Internet Usage of Chinese University Students: A Novel Pattern Recognition Method Based on Large-Scale Log Analysis (Postprint)

Authors: Yan Chengxi, Wang Jun, Wang Ke

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

Abstract

[Purpose/Significance] Thoroughly mining and accurately understanding the daily online behavior patterns of Chinese university students not only holds significant theoretical importance for advancing the fields of user behavior and information retrieval, but also possesses potential social value and practical significance in enhancing enterprises' personalized services and information recommendation capabilities for university student users. [Method/Process] This paper proposes a novel method for identifying behavior patterns of university student users based on large-scale log analysis. The method includes a semi-supervised learning algorithm called "MaxMatching" based on deep learning and text analysis techniques, as well as a clustering model that hybridizes two types of feature entropy (Shannon entropy and true entropy). [Results/Conclusion] Empirical results demonstrate that the proposed method not only possesses certain advantages in algorithmic performance and result interpretability, but also can summarize and present comprehensive patterns of Chinese university students' online behavior from three dimensions: network usage capability, access temporality, and thematic preference. The method and conclusions effectively expand the methodological system for semantic understanding of query terms in the information retrieval domain, and also provide certain references and feasible recommendations for enterprises to enhance personalized information recommendation services for university student users.

Full Text

Preamble

ChinaXiv Cooperative Journal
Vol. 63, No. 14, July 2019

Chinese College Students' Internet Use: A New Method of Pattern Recognition Based on Large-Scale Log Analysis

Yan Chengxi, Wang Jun, Wang Ke
Department of Information Management, Peking University, Beijing 100871

Abstract

[**Purpose/Significance**] Deeply mining and accurately understanding the daily online behavior patterns of Chinese college students holds substantial theoretical significance for advancing user behavior and information retrieval research, while also offering potential social value and practical implications for improving personalized services and information recommendation capabilities for college student users. [**Method/Process**] This paper proposes a novel method for identifying college student user behavior patterns based on large-scale log analysis. The method includes a semi-supervised learning algorithm called "MaxMatching" that leverages deep learning and text analysis techniques, along with a clustering model that hybridizes two types of feature entropy (Shannon entropy and real entropy). [**Result/Conclusion**] Empirical results demonstrate that this method offers advantages in both algorithmic performance and result interpretation, and can comprehensively characterize Chinese college students' online behavior patterns from three dimensions: network usage capability, temporal sequence of access, and thematic propensity. This method and its conclusions effectively expand the methodological framework for semantic understanding of queries in information retrieval and provide valuable references and feasible recommendations for enterprises seeking to enhance personalized information recommendation services for college student users.

Classification Number: G250
Keywords: Chinese college students; online behavior; pattern recognition; large-scale log analysis

1. Introduction

The development of information technology has made the Internet an integral part of people's daily lives. According to the 2018 statistical report from the China Internet Network Information Center (CNNIC) [1], Internet users aged 20-29 account for 30% of China's total netizen population, accessing the Internet through various channels including PCs and mobile devices. Those with college degrees or higher represent 21.1% of the user base. Unlike other age groups, college students—predominantly the post-90s generation—are more receptive to new cultures, ideas, and technological influences in this era of continuous innovation, which is reflected in their daily online behavior patterns and thematic preferences, such as strong search capabilities, pursuit of subcultures and gaming culture, and extensive use of social media. Therefore, understanding and identifying useful user patterns, detecting meaningful events, assessing potential

risks, and informing strategic decision-making hold profound social significance [2]. Large-scale web log analysis represents an effective technique for analyzing the macro-structures and micro-features of different user groups through data mining and machine learning algorithms applied to massive records of user online behaviors [3].

In this context, this study proposes a new method for identifying college student user behavior patterns based on large-scale log analysis. This method includes a semi-supervised learning algorithm called "MaxMatching" based on deep learning and text analysis, and a clustering model that combines two types of feature entropy (Shannon entropy and real entropy) to address two research questions: (1) How can we perform semantic understanding and accurately identify user intent and thematic preferences from query terms in log data? (2) How can we understand the online behavior patterns of different college student groups by comprehensively considering their network usage capability features, temporal features, and thematic features?

2. Literature Review

Existing research has explored various aspects of college students' information search and usage behaviors, thematic preferences, and socio-psychological changes, including online music usage [4], Internet usage behavior and psychological factors [5-6], and learning-oriented search behaviors [7]. M. Madden et al. found that college students enjoy downloading and listening to music, as well as online chatting and socializing, but rarely for entertainment purposes [8]. Zhang Pengyi et al. investigated Chinese college students' personal information management practices on mobile devices, discovering that an increasing number of students use smartphones for information storage, with nearly half accessing call logs, photos, social media, email, notes, clocks, and work or personal documents [9]. Wu Dan et al. focused on follow-up behaviors triggered by mobile searches among college students, recording 15 days of smartphone usage data from 30 students in an uncontrolled experimental setting, and conducted a comprehensive qualitative and quantitative analysis combined with structured diaries and interview data. The results revealed three types of follow-up behaviors: continuous searching, shopping decision-making, and information sharing, with most follow-up actions occurring within one hour after the initial search session. Most participants adopted different strategies based on the App used—only when search results met their needs would they proceed with shopping and sharing; otherwise, they would use different Apps or modify query terms for further searches [10].

Query term semantic understanding based on log analysis has long been a research focus in computer and information science. The most classic user intent classification method is A. Z. Broder's INT taxonomy [11], which includes "informational," "navigational," and "transactional" intents. O. Alonso et al. found through crowdsourced annotation of queries that informational queries account for over 90% of current query terms [12]. C. Gonzalez-Caro et al. classified query

intent into “informational,” “not-informational,” and “ambiguous” categories, and proposed a multi-faceted query intent classification method based on N. J. Belkin’s search task context theory, dividing user query intent into nine facets including type, topic, task, objectivity, concreteness, scope, authority sensitivity, spatial sensitivity, and temporal sensitivity, where the task facet relates to the resource type of the query term [13-14].

Recent studies have recognized the importance of multi-dimensional query features for deep user query intent classification tasks, such as result records [15], query length [16], part-of-speech and position features of query terms [17-18], and mouse browsing features [19]. R. V. Pujeri et al. argued that query ambiguity generally arises from overly short search terms that fail to provide sufficient background knowledge [20], consistent with H. Cao et al.’s view that “queries need to be context-aware” [21]. J. Teevan et al. constructed a Bayesian dependency network classification model using multi-query features including result quality clarity, click entropy, and query term attributes (character length, URL inclusion, geographic information inclusion), achieving approximately 80% classification accuracy [22]. Additionally, semi-supervised query classification based on taxonomy [23-24], LDA topic modeling [25-26], and deep neural network models [27-28] have achieved notable research results in recent years. S. Dou et al. proposed a taxonomy-based classification algorithm using query term mapping bridges, employing the Open Directory Project (ODP) as an intermediate taxonomy to obtain candidate relationships by maximizing a matching score function between query terms and facet vocabularies, followed by SVM classification modeling. Experiments showed this method improved F1 and Precision metrics by approximately 3% and 9% respectively compared to the first-place algorithm in the ACM KDD CUP 2005 competition [29]. T. Konishi et al. noted the strong sparsity assumption limitation of LDA models and incorporated pairwise topic co-occurrence relationships into the topic model, proposing a Pairwise Coupled Topic Model (PCTM) [30]. This model uses collapsed Gibbs sampling based on pairwise topic co-occurrence probabilities for each word to address sparse associations between topics, achieving 3% higher precision than traditional models like LDA. Guo Cheng et al. combined HowNet and the ATF*PDF model to propose an unsupervised topic mining model for query terms, which effectively identifies important thematic words with low frequency [31]. B. Wu et al. used clicked and skipped pages as positive and negative feedback document sets respectively, combined content and position embedding vectors of pages, and constructed a deep feedback memory network with attention mechanisms, achieving optimal performance in query suggestion tasks and query intent identification tasks of varying lengths and sessions [32]. Other query intent identification methods such as query sub-intent decomposition [33-34] and key entity recognition [35-36] also demonstrate advantages for specific tasks.

Based on the analysis and summary of the above research, we argue that: (1) Existing studies on college students’ online behavior primarily employ local questionnaire surveys or user interviews, focusing on mobile device usage behavior

and psychological factors. Due to limitations in small-scale data volume and device scenarios, their conclusions may suffer from bias (data bias and subjective cognitive bias). (2) In user intent identification analysis, most research remains coarse-grained extensions based on "Broder's user intent classification," with few conducting multi-dimensional online behavior pattern analysis combining time, theme, and behavioral hierarchies for specific groups (college student users). As Y. K. Seock stated, "The penetration of the Internet into college students' lives has changed their behaviors, habits, and preferences, not just device usage patterns" [37]. (3) Accurately understanding query semantics and user intent in log analysis remains a challenge in information retrieval and pattern recognition. Current query expansion methods based on taxonomy, topic models, or deep learning are primarily supervised learning algorithms that rely heavily on large amounts of high-quality annotated training samples, while also being computationally complex and difficult to implement, making their high cost prohibitive for small and medium-sized enterprises to deploy in practice. Therefore, designing simple yet effective unsupervised (weakly supervised) learning models to accurately identify Chinese college students' online behavior patterns is a worthwhile exploration.

3. Research Methodology

3.1 Research Framework

This section describes the framework and steps of the proposed method for college student user behavior pattern recognition based on large-scale log analysis, as shown in Figure 1 [Figure 1: see original paper]. We first designed a prototype navigation website for college student users' online information needs (see Figure 2 [Figure 2: see original paper]), deployed it within campus gateway services across multiple provinces in China, and collected college students' web log records including login time, clicked URLs, and search queries. Considering college student information needs and interest preferences, we constructed a thematic classification table based on query terms and log records, and performed manual semantic indexing of website URLs.

For search query data, this study designed a semi-supervised matching learning algorithm called MaxMatching based on external corpus knowledge and machine learning theory to achieve thematic mapping and transformation of query words, which were then merged with thematically transformed URL records. By introducing feature entropy composed of a "time-behavior-theme" triplet, this study represents and extracts features of users' online behaviors and implements college student user group behavior pattern recognition based on a clustering analysis model.

3.2 Navigation Website Design and Deployment

To obtain daily online log data from college student users, this study selected websites from Alexa's 2016 rankings that cover nine aspects of college students'

daily online life ("eating," "playing," "entertainment," "earning," "chatting," etc.), filtered them (76 websites total), and constructed a user-friendly navigation website. In partnership with network service providers, the navigation website was deployed across more than 20 provinces and 79 cities nationwide, including Hubei, Jiangxi, Guangdong, Zhejiang, and Hebei, covering nearly 150 universities. To ensure stable and high usage rates, the network service providers embedded the website (including data usage privacy agreements) at the entrance of campus network gateway systems. College students at these institutions could see the website immediately after logging into the campus network and freely use or close the navigation service. Additionally, based on our statistics of usage records and frequency across different student groups, the network providers offered internet package discounts and promotions to high-frequency users to encourage broader usage.

This study selected one full year of user data from March 10, 2017, to March 10, 2018, as the dataset, comprising over 400,000 log records from more than 3,500 users. Statistical analysis of this dataset revealed that the website's operational metrics—unique visitor count (UV) and page view count (PV)—reached daily averages of 36,897 and 73,727 respectively, with a conversion rate maintained at a relatively high level of approximately 3%. Given the data collection coverage and user engagement, we believe this dataset can to some extent represent Chinese college students' online behavior. We collected website visit data through JavaScript tracking on the navigation site and the authoritative third-party platform Baidu Statistics, and wrote automated R scripts for scheduled data downloads stored in a local database (see Figure 2). To limit the sampling scope, we filtered user identity character identifier segments (the first three digits of uid) in Python scripts to confirm that visitors were college student groups (including undergraduate and graduate students).

3.3 User Log Preprocessing

The user log preprocessing stage consists of two main components: (1) Data cleaning of existing logs, including removal of invalid query terms, filtering of erroneous and missing user attribute fields, and elimination of error entries, resulting in 347,387 records from 3,550 users. The data fields include six attributes: user account (uid), access date (date), click or search behavior time (acttime), website login time (logintime), user behavior type (type), and item (item). User behavior types include searching (search) or link clicking (link), while items include query terms (query) or website URLs, as specifically shown in Figure 2. (2) Based on existing research and preliminary investigations, we thematically classified college students' online preferences and intentions, manually indexing websites for semantic mapping and providing conceptual vocabulary for different categories as seed word sets for subsequent query term semantic matching. The thematic classification includes the categories shown in Table 1 .

As mentioned in the literature review, most query terms are short and often con-

sist of non-standard natural language with multiple ambiguities or newly coined terms, such as "84," "Running Man," and "Dragon Ball," making them difficult for computer systems to understand. This study similarly adopts a query expansion strategy, introducing metadata expressions from high-ranking search engine results as background semantic knowledge for query terms, primarily selecting the top 10 records. A. Malik et al.'s research indicates that for most users (particularly college students), they are satisfied with only the top 10 or so web pages returned by search engines [38]. To achieve reasonable semantic representation of these returned records' metadata, this study introduces distributed embedding representation of word vectors, specifically the Word2Vec deep neural network model [39] for open corpus pre-training, and then designs a new semi-supervised heuristic matching algorithm called MaxMatching for thematic classification and identification of query terms.

In the pre-training stage, we crawled nearly 13,000,000 text resources (130GB) from Baidu Baike, Sohu News, and Sogou corpora, trained word vectors using jieba segmentation and the CBOW model (window size = 5, minimum word frequency = 5, vector dimension = 64), and obtained an ultra-large dictionary covering 6,100,000 word vectors.

3.4 Feature Entropy Representation

This study constructs a triplet representation comprising behavior, time, and thematic features, denoted as <'behavior', 'temporality', 'topicality'>. Previous research has demonstrated that users with different network search capabilities exhibit differences in clicking and searching habits. For example, D. Tabatabai et al.'s findings show that users with poorer search skills tend toward impatient trial-and-error strategies, leading them to select and click navigation links before spending sufficient time on evaluation and planning [40]. R. Mihalcea proposed the concept of network competence to describe and characterize features of user search behavior, such as preferences for ICT tool usage [41]. Based on this, this study measures this network capability (or behavioral usage preference) as SCratio using the ratio of search frequency in access records, as shown in formula (2):

$$SCratio = \frac{SearchingNum}{ClickingNum + SearchingNum}$$

Information entropy (Shannon entropy, SE) essentially characterizes the uncertainty of random variables—the more uncertain we are about information content, the more information is needed to clarify it. Similarly, if users' selection probabilities across different thematic categories are similar, their information entropy value will be larger, reflecting the absence of clear thematic preferences. Here we use it to measure thematic specificity (topicality). Since user access times have sequential order, using information entropy alone to measure temporal features poses problems. A. Barabasi et al. proposed the concept of actual

entropy (AE) to effectively address sequential entropy prediction issues [42]. This study uses AE to calculate users' access behavior time sequences to determine the degree of order or regularity in user access. A larger AE indicates more irregular (disordered) user access behavior.

Assuming P(x_j) is the occurrence probability of theme x_j, ϕ_i represents the shortest substring length starting at time sequence position i that has not appeared in positions 1 through i-1, Z and n represent the number of independent user-accessed theme categories and sequence length respectively. The calculation methods for SE and AE are shown in formulas (3) and (4). Note that this study adopts a 24-hour interval with 15-minute intervals as the time sequence segmentation standard—for example, 00:00-00:15 is recorded as time point 1, 00:15-00:30 as time point 2, yielding 96 time intervals total.

$$SE = -\sum_{j=1}^{Z} P(x_j) \cdot \log(P(x_j))$$

$$AE = \left(\sum_{i} \phi_i\right)^{-1} \cdot \ln(n)$$

4. Empirical Results

4.1 MaxMatching Algorithm Evaluation

The MaxMatching algorithm aims to map query terms to given themes, and its quality significantly affects subsequent clustering model accuracy. Since this method calculates based on metadata text obtained through query expansion strategies, parameter settings directly impact MaxMatching. Therefore, this study considers two important parameters: (1) Number of records returned by Baidu (NTP), with a range of [1, 10]; (2) Metadata keyword extraction algorithm (SKE), for which we evaluate three common text feature extraction methods: frequency, TF-IDF, and TextRank [43]. Additionally, we randomly selected 2,000 query terms and assigned them to seven annotators for manual labeling (the category with maximum probability from manual annotation is considered the query's thematic category). To demonstrate MaxMatching's advantages, we used a rule-based matching algorithm [44] as a baseline for comparison.

Figure 4 [Figure 4: see original paper] shows that the "TF-IDF + W2V" based MaxMatching algorithm is optimal, achieving 84.76% accuracy with the best parameter NTP_{best} = 3. This demonstrates that MaxMatching, combined with Word2Vec deep learning-based query expansion, is more efficient and accurate than traditional rule-matching methods for identifying user search intent and thematic preferences.

4.2 Optimal Clustering Model

This study employs two classic clustering models (K-means & DBSCAN) for user feature clustering and evaluates model quality using silhouette coefficient (SC) [45], as shown in Figure 5 [Figure 5: see original paper]. We examined DBSCAN parameters (scanning radius Eps and minimum neighborhood size Msn) and K-means cluster numbers. Testing revealed that K-means generally outperforms DBSCAN, with the highest SC value achieved when the number of clusters = 3, indicating optimal clustering performance.

4.3 Clustering Result Analysis

The clustering model yields three distinct user groups (cluster0, cluster1, cluster2). Overall, the vast majority of college student users have low search engine usage frequency (73.21% of SCratio values below the mean of 0.15, see Figure 6 [Figure 6: see original paper]), with nearly half (50.79%) having never used search engines. Analysis of cluster centroids and means (see Figure 6) shows that cluster0 has high feature entropy values (both SE and AE) with SCratio values uniformly distributed across the [0, 1] interval; cluster1 has the lowest SE but highest AE, with SCratio in a low range (95% of users' SCratio in [0, 0.3]); cluster2 has the lowest AE, average SE (SE_{mean} = 0.73), and similarly low SCratio (95% of users' SCratio in [0, 0.2]). To determine statistical significance, we conducted Mann-Whitney U non-parametric rank-sum tests (due to unequal variances). Table 2 shows that despite similar AE centroid means between cluster1 and cluster0, significant differences exist in entropy features (SE and AE) across the different user groups.

To better illustrate feature entropy characteristics in temporal (temporality) and thematic (topicality) dimensions, we projected the three user groups onto clock visualizations and thematic distributions, as shown in Figure 7 [Figure 7: see original paper]. The distributions of thematic indicator SE and temporal indicator AE reveal clear clustering distinctions among the three college student groups, demonstrating good discriminative power of entropy features. Users in cluster2 with the smallest AE values exhibit the most regular online behavior, accessing the Internet during specific periods: 13:15-13:30, 17:15-18:00, 19:15-19:30, and 21:45-22:00 (green segments). However, the other two groups show significantly longer online activity periods, covering approximately one-third of the entire day (12:00-02:30 and 16:00-22:30), indicating disordered network access patterns that are difficult to predict precisely. Notably, cluster1 shows significantly higher average access intensity (65.3 visits per time period) compared to cluster2 (38.3) and cluster0 (28.2). In terms of thematic preference distribution, cluster1 with the lowest SE value exhibits clear thematic specificity, showing significantly stronger preference for video live streaming ("Livevideo") than other categories (red segments). Although the other two groups also use video live streaming more than other categories, the differences are less pronounced, particularly for cluster0, which shows no significant preference for any particular theme.

Based on the above triplet feature analysis, we identified three segmented patterns of college student user groups in daily online life: Conjoint-Utilizing Users (CU), Single-Utilizing Users in Disorder (SUD), and Single-Utilizing Users in Order (SUO), corresponding to cluster0, cluster1, and cluster2 respectively. The specific group characteristics are detailed in Table 3 .

Table 3. Segmented Characteristics of College Student Groups

College Student Group	Network Usage Capability (behavior)	Access Temporality (temporality)	Thematic Propensity (topicality)
Conjoint-Utilizing Users (CU)	Use both URL link clicking and searching for information acquisition; strong network usage capability	Disordered access patterns; long, sustained online activity periods; high activity intensity	Diverse thematic preferences; no significant thematic specificity
Single-Utilizing Users in Disorder (SUD)	Primarily clicking behavior; rarely use search tools; weak network usage capability	Disordered access patterns; long, sustained online activity periods; low activity intensity	Strong, singular preference for video content
Single-Utilizing Users in Order (SUO)	Primarily clicking behavior; rarely use search tools; weak network usage capability	Highly ordered and regular access patterns; very short, sustained online activity periods; low activity intensity	Diverse thematic preferences; relatively weak preference for video content

5. Discussion and Conclusion

This study constructs a novel method for online user behavior pattern recognition using large-scale web log data from Chinese college students, examining three dimensions: behavioral (network capability), temporal (temporality), and thematic (thematic specificity). The method's core components include the "MaxMatching" matching algorithm based on deep learning and query expansion strategies, and multi-dimensional feature entropy measurement algorithms. Algorithm evaluation results show that, compared to traditional rule-matching methods, this algorithm performs excellently in identifying user intent from query terms, offering theoretical and practical value for expanding and enriching user intent understanding tasks in information retrieval—representing a methodological contribution.

The introduction of multi-dimensional feature entropy provides a new perspective for understanding college students' online behavior patterns. Empirical results indicate: (1) When using comprehensive navigation websites daily, college student users rarely utilize internal search tools and components (e.g., search boxes), preferring navigation link functions (clicking popular website links). This reflects that college students' daily online lives are not heavily dependent on search tools. The possible reason for this "surprising" phenomenon lies in the simple and clear intentions of college students using navigation websites. For example, when seeking flight and travel information, users immediately think of Ctrip and Qunar; when purchasing clothing, they easily think of Taobao and JD.com. They only need to click navigation links to quickly access popular third-party platforms and find needed information resources, logically avoiding more complex search strategies—consistent with the "principle of least effort" [46]. (2) This study's large-scale dataset experiment segmented Chinese online college student users into three groups: Conjoint-Utilizing Users, Single-Utilizing Users in Disorder, and Single-Utilizing Users in Order. Although video websites represent the primary theme category for college student users, the three groups show distinct characteristic differences—Conjoint-Utilizing Users fully utilize both navigation links and search tools for information access, demonstrating strong network usage capability, long active periods with high intensity, but no significant thematic specificity; Single-Utilizing Users in Disorder primarily click popular websites, have long online activity periods with weak intensity, and show specific preferences for video content; Single-Utilizing Users in Order also primarily click links but exhibit very regular network usage times with short, low-intensity activity periods and no clear thematic preferences. These conclusions help us better understand the patterns and characteristics of this special group's Internet usage behavior, providing references for related research targeting this specific user population.

For service providers targeting college student users (especially small and medium-sized enterprises), user market segmentation and behavior pattern mining can effectively help enterprises understand user group needs to support more personalized information recommendation services, expand potential user groups, and develop new service models for data value-added. This study introduces a user market segmentation method based on enterprise access logs, which is relatively straightforward at both the data and model application levels. For these three user groups, enterprises can develop personalized information push strategies. For instance, Single-Utilizing Users in Disorder can receive long-term, single-type information content pushes covering only "video live streaming" resources. For Single-Utilizing Users in Order, enterprises should adopt timed mixed recommendation strategies—pushing generalized thematic information content within fixed time windows (e.g., four short periods identified in this study), covering learning tools, gaming/anime, entertainment, and video live streaming resources. This approach enables precise capture of user group needs as an intermediate processing step for more accurate personalized services, while enabling timed, targeted automated push

services for different groups, thereby improving enterprise computing resource utilization and reducing unnecessary server overhead and maintenance costs.

However, this study has several limitations: (1) Experimental data is based on a constructed virtual navigation platform and does not record complete daily user network usage. For example, users might bypass the navigation website and directly use search engines, whose logs we cannot obtain. Therefore, despite embedding the navigation site at gateway entrances and implementing discount policies to maximize platform usage for more complete log data, whether the results are completely unbiased regarding low search tool usage efficiency remains debatable and requires further verification, particularly regarding underlying causes like the "principle of least effort" and other socio-psychological factors that need analysis through questionnaires and in-depth interviews. (2) Although relevant college student access datasets are scarce, this study's dataset and dimensions still require further expansion to more robustly demonstrate the proposed method's effectiveness and generalizability. Future work will optimize navigation website design and expand promotion and deployment scope to attract more college student traffic. (3) The MaxMatching algorithm depends on manual annotation quality of seed word sets and belongs to hard clustering models, requiring further accuracy improvement. Next steps will consider soft learning approaches, incorporating constraints and entity recognition algorithms for more precise query identification, while conducting multiple repeated experiments with different manual annotation levels and seed set quantities for better model results. Finally, this study's experiments primarily used PC-based testing without considering mobile device usage. Future work will incorporate different device channels (e.g., mobile phones, tablets) combined with corresponding demographic features for more refined feature analysis and statistical description to comprehensively and deeply mine and display college student users' online behavior pattern characteristics and regularities.

References

[1] China Internet Network Information Center. The 41st Statistical Report on China's Internet Development [EB/OL]. [2018-03-05]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/P020180305409870339136.pdf.

[2] Hassan M T, Karim A. Impact of behavior clustering on Web surfer behavior prediction [J]. Journal of information science & engineering, 2011, 27(6): 1855-1870.

[3] Jamali H R, Nicholas D, Huntington P. The use and users of scholarly e-journals: a review of log analysis studies [J]. Aslib proceedings, 2005, 57(57): 554-571.

[4] Kinnally W, Lacayo A, McClung S, et al. Getting up on the download: college students' motivations for acquiring music via beyond P2P [M]. Washington, DC: Pew Internet and American life project, 2005.

[5] Fortson B, Scotti J, Chen Y C, et al. Internet use, abuse, and dependence amongst students at a southeastern regional university [J]. Journal of American college health, 2007, 56(2): 137-144.

[6] Wang Y, Niiya M, Mark G, et al. Coming of age (digitally): an ecological view of social media use among college students [J]. New media & society, 2008, 10(6): 893-913.

[7] Tenopir C. Use and users of electronic library resources: an overview and analysis of recent research studies [M]. Washington, DC: Council on library & information resources, 2003: 72.

[8] Madden M, Rainie L. Music and video downloading moves beyond P2P [M]. Washington, DC: Pew Internet and American life project, 2005.

[9] Zhang P, Liu C. Personal information management practices of Chinese college students on their smartphones [C]//The third international symposium of Chinese CHI. New York: ACM, 2015: 47-51.

[10] Wu D, Liang S. Research on the follow-up actions of college students' mobile search [C]//Proceedings of the 16th ACM/IEEE-CS joint conference on digital libraries. New York: ACM, 2016: 59-62.

[11] Broder A Z. A taxonomy of Web search [C]//Proceeding of ACM SIGIR forum. New York: ACM, 2002, 36(2): 3-10.

[12] Alonso O, Stone M. Building a query log via crowdsourcing [C]//Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval. New York: ACM, 2014: 939-942.

[13] Gonzalez-Caro C, Baeza-Yates R. A multi-faceted approach to query intent classification [C]//Proceedings of the 18th international conference on string processing and information retrieval. Berlin: Springer-Verlag, 2011: 368-379.

[14] Baeza-Yates R, Calderon-Benavides L, Gonzalez-Caro C. The intention behind Web queries [C]//Proceedings of the 13th international conference on string processing and information retrieval. Berlin: Springer-Verlag, 2006: 98-109.

[15] Khudabukhsh A R, Bennett P N, White R W. Building effective query classifiers: a case study in self-harm intent detection [C]//Proceedings of the 24th ACM international conference on information and knowledge management. New York: ACM, 2017: 1735-1738.

[16] Mansouri B, Zahedi M S, Campos R, et al. Online job search: study of users' search behavior using search engine query logs [C]//Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval. New York: ACM, 2018: 1185-1188.

[17] Kang I H, Kim G C. Query type classification for Web document retrieval [C]//Proceeding of the 26th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2003: 64-71.

[18] Sun J, Xu J, Zheng K, et al. Interactive spatial keyword querying with semantics [C]//Proceedings of the 2017 ACM conference on information and knowledge management. New York: ACM, 2017: 1727-1736.

[19] Guo Q, Agichtein E. Exploring mouse movements for inferring query intent [C]//Proceeding of the 31st annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2008: 707-708.

[20] Pujeri R V, Karthik G M. Constraint based frequent pattern mining for generalized query templates from Weblog [J]. International journal of engineering science & technology, 2011, 2(11): 17-33.

[21] Cao H, Jiang D, Pei J, et al. Context-aware query suggestion by mining click-through and session data [C]//Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2008: 875-883.

[22] Teevan J, Dumais S T, Liebling D J. To personalize or not to personalize: modeling queries with variation in user intent [C]//Proceeding of the 31st annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2008: 163-170.

[23] Chuang S L, Chien L F. Towards automatic generation of query taxonomy: a hierarchical query clustering approach [C]//Proceedings of the 2002 IEEE international conference on data mining. Washington, DC: IEEE Computer Society, 2002: 75-82.

[24] Park J Y, O'Hare N, Schifanella R, et al. A large-scale study of user image search behavior on the web [C]//Proceedings of the 33rd annual ACM conference on human factors in computing systems. New York: ACM, 2015: 985-994.

[25] Led T, Bernardi R. Query classification using topic models and support vector machine [C]//Proceedings of ACL 2012 student research workshop. Stroudsburg: Association for Computational Linguistics, 2013: 19-24.

[26] Zhai H, Guo J, Wu Q, et al. Query classification based on regularized correlated topic model [C]//Proceedings of the 2009 IEEE/WIC/ACM international joint conference on Web intelligence and intelligent agent technology. Washington, DC: IEEE Computer Society, 2009: 552-555.

[27] Zhang C W, Fan W, Du N, et al. Mining user intentions from medical queries: a neural network based heterogeneous jointly modeling approach [C]//Proceedings of the 25th international conference on World Wide Web. The Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016: 1373-1384.

[28] Hashemi S H, Williams K, Kholy A E, et al. Measuring user satisfaction on smart speaker intelligent assistants using intent classification [C]//Proceeding of the 29th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2018: 1183-1192.

[29] Dou S, Sun J T, Yang Q, et al. Building bridges for web query classification [C]//Proceedings of the 9th ACM international conference on Web search and data mining. New York: ACM, 2016: 655-664.

[30] Konishi T, Ohwa T, Fujita S, et al. Extracting search query patterns via the pairwise coupled topic model [C]//Proceedings of the 26th international conference on neural information processing systems. New York: Curran Associates Inc., 2013: 3111-3119.

[31] Guo Cheng, Bai Yu, Zheng Jianxi, et al. An unsupervised sub-topic mining method [J]. Journal of Chinese Information Processing, 2016(1): 50-55.

[32] Wu B, Xiong C Y, Sun M S, et al. Query suggestion with feedback memory network [C]//Proceedings of the 27th international conference on World Wide Web. The Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2018: 1563-1571.

[33] Wang Z, Wang F, Wang H, et al. Unsupervised head-modifier detection in search queries [J]. ACM transactions on knowledge discovery from data, 2016, 11(2): 1-28.

[34] Duan H, Zhai C X. Mining coordinated intent representation for entity search and recommendation [C]//Proceedings of the 24th ACM international conference on information and knowledge management. New York: ACM, 2015: 333-342.

[35] Feng Xiaohua, Lu Wei, Zhang Xiaojuan. A review of search result diversification research [J]. Journal of Intelligence, 2015, 34(7): 776-784.

[36] Liu P Q, Azimi J, Zhang R f, et al. Contextual query intent extraction for paid search selection [C]//Proceedings of the 24th international conference companion on World Wide Web. New York: ACM, 2015: 71-72.

[37] Seock Y K, Chen Y. Website evaluation criteria among US college student consumers with different shopping orientations and Internet channel usage [J]. International journal of consumer studies, 2007, 31(3): 204-212.

[38] Malik A, Mahmood K. Web search behavior of university students: a case study at university of the Punjab [J]. Webology, 2009, 6(2): 1-13.

[39] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 26th international conference on neural information processing systems. New York: Curran Associates Inc., 2013: 3111-3119.

[40] Tabatabai D, Shore B M. How experts and novices search the Web [J]. Library & information science research, 2005, 27(2): 222-248.

[41] Savolainen R. Network competence and information seeking on the Internet: from definitions towards a social cognitive model [J]. Journal of documentation, 2002, 58(2): 211-226.

[42] Song C, Barabasi A L. Limits of predictability in human mobility [J]. Science, 2010, 327(5968): 1018-1021.

[43] Mihalcea R. TextRank: bringing order into texts [C]//Proceeding of 2004 conference on empirical methods in natural language processing. Barcelona: ACL, 2004: 404-411.

[44] Koutrika G, Ioannidis Y. Rule-based query personalization in digital libraries [J]. International journal on digital libraries, 2004, 4(1): 60-63.

[45] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis [J]. Journal of computational & applied mathematics, 1999, 20(20): 53-65.

[46] Wang Xiaona. The principle of least effort and accessibility of information retrieval systems [J]. Information science, 2000, 18(2): 135-136.

Author Contributions:
Yan Chengxi: Paper writing, experimental coding, and data analysis;
Wang Jun: Conceptualization and paper revision;
Wang Ke: Experimental evaluation and paper revision.

Chinese College Students' Internet Use: A New Method of Behavior Pattern Recognition with Massive Log Analysis

Yan Chengxi, Wang Jun, Wang Ke
Department of Information Management, Peking University, Beijing 100871

Abstract: [Purpose/significance] It is of great significance to analyze and understand users' daily Web behavior patterns, which not only makes progress in the domain of user behavior analysis and information retrieval theoretically, but also has potential social values and practical significance in promoting personalized service and information recommendation for the undergraduate-oriented enterprises. [Method/process] In this paper, a new method for college students' behavior Web pattern recognition based on large-scale log analysis was proposed. It included a semi-supervised learning algorithm "MaxMatching" based on deep learning and text analysis, and a hybrid model combined with two characteristic entropy (Shannon Entropy and Real Entropy). [Result/conclusion] The empirical results showed that this method has the excellent performance in the algorithm and the result interpretation. Also, it can generalize and present all-round Chinese college students' Web behavior pattern in three aspects of network ability, temporality and topicality. The method and conclusion can effectively expand the methods about semantic understanding of queries in information retrieval, and provide some reference and feasible suggestions to undergraduate-oriented enterprises on personalized recommendation service.

Keywords: Chinese students; online behavior; pattern recognition; massive log analysis

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