Academic Blog User Profile Model Construction and Empirical Study: A Case Study of ScienceNet Blog Postprint

Authors: Yuan Run, Wang Qi

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] User persona theory can be employed to label the behavioral characteristics of academic communities, providing a basis and reference for the precise identification of users, precision marketing on academic social platforms, and improving user experience during cold-start periods.

[Method/Process] Programs were developed using Python and R to acquire and process publicly available user behavior data. A conceptual model for user personas was constructed from five dimensions: basic blog attributes, activity level, authority, blog post influence, and interest preferences. Empirical research was conducted using user behavior data from ScienceNet blogs as a case study.

[Results/Conclusions] Specific indicators and calculation methods for characterizing features of academic blog users were proposed, demonstrating that the user persona model possesses certain theoretical significance and application value for the management and operation of academic social platforms.

Full Text

Construction and Empirical Study of User Portrait Model of Academic Blog: Taking ScienceNet Blog as an Example

Yuan Run[1], **Wang Qi**[2] [1]Library of Jiangsu University, Zhenjiang 212013 [2]Institute of Science and Technology Information, Jiangsu University, Zhenjiang 212013

Abstract: *[Purpose/Significance]* User portrait theory can be used to characterize the behavioral features of academic groups, providing a basis and reference for precise user identification, targeted marketing of academic social platforms, and improving user experience during cold start periods. *[Method/Process]* Using Python and R language programs to obtain and process publicly available

user behavior data, this study constructs a user portrait conceptual model from five dimensions: basic blog attributes, activity level, authority, blog post influence, and interest preferences. Taking ScienceNet Blog user behavior data as an example, an empirical study is conducted. *[Result/Conclusion]* The paper proposes specific indicators and calculation methods for characterizing academic blog user features, demonstrating that the user portrait model has theoretical significance and application value for the management and operation of academic social platforms.

Keywords: academic blog; user portrait; R language; case analysis

1. Introduction

The concept of user portrait (User Profile) was first proposed by A. Cooper, the "father of interaction design"[1]. A user portrait is a virtual representation of real users and a target user model built upon a series of authentic data. Real data primarily refers to user information data, including static data (relatively stable user attribute data) and dynamic data (continuously changing user behavior data). User portraits extract tags for individual or groups of users based on user attributes and behaviors, providing a structured description of user information. The process of constructing user portraits is also a process of understanding users. By abstracting concrete information data into user characteristics, we can accurately target user groups and predict both their actual and potential needs.

Currently, user portraits are widely applied in precision marketing [2], intelligent recommendation [3-5], product development [6], and other fields, employing methods such as statistics [7], Bayesian networks [8], machine learning [9-11], topic models [12], clustering analysis [4], hierarchical classification [13], and others. Existing research includes many user portrait studies based on social network platforms. A. Raghuram et al. [14] proposed an efficient supervised machine learning method to classify Twitter users into six interest categories. R. Jiamthapthaksin et al. [15] constructed a user interest feature model based on Naive Bayes, ANN, and SVM, and validated its effectiveness using Facebook datasets. Han Meihua et al. [16] combined user portraits with reading therapy, analyzing Weibo texts to calculate depression sentiment indices to obtain user portraits. Wang Lingxiao et al. [17] took Zhihu as an example to construct user portraits from four aspects: user qualifications, participation, answer quality, and development trends. Liu Haiou et al. [18] deeply mined social platforms such as QQ groups, Tianya Forum, and Renren Network to construct user portrait models. Cui Chao et al. [19] proposed ideas for data collection and model construction for knowledge community user portraits based on user portrait theory and the relationship between knowledge and users. Yu Chuanming et al. [11] conducted deep learning on user posting content in stock forums, combining follower data of stock forum users, which has certain academic value and

practical significance.

Social network platforms can be divided into academic and non-academic social platforms based on whether they are specifically designed for academic communication. Existing user portrait research focuses on non-academic social platforms, with less attention paid to academic social platforms. The user groups of academic social platforms are researchers interested in scientific work. They use platforms to create personal information, publish research results, and conduct academic exchanges, achieving knowledge exchange, dissemination, and sharing, embodying the open, shared, and collaborative concepts of Science 2.0. This paper takes ScienceNet Blog as an example, summarizes the obtained user attribute and behavior data into five dimensions of indicators: basic blog attributes, activity level, authority, influence, and topic preference, and conducts empirical research on academic blog user portraits to supplement and improve existing research.

2. Academic Blog User Portrait Model

Blogs and blog posts are two main entities on network social platforms. Any scholar can register with real name on ScienceNet to become a blogger and participate in academic exchanges by publishing posts. All network behaviors of bloggers—such as registration, publishing, categorization, tagging, reading, recommending, commenting, downloading, citing, visiting, messaging, and establishing friend relationships—are recorded by the platform. This paper refers to these records as user behavior data. The richer the user behavior data, the more precisely we can characterize user features. However, due to limitations in collection costs, technology, and privacy protection, some user behavior data is difficult to obtain. To conduct empirical research, this paper collected 20 data items covering major user behavior data, as shown in Table 1 .

User behavior data is divided into numerical and character types. To utilize this data for academic blog portraits, this paper proposes a five-dimensional user portrait model (UPM), expressed by formula (1) as follows:

$$\mathrm{UPM} = \{B, V, Q, I, G\}$$

where B represents basic attributes, V represents activity level, Q represents authority, I represents blog post influence, and G represents interest preferences.

2.1 Basic Attributes This paper treats research field (a_3) as a basic blog attribute, which can accurately describe the user' s discipline area. Additionally, objective parameters such as user ID, education, and title are static or relatively stable over a period and are publicly available data that can be obtained through platform public channels. The title and education information filled in by blog users during registration are uniformly categorized as "Title" (data item a_4).

Due to poor standardization of this data, this paper performed hierarchical quantification processing, with the method shown in Table 2 .

Thus, basic blog attributes can be expressed as:

$$B = \{\text{EduPos}, \text{ResF}\}$$

2.2 Activity Level The activity level indicator (V) is positively correlated with blog behavior data including number of blog posts, activity points, shares, topic posts, replies, and online duration. This paper defines the entropy weight value of these six data items as the blog activity level indicator, expressed by formula (3):

$$V = \sum \omega \cdot \alpha$$

where α represents the normalized value of behavior data items a_5-a_{10}, and ω is their weight coefficient. The normalization calculation formula is as follows:

$$\alpha = \frac{a}{a_{\max}}$$

The entropy weight method utilizes the uncertainty of information provided by each data item to determine its weight, making it suitable for various assignment problems. The method for calculating the weight coefficient of behavior data items is shown in formula (5):

$$\omega_i = \frac{1 - e_i}{\sum_{i=1}^{n}(1 - e_i)}$$

where e_i is the information entropy of behavior data item i, calculated as shown in formula (6):

$$e_i = -\frac{1}{\log(n)} \sum_{j=1}^{n} p_{ij} \log(p_{ij})$$

In formula (3), the top 25% of indicator V values are defined as high activity group (H), top 50% as medium activity group (M), top 75% as normal activity group (C), and the rest as low activity group (L). Threshold settings are determined by platform management according to actual needs, and different activity groups can be obtained by adjusting them.

2.3 Authority Authority is primarily influenced by the user' s academic status and blog content authority. Data item a_4 contains user education and title information that can reflect the user' s academic status. Blog content authority can be measured by blog content dissemination coverage. As the number of people following blog dynamics increases, the dissemination speed of blog posts accelerates, and blog authority increases [22]. Dissemination coverage is positively correlated with blog friend count, homepage visits, featured blog posts, and total recommendations. This paper defines the entropy weight value of these five data items as the blog authority indicator, expressed by formula (7):

$$Q = \sum \omega \cdot \beta$$

Formula (7) can process data items and determine weight coefficients by referring to formulas (4)-(6). The threshold division standard for Q indicator is the same as for V indicator.

2.4 Blog Post Influence Blog post influence can be quantified as the ability of blog content to change other users' thoughts and behaviors (reading, recommending, commenting, etc.) [22]. Zhang Xiaoyang et al. [23] and Zheng Chao et al. [24] expanded the application scope of the h-index, evaluating academic blog influence based on blog post reading counts and comment counts, comprehensively considering both quality and quantity of blog content. Building on previous work, this paper further improves the blog post influence evaluation index system, quantifying blog post influence from two perspectives: blog post reading count and blog post interaction count (comment count and recommendation count).

Based on the inference of the h-index, we define observed statistics for blog post reading count (c) and blog post interaction count (q):

$$c = \sqrt{a_{15}} \quad \text{and} \quad q = \sqrt{a_{17}^2 + a_{18}^2}$$

The mathematical formula for the h-index is as follows:

$$h_c = \max\{r_1 : r_1 \leq c\}; \quad h_q = \max\{r_2 : r_2 \leq q\}$$

where r_1 is the rank of blog posts in descending order of observed quantity c, and r_2 is the rank of blog posts in descending order of observed quantity q. In practical application, it was found that the h-index has the same value and relatively low value levels. Due to the social attributes of academic blogs, user behavior data is sparse, making the above phenomenon more significant.

To address the shortcomings of the h-index, Jin Bihui et al. [25] proposed the R-index. The R-index is the square root of the total citation frequency of papers in the h-core, and its measurement results can effectively distinguish the same

h-index value without changing the shape of the h-core. The mathematical formula for the R-index is as follows:

$$R = \sqrt{\sum_{j=1}^{h} c_j}$$

where c_j represents the citation frequency of the j-th paper in the h-core, and $c_j \geq h$. Here, c_j represents the reading count or interaction count of the j-th blog post in the h-core and $h \in \{h_c, h_q\}$. Using the h-index and R-index in combination can effectively compensate for the shortcomings of the h-index and better evaluate and distinguish blog post influence, expressed by formula (12):

$$I = \{(h_c, R_c), (h_q, R_q)\}$$

where the top 25% of h_c values are defined as high reading influence group (H), top 50% as medium reading influence group (M), top 75% as normal reading influence group (C), and the rest as low reading influence group (L). Similarly, the top 25% of h_q values are defined as high interaction influence group (H), top 50% as medium interaction influence group (M), top 75% as normal interaction influence group (C), and the rest as low interaction influence group (L).

2.5 Interest Preferences Interest is a tremendous driving force for people' s activities. Interest preference represents the rational, 倾向性 choices users make when creating blog posts. Academic blog interest preferences reflect scholars' academic interest directions and can typically be described by several thematic words (keywords). ScienceNet Blog platform provides two categorization approaches for blog posts: "system category" and "personal category." When bloggers categorize their posts into a system category, they can further subdivide them using personal categories. If system categories are treated as one type of node and personal categories as another type of node, these two types of nodes form a bipartite network relationship.

Generally, if system category phrases are denoted as V_1 and user category phrases as V_2, the bipartite network is denoted as:

$$G = (V_1, V_2, E)$$

For a given undirected graph $G = (V, E)$, in this paper, the vertices V are a_{19} (V_1) and a_{20} (V_2). Clearly, $V = V_1 \cup V_2$, $V_1 \cap V_2 = \emptyset$, and $\forall e = (u, v) \in E$, we have $u \in V_1$, $v \in V_2$, satisfying the bipartite network condition.

This paper uses the R language `bipartite` package to create a blog "system category - personal category" bipartite network, uses the `computeModules` function to divide network communities, and extracts classification phrases by weight sorting to describe blog interest preferences.

3. Empirical Analysis of Academic Blog User Portraits

3.1 Data Source and Acquisition Under the premise of emphasizing user privacy protection, this paper uses Python language to write programs to collect academic blog user behavior data. The collection targets users with featured blog posts. User URL collection was conducted on December 12, 2018, collecting 3,799 non-duplicate user URLs. Before collecting blog post data, simple manual processing was performed on the original URL data, removing 146 URLs with missing data due to privacy settings and other factors. Blog post data items were obtained on December 19, 2018.

During the crawling process, two original datasets were constructed: `BlogUsers` and `BlogContents`. After collection, necessary processing was performed on the collected data: blog-related data in `BlogUsers` was based on the actual data obtained from `BlogContents`, with centralized statistics for total reads, total recommendations, and total comments; users with extreme anomalies or large amounts of missing data in blog post data were removed. Finally, 2,339 valid user data and 437,832 blog post data were obtained. Among the 400,000 collected blog post records, the cumulative number of valid comments reached over 3.13 million, valid recommendations over 2.83 million, and total reads exceeded 1.429 billion, indicating that ScienceNet Blog has significant influence and is meaningful for academic exchange and dissemination.

3.2 Results Calculation and Analysis This paper uses self-written R language functions to calculate the weight coefficients of each data item for indicators V and Q, with results shown in Table 3 and Table 4 .

The entropy weight method determines weights based on data dispersion. Greater data dispersion means more information content, less information uncertainty, smaller information entropy, and correspondingly larger weight coefficients. As can be seen from Tables 3 and 4, the data items contributing most to indicators V and Q are share count and total recommendations, respectively, while the least contributing data items are activity points and title. This phenomenon indicates that blog share count and total recommendation count have greater dispersion compared to other data items, contain the most information with the least uncertainty, while activity points and title show the opposite data characteristics. As the management party, ScienceNet has always encouraged users to generate and share various content to build ScienceNet Blog into an active academic exchange community. The above weight distribution characteristics can better identify active users and authoritative users who continuously generate and share content.

Based on the above weight coefficients, the blog activity level and authority can be calculated using the quantitative model, with partial results shown in Table 5 and Table 6 .

Affected by data normalization, the values of activity level indicator and authority indicator fall within the [0,1] range. From the calculation results, the overall value levels of both indicators are relatively low. Analysis shows that this phenomenon can be mainly attributed to three reasons: (1) According to the characteristics of the entropy weight method, if data items with larger weight coefficients (such as share count and total recommendations) have poor overall user behavior data performance, it may result in low indicator values. (2) As an informal academic exchange platform, academic blog user behavior is highly random and uncertain. Influenced by user preferences and platform function settings, a small number of users with missing single indicators or outstanding performance may cause low indicator values. (3) In actual use, only a small portion of users actively use various functions and continuously contribute high-influence blog posts to the platform.

Observing the calculation results and calculating the correlation between activity attributes and authority attributes ($r = 0.484$), it was found that some users with higher activity levels also have relatively higher authority. This phenomenon indicates that active use of the blog platform helps enhance authority, and blogs with higher authority are relatively active on the platform. The standardized values of blog post influence calculation results based on the h-index and R-index concept are shown in Table 7 .

From the calculation results in Table 7, the h-index overcomes the mathematical logic flaws of simple summation, and the R-index compensates for the defect that the h-index cannot distinguish blog influence when values are the same. The R-index serves as a supplementary indicator, only requiring comparison of R-index values when h-index values are the same. After calculation and observation, although there is a significant positive correlation between h_c index and h_q index ($r = 0.743$), there are still certain differences between them. Blogs with high h_c index do not necessarily have high h_q index. Therefore, using them in combination can more comprehensively evaluate blog post influence.

Taking user A (ID=1557) as an example, this paper generates a system-personal category network, as shown in Figure 1 [Figure 1: see original paper]. Figure 2 [Figure 2: see original paper] shows the clustering results based on the `computeModules` function. Blog interest preferences are represented by classification phrases.

Academic blogs are one of the main forms of informal academic exchange. From Figures 1 and 2, it can be seen that user A' s blog content is divided into five categories by interest preference: research, science popularization, learning, teaching, and life, simultaneously meeting A' s academic and social needs. After weighting and sorting by weight, the categories with the most blog posts are scientometrics research and daily life. Scientometrics research is mainly presented in forms such as blog information, opinion reviews, research notes, overseas observations, and paper exchanges. Based on community division results, classification phrases "scientometrics, daily life" can be extracted as blog interest preference tags.

Based on the above research, this paper randomly selects three users (A, B, and C with IDs 1557, 5430, and 287179) as examples, and uses the user portrait model UPM to obtain academic blog user portraits as shown in Table 8 . Table 9 shows the evaluation thresholds for each indicator.

From Table 8, it can be seen that differences in research fields affect blog interest preferences to a certain extent, especially academic interest points. Users A and C have accumulated high authority and significant influence through long-term, continuous contribution of quality content on the blog platform, and can be considered quality users of ScienceNet Blog, while user B' s performance is relatively ordinary.

The results of these user portraits can serve multiple purposes. In precision marketing scenarios, platform administrators can identify users with high, medium, and normal characteristics in different dimensions based on operational needs and user portrait results, and carry out differentiated marketing to enhance the platform' s core competitiveness. By combining user research fields and interest preference tags, the platform can recommend quality blogs and posts in respective fields to users, especially new users, in a targeted manner, achieving the goals of precision recommendation and improving user satisfaction during cold start periods. Users can also directly search for tags of interest to find information resources of relevant users, laying a foundation for establishing friend relationships, academic exchanges, and knowledge sharing, and improving user satisfaction with academic blog platform services.

---

4. Conclusion

Conducting user portrait research on online academic social platforms represented by academic blogs has certain academic value and practical significance. This paper selects ScienceNet Blog user behavior data as the research object, builds an academic blog user portrait model from five dimensions—basic attributes, activity level, authority, blog post influence, and interest preferences —based on user portrait theory, and demonstrates representative user portrait examples. The paper proposes several methods for characterizing academic blog user features: (1) selecting the entropy weight method to determine data item weight coefficients; (2) improving the blog post influence evaluation index system from two perspectives of blog post reading and interaction, maintaining methodological consistency with scientometric evaluation of formal literature exchange, and using the R-index to compensate for the h-index' s inability to distinguish same-value situations; (3) generating a system category-personal category weighted bipartite graph and dividing communities based on the bipartite network relationship between system categories and personal categories to extract blog interest preference tags. Through blog user portraits, the platform can effectively identify user characteristic differences, serve platform precision marketing, and improve cold start period user experience.

Due to dataset limitations, this paper did not extract user topic preferences from blog post content and did not consider the impact of time on topic preferences and other indicator features. In future research, the author will incorporate the concept of time domain to analyze user behavior data features in different time windows, extract user topic preferences based on blog post content, and obtain more meaningful academic blog platform user portraits.

References

[1] COOPER A. About face 3: the essentials of interaction design[C]//John Wiley & Sons, 2007. [2] Zeng Hong, Wu Suni. Precision marketing based on big data user portraits from Weibo[J]. Modern Economic Information, 2016(16): 306-308. [3] Ran Deng. Research on personalized content recommendation for mobile game users based on user portraits[D]. Xi' an: Xi' an University of Technology, 2018. [4] Li Bing, Wang Yue, Liu Yongxiang. Application of user portraits and intelligent recommendation based on K-means in big data environment[J]. Modern Computer (Professional Edition), 2016(24): 11-15. [5] Bi Runfang. Research on personalized movie recommendation based on SVR collaborative filtering and user portrait fusion[D]. Zhengzhou: Zhengzhou University, 2018. [6] Yu Mengjie. Data modeling of user portraits in product development—from concrete to abstract[J]. Design Art Research, 2014, 4(6): 60-64. [7] XU G, ZHANG Y, ZHOU X. Towards user profiling for Web recommendation[J]. Lecture notes in computer science, 2005, 3809: 415-424. [8] Zhang Xiaoke, Shen Wenming, Du Cuifeng. Research on Bayesian networks in user portrait construction[J]. Mobile Communications, 2016, 40(22): 22-26. [9] Zhou Meixuan. Research on user portraits based on deep neural networks[D]. Changsha: Hunan University, 2018. [10] Yu Chuanming, Tian Xin, Guo Yajing, et al. Research on user portraits based on a behavior-content fusion model[J]. Library and Information Service, 2018, 62(13): 54-63. [11] Xin Juqin, Jiang Yan, Shu Shaolong. Personalized recommendation based on integrated user preference model and BP neural network[J]. Computer Engineering and Applications, 2013, 49(2): 57-60. [12] Ma Chao. Analysis method for social network user portraits based on topic models[D]. Hefei: University of Science and Technology of China, 2017. [13] Yao Yuan. User portrait construction method based on ontology[C]//China Computer Users Association Network Application Branch. Proceedings of the 22nd Annual Conference on New Network Technologies and Applications. Beijing: Beijing Union University Beijing Information Service Engineering Key Laboratory, 2018. [14] RAGHURAM M A, AKSHAY K, CHANDRASEKARAN K. Efficient user profiling in twitter social network using traditional classifiers[EB/OL]. [2019-05-20]. https://doi.org/10.1007/978-3-319-23258-4_{35}. [15] JIAMTHAPTHAKSIN R, AUNG T H. User preferences profiling based on user behaviors on Facebook page categories[C]//International conference on knowledge & smart technology. Chonburi, Thailand: IEEE, 2017: 248-253. [16] Han Meihua, Zhao Jingxiu. Research on reading therapy model

based on user portraits—taking depression as an example[J]. Journal of Academic Libraries, 2017, 35(35): 110. [17] Wang Lingxiao, Shen Zhuo, Li Yan. Construction of user portraits in social Q&A communities[J]. Information Theory and Practice, 2018(1): 129-134. [18] Liu Haiou, Sun Jingjing, Zhang Yaming, et al. Research on information dissemination behavior of user portraits in online social activities[J]. Information Science, 2018, 36(12): 17-21. [19] Cui Chao, Luo Ou. Implementation ideas for user portraits in scientific research knowledge communities[J]. Information and Communications Technology and Policy, 2018(6): 75-78. [20] Chen Tiange. Research on factors influencing brand selection based on social media user portraits[D]. Guangzhou: South China University of Technology, 2018. [21] Zhou Wenjing. Research on user portrait construction methods for campus forum user interests[D]. Beijing: Beijing University of Posts and Telecommunications, 2018. [22] Wang Chen. Research on academic blog influence evaluation[D]. Taiyuan: Shanxi University of Finance and Economics, 2018. [23] Zheng Chao, Chen Feng. Correlation analysis between scientist blog h-index and scientist h-index[J]. Library Science Research, 2013(3): 53-57. [24] Zhang Xiaoyang, Li Xiaoliang. Evaluation and correlation analysis of scientist blog h-index[J]. Library and Information Service, 2010, 54(2): 66-69. [25] Jin Bihui, ROUSSEAU R. R-index, AR-index: supplementary indicators for h-index function expansion[J]. Science Focus, 2007(3): 1-8. [26] Liu Chen, Ji Li, Tang Li. Research on product review feature-opinion word pair extraction based on bipartite network central node identification[J]. Computer Systems & Applications, 2018, 27(11): 9-15. [27] Wan Guo, Zhang Guiping, Bai Yu, et al. News topic sentence extraction based on feature weighting[J]. Journal of Chinese Information Processing, 2017, 31(5): 120-126.

Author Contributions: Yuan Run proposed the research idea, designed and conducted the experiments, and revised the paper; Wang Qi collected experimental data, conducted experiments, and wrote the paper.

Note: Figure translations are in progress. See original paper for figures.