[1]Leiming Shen

Application of Lightweight Freshmen Group Portrait Based on Echarts

JES Journal of Electrical Systems

Abstract: - This study designs a general analysis model for fusion method, big data portrait of young users using map engine data, and tries to solve the problem of fusion of qualitative and quantitative methods in user portrait. This paper determines the necessity of freshmen group portrait in the practice of ideological education by studying the freshmen group portrait and its objectives, connotations and roles; through the various aspects of data collection, cleaning, structuring, data analysis and modeling, the Echarts platform is used to establish a data visualization example, to achieve a three-dimensional, multi-dimensional and interactive visualization output of the freshmen group portrait, and with the help of the combination of qualitative and quantitative methods, the design of the Research model. Based on the theory of sociological psychology, we construct a user value map, use Look-alike algorithm to construct the map data labeling system, use K-Means clustering algorithm to get the data results, and analyze the data results for business. Using the model for big data empirical demonstration, the results show that young users can be divided into 20 categories of groups, and the total number of data results reaches 170 million, and the number of preference labels reaches 606, which is better than the results of survey data. And it is applied in the initial education and teaching practice process, obtaining better results, with replicable and generalizable practical application scenarios.

Keywords: user profiling; Echarts; freshmen group; user profile; visualization

I. INTRODUCTION

In the era of big data, there are many methods for user research and user profiling, but there has not been any profound research on general users in the literature based on billion-scale data volume and multi-dimensional labeled data from data sources, and there is also a certain gap in the design of the corresponding methodology. The so-called general users are mainly consumers with product and service consumption needs, where "general" refers to the generality of the group, and "user" refers to the dual demand for products and services [1]. In this paper, we take the portrait of young users as the basic goal, based on the unique map engine big data, the establishment of qualitative-quantitative fusion of big data composite dimensions, data analysis composite tools of the basic method of user portrait, and the application of this method of empirical evidence, in order to demonstrate the method, study the differences in the characteristics of the different populations.

Freshmen group portrait, on the other hand, is the application of user image in the field of education, but also a visual learning analysis technology, this paper specifically refers to the technical processing tools that can reflect the overall behavioral characteristics and information of university freshmen in the early stage of enrollment in general.

The budding new data, which has been a big hit in the school season in recent years, belongs to a typical application of freshmen group portrait in the field of big data. Thanks to the popularization and application of big data,

[1] Department of Engineering, Shanghai Customs Colloge, Shanghai 201204, China

visualization and other new technologies, Shanghai universities generally from 2017 onwards, on gender, geography, height, constellation, surname, birthday and other dimensions of the analysis, so that the original two-dimensional structure of the data to get three-dimensional, multi-dimensional presentation, for the college freshman group of the user portrait has become increasingly mature [2].

New students in colleges and universities, wide geographical distribution, distinctive personality, and a large number, how to quickly close the distance between teachers and students, students and students, "ice-breaking" activities are undoubtedly simple and effective, and social phobia and sensitive temperament to a certain extent also plagued some of the students, if the embarrassment of direct contact before a student group portrait has been pre-understood and familiar with, the situation may be very different. familiarized with, the situation may be very different [3]. Returning to the theme of freshmen group portrait, we analyze its advantages:

First of all, by clearly identifying and defining information about the new students, the new student group portrait forms a labeling language used for effective interaction, making it easier to see the whole picture rather than the local, and the statistical level of reality rather than specific bias, ensuring that the adaptation period of multiple parties involved is unified in a rational, authentic, and good transition. Freshman group portraits allow both parties in education to focus on and design for a particular freshman group, helping both parties in education make better decisions, circumventing group blindness and powerlessness, and facilitating closer proximity to each other.

Freshmen group portraits are more likely to generate interest and empathy for freshmen groups, as opposed to the old two-dimensional forms or a simple string of numbers. In other words, the Freshman Profile will help teachers, class members, and schools become more attentive to specific groups of freshmen.

The lightweight visualization application of neonatal group portrait explored in this paper is realized by Echarts, which is suitable for visualization application of user portrait data as Echarts is a JavaScript-based opensource charting library for data visualization, and relies on a lightweight 2D charting engine at its bottom for personalization of data visualization charts. Affinity for non-professional users is reflected in the regular graphical examples provided on its official website , covering line charts , bar charts , scatter plots , pie charts , K-line charts , box plots , geographic data visualization maps , heat maps , line charts , relational data visualization relationship charts , tree diagrams , Rising Sun charts , multidimensional data visualization parallel coordinate charts , as well as funnel charts for BI , dashboards and so on. It also supports mixing and matching between charts, totaling no less than a hundred types, suitable for a variety of application scenarios; at the same time, the open data interface provides the possibility for professional users to customize and develop more complex visualization integration applications.

II. RELATED WORK

User profiling is essentially the process of studying users and analyzing information about them. This content originated from the earliest business analysis, business out of sensitivity to the market and the pursuit of profit, promoting the development of user analysis [4]. Early user analysis is a concept in the field of marketing, which is deeply influenced by the rational economic man theory of economics, but the study [5] points out that this process should be based on the analysis of user behavior. Study [6] concluded that user analysis is mainly about studying user preferences and examining the factors that influence them. Study [7] summarizes several important stages,

containing key evolutions and points in time. From a methodological perspective, user profiling has gone through survey research methods (including ethnographic research), model analysis methods (with causal and structural equation modeling predominating), experimental methods, and also derives models for user segmentation, models based on sociology and psychology (e.g., Profiler, SIGMA Milieus®), and models based on the family lifecycle and analytical models of lifestyles (e.g., models and improvements of Price et al.), VALS (Values and Lifestyles), LOV (List Of Values), etc., and the data source is based on survey data [8-9].

Under the influence of business digitization, the concept of user profiling based on big data has gradually emerged. Study [10] considers user profiling as the process of fully depicting and characterizing the virtual image of the user built on data. Study [11] believes that user profiling is a modeling and analysis process that extracts the corresponding features with the help of big data in order to describe the user's needs, preferences, and so on. Study [12] believes that user profiling focuses on algorithms and technologies. In terms of understanding, user profiling favors conceptual and theoretical exploration, and it is also more often strongly associated with user profiling and big data, which is considered to be a product of the big data era. In terms of application, it is mainly in the fields of healthcare, finance, retail and consumption, industry and automobile manufacturing, etc., and the core is to explore the opportunities of application in specific fields. However, empirical studies using big data with a volume of more than 100 million for application are not rich enough. In terms of methodology, due to the higher attention to data and algorithmic models, especially in computer science to explore new algorithmic models and improvements, such as clustering algorithms mainly include K-Means, hierarchical clustering (e.g., Hierarchical Method), density clustering (e.g., Density-Based Spatial Clustering of Applications with Noise, DBSCAN) and other methods, highly dependent on data magnitude and data dimensions; in the design of the method model, there is a design based on a single source of big data [13], but there is no research model design that explicitly proposes the integration of qualitative and quantitative methods. In terms of data sources, although with the application of business big data, many data sources have begun to appear and be applied, but fewer data points are used, which fails to fully reflect the characteristics of large data magnitude, and the data dimensions are also slightly insufficient [14].

In summary, the research history of user image can be traced back to the 20th century, but the research literature on user image is mostly theoretical discussions and overview studies, and the combination of research with business big data is relatively less; in empirical evidence, there is a greater need for research that can reflect the characteristics of big data with large data magnitude and high data dimensions, in order to more accurately reflect the multidimensional characteristics of users. Overall, it is of some research significance to explore the big data user portrait model under the convergence perspective and empirically validate it.

III. RESEARCH METHODOLOGY DESIGN

3.1 Research Overview and Characteristics

This paper systematically integrates and solidifies discrete big data analysis methods with the help of empirical evidence, and integrates big data algorithms, sociological and psychological theories, and results of research and interviews in order to construct a general model of big data user profiling and crowd analysis (referred to as the "analytical model" or "model") that integrates qualitative and quantitative methods. model"), on the basis of this

model, to carry out empirical evidence and get the conclusion, to realize the theory and application of practice of the closed loop of the research. The big data used in the study provides full coverage of the users to be described, utilizing all 200 million data points and corresponding labels of young users in the server, and utilizing cloud clusters to directly perform calculations and output results.

The research hopes to establish a basic methodology for user profiling with composite dimensions and composite tools, and to apply this methodology for empirical demonstration, characterized by the following two aspects:

(1) Convergent approaches. On the one hand, the fusional approach is mainly the fusion of qualitative and quantitative, i.e., the fusion of sociological demographic research model and big data algorithmic model and the validity of its results. Sociological research models and big data algorithmic models seem to be theoretically incommensurable, and this paper combines empirical evidence to try to solve this fusion problem, enhance the strengths and shortcomings, and improve the efficiency. Specifically, it is reflected in the fusion of qualitative and quantitative when determining user labels and the qualitative screening of clustered big data results. When determining user labels, the qualitative method is chosen to establish a value map and the quantitative method is fused to obtain a comprehensive collection of user data labels; after obtaining the big data clustering results by using the quantitative method, the qualitative method is combined to exclude the data results that are not sociologically significant, and the valuable data are finally interpreted in detail, which is the complementary feature of the fusional approach. On the other hand, the fusional approach is reflected in the environment of static and dynamic big data combination to establish a user profiling method that fits with it. The data source used in this paper itself has the characteristics of static data and dynamic data combination, and at the same time, in the analysis process, the different meanings of the user's static label combined with the dynamic location are also fully taken into account, such as the difference between overnight (business travel-oriented) and non-overnight (entertainment and leisure-oriented) in the positioning of high-end hotels.

(2) Integration and application of technology. This paper, as a design and empirical study of analytical model, is different from the empirical study of single method, not highlighting the optimization process of a single algorithm, but focusing on the design of analytical model, which is a kind of integration and application of multiple techniques. Specifically, different algorithms are used in the design process of analytical models to solve different problems, and the reasons and results of using the corresponding algorithms are evaluated qualitatively or quantitatively. For example, the selection of seed populations and the evaluation of K-value selection reflect the focus on technology integration and the replicability of the overall application method. The results represent an innovation at the application level, especially in the application methodology of technology integration.

3.2 Analytical Model Design

The analysis model is mainly composed of one data input, one data output and three core links. The data input part is mainly to confirm the data; the three core links are user value map construction, labeling system construction, and algorithm construction; the data output is mainly the analysis of the results and the output of the user portrait.

The core of data input confirmation is to confirm the range of data points and data labels, and to determine the number and quality of user data points, data dimensions (labels), etc. to be included in the calculation according to

the selected research objectives. This is the most time-consuming element and the most tedious groundwork. Data and labels whose quality is not up to standard can hardly reflect the objective user status, so a lot of basic data quality control work is needed. In this study, the big data of map engine is selected for analysis, also considering that this data source is of higher quality in the processing of preoperational datatization, which can save the subsequent operation time.

Among the three core aspects, the purpose of the construction of the values map is to establish the initial categorization assumptions of the population, to form a user values map with research significance, and to drop user data points into it based on the validated map; at the same time, the construction of the values map is also used to support the scientific selection of the data labels. The theoretical basis of user value map is Adler and Jung's individual psychology and social personality theory [15]. When people deal with personal and social relations, they choose to show their energy or find social belonging in the society in the theory of social consciousness, and in this selection process, two bases of personal value orientation and group value orientation are formed. Synovate, a British research institute, used this basis to construct the Censydiam model and widely used it in user analysis [16], and TNS market research company also used this theory to establish the NeedScope model [17].

The purpose of the labeling system is to optimize a large number of feature labels and select accurate inputs for the algorithm. The input labels greatly affect the results of the algorithm, and there are more than 600,000 feature dimensions in the map engine, so in order to ensure that the algorithm is efficient and accurate, it is necessary to construct a streamlined labeling system in advance. The author found that using the Look-alike method [18] to establish a generalized labeling system and combining the labels in the value map to form a concatenation set is an efficient way. 2016, the Yahoo team made a corresponding discussion of the model, method and results [19]; Tencent's WeChat team also made a discussion of the RALM (Real-time Attention Based Look-alike Model) technology in 2019; the Tencent WeChat team also made a discussion of the RALM (Real-time Attention Based Look-alike Model) technology in 2019. Based Look-alike Model) technology was described and the system framework for implementation was built [20]. The construction of user value maps and labeling systems highlights the process of combining sociological research methods with big data research methods. On the one hand, the two are based on the classical sociological methodology, describing users from the social and individuality dimensions; on the other hand, the feature labels are streamlined and key contents are extracted by combining the Look-alike approach, and together with the sociological model, the final labeling system is formed.

The algorithm construction process is mainly based on the selected data dimensions and labeling system, confirming the algorithm with the implementation of the calculation process. The idea of algorithm construction is filtered from the binary classification perspective of hierarchical clustering and non-hierarchical clustering. Hierarchical clustering dimensions of each sample as a cluster for processing, according to the distance between the clusters and similarity of the level by level merger, repeated calculations, and the output of a clear hierarchical structure; non-hierarchical clustering based on the assessment of the individual samples in the spatial distribution of the dispersion, the use of the function to assess the degree of dispersion, the function itself is also in the process of continuous optimization, the most representative is the K-Means algorithm [21]. Although the evaluation of algorithms has more indicators, including the amount of computation, arbitrariness, softness, etc., but combined with the actual

computing environment, the amount of data analyzed this time reaches hundreds of millions of dollars, and it is necessary to choose algorithms that can minimize the burden of computation, and therefore focus on the consideration of the algorithm K-Means, which is a small amount of computation, the process of less resource consumption, and the output of the algorithm is good.

Data results output and application is to verify and judge the data results and effects, match the data results with the actual situation to verify and ensure that the data reflect the actual situation; at the same time, based on a reasonable visualization process, a large number of multi-dimensional data results for the secondary analysis, simple and clear presentation, the output of the portrait. The specific model design is shown in Figure 1.

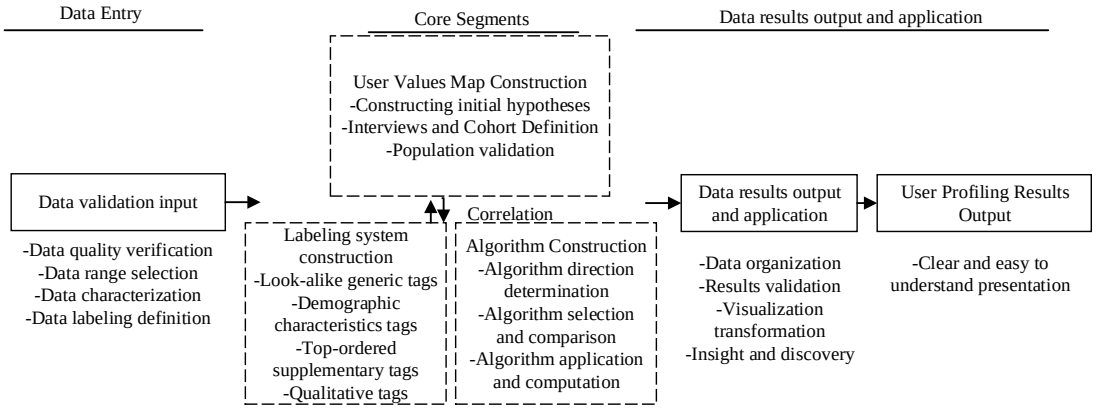

Figure 1 Overall model design for fusion analysis

IV. KEY RESEARCH PROCESSES

4.1 Data validation

Table 1 Connotation table of new student group portrait

Serial No.	Project	Connotation	Role
1	Sex	Gender structure of the embodied group	Different proportion structures have an implicit effect on the ecological construction of collective harmony, and the corresponding educational strategies and communication methods are adjusted accordingly.
2	Age	Age distribution, minimum and maximum age	Most of the students in the same grade do not differ much in age, so it is easier to highlight distinctive individuals and achieve a focusing effect, which will leave a deep impression on the group and bring sustained attention in the subsequent common learning and life.
3	Birthdate	Distribution of	Providing material for group birthdays, individuals

		birth year and month	will also get along better in an atmosphere of relative consistency.
4	Born on the same day	Same month and same birthday	Finding accidental similarities reinforces the preciousness and specialness of such serendipity.
5	Surname	Sorting the number of surnames	Gain a sense of identity in terms of the tribal or clan aspect of the surname in a particular cultural context.
6	Birthplace	Counting the number of birthplaces by province	Marking the map and reflecting the number in shades of color allows for a visual image of the information and strengthens the sense of identity between individuals.
7	Distance	Distance between two birthplaces on the map	Distance is not an obstacle to progress, but rather creates a rare opportunity to come together across geographic barriers, enhancing cohesion among individuals; at the same time, the novelty and curiosity brought about by geographic differences will promote faster integration.
8	Constellation	Constellation distribution, which is a derivative of the month distribution	Finding common interests in popular culture among some groups of people and catering to the desire to explore the mysteries of astrology
9	Ethnicity	Ethnicity distribution	Emphasize respect for the living customs of ethnic minorities, and strengthen the identity of great national unity and Huaxia as one family.
10	First name	High-frequency words in names	Names are unique symbols for each person, and the words contain hopeful and beautiful meanings. Analyzing and summarizing the frequency of these words can reinforce such hopefulness and beauty, and find tacit understanding among groups.
11	Hobby	Hobbies and Interests	Hobbies, which can be either an independent, idiosyncratic individual tendency or the unanimous choice of a group, are the easiest way to form long-term, stable rapport, and the aggregation of teams, bands, and clubs, for example, can easily lead to lifelong friendships.

12	Specialties	Comparison of specialties	Promote positive competition and learning from each other's strengths and weaknesses.
13	Awards	Awards in secondary school	Enhance the sense of achievement of individuals, and motivate the positive atmosphere of the group.
14	Height	Connotation of individual height ranking	Beneficial to the understanding of the group as a whole, the Post 00s generally have the advantage of height, but may also have a negative impact on individual individuals, and need to be used with caution.

Through vivid and interesting copy, color scheme and graphic display of these data analysis, we can get good feedback on the use of the group, effectively enhance the cohesion and visibility of the group, and improve the public attention.

As shown in Figure 1 after clarifying the goal and scope of the portrait, we can start to construct the portrait of the newborn group, which can generally be divided into several steps such as data collection, data preparation, data analysis, data modeling, visualization implementation, browser debugging, etc.

The labeling system required for the study is formed through four layers: demographic labels, generic labels, supplementary labels, and qualitative labels. For the characteristics of the data source and the purpose of the study, the demographic labels mainly contain 11 categories, including resident city, gender, age, education, occupation, marriage, children, pets, assets, and driving preferences on weekdays and holidays. The goal is to portray the basic characteristics of the population and form accurate attribute data.

The goal of generalized labels is to filter out scientific, objective and suitable labels to be included in the calculation. Lookalike model extension based on the core labels of the seed group is the key process to obtain the generic labels.

First, based on the user value map, find the seed population. Find seed populations with extreme characteristics, i.e., the 4 vertices in the above two-dimensional division. Combining the descriptions of the value map and the seed population, the labels of the four seed populations are sketched, as shown in Table 2.

Table 2 Core Distinguishing Labels of Seed Crowds

Tags. / Math.		Appearance		introversion
Group values	Apparel Consumption	Acquisition of luxury brands (Prada, LV, Hermes, etc.)	Passionate about research	Going to tech shows; Acquiring VR glasses: Using geeky apps for a long time (more than two hours a day)
	Content	Deep users of	Enjoyment	Fishing garden more

	Consumption	Meitu and selfie apps		than twice a month; using takeaway software on weekends (more than twice a week); buying instant noodles and instant hot pot online
	Entertainment	access to specific entertainment venues (bars, equestrian clubs); no overnight stays in premium hotels	Self-study and self-enjoyment	Deep users of educational APPs who can't leave home (Tencent Classroom, Netease Open Class, etc.)
Tags. Math.		Seeking new things		Stable
Individual values	Niche Pursuits	Travel to niche countries (Morocco, Czech Republic, etc.)	No overtime	Getting off work at 6:00 on weekdays and not working overtime on weekends and holidays.
	Offline Activities	Frequent business trips (1~2 times per month); work overtime on weekends; use shared office apps (We Work, etc.) with high frequency	Offline activities	Single city; traveling city
	Information Gathering	Heavy user of financial APP (more than two hours per day)	Two points and one line	Always stay at home or go to shopping centers on weekends.

Secondly, the labels of the above new-seeking, stable-seeking, exoteric, and introverted users are used as positive

and negative samples respectively, and the Look-alike model is used to obtain a generic label set, and the eigenvalues in the model are sorted according to their weights. Look-alike results are evaluated using the AUC (Area Under Curve) metric [22], the larger the AUC value, the more likely that the current classification algorithm will rank the positive samples in front of the negative samples, i.e., better classification. The results show that in the selected 300~500,000 users, the intersection of positive and negative samples under the condition of "or" is small, and the AUC is very high, such as the AUC of epistemic and convergent classification > 0.96, and the AUC of new and stable classification > 0.82. The final judgment model distinguishes the positive and negative samples very well, as shown in Table 3.

Table 3 Seed population selection and result evaluation

Math.	Positive and Negative Classification	Number of seed population	Intersection number	AUC
Group values	Externally Expressed	476855	1422	0.96
	introverted	460653		
Individual values	seeking newness	301901	8	0.82
	Stability	315411		

The supplemental tags are selected from the Top tag sets that have a high number of crowd categorization applications in the map engine in order to leverage the historical experience. The Top500 tags from the data source were complemented with the above types of tags, and finally 24 tags were added to the tag system. Qualitative labels were based on the initial population study mentioned above, and 155 labels were identified and finally included in the calculation.

In the end, the total number of labels included in the calculation was 617, of which 606 were behavioral preference labels. The results of the main labels by category and time are shown in Table 4. In addition, if divided by static and dynamic data, among the following four types of data, except for the demographic characteristics labels, which are static, the rest of the labels are dynamic labels combining the user's location information and time information.

Table 4 Labeling system classification and composition

Generic labels (427)		Demographic Characteristics Tags (11)		Supplementary labels (24)		Qualitative labels (155)	
Serial No.	ftr_name	Serial No.	tag	Serial No.	ftr_name	Serial No.	ftr_name
1	action-0_0_0_24-Service Area	1	Resident City	1	app-1_8_24-Map-navigation	1	action-0_0_0_24-Scenic Spots
2	action-	2	Sex	2	app-0_8_24-	2	action-

	0_0_0_24-Highway Exit				Map-navigation		0_0_0_8-Rail Station
3	action-0_0_0_24-Public Prosecution and Law Institutions	3	Age	3	app-0_0_8-Map-navigation	3	action-0_0_0_8-Kindergarten
4	action-0_0_0_24-Convention and Exhibition Centers	4	Education	4	app-1_0_8-Map-navigation	4	action-0_0_8-High school
5	action-0_0_0_24-Churches	5	Occupatio..n	5	app-0_0_8-Audio Entertainment	5	action-0_8_24-Fast food restaurant
…	…	6	Married or not	…	…	…	…
423	arrival-1_8_24-Coach Ticket Vending Center	7	Have children	20	action-0_0_0_24-Public Prosecutors and Law Enforcement Agencies	151	arrival-1_8_24-School-other
424	arrival-1_8_24-Coach station	8	Pets	21	actor-0_8_18-Clothing, Shoes, Hats, and Leather Goods Stores	152	arrival-1_8_24-School-other
425	arrival-1_8_24-Government Offices	9	Asset class	22	action-0_8_18-High school	153	arrival-1_8_24-Mid-range-hotel
426	Arrival-1_8_24-Mid-range hotels	10	Weekday Driving Preference	23	arrival-1_8_24-Commercial and Residential Buildings	154	arrival-1_8_24-Mid-range-hotel
427	Arrival-1_8_24-	11	Holiday Driving	24	action-1_8_24-Chinese	155	arrival-1_8_24-Mid-range-

	Residential neighborhoods		Preference		restaurant		hotel

4.2 Clustering Algorithm Construction

Considering the differences of different age groups, the clustering algorithm is used to cluster the Post-80s, Post-85s, Post-90s and Post-95s, and the final clustering results are combined into the user value map for comprehensive judgment.

The clustering algorithm mainly takes into account the development cost of the algorithm and the computational efficiency of the algorithm. Since the calculation volume reaches hundreds of millions of user data points, in order to improve the calculation efficiency and reduce the cost, this paper chooses the classic K-Means algorithm for clustering, and the core expressions of the K-Means algorithm are shown in Equation (1) to Equation (3).

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{in} - x_{jn})^2} \quad (1)$$

Where, $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ are two n-dimensional data objects. $m(i = 1, 2, \ldots, k)$ are the initial cluster centers; and $d(p, m_i)$ is the distance from each p to k cluster centers.

$$m_k = \sum_{i=1}^{N} x_i \; / \; N \quad (2)$$

Where, m_k represents the cluster center of the k th cluster; N represents the number of data objects in the k th cluster.

$$E = \sum_{i=1}^{N} \sum_{p \in C_i} \left| p - m_i \right|^2 \quad (3)$$

Where, E is the sum of squared errors of all objects; p is the objects in the space; m_i is the mean value of C_i.

For the K-Means algorithm, the selection of K value largely determines the model effect. It is found that the results of initial value taking can be judged by overall interpretability and label interpretability. For overall interpretability, three factors are mainly considered: the highest degree of fit with the pre-initial judgment population; the highest number of newly discovered populations and the conclusion that the newly discovered population is an interpretable population; and the lowest number of unexplainable populations.

Interpretability mainly considers the principle that the demographic characteristics of the clustered-out populations

are consistent. For example, if a group of unmarried men in a category contains a large number of married men, that portion of the data will be excluded as an unexplained result (resulting in a final data result of less than 200 million).

Based on the above judgment principle and process, the selection is made within the range of preliminary values of K-value. The preliminary values for the clustering algorithm are: 17 groups for Post-80s (K=17), 16 groups for Post-85s (K=16), 15 groups for Post-90s (K=15), and 14 groups for Post-95s (K=14), and then groups with consistent characteristics in different age groups are combined for calculation. Since the data source provides a mature clustering algorithm module, the clustering module is called directly, and the aforementioned labeled data are imported into the module for clustering. Considering the server load distribution, the final computation is carried out in the low load period (0:00 to 5:00), and the average computation time for each age group is 1 hour, totaling 4 hours for the 4 age groups. The clustering results are combined with the groups that have the same label performance in different age groups, and finally 20 groups are obtained. The results of other differences in K-value values were analyzed in detail in the evaluation section.

V. EVALUATION OF RESULTS

5.1 User Profiling Results

The results are re-mapped into the user value map according to the dimensions of age, income class, etc., and the results are obtained to form the distribution and proportion of 20 types of user groups according to the combination of group values and generations, individual values and social class (where the height of the rectangle indicates that the crowd covers this income class). The specific categorization is shown in Figure 2.

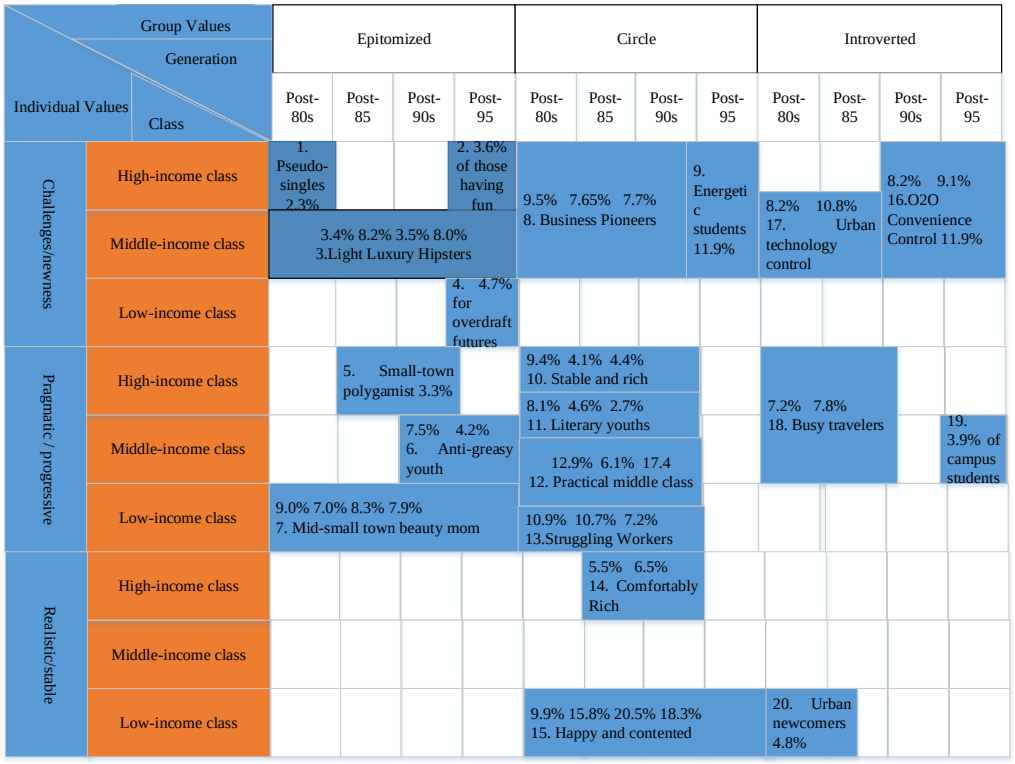

Figure 2 Overview of 20 categories of people

(1) There are seven categories of people whose social attributes are epiphenomenal, named as follows (data rounded to the nearest whole number for ease of understanding, same below): pseudo-single nobility, about 1 million people; fun-loving people, about 1.3 million people; light-luxury hipsters, about 11 million people; overdrawn futurists, about 1.7 million people; small-town gold-drawing women, about 2 million people; anti-grease youths, about 5.9 million people; and small and medium-sized cities' beautiful mothers, about 16 million people.

(2) There are mainly 8 types of people whose social attributes are circle-type, named as follows: business pioneers, about 13 million people; energetic students, about 4.3 million people; stable and rich people, about 9.1 million people; literary youth, about 7.9 million people; affordable middle-class people, about 19 million people; laborers and strugglers, about 15 million people; comfortable rich second-generation people, about 7.1 million people; and joyful and contented people, about 32 million people.

(3) There are five categories of people with introverted social attributes: O2O convenience control, about 8 million people; urban technology control, about 10.2 million people; busy travelers, about 7.9 million people; campus study group, about 1.4 million people; and urban new immigrants, about 2 million people.

In order to further visualize the full display, the demographic characteristics are streamlined, the summary of app usage is transformed into a bar chart and sorted, and the offline activity labels are transformed into a mesh diagram, so that the 20 categories of people represent the current main user profiles.

5.2 Evaluation results

The results of the study were evaluated through two dimensions: vertical and horizontal. Longitudinal mainly evaluates the comparative advantages and disadvantages of the results corresponding to different core parameters; horizontal mainly evaluates the comparative advantages and disadvantages of the research results with other methods.

(1) Longitudinal evaluation of the portrait results. The difference of K-Means results depends on the different values of the core parameter K under the premise of data consistency [23]. Therefore, inter/intra-cluster distances combined with profile distances are mainly used to evaluate the results. Referring to the preliminary cluster classification, the ±2 range of the number of cluster results and related data for the four age groups of Post-80s, Post-85s, Post-90s, and Post-95s were supplemented with calculations, e.g., if Post-85s was 16, the total of five groups of classification results data were calculated as 14, 15, 16, 17, and 18. The number of superior clusters was determined by the inter/intra-cluster distance and contour distance. The calculation of inter/intra-cluster distance is relatively simple and will not be repeated. The contour distance indicates the degree of closeness between the samples of each category and the degree of dispersion between the categories after clustering [24]. Assuming that the data to be classified have been clustered, the data to be classified are divided into n clusters, and for each vector in the clusters, their contour distances are calculated separately. The interval of contour distance is [-1, 1], and the closer to 1, the better the degree of cohesion and separation are, as shown in Equation (4).

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (4)$$

Where $a(i)$ is the average distance from sample i to all samples in the category to which sample i belongs; $b(i)$ is the average distance between sample i and its nearest samples in different categories.

The results of the two indicators are evaluated comprehensively, and the data points with larger inter/intra-cluster distances, inflection points of contour distances, or high proximity to inter/intra-cluster distances are selected as the optimal results with the help of the "elbow method". Therefore, the optimal results were obtained: K-value of 17 after 80, K-value of 15 and 16 after 85, K-value of 15 and 17 after 90, and K-value of 14 after 95. The optimal clustering results were consistent with the previous clustering results, as shown in Table 5.

Table 5 Evaluation results of different parameters for 4 age groups

K-value category	Post-80s	Post-85	Post-90s	Post-95
K(inter/intra cluster distance)	14(0.9520)	14(0.9712)	13(0.9622)	10(0.9041)
	15(0.9639)	15(0.9799)	14(0.9618)	11(09199)
	16(0.9736)	16(0.9921)	16(0.9777)	12(0.9392)
	17(0.9912)	17(0.9990)	16(0.9952)	13(0.9712)
	18(1.000)	18(1.000)	17(1.000)	14(1.0000)
K(contour distance)	14(1.000)	14(0.8311)	13(0.7451)	10(1.0000)
	15(0.7742)	15(1.0000)	14（0.8520）	11(0.8559)
	16(0.8019)	16(0.9551)	15(1.0000)	12(0.8439)
	17(0.9587)	17(0.7119)	16(0.9059)	13(0.9191)
	18(0.7283)	18(0.9013)	17(0.9958)	14(0.9569)

(2) Cross-sectional evaluation of user profiling results. The results of the 2018 NCBS user survey data from the National Information Center, which also uses social and personal values for classification, are selected for comparison. NCBS (New Car Buyer Survey) is a cyclical study of user profiling by the National Information Center, and the related results have been applied to academia and industry [25]. Three key metrics were selected for comparison, including the total data volume covered, the number of populations into which the results are divided, and the number of labels/dimensions parsed. A comparison of the data volume and data results reflected in the results reveals that they are similar in the number of groups into which the population is divided (20 and 18 categories, respectively). However, the results formed based on qualitative-quantitative fusion of big data are richer in terms of data magnitude, preference data, and data depicted by groups, while the questionnaire-based NCBS data are richer in terms of reflecting the basic demographic characteristic information of users, as shown in Table 6. In terms of parsing dimensions, they are all in three main categories: big data portrayal for basic information, online information, and offline information; and NCBS data for basic information, group needs, and competitive relationship information. Overall, the comparison of research results, based on the combination of qualitative and quantitative map engine big data derived from the user portrait can reflect the objective state of the user, and in the total amount of user data and user preference labels on the data is richer, and the visualization results are shown in Figure 3,4,5.

Table 6 Cross-sectional comparison of research results

Key Indicators	Big Data Portrait Based on Qualitative-Quantitative Fusion Approach	User profiling based on survey data
Total data volume/$\times 10^3$	178992 (output)	11
Number of segmented groups	22	19
Number of tags/questions	615	77
Of which: Basic information	12	18
Of which: Preference information	605	58
Parsing dimension broad categories	3	3

Figure 3 Surname word cloud map

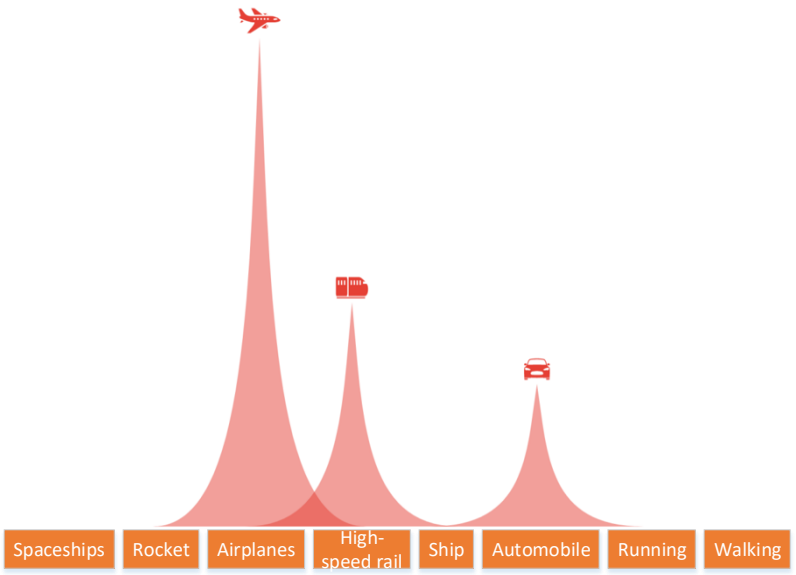

Figure 4 Transportation to Shanghai

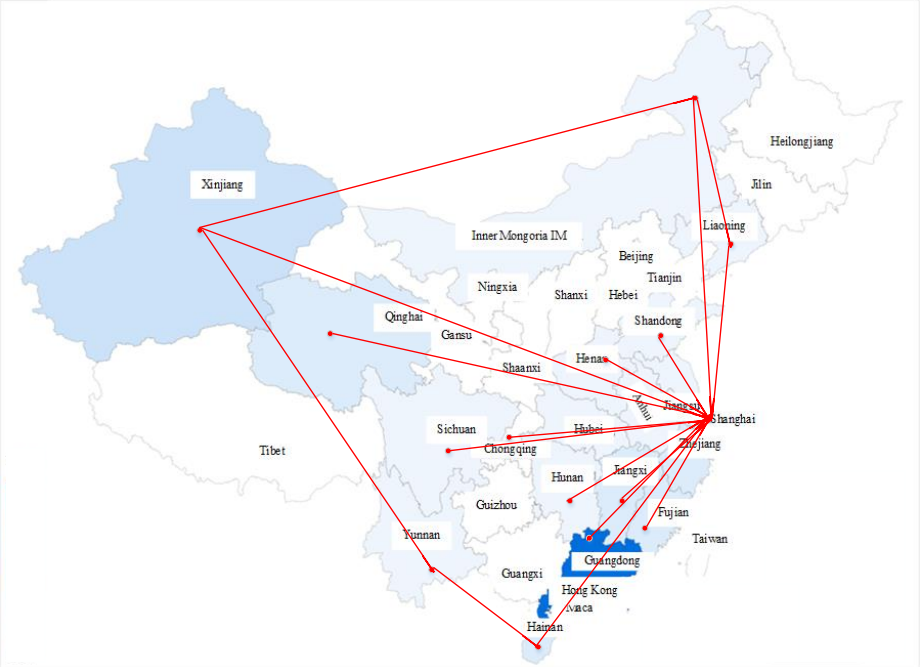

Figure 5 Geographical distance of the student population

Evaluation of user profiling methods. At the methodological level, big data portraits based on qualitative-quantitative fusion methods are also different compared with NCBS survey methods, mainly in the data level and technical level. On the data level, big data portraits include not only static user information, but also dynamic location changes; at the same time, the formation of data is also highly related to the degree of business dataization. On the technical level, the key technology also consists of the business design and algorithm design of the research process especially focusing on user objective data, which is different from questionnaires and quantitative techniques. Details are shown in Table 7. Through the methodology level comparison, we can have a deeper understanding of the differences between big data user profiling and survey data user profiling, and also be able to better allocate resources in the process of big data user profiling and focus on breaking through the key research, and the visualization results are shown in Figures 6, 7, 8,9.

Table 7 Methodology level comparison

Comparison Dimension	Big Data Portrait Based on Qualitative-Quantitative Fusion Approach	User profiling based on survey data
Data Types	Static data combined with dynamic location	Static Data
Data Formation	User Behavior Datamining	Quantitative research
Key Technology	Research process and algorithm design	Questionnaires and interviews
Research Time Consumption	Relatively time-consuming (spreadable)	Time-consuming, mainly in finding research subjects

Research Perspective	Objective user data dominated	Users' subjective responses
Presentation	Dynamic graphics predominant	Tabular data

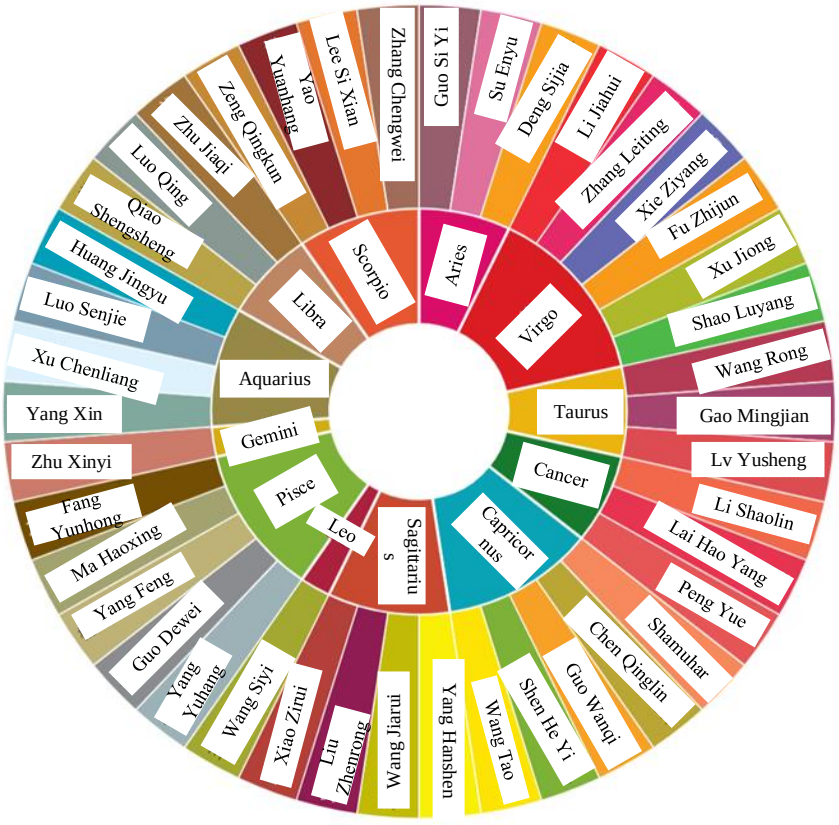

Figure 6 Distribution of constellations

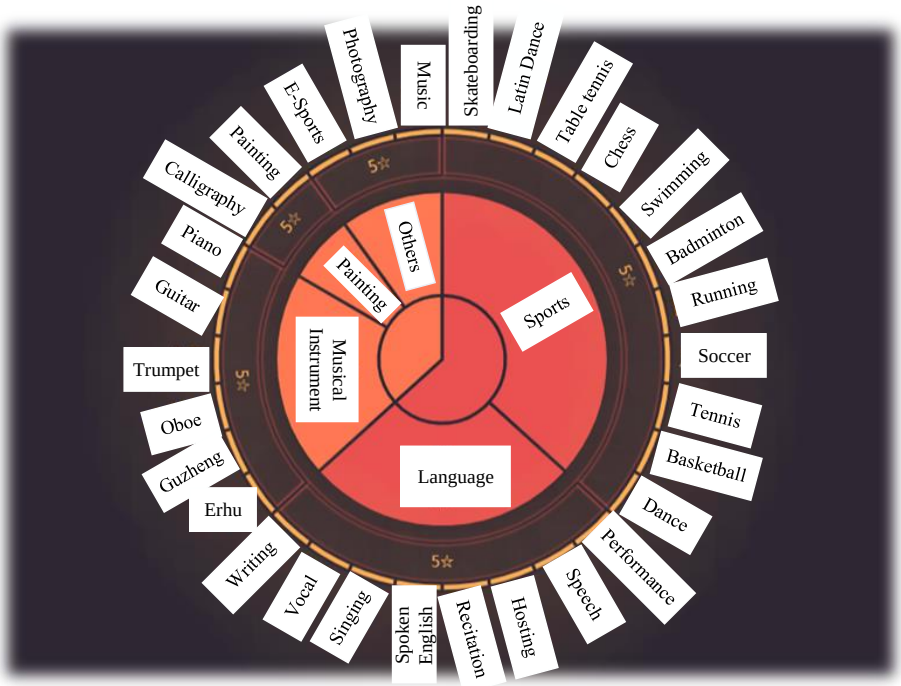

Figure 7 Specialties and Hobbies

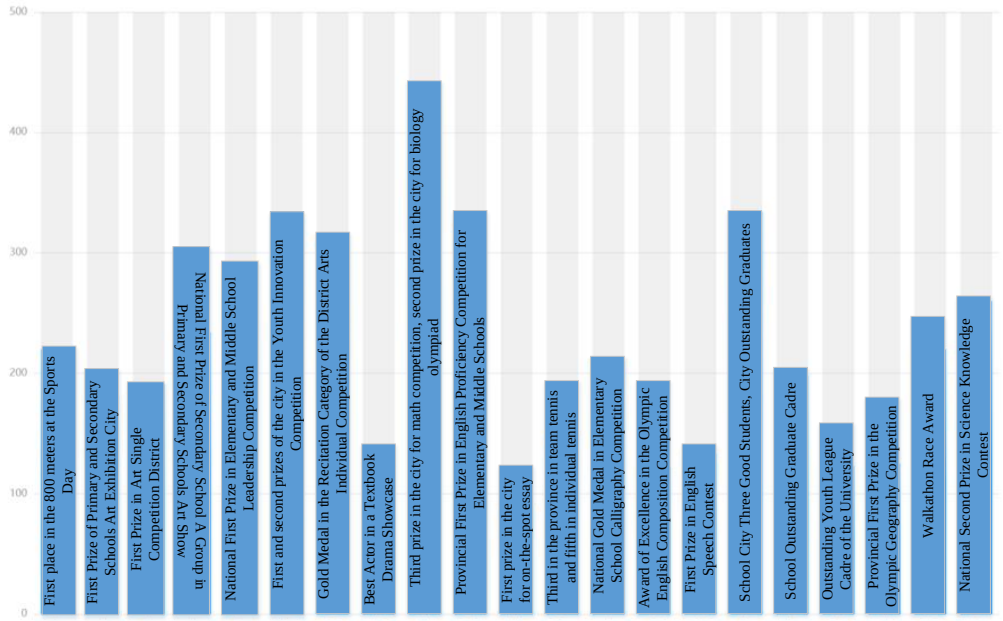

Figure 8 Selected Awards at Secondary Level

Figure 9 Map of birthplace by province

5.3 Deficiencies

First of all, limited to the inadequacy of the freshmen data, the statistical dimension is bound to be relatively single, unable to grasp further information such as economic information, growth history information, life information, health status, social skills, study habits and abilities. It is only a generalized display of the whole picture of the group,

and cannot reflect the depth of personalized revelation.

Secondly, data collection and pre-processing, as well as data file production and other preparatory processes, require too much manual intervention, still need more time and energy to do the initial work, for the busy counselor group is not very friendly, simple data or can be made in the same way, the amount of big data is the amount of big data should be analyzed with the help of big data analysis and processing systems to achieve automated data processing and get the correct results, is the subject of the meaning.

Again, on the choice of visualization platform, it is best to have a more humane and rich optional custom interactive features, Echarts is relatively suitable for lightweight applications, winning in the compact and flexible and foolproof configuration mode, both advantages and disadvantages, but also depending on the project needs. Such as through the virtualization and abstraction of a particular model, you can predict the future state of a certain stage and its development trend is better.

Finally, the presentation should be as compatible as possible with a variety of heterogeneous media, whether cell phones, tablets, smart terminals, or desktops, or Mac OS/Android/windows/Linux and other system-supporting browsers, there should be no obstacles to access. Although Echarts is based on JavaScript application architecture, compatibility should not be a problem, but at present still can not get full support in the cell phone. In order to display the charts perfectly, you need to configure different parameters, which brings troubles to common users [26].

VI. CONCLUSION

This paper determines that the freshmen group portrait has irreplaceable advantages in education and teaching practice by studying the freshmen group portrait and its objectives, connotations and roles; through the various aspects of data collection, cleaning, structuring, data analysis and modeling, the Echarts platform is used to establish a data visualization example, to realize the three-dimensional, multi-dimensional and interactive visual output of the freshmen group portrait and to apply it during the preliminary education and teaching practice. It has been applied in the initial education and teaching practice, and has obtained good results, and has the practical application scenarios that can be replicated and popularized.

Excluding the discussion of the technical level, the drawing of the freshmen group portrait should also be supported by the leadership team, in addition to guaranteeing the smooth access to data, it will also involve part of the data or the existence of privacy and ethical questioning, and should be careful to use the convenience of massive data. The imperfections of the nascent data happen to circumvent some of the ethical dimension risks, but can lead to the existence of problems such as inadequate and unreliable portraits and lack of credibility, which are not related to methodological rigor and data accuracy.

The next step envisages that a digital portrait of individual students with continuity mining can be established under the premise of data accessibility to facilitate in-depth research and form countermeasures on ideological topics such as academic crises, personal behavior, psychological warning, withdrawn behavior, and Internet addiction.

REFERENCES

[1] Hahn, G., Ponce-Alvarez, A., Deco, G., Aertsen, A., & Kumar, A. (2019). Portraits of communication in neuronal networks. *Nature Reviews Neuroscience*, *20*(2), 117-127.

[2] CMS Collaboration cms-publication-committee-chair@ cern. ch. (2022). A portrait of the Higgs boson by the CMS experiment ten years after the discovery. *Nature*, *607*(7917), 60-68.

[3] Sheng, Y., Chen, W., Wen, H., Lin, H. J., & Zhang, J. J. (2020). Visualization research and application of water quality monitoring data based on ECharts. *J. Big Data*, *2*(1), 1-8.

[4] Xue, N., Yao, F., & Xie, Z. (2023). Python-based Epidemic Data Visualization System. *Academic Journal of Science and Technology*, *6*(3), 111-113.v

[5] Straka, A. (2020). Structuring arts-based analysis in portraiture research. *Qualitative Research Journal*, *20*(1), 76-85.

[6] Bruhn, S., & Jimenez, R. L. (2020). Portraiture as a Method of lnquiry in Educatianal Research. *Harvard Educational Review*, *90*(1), 49-53.

[7] Travis, S. (2020). Portrait of a methodology: Portraiture as critical arts-based research. *Visual Arts Research*, *46*(2), 100-114.

[8] Lahman, M. K., De Oliveira, B., Cox, D., Sebastian, M. L., Cadogan, K., Rundle Kahn, A., ... & Zakotnik-Gutierrez, J. (2021). Own your walls: Portraiture and researcher reflexive collage self-portraits. *Qualitative Inquiry*, *27*(1), 136-147.

[9] Sheard, L., & Marsh, C. (2019). How to analyse longitudinal data from multiple sources in qualitative health research: the pen portrait analytic technique. *BMC Medical Research Methodology*, *19*(1), 1-10.

[10] Fry, R., & Parker, K. (2019). A demographic portrait of today's 6-to 21-year-olds, from the Pew Research Center. *Phi Delta Kappan*, *100*(7), 13-16.

[11] Curammeng, E. R. (2023). Portraiture as collage: Ethnic studies as a methodological framework for education research. *International Journal of Qualitative Studies in Education*, *36*(2), 186-202.

[12] Yeh, Y. Y., Nagano, K., Khamis, S., Kautz, J., Liu, M. Y., & Wang, T. C. (2022). Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, *41*(6), 1-21.

[13] Huang, Z. M. (2022). The use of blind-portrait: an opportunity to de-essentialise intercultural, educational research. *Language and Intercultural Communication*, *22*(2), 176-190.

[14] Holappa, A., Lassila, E. T., Lutovac, S., & Uitto, M. (2022). Vulnerability as an emotional dimension in student teachers' narrative identities told with self-portraits. *Scandinavian Journal of Educational Research*, *66*(5), 893-906.

[15] Du, J. (2020). Research on optimization of portrait sculpture data based on 3D image and mobile edge computing. *IEEE Access*, *8*, 224452-224460.

[16] Afiah, N., Arafah, B., & Abbas, H. (2022). Burmese women portrait under the British imperialism in Orwell's Burmese

Days. *Journal of Language Teaching and Research*, *13*(1), 213-219.

[17] Pentón Herrera, L. J. (2021). An Ixil portrait: Exercising resilience amidst inequity,(dis) interest, and self-discovery. *Diaspora, Indigenous, and Minority Education*, *15*(1), 22-33.

[18] Chen, A., Liu, R., Xie, L., Chen, Z., Su, H., & Yu, J. (2022). Sofgan: A portrait image generator with dynamic styling. *ACM Transactions on Graphics (TOG)*, *41*(1), 1-26.

[19] Sulistiyo, U., Haryanto, E., Widodo, H. P., & Elyas, T. (2020). The portrait of primary school English in Indonesia: policy recommendations. *Education 3-13*, *48*(8), 945-959.

[20] Sun, T., Barron, J. T., Tsai, Y. T., Xu, Z., Yu, X., Fyffe, G., ... & Ramamoorthi, R. (2019). Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, *38*(4), 1-12.

[21] Khorob, S. (2022). VASYL STEFANYK'S PUBLICIST WORKS: THE PORTRAIT DETAILS. *PRECARPATHIAN BULLETIN OF THE SHEVCHENKO SCIENTIFIC SOCIETY Word*, (16 (63)), 264-278.

[22] Yaniv, J., Newman, Y., & Shamir, A. (2019). The face of art: landmark detection and geometric style in portraits. *ACM Transactions on graphics (TOG)*, *38*(4), 1-15.

[23] Duong, B. H., & Silova, I. (2021). Portraits of teachers in neoliberal times: projections and reflections generated by shadow education research. *Globalisation, Societies and Education*, *19*(5), 696-710.

[24] Shang, Y., & Wong, H. C. (2021). Automatic portrait image pixelization. *Computers & Graphics*, *95*, 47-59.

[25] Lu, Y., Chai, J., & Cao, X. (2021). Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, *40*(6), 1-17.

[26] Jingchun Zhou, Jiaming Sun, Weishi Zhang, Zifan Lin. Multi-view underwater image enhancement method via embedded fusion mechanism. Engineering Applications of Artificial Intelligence, 2023, 121, 105946.