Based on CHAID's Knowledge Base Uses Expectations and Intelligent Recommendations

Kuotai Tang[1(✉)], Hailing Duan[2], Lin Fang[1], and Xiaowei Huang[2]

[1] Department of Business Management, Fujian Polytechnic of Information Technology, Fuzhou, China
`3224243499@qq.com`

[2] School of Internet Economics and Business, Fujian University of Technology, Fuzhou, China

Abstract. Knowledge base is an important part of cultural soft power, and the expected use of knowledge data is often the concern of knowledge base operators, and also the expectation of college students when using it. In this paper, a questionnaire was constructed based on the technology acceptance model (TAM). A total of 319 valid questionnaires were received. Cronbach $\alpha = 0.965(>0.9)$, $KMO = 0.576(>0.5)$, indicating that it has the significance of in-depth research. In this paper, the CHAID algorithm is used to mine the data and explore the matching mode between college students' data usage demand and usage expectation. The results show that the use of knowledge base is often reflected in the acquisition of skills and knowledge ($P = 0.001$); Most students are more dependent on paid knowledge ($P = 0.002$); Real-time information apps are favored by college students ($P = 0.005$); Video knowledge is not only favored by college students but also willing to share ($P = 0.023$). Finally, it is suggested that knowledge base operators can strengthen the application of intelligent recommendation, and match real-time according to user characteristics to achieve the purpose of knowledge service.

Keywords: CHAID · TAM · *Knowledge Base*

1 Introduction

Knowledge base is an important part of cultural soft power. With the continuous development of Internet technology, the expected use of knowledge data is often the issue that knowledge base operators are concerned about, and it is also the use expectation demand of college students. The Technology Acceptance Model (TAM) shows that users perceive the usefulness of the new technology for their work, and users perceive the ease of using the new technology. However, we do not know the specific content of college students' perception of usefulness and ease-of-use in knowledge base.

This study tries to understand the situation of college students' use of knowledge base, as well as the specific content of the perception of usefulness and ease-of-use of knowledge base, and attempts to explore the matching mode of college students' data use needs and usage expectations by means of intelligent recommendation. This paper

first establishes a questionnaire through literature discussion, and then uses CHAID algorithm to mine data to bring us the enlightenment of college students' expected use of knowledge base, and finally makes a discussion to provide reference for relevant industry.

2 Literature

2.1 Forms of Knowledge

Knowledge form refers to the internal thinking mode and expression mode of knowledge, and refers to the knowledge form (thinking mode) and expression structure (knowledge form) of argumentation process. The form of knowledge includes the way of thinking, the way of testing and the way of expression of knowledge, and the form contains human wisdom and ability. Learning and mastering the form of knowledge is an important way to transform knowledge into wisdom and knowledge into energy [1].

Knowledge refers to the practical experience and spiritual wealth accumulated by people in production, life and practice for a long time. Knowledge form is the form and method of knowledge production, testing and expression. Paul Hurst (1993), a British educator, proposed the concept of "forms of knowledge". He believed that human knowledge has eight forms (mathematics, natural science, humanities, history, religion, literature, philosophy and morality), each of which has a different concept, logic and expression [2].

2.2 TAM

Technology Acceptance Model (TAM) is a model proposed by Davis (1993) to study users' acceptance of information systems by using the theory of rational behavior. Davis posits that whether a user uses a new technology is determined by the user's Behavioral intention (BI). The willingness to use is determined by the user's Attitude towards using the system (A). ATM has two determining factors: one is Perceived usefulness (PU), which refers to whether the user's role brought by the new technology is helpful to his work or life, and whether the performance is improved as a result. Another is Perceived ease-of-use (PEOU), which refers to the subjective degree of how easy it is for users to use the new technology. Finally, these two factors are also influenced by some external variables [3].

2.3 CHAID Decision Tree

CHAID (Chi-Square Automatic Interaction Detection) is a classification algorithm proposed by Mr. Koss in 1975. Case proposed the CHAID algorithm in 1980, which was based on the target variable, classified each attribute according to the basic node of the variable, and calculated its square value [4]. The CHAID algorithm is a method that combines machine learning and statistics to classify samples according to pre-set decision tree classification criteria. CHAID algorithm is designed to solve the decision problem of multiple independent variables and single factor. In the case of multiple variables, this

method has high computational efficiency [5]. The CHAID system adopted in this paper is not limited by explanatory variables and will not be interfered by multiple collinear, and the analysis results are clear and intuitive.

CHAID determines explanatory variables and response variables, that is, dependent variables and independent variables, based on sample data, takes Chi-square value statistics as branch criteria [6], constantly merges and splits explanatory variables and response variables in the process of comparison with the original set merging level, and classifies them until the condition of stopping the growth of decision tree is met, and decisions are made [7].

In this paper, SPSS Modeler machine learning software is used to model the decision tree, the key is to choose the best "node". In the process of node splitting, the samples contained in the branch nodes should belong to the same category as much as possible, so that the "purity" of the node is higher. The three commonly used feature selection methods are as follows:

1) Information Gain: According to the principle of maximum Information gain, starting from node 0 with feature split, Makes the "Information Gain = Original Information - split Information." D_p D_n: Total information; N_p N_n: Total number of data; D_j: jth split information; N_j N_j: jth number of data:

$$IG(D_p, f) = I(D_p) - \sum_{i=1}^{m} \frac{N_j}{N_p} I(D_j) \tag{1}$$

2) Entropy is often used to measure the amount of information gain, In this $P(i|t)$: Probability of attribute; c: category; t: total number of categories:

$$I_H(t) = -\sum_{i=1}^{c} P(i|t) \log_2 P(i|t) \tag{2}$$

3) Chi-square Test: When the observed value is exactly the expected value, $x^2 = 0$; The observed value is closer to the expected value, the less different on two sides, then the smaller the x^2 value is; On the contrary, the bigger the difference between the observed value and the expected value, the bigger the difference between the two, the bigger the x^2 value. A_i is the observed value of i, n is the total number, and P_i is the expected value of i. The formula is as follows:

$$x^2 = \sum_{i=1}^{k} \frac{(A_i - np_i)^2}{nP_i} \tag{3}$$

3 Method

This paper first discusses the literature, collects and classifies the literature, and selects important factors such as free, useful, easy to use and social value demand according to TAM theory. Push time, too much information and real-time information are the main reasons for reducing customers' use of knowledge base apps. Secondly, a model of influencing factors of college students' knowledge acceptance is established.

3.1 Perceived Usability Variables

Scholars believe that factors affecting perceptual use include network externalities, social effects, system characteristics, interface characteristics, individual innovation and individual differences (domain knowledge, self-efficacy) [8]. Mobile phone reading is beneficial to their work and study, and easy to use, can improve the interest in mobile phone reading; However, users believe that the easier mobile phone reading is to use, the better their cognition of mobile phone reading [9].

3.2 Perceived Usefulness Variables

Scholars have pointed out that users will decide whether to use any new thing according to their usage habits and psychological tendency, and users will be influenced by mass media and the surrounding environment, thus affecting their cognitive usability and ease of use [10]. The factors affecting perceived usefulness are summarized in this paper, including network externalities, social effects, system characteristics, information quality, etc.

3.3 Research Model

This paper establishes a model of influencing factors of college students' knowledge acceptance and explains the influencing factors of knowledge base acceptance variables, as shown in Table 1. Among these factors, the perceived effectiveness depends on the ease of use and external factors, and the ease of use depends on external factors; External variables mainly include system design characteristics, user characteristics, task characteristics, development nature, policy impact, organizational structure and so on. A questionnaire survey was conducted on college students and Exploratory Factor Analysis (EFA) was carried out to identify the essential structure of various observed variables and conduct dimensionality reduction processing. The complex variables were summarized into a few core factors, as shown in Table 1.

3.4 Questionnaire Design

The purpose of the questionnaire is to understand the factors influencing college students' use of knowledge base. The questionnaire is divided into five parts: basic information, habit, feeling, technology and willingness, with a total of 24 questions and 176 sub-items. The questionnaire was distributed online from March 28–22 to April 18, 2022, targeting college students. A total of 325 questionnaires were collected through online questionnaire delivery. Secondly, verify and inspect the data and delete the redundant information, that is, delete the redundant information in the invalid questionnaire and the valid questionnaire to ensure the validity of the data. In this paper, the criteria for determining invalid questionnaires are as follows: (1) Respondents choose the questionnaire of "unused knowledge base"; (2) Respondents who fill in the questionnaire for less than 20s; (3) Questionnaire with the same questionnaire options; (4) Questionnaires with missing options; (5) Eliminate overseas IP; (6) non-college students. Through screening and sorting out the invalid questionnaires, 319 valid questionnaires were obtained, with an effective rate of 98.15%. Subsequently, the questionnaire data were sorted out and analyzed.

Table 1. Model study variables

Category	Variable
Basic attribute	There are differences in the lifestyle and characteristics of users based on gender, age and education.
Habit	Frequently use APP users' reading habits, knowledge base preferences, communication and communication methods.
View	Users' suggestions on the management of knowledge base and the improvement of knowledge base APP.
Expect	Knowledge base user's expectation of knowledge base display form and communication content.
	Users can use knowledge base APP better for its features and supporting technologies.
Will	Survey of users' willingness to read and share the content and types of articles to provide users with better knowledge intelligent recommendation.

3.5 Reliability and Validity

The data obtained by the questionnaire in this paper mainly focus on the influencing factors of college students' use of knowledge base, and the two methods of KMO (Kaiser-Meyer-Olkin) value and Bartletts sphericity test are used for structural validity test analysis. Cronbach α was used for reliability analysis. SPSS results showed that the overall reliability was 0.964, higher than 0.9, indicating that the reliability of the survey data was very high. KMO = 0.58 (>0.5), the validity is acceptable, the data analysis with this set of questionnaires can be carried out, and the significance level P value of Bartletts sphericity test is 0.000 (<0.05), indicating that there is a significant difference between the correlation matrix and the identity matrix, as shown in Table 2.

4 Results and Discussion

After the statistical analysis of the data, this paper uses the decision CHAID decision tree algorithm to deeply explore the significance of the data.

The research finds that the use of knowledge base by college students often lies in the acquisition of skills and knowledge. Most students rely more on paid knowledge, apps with real-time information are more favored by college students, and college students prefer video knowledge and are more willing to share it. The explanation is as follows:

Table 2. Cronbach Alpha and KMO

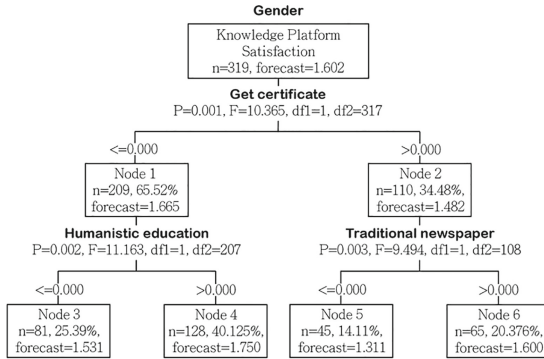| Items	Validity Test		Reliability Statistics	
	KMO	Bartlett's Test P-value	Cronbach α	Cronbach's α based on standardized items
176	0.58	0.000	0.922	0.924

Fig. 1. Focus on skills credentials decision tree

4.1 Focus on Skills Credentials

Taking gender as the target variable, there is a significant correlation with Get certificate knowledge (P = 0.001), and the proportion of male is slightly higher than that of female (52% > 48%), which is the biggest influencing factor of the model. 34% of the respondents wanted to obtain the knowledge of textual examination, 20% paid more attention to traditional newspapers (P = 0.003), and 14% did not pay attention to them. 65% of them do not want to obtain the knowledge of the certificate examination, 40% of them are more concerned about the humanities education (P = 0.001), 25% are not concerned.

The discussion in this section has inspired us that among college students who are more in need of knowledge about life application skills (such as driving license and skill certificate), 20% of them still come from traditional newspaper media, and 40% of them still maintain a high interest in humanities education, as Fig. 1.

4.2 More Trust in Paid Knowledge

Taking paid reading as the target variable, the most significant correlation is reading knowledge base articles (P = 0.002), reading knowledge base articles accounts for 26%, and being more willing to share public culture articles accounts for 19.43%. 73.35% read other types of articles, and 40.12% hope that articles can provide decisions (P = 0.008). Among them, 46.7% are willing to pay 0–100 yuan a year, followed by 300–500 yuan accounted for 21%, and free readers accounted for 19.7% (P = 0.020).

According to the research and discussion in this section, 57.7% of college students are more likely to trust paid articles and are willing to pay less than $500 per year. They are more willing to share cultural articles and are more likely to read non-professional articles., as Fig. 2.

4.3 Real-Time Information Attracts Users

The willingness to publish was taken as the target variable. The more significant correlation was between knowledge base APP with real-time information (P = 0.005), which accounted for 60.5%, and regular users of Weibo (P = 0.026) accounted for 40.12%,
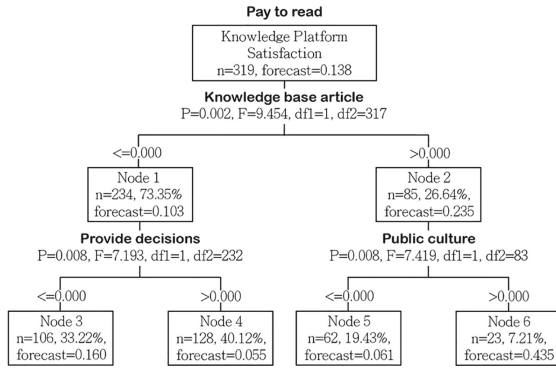
Pay to read

Knowledge Platform
Satisfaction
n=319, forecast=0.138

Knowledge base article
P=0.002, F=9.454, df1=1, df2=317

<=0.000

>0.000

Node 1
n=234, 73.35%
forecast=0.103

Node 2
n=85, 26.64%,
forecast=0.235

Provide decisions
P=0.008, F=7.193, df1=1, df2=232

Public culture
P=0.008, F=7.419, df1=1, df2=83

<=0.000

>0.000

<=0.000

>0.000

Node 3
n=106, 33.22%,
forecast=0.160

Node 4
n=128, 40.12%,
forecast=0.055

Node 5
n=62, 19.43%,
forecast=0.061

Node 6
n=23, 7.21%,
forecast=0.435

Fig. 2. More trust paid knowledge decision tree

among which 28.52% wanted to share videos through social apps (P = 0.023). Those who do not frequently use Weibo do not want to be published in professional journals (P = 0.001), accounting for 16.61%, and those who hope to do so account for 3.76%. 39.4% are less concerned about real-time information, significantly related to the knowledge base APP with political views (P = 0.003), will be more willing to use 20.06%, willing to use 19.43%, as Fig. 3.

According to the research and discussion in this section, college students prefer to use the knowledge base APP with real-time information. Those who frequently use Weibo are willing to share and share knowledge in video form through social apps, and they are less likely to be published in professional journals. Users who do not care about real-time information and want articles with political views will prefer to use them.
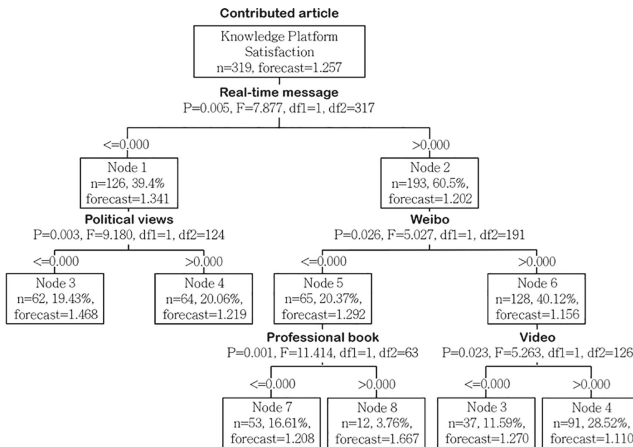
Contributed article

Knowledge Platform
Satisfaction
n=319, forecast=1.257

Real-time message
P=0.005, F=7.877, df1=1, df2=317

<=0.000

>0.000

Node 1
n=126, 39.4%
forecast=1.341

Node 2
n=193, 60.5%,
forecast=1.202

Political views
P=0.003, F=9.180, df1=1, df2=124

Weibo
P=0.026, F=5.027, df1=1, df2=191

<=0.000

>0.000

<=0.000

>0.000

Node 3
n=62, 19.43%,
forecast=1.468

Node 4
n=64, 20.06%,
forecast=1.219

Node 5
n=65, 20.37%,
forecast=1.292

Node 6
n=128, 40.12%,
forecast=1.156

Professional book
P=0.001, F=11.414, df1=1, df2=63

Video
P=0.023, F=5.263, df1=1, df2=126

<=0.000

>0.000

<=0.000

>0.000

Node 7
n=53, 16.61%,
forecast=1.208

Node 8
n=12, 3.76%,
forecast=1.667

Node 3
n=37, 11.59%,
forecast=1.270

Node 4
n=91, 28.52%,
forecast=1.110

Fig. 3. Real-time information attracts users to the decision tree

5   Conclusions

According to the above analysis, we can get some habits and preferences of college students for knowledge base platform, through these habits and preferences, we can make intelligent recommendation to them. According to the decision tree analysis, the intelligent recommendation is obtained: boys tend to understand the knowledge of certificates, so the knowledge base platform can recommend the relevant content of the knowledge of certificates to male users; College students have more trust in paid knowledge, we can recommend more valuable knowledge to users with higher education, and charge a certain fee; People who have the habit of reading current political articles hope that the knowledge base can provide public cultural knowledge, so we can extract the frequency of users' reading current political articles to make knowledge recommendation; Real-time information attracts users to the knowledge base, so the knowledge base platform can recommend the latest news to users.

References

1. Pan Hongjia, Knowledge Form: Basic Implication, Educational Value and Teaching Strategy [J], 2014 (11): 40–45, https://doi.org/10.19877/j.carolcarrollnkiKCJCJF.2014.11.
2. Paul Hurst, Liberal Education and the Nature of Knowledge, Beijing: People's Education Press,1993.
3. Davis F D. User acceptance of information technology: system characteristics, user perceptions and behavioral impacts[J]. International Joumnal of Man-Machine Studies, 1993, 38(3):475–487.
4. Liu Yan, Fan Yu, ZENG Zili., Research on final paper grades based on CHAID algorithm in Modeler environment [J]., Computer programming skills and maintenance, 2018(01):67–68, https://doi.org/10.16184/j.cnki.com.prg.2018.01.
5. Qin Guolan., Empirical research on risk segmentation of engineering project cost control based on Decision tree CHAID algorithm [J]., Project Management Technology, 2019,19(08):136–140.
6. Sun Xu., Research on Human Resource Recommendation Based on Decision Tree Algorithm [J]., Microcomputer Applications, 2021, 37(07):140–143.
7. Che Minshi, Nie Chunyan, Fan Rujun et al., An emotion recognition method based on chaotic features and optimized CHAID decision tree [J]., Applied Research of Computer, 2020, 37(S2):105–107.
8. Ming Jun-Ren, GUO Cai-qiang, ZHANG Jun., A review of research on mobile Library user perception behavior based on technology acceptance model in China [J]., Library Forum, 2017, 37(07):125–131.
9. Han Chaoqun, Yang Shuiqing, Cao Yuzhi., Integration model of mobile service users' Adoption behavior: An empirical Study based on mobile reading [J]., Soft Science, 2012, 26(03):134–139.

10. Wang Shuang., Research on user acceptance Model of mobile library [J]., Information Science, 2013 (4):39–44.