully analyzed, it can provide a basis for communication to help teachers assist students efficiently. At the same time, teacher-student communions appear in many student training sessions, such as career planning, deep counseling, and subject researching [3], that is, the needs for specific scenarios are different. However, there is still a certain similarity in the data and paradigm. Consequently, it is necessary to design a systematic framework based on such scenarios, which can help analyze and apply these data. It is in this perspective that this article sheds light on the problem, focusing on analysis and application of data generated during teacher-student communication process. This article is mainly dedicated to present a framework to support the process of analysis and application for data generated by teacher-student communication. More specifically, the main contributions of this research are as follows:

- A practical method of analyzing data related to teacher-student communication is developed, including scene pattern acquisition and topic modeling based on text data.
- A feasible perspective of applying this kind of data can be summarized, which aims to achieve a closed loop from data analysis to application optimization.

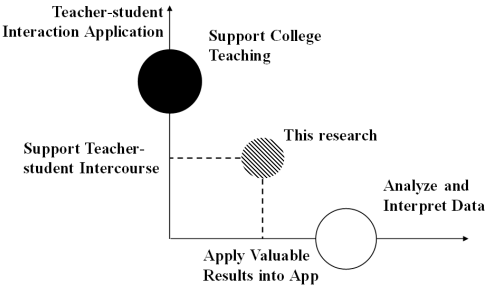

Fig. 1. Positioning of this research.

The positioning of this research is shown in Fig. 1. The paper is composed of five parts: i) the background of the study; ii) related work; iii) the general idea of the research framework; iv) illustration of the framework specifically; v) a real case. And then conclusions and work prospects are drawn.

II. Related Work

In recent years, with the development of smart campus and the emergence of lightweight applications based on scene thinking, many applications have been built for the interaction between teachers and students in colleges, mainly for the purpose of teaching and learning. It includes 3 types applications: applications based on smart classrooms [4], application based on online teaching [5] and applications based on classroom communication [6]. The above applications are involved in the teacher-student intercourses, such as the teacher-student discussion module in online teaching, the results-sharing module in classroom teaching, and the problem-discussing module

Supported by "Natural Science Foundation of China", named "Context based Multi-dimension ontology modeling and alignment"; Supported by "Ministry of Education, Science and Technology Development Center", named "Research on College Students' Behavioral Intelligence Based on Collaborative Scenario"

978-1-7281-1566-5/19/$31.00 © 2019 IEEE

in after-class communication, all of which are only about the teaching, lack of attention to teacher-student interaction.

The teacher-student communication data generated in the above applications are mostly text data with time and individual attributes. Therefore, the methods used in the analysis mainly include statistical analysis of interaction-related attributes and text mining for interactive content. The former mainly analyzes the interaction related attributes such as time, frequency, situation and interaction participants of interaction, and uses descriptive statistics to explore the changing characteristics and trends of related attributes [7]. The text mining research on interactive content mostly uses text mining related technology to segment the text content, word frequency calculation and other operations. Some also use emotion analysis [8], term extraction [9] and other algorithms to analyze the deep information of content. In summary, the existing research of teacher-student intercourse data stays at the theoretical level, mostly used to verify the effectiveness of the platform, or to conduct simple analysis and interpretation based on data, lack of closed loop to apply the results of data analysis in the corresponding teaching links. That means the valuable information obtained through communication data analysis should be used effectively.

III. Research Framework

According to the related work above, it can be seen that some teacher-student communication data has been collected through the online learning platform or other applications, and are mostly used as an auxiliary link to improve teaching, making the application of these data insufficient. It is reflected in two aspects. First, the data analysis perspective tends to focus on evaluating teaching effects rather than improving the communication link itself, the second is that the data analysis results are not fully applied or reflected in the relevant modules. Based on this, the two key questions to fully use data of teacher-student communication is how to perform analysis on the collected data and how to apply the results to the scene.

A. How to Analyze

For the first problem, the purpose of data analysis is to improve efficiency and reduce cost of communication. Based on the characteristics of these data, the key step to solve the problem is to find similarities and differences in such scenarios. The same point can help understand the operation mechanism and data generation process of such scenarios. And according to differences, the method of analyzing and modeling can be selected. The scene pattern summarized by data is able to represent similarities of scenarios, describing all the links of teacher-student communication and the data required in each link. The modeling and analysis methods of these data are mainly selected based on the content and scene features. Since the theme is the key point of teacher-student communication, the topic modeling is an appropriate method. So how this method applied in analyzing of these scene data is described specifically in next section.

B. How to Apply

For the second problem, it is necessary to transfer the results of data analysis into corresponding services in a specific part of the scene based on its pattern. Before the communication, the teacher needs to understand the background information of the students interviewed and prejudge the problems they want to consult. After the intercourse, the teachers hope to get feedback on the communication effect. Therefore, the pre-judgment of students and the analysis of communication records are the core applications during this process.

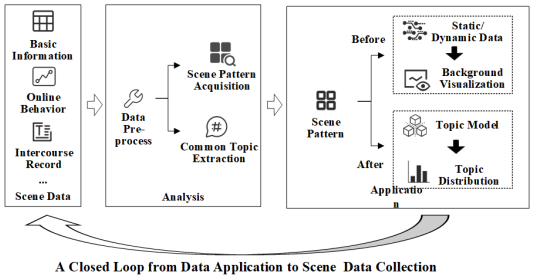

Fig. 2. Analysis and application framework for teacher-student exchange data.

The general idea of the analysis and application of teacher-student communication scene data is shown in Fig. 2.

IV. Data Analysis Level

A. Scene Pattern Acquisition

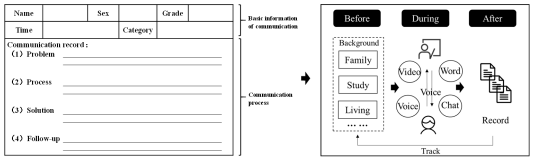

Fig. 3. Scene pattern acquisition.

Some historical communication records were obtained mainly from the teacher's archive, network and related literature, a total of 102 articles, 111,667 words. The record template is roughly as shown in Fig. 3(left). The template is mainly divided into two parts, one is to collect structured data, including the basic information and communication purpose of the students interviewed, and the other is to record unstructured data, including the pre-interview questions, interview process, solutions and follow-up progress. Based on this, the teacher-student communication scene can be further divided into three stages. Among them, the basic information of the interview that the teacher needs to know before communication, including the family, learning, and living situation. When communicating, the process of communication may be direct or indirect, so the record may be online video, voice, or chat history. Regardless of the form of communication, the teacher or related application will keep a record of the process, and may also include the solution to the problem and the follow-up progress.

The communication record is the core data in the scene, this paper chose the NLPIR Chinese word segmentation system (2016 version) to preprocess them. Since the results of the initial word segmentation have invalid content such as vocabulary misclassification and punctuation, it is necessary to apply the stop-word dictionary and the user dictionary to optimize the result. After performing word frequency statistics, the results show some characteristics. The high-frequency vocabulary is mostly concentrated on study, economy, work, living, grades, classmates, etc. This shows that teacher-student intercourse is mostly developed around one or several themes, so the records also reflect certain themes. Meanwhile, there are often words such as stress, encouragement, and negative in the records. It shows that the emotional state of the students, which will also be reflected in the process of communication, so there is a certain emotional tendency in the record. In summary, the scene pattern based on the preliminary results of data analysis is also shown in Fig. 3(right).

B. Topic Modeling based on Communication Records

1) Topic modeling process: Based on the scene pattern above, it can be concluded that the communication record has certain themes and emotional tendencies, so it is necessary to analyze the subject and emotional tendency of the communication record. This discovery of deep semantic knowledge is currently using topic modeling [10] in the field of Chinese information analysis. Some scholars have practiced in different area [11], [12], and the related methods are also widely applied in the field of education [13].

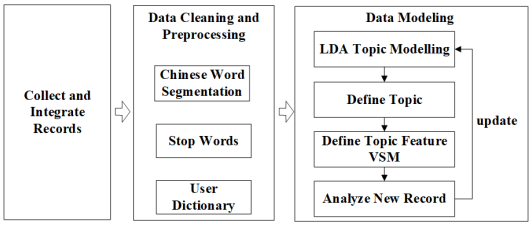

Fig. 4. Topic modeling process.

This paper will apply the LDA topic modeling algorithm to analyze the content of these data. The algorithm considers the document to be composed of several topic distributions, and the theme is composed of several vocabulary distributions. It is a three-layer Bayesian model, which can be used to identify the semantic topic information in the document [14]. The specific modeling process based on the record is shown in Fig. 4. Firstly, the common themes of communication records are mined and defined, and then the feature vector space model (VSM) of each topic is constructed. Finally, the topic information is stored, and the topic model is updated according to the subsequent communication records.

2) Construction of the topic model: Based on the preprocessing of history records in the scene pattern acquisition phase, all the records that have been well-written are integrated into one document (.txt file), and each row in the collection represents an original record. This document collection is an initial set of words to be used as the input (.dat file) for the LDA model construction. Then the LDA algorithm written in Python is used to analyze and process the above files, and the related parameters are set based on the general experience. The parameters are:

k: the preset number of topics, based on pre-analysis of communication records;

With:$k = 6$;

α : hyperparameter, taking the general experience value;

With:$\alpha = 0.1$;

β : hyperparameter, taking the general experience value;

With:$\beta = 0.1$;

iter_times: the number of iterations, taking the general experience value;

With:$iter\_times = 1000$;

top_words_num: the number of high-frequency vocabulary displayed under each topic, where approximately 10% of the total number of vocabularies of each document in the training word set;

With:$top\_words\_num = 100$;

TABLE I
OUTPUT DOCUMENTS

File Name	Content
wordidmap.dat	Unique ID of all words in the word set
model_theta.dat	Probability distribution of each document on each topic
model_tassign.dat	Each topic's individual word is assigned a topic number
model_phi.dat	Probability distribution of words under each topic
model_twords.dat	Characteristic words of TOP N under each topic

The LDA algorithm first encodes the vocabulary of the complete word set, and then clusters it into six topics. Each topic has several vocabularies. And there are 5 output documents, which shows the relationship between words and topics in different forms. The contents of each document are shown in TABLE I. The original data is in Chinese, and the analysis results are translated into English for easy understanding.

When k=6, by looking at the output results, it is found that the feature words under each topic can more clearly determine the meaning of each topic, thereby determining major learning, daily life, mental health, social activities, job hunting, and native family as six topics. Further, in order to filter out vocabulary that is less relevant to the topic, this paper excluded the characteristic words with probability ¡0.003 under the topic. At the same time, in order to facilitate display, the probability of vocabulary correspondence is kept 5 digits after the decimal point. The characteristic vocabulary and probability under each theme are as shown in TABLE II.

Thus, based on Top N vocabulary and their probability, the vector space model of each topic (the topic model) can be built. The Vector Space Model (VSM) is a common method for

TABLE II
FEATURE VOCABULARY AND PROBABILITY OF EACH TOPIC (PART OF)

Major Learning		Daily Life		Mental Health		Social Activity		Job Hunting		Native Family	
study	0.02253	dormitory	0.02318	problem	0.02449	internet	0.01693	job	0.02028	economic	0.00983
school work	0.00997	classmate	0.01926	psychological	0.01030	event	0.01044	find	0.01755	parent	0.00953
time	0.00983	dormitory	0.01463	thought	0.00784	work	0.00914	job hunting	0.01154	family	0.00797
warning	0.00910	life	0.01283	depth	0.00981	handle	0.00741	major	0.01099	encourage	0.00750
score	0.00750	roommate	0.01036	appear	0.00957	start	0.00697	university	0.00716	parent	0.00738
method	0.00736	university	0.00965	mood	0.00932	class	0.00590	employment	0.00607	state	0.00657
work hard	0.00564	understand	0.00787	situation	0.00565	guide	0.00524	industry	0.00607	hard	0.00614
course	0.00522	intercourse	0.00644	anxiety	0.00418	school	0.00524	graduation	0.00607	mother	0.00564
homework	0.00492	expenses	0.00607	positive	0.00296	class committee	0.00467	graduate	0.00607	happen	0.00436
grasp	0.00481	communication	0.00573	confidence	0.00247	cadre	0.00467	plan	0.00552	situation	0.00393

transforming unstructured data such as text into a structured form. The topic model is as follows:

$$Topic_k = [FS_k, FP_k]_n \tag{1}$$

With:$Topic_k$ is kth topic, $k \in \{1, 2, 3, 4, 5, 6\}$;

$FS_k = \{w_1, w_2, ..., w_n\}$ is the Top N word set under this topic;

$FP_k = \{p_1, p_2, ..., p_n\}$ is the probability set corresponding to each feature word in the word set;

$n \leq N$, with: n is the number of words in the vector space model of the topic; N is the number of initial feature words output, equals to the value of top_words_num.

The construction of the topic model is the basis for the analysis of the subsequent communication records, so it is necessary to store the training corpus and the relevant information of the topic model, including the output documents, and some dictionaries predefined during the preprocessing. Since the content of intercourse changes with the scene, the corresponding topic model also needs to be updated continuously.

V. DATA APPLICATION LEVEL

A. Pre-judgement of problem

Before the intercourse, the teacher will find the basic information, family situation, learning situation and other information of the student as the reference in order to preliminarily determine the cause of students problem. Therefore, integrating the multi-dimensional behavior data of the students to visually present information to the teacher is necessary. The background information can be divided into two types: static and dynamic. Static information refers to information including basic student information, family information, historical achievements, etc., which are recorded and generally do not change. Dynamic information refers to behavior data related to consumption, learning and work.

The application model of pre-judgement of problem is as follow: firstly, combined with the purpose of students' communication, the corresponding static information indicators are selected and pre-processed. And then appropriate descriptive statistics methods will be used to visualize the relevant data, obtaining the changes and their development trends. Based on the above results, the dynamic behavior information is selected to construct the analysis model, mainly using correlation analysis, multiple regression statistical analysis and other methods to deduct the potential causes that may affect the static information indicators. The application model is shown in Fig. 5.

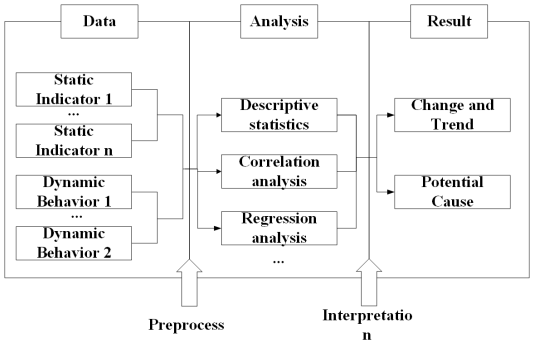

Fig. 5. Application model of pre-judgement of problem.

B. Post-analysis of records

Communication records are the key to provide teachers with basis of follow-up communications. Therefore, through the analysis and visualization of the records, the focus of the intercourse can be drawn, providing reference for teachers to provide suggestions for the future intercourses. Based on the common communication topics identified in the previous section, the application model mainly consists of two parts. One is the construction of a new document vector space model, and the other is to calculate the similarity between the new document and each topic. First, the communication records are cleaned and preprocessed. Then TF-IDF algorithm is used to structure them, and construct the corresponding vector space model (VSM). Finally, the cosine similarity algorithm is applied to calculate the semantic similarity between the record and each topic. TF-IDF algorithm is used to calculate the weight of vocabulary in documents, and is widely used in related fields such as information retrieval and document

classification [15]. Therefore, the VSM of the ith new record can be expressed as:

$$Doc_i = [NS_i, NP_i]_m \quad (2)$$

With: Doc_i is a new record;

$NS_i = \{u_1, u_2, ..., u_m\}$ is a set of words for this new record with pre-processing;

$NP_i = \{q_1, q_2, ..., q_m\}$ is a set of weights for the word set, calculated by TF-IDF weighting strategy; m is the number of words remaining after the new document has been filtered by word segmentation and stop words.

Subsequently, in order to obtain the topic distribution, the cosine similarity is calculated based on the new record and the vector space model of each topic. Cosine similarity reflects the similarity between texts by calculating the cosine of the angle of the vector. The closer the cosine value is to 1, the more similar the two vectors are. So, unifying the dimensions with new record and each topic is necessary. Take (1) (2) as an example, NS_i and FS_k are the characteristic word sets, NP_i and FP_k are the corresponding weight sets:

when $i \neq k$,

$CS_j = NS_i \bigcup FS_k$, $j = \{i, k\}_{max}$; So NS_i transfer into NS_j;

When a word belongs to CS_j, but not belongs to NS_i, its weight values 0;

The corresponding weight set FP_k turns into $FP_j = \{p_1, p_2, ..., p_n, 0, ..., 0\}$; And the weight set NP_i turns into $NP_j = \{q_1, q_2, ..., q_m, 0, ..., 0\}$;

Finally, the cosine similarity of two vectors is as follows:

$$\cos(\theta) = \frac{FP_j \cdot NP_j}{|FP_j| \times |NP_j|}$$

By analogy, the similarity of the document to other topics can be calculated, and the distribution of the document under the six themes is obtained.

VI. Experiments of In-depth Counseling Scenario

In-depth counseling refers to the in-depth understanding of the actual situation of college students, according to the needs of students' growth and development, using scientific knowledge and methods to counsel students purposefully. So, this paper takes it as an example, collecting the records by an application, and finally integrates the process and results into it. Due to space limitations, this paper selects only one student who participated in the whole process of in-depth counseling as a sample, and apply the above framework to it.

TABLE III
Topic distribution of the example

Learning	Life	Mental	Social	Job	Family
0.383	0.204	0.140	0.042	0.069	0.074

The interviewee was a sophomore student and the purpose of communication was related to learning. In the pre-judgement of problem, based on the family information, the teacher selected learning situation and daily life as two key indicators of concern. Then the application analyzed his grades with the student's learning behavior data on the curriculum platform by correlation analysis. The consumption curve based on the student card data was also shown in the application. All results are visually presented. After analyzing the communication record, the distribution of the topics in this intercourse is shown in TABLE III. It shows the theme of this communication focuses on learning and daily life. This result is basically consistent with the analysis result of the problem prediction.

VII. Conclusion

The presented framework of analysis and application on teacher-student communication scene data might be considered as a practice of big data on ideological and political education in college, which could contribute to a deeper understanding of teacher-student intercourse. The future works would focus on: i) improve the storage scheme of related scene data, including the inputs and the outputs of topic modeling; ii) design the update mechanism of the topic model using ontology; iii) refine the process of pre-judgment, including the selection of static indicators and the extraction and pre-processing of dynamic data.

Acknowledgment

The presented research works have been supported by "Natural Science Foundation of China", named "Context based Multi-dimension ontology modeling and alignment", and "Ministry of Education, Science and Technology Development Center", named "Research on College Students' Behavioral Intelligence Based on Collaborative Scenario". The authors would like to thank the project partners for their advice and comments.

References

[1] Z.Biao, "Research on the Problems of College Teachers and Students' Communication Inefficiency and Improvement Countermeasures," in Contemporary Education Sciences, vol.1, pp. 17-19, 2014.

[2] Y.X.Min, W.L.Hui, T.S.Si, "Application Model and Policy Suggestions of Education Big Data," in E-education Research, vol.9, pp. 54-61, 2015.

[3] Z.W.Hui, W.X.Bing, L.M.Lei, "Investigation and Research on Academic Exchange between Tutors and Postgraduates," in Journal of National Academy of Education Administration, vol.6, pp. 82-86, 2012.

[4] L.B.Qi, L.Xin, "Empirical Study on Analysis and Application of Smart Classroom Data Mining," in E-education Research, vol.6, 2018.

[5] D.X.Xing, W.Y.Dong, L.Q.Sheng, J.H.Yu, S.J.Yan, P.Hong, "Research on the Visual Analysis of Students' Behavior Data in Network Courses," in Journal of Southwest University of Science and Technology, vol.21, no.2, pp. 93-98, 2016.

[6] Z.S.Chao, Y.Z.Min, R.W.Zhong, "Blog-based teacher-student interaction platform," in Experiment Science and Technology, vol.4, no.5, pp. 98-100, 2006.

[7] H.X.Bin, M.Jing, C.J.Gang, "Analysis of Relationships among Group Behaviors of Teachers and Students under Different Implementation Strategies of Blended Learning in Colleges and Universities," in E-education Research, vol.12, pp. 37-43, 2017.

[8] Z.J.Jing, Y.Y.Hong, A.Xin, "Enabling learning interaction through 'bullet screen' videos," in Distance Education In China, vol.11, pp. 22-30, 2017.

[9] W.S.Ping, H.Y.Hui, W.L.Na, "Reflections on Online Teaching and Learning Based on Data Mining and Analysis of Learning Process Data," in Modern Educational Technology, vol.25, no.6, pp. 89-95, 2015.

[10] L.S.Hao, C.L.Men, "The Review on the Application of Text Mining in Chinese Information Analysis," in Information Science, vol.34, no.8, pp. 153-159, 2016.
[11] C.G.Hui, Study on Opinion Mining Based on Semantic Analysis, Wuhan University, 2010.
[12] C.X.Mei, Study of Knowledge Discovery of Opinions from Web Review, Jilin University, 2014.
[13] C.Zhi, H.R.Huai, "Text mining and its application in education," in Modern Distance Education, vol.2, pp.71-73, 2008.
[14] W.Hong, Z.Hao, S.J.Chuan, "Research on domain ontology concept acquisition method based on Latent Dirichlet Allocation," in Computer Engineering and Applications, vol.54, no.13, pp. 252-257, 2018.
[15] W.S.Peng, P.Yan, W.Jie, "Research of the text clustering based on LDA using in network public opinion analysis," in J Shandong Univ Nat Sci, vol.49, no.9, pp. 129-134, 2014.