Construction and Analysis of an Academic Search Intent Taxonomy: An Empirical Postprint Based on Baidu Scholar Query Logs

Authors: Wang Ruixue, Fang Jing, Li Xin, Lu Wei, Zhang Xian

Date: 2023-04-01T16:02:46+00:00

Abstract

[Purpose/Significance] Understanding, analyzing, and identifying the information needs expressed by users during academic search constitutes the primary step for optimizing query results and enhancing user experience in academic search engines. The explicit and latent information needs expressed by users through query formulations during academic search can be termed academic search intent. This paper's synthesis of an academic search intent taxonomy facilitates academic search intent recognition and the presentation of search result pages. [Method/Process] Building upon A. Broder's query intent taxonomy and incorporating query formulation instances from Baidu Academic search logs, we construct a taxonomy of academic search intent. From this foundation, we summarize different categories of academic search intent and analyze the characteristics of query formulations under each category. [Results/Conclusion] Academic search intent is primarily categorized into five types: academic literature, academic entities, academic exploration, knowledge Q&A, and non-academic literature; we derive the approximate proportional distribution of different categories of academic search intent in academic search; and we present the query formulation features, search scenarios, and search result page characteristics for each category.

Full Text

Construction and Analysis of Academic Query Intent Taxonomy: An Empirical Study of Baidu Academic Search Logs

Wang Ruixue[1], **Fang Jing**[1], **Li Xin**[1], **Lu Wei**[1,2], **Zhang Xian**[3] [1]School of Information Management, Wuhan University, Wuhan 430072 [2]Information

Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072
[3]Baidu Times Network Technology (Beijing) Co., Ltd., Beijing 100085

Abstract: [Purpose/Significance] Understanding, analyzing, and identifying the information needs expressed by users during academic search is the first step toward optimizing query results and improving user experience in academic search engines. The explicit and potential information needs expressed by users through query formulations during academic search can be termed academic query intent. Summarizing the academic query intent taxonomy facilitates intent identification and the presentation of search result pages. [Method/Process] Based on A. Broder's query intent taxonomy and combined with query expression examples from Baidu Academic search logs, this study constructs a taxonomy of academic query intent. Using this foundation, we summarize different categories of academic query intent and analyze the characteristics of query expressions under each category. [Result/Conclusion] Academic query intent is primarily divided into five categories: academic literature, academic entity, academic exploration, knowledge Q&A, and non-academic literature. We determine the approximate proportions of each category in academic search and provide query expression features, search scenarios, and result page presentation recommendations for each category.

Keywords: academic search; query intent; taxonomy; query logs; Baidu Academic **Classification Number:** G254 **DOI:** 10.13266/j.issn.0252-3116.2021.04.008

Academic search represents a specialized vertical domain of search with unique characteristics and professional user bases. Current academic research on academic search primarily focuses on two aspects: (1) User academic search behavior studies based on questionnaires, which investigate user preferences, search frequency and timing, and basic strategies that users subjectively hope search engines will employ when constructing queries; and (2) User behavior analysis studies based on search engine query logs, which typically use data from specialized search engines such as CiteSeerX, Springer, PubMed, and library retrieval logs. Although both types of analyses reveal academic search behavior to some extent, their findings may be difficult to generalize to other domains or larger user populations due to data limitations (questionnaires typically contain only hundreds of data points) and user group constraints (users of specialized academic search engines represent only a portion of the overall academic search user population).

The concept of "query intent" was first proposed by A. Broder in 2002, defined as the user information needs, search goals, and motivations contained within queries [1]. Other scholars have proposed alternative definitions. For example, B. Jansen et al. [2] defined query intent as the emotional, cognitive, and situational goals expressed during user interaction with search engines, while Jiang Xue et al. [3] defined it as the basic strategies users hope search engines will employ when constructing queries. Although no unified definition exists in academia, query intent generally refers to user needs contained within queries.

Based on the aforementioned research, we define academic query intent as the explicit and potential information needs expressed by users through query formulations during academic search.

Current query intent research primarily focuses on general search contexts, with limited attention to the academic search domain. Among existing studies on academic query intent, few have systematically organized academic query intent or mined it from real user query logs. Most research only identifies partial user intents and provides academic search functions covering limited needs, lacking a comprehensive taxonomy of user query intent in academic search to provide necessary theoretical foundations for meeting all search requirements.

2 Related Research

2.1 Current Status of Query Log Analysis Research

User query logs record all interaction information between users and search engines [4]. Research utilizing query logs for information retrieval has yielded abundant results addressing diverse practical problems through various methods, which can be broadly categorized into basic research and extended research [5]. The former primarily analyzes fundamental log characteristics, including query length, search time/periods, and keyword frequency. The latter combines query logs with other data sources such as anchor texts, corpora, and knowledge bases to improve other information retrieval research problems, including query expansion, query recommendation, and named entity recognition.

Researchers currently use primarily general search engine log data. English query research data sources mainly include AltaVista, AOL, and Excite query logs, while Chinese query research data sources primarily include Sogou, Peking University's "Tianwang," and Baidu search query logs. Many researchers have utilized these publicly available log datasets. B. Jansen et al. [6] used Excite search engine query logs to analyze user query length, finding that users averaged 2.21 terms per query, with fewer than 4% of queries exceeding six terms. Yu Huijia et al. [7] analyzed 45 million Sogou query records, discovering that Chinese queries averaged 1.85 terms, with 93.15% containing no more than three terms. Tong Guoping et al. [8] analyzed a Sogou query log from a day in 2011, examining query length, search methods, query topics, click behavior, and user types, finding that users preferred query strings composed of 2-5 Chinese noun phrases.

Academic search engine user query logs record user academic search behavior, and analyzing these queries can relatively accurately and comprehensively infer user intent. X. Li et al. [9] analyzed five months of query data from the academic database ScienceDirect, finding that academic queries averaged 3.77 terms, higher than the 1.4 terms in general search, and that 92.37% of queries contained "named entities," higher than the 70% in general search [10]. Commercial academic search engines like Google Scholar and Baidu Academic make user log data difficult to obtain, leading many scholars to use university library

OPAC system query logs for academic search behavior analysis. For example, Jiang Tingting et al. [11] used 18 days of query logs from Wuhan University Library OPAC to analyze research retrieval types, faceted retrieval types, query length, and session duration. S. Chapman et al. [12] manually screened and processed 996 queries from the University of Michigan Library query logs to annotate search needs. However, whether academic databases or university library OPAC systems, their user profiles and academic search behaviors differ from those in commercial academic search engines, resulting in different query log analysis outcomes. Therefore, this study uses query logs from the commercial academic search engine "Baidu Academic" as the data source for researching query intent in academic search.

2.2 Current Status of Query Intent Taxonomy Construction Research

Current query intent research primarily focuses on query intent identification within given classification systems, encompassing four aspects: query intent identification methods, taxonomy construction, feature recognition, and datasets/evaluation methods [13]. Taxonomy construction forms the foundation for query intent research.

Although different researchers have proposed various classification systems, most can be considered improvements upon the taxonomies constructed by A. Broder [1] or D. E. Rose et al. [14]. Their systems divide query intent into three categories: informational, navigational, and transactional (resource). Informational queries involve users statically searching for information believed to be available online, with no interaction beyond reading or browsing. The content can be any form, such as data, documents, or multimedia, with information needs ranging from precise to vague. Navigational queries involve users aiming to reach a specific website or webpage, whether personal or organizational, meaning the search intention is already expressed in the user's mind before searching due to prior knowledge of the URL's existence. Transactional (resource) queries involve potentially more interaction with search engines, generally for purchasing goods, downloading images/music/videos, or using web services. A. Broder termed this "transactional," while D. Rose & D. Levinson called it "resource," with consistent meaning.

Many scholars have expanded upon this classification system, but D. Rose et al.'s taxonomy remains mainstream. Based on this system, scholars have analyzed the proportional distribution of query intent categories across different datasets, as shown in Table 1 .

Table 1 shows that dataset differences or search domain differences lead to variations in query intent category distribution. Compared with general search engines, the "navigational" query proportion is significantly lower in news domains, while "informational" queries increase. In academic search datasets, the "navigational" query proportion is slightly lower, and the "resource" query proportion remains below 1%. Academic search and general search exhibit different

query intent distributions, and existing query intent classification systems are insufficient to summarize academic users' query intents. This study will construct and analyze an academic search query intent taxonomy using Baidu Academic query logs based on general search query intent research methods.

3 Data Sources and Research Methods

3.1 Data Sources

This study uses Baidu Academic search query logs as the dataset. These logs record all interaction information in Baidu Academic's search box, with each record containing (from left to right) user unique identifier (UID), query time, query expression, and IP address, as shown in Figure 1 [Figure 1: see original paper]. The first line in Figure 1 is interpreted as follows: A user with UID "1cc6aac735c0285f62d345fb00d3e4f" and IP address "101.94.11.33" submitted the query "Liu Yang (2005)" at "2018-03-01 22:13:18."

The obtained Baidu query logs cover March 1 and 7, 2018, both weekdays. Statistical comparisons revealed no significant differences between the two days, so the logs from both days were combined for analysis. Table 2 shows basic statistics for the two-day period:

3.2 Research Methods

To develop a scientifically sound taxonomy, this study invited two master's students and two doctoral students specializing in information retrieval to determine the academic query intent taxonomy through a "browse-classify-discuss-formulate" process:

(1) First, the four students were introduced to the research background and Broder's query intent taxonomy components;
(2) Next, 4,000 records were randomly extracted from the processed dataset and evenly distributed to the four researchers, who independently browsed and categorized user query intents to initially formulate category names and divisions;
(3) Then, the four students discussed classification results through brainstorming, reviewing query statements in disputed categories to finalize the taxonomy: academic literature, academic entity, academic exploration, knowledge Q&A, and non-academic literature;
(4) Finally, 1,000 new records were randomly extracted and divided into two groups, with the four students split into two groups (Group A and Group B). Each group processed 500 records: each student independently annotated their 500 logs according to the established taxonomy. The author conducted inter-rater reliability Kappa tests on the results, as shown in Table 3 .

The Kappa results show that both groups achieved strong agreement ($0.81 \leq$ Kappa < 1 indicates very strong or perfect agreement), confirming that the

taxonomy can be used in academic search query intent research.

4 Academic Query Intent Taxonomy

This study divides academic query intent into five categories based on the following principles: (1) Retaining Broder's "navigational" category and redefining it as "academic literature" in the academic search context; (2) Eliminating Broder's "transactional" category and incorporating relevant academic search intents into "academic literature," explained in detail in subcategories; (3) Splitting Broder's "informational" category into four categories—"academic entity," "academic exploration," "knowledge Q&A," and "non-academic literature"—based on log content and academic search scenarios.

The final "Academic Search Query Intent Taxonomy" comprises five categories: academic literature, academic entity, academic exploration, knowledge Q&A, and non-academic literature. Classification criteria and examples are summarized below. Table 4 shows the classification results for the 1,000-query validation set.

4.1 Category 1: Academic Literature

"Academic literature" query intent refers to users seeking to obtain a specific publication, accounting for 26.44% of queries. Users conducting this type of search already know a specific document meets their needs and hope to navigate to it through search. This intent aligns with Broder's "navigational" category in general search. Table 5 shows subcategories of query expressions.

As summarized in Table 5, "academic literature" queries have approximately seven types. Through browsing and annotation, Type 1 (document title) appears most frequently. The search scenario typically involves users encountering document titles in reference sections, academic exchanges, or scholar homepages. Incomplete titles and typos commonly appear because users cannot recall full titles. Type 2 (citation format) often involves users copying reference information, resulting in formatting errors (spaces, unrecognized characters) and incomplete citations due to text editor differences. Type 3 represents a format between Types 1 and 2. Type 4 (DOI number) frequently appears in foreign literature searches. DOI (Document Object Identifier) serves as a document identifier. Types 5-7 involve academic entities (scholar, institution, publisher—see "academic entity" category discussion). When users combine "academic entity + time/research topic" or "multiple scholar names," the results appear as paper lists, but users typically seek one specific document, submitting incomplete queries when unable to obtain full information.

For this category, since users' intent is "academic literature," the optimal result presentation is "direct navigation to the required document detail page." Historical user browsing/click logs can determine module arrangement on result pages. Personalized displays are possible based on historical behavior: for documents with high download rates, the download module can be prioritized with free

access channels indicated based on user IP; for highly cited documents, the citation module can be emphasized, even highlighting the document's paragraphs in citing literature. For Types 5-7, where single queries cannot determine the needed document, user search sessions and historical records can mine intent to recommend likely needed documents while presenting other matching documents as lists. Currently, Baidu Academic only navigates to document detail pages when users fully input document titles (Type 1), achieving direct navigation for 53.8% of "academic literature" queries (14.22% of all searches). Other cases still display literature lists requiring secondary user judgment.

4.2 Category 2: Academic Entity

"Academic entity" query intent refers to users seeking information about a specific academic entity, accounting for 3.53%. "Academic entity" extends general named entities (person, location, organization, time, date, currency, percentage) to the academic search context [18], specifically including scholars, institutions, journals, and conferences involved in research. Table 6 shows subcategories.

As Table 6 shows, "academic entity" queries have approximately five types, with "single scholar name" queries being most frequent. Unlike multiple scholar name queries in "academic literature," these focus on the scholar themselves, seeking information about the scholar, research fields, recent work, and core publications. Type 5 (scholar name + institution) represents user disambiguation efforts to identify a specific scholar among those with identical names. Types 2-4 represent queries for journals, institutions, and conferences, seeking information like impact factors, submission links, publication cycles, conference ratings, venues/dates, and proceedings. Types 6-7 could also be academic entities but rarely appeared in manual annotation and log data, suggesting such queries are uncommon.

For this category, result pages should comprehensively display relevant information, optimizing module order based on click history. Since academic entity information updates in real-time, with URLs and content changing over time, academic search engines should maintain information accuracy through official links and "invite users to improve" features while managing presentations. Currently, Baidu Academic provides one-click navigation for "scholar and journal" entities and conference proceedings links for "conference name" entities, though incomplete databases often yield empty results. No navigation responses exist for "institution" or "scholar + institution" queries.

4.3 Category 3: Academic Exploration

"Academic exploration" query intent refers to exploratory searches in an academic field where users seek academic resources related to their query but without clear intent 指向 a specific document or entity, requiring continuous selection or multiple interactions to identify needed resources. Due to the inherently exploratory nature of academic search, this category has the highest proportion

(49.34%). Table 7 shows subcategories.

Unlike other categories with clear type distinctions, academic exploration lacks explicit query type boundaries, making the above classification somewhat ambiguous. For example, abbreviations are also a form of academic concept terms. Through subjective perception during browsing and annotation, Type 1 (academic concept terms) appears most frequently. Some team members proposed making "academic concept" a separate category, but overlap exists between "academic concept" and "academic exploration": (1) When querying academic concept terms, users may seek either basic definitions or core literature/recent research in that concept's field, still requiring exploratory, multi-interaction confirmation; (2) Academic concepts themselves are difficult to define, and other "academic exploration" query types typically contain academic concepts. For these reasons, this study retains academic concepts within the academic exploration category.

For Type 3 (academic field progress), queries typically include terms like "status, progress, literature, papers, cases" in formats such as "review of XXX" or "progress in XXX." Users seek review literature but, unlike similar titles in "academic literature," are uncertain whether such papers exist.

For "academic exploration" queries, since search goals are unclear and exploratory, result pages can display definitions, research hotspots, and trends when queries contain academic concept terms. Since analyzing query information alone cannot easily satisfy user needs, we recommend: (1) Query sessions—mining user needs from query history in logs and providing session completion suggestions to help clarify intent; (2) Similar queries—finding similar queries or domains in large-scale log data and displaying results in order of click probability based on mass user browsing behavior. Currently, Baidu Academic still uses keyword matching to provide literature lists for academic exploration queries, displaying "encyclopedia entries, research point analysis, and related hot search terms" on the right side for some popular concepts. The "academic concept definition" display as "encyclopedia entries" benefits non-research users by enabling quick understanding but suffers from non-academic sources that may be too superficial for research work or proper citation.

4.4 Category 4: Knowledge Q&A

"Knowledge Q&A" query intent refers to users seeking specific answers to questions, accounting for 13.82%. Note that we do not define whether questions are academic; classification depends on whether queries express desire for specific answers. Knowledge Q&A is a mature research field [19,20], typically using platforms like Yahoo! Answers, Baidu Zhidao, and Sina iAsk [21] for user behavior studies. This category was not found in existing academic search research but was identified as non-negligible during log browsing and annotation. Table 8 shows subcategories.

Re-examining some knowledge Q&A queries in Baidu Academic revealed two issues: (1) Some queries yield no answers, particularly everyday life questions outside academic research scope with few relevant results; (2) Results remain literature lists, with relevant content possibly only in small document sections requiring download and reading. For non-research users, reading specialized literature may be difficult. Analyzing the scenario, knowledge Q&A queries could retrieve suitable results on professional Q&A platforms like "Baidu Zhidao" or "Zhihu." Since Baidu Zhidao has a more prominent entrance than Baidu Academic on Baidu's homepage, we speculate these users choose Baidu Academic for professional domain results. Based on this speculation, result presentation could include: (1) Excerpted discussion sections from multiple relevant documents addressing the query question; (2) Reading guidance for relevant papers with download guidance; (3) Baidu Zhidao navigation as supplementary material for non-research users seeking quick, 通俗 understanding.

4.5 Category 5: Non-Academic Literature

"Non-academic literature" query intent refers to users seeking documents that are not academic literature, mostly not included in Baidu Academic's database. This category was not found in existing academic search research but was identified as significant during log browsing and annotation (accounting for 6.86%). Table 9 shows subcategories.

Re-examining the two query types in "non-academic literature" on Baidu Academic shows most results are reports rather than direct documents. For example, the query "home appliance and automobile 下乡 policies" likely seeks policy documents, but current results show review academic literature like "Reflections Behind Automobile 下乡 Policies." While some research users may study the impact of such non-academic documents, current result pages and queries only reflect intent for the documents themselves. Such documents are rarely included in academic databases (though 极少数 disciplinary summaries and reflections appear in academic journals). Conversely, searching on "Baidu Wenku" would increase success rates. Based on this speculation, we recommend that Baidu Academic, as a comprehensive academic search engine, could include policy documents in its database or provide 跳转 services to Baidu Wenku for such searches.

5 Discussion and Conclusion

This study, grounded in A. Broder's query intent taxonomy and supported by Baidu Academic query logs, analyzes query expressions to construct a five-category academic query intent taxonomy: academic literature, academic entity, academic exploration, knowledge Q&A, and non-academic literature. Through manual annotation, we determine approximate proportions of each category in academic search and present typical query examples for each category, discussing result page presentation recommendations based on user search scenarios.

Overall, this study provides a foundational taxonomy for academic search query intent, laying theoretical groundwork for automatic intent identification to improve academic search engine results and better serve researchers. However, limitations exist: dataset restrictions only provide UID, query time, query statement, and IP address, lacking click data and search result documents needed for deeper interaction-based intent determination. Time and labor constraints limited manual annotation to 5,000 queries, which could be expanded in future research. Additionally, application to other academic search platforms like library discovery systems and CNKI requires further exploration and analysis as potential extensions of this research.

References

[1] BRODER A. A taxonomy of web search[C]//ACM sigir forum. ACM, 2002, 36(2): 3-10. [2] JANSEN B J, BOOTH D L, SPINK A. Determining the informational, navigational, and transactional intent of Web queries[J]. Information processing & management, 2008, 44(3): 1251-1266. [3] Jiang Xue, Sun Le. Research on user query intent segmentation[J]. Chinese Journal of Computers(3): 210-216. [4] JANSEN B J. Understanding user-web interactions via web analytics[J]. Synthesis lectures on information concepts, retrieval, and services, 2009, 1(1): 1-102. [5] Tang Xiangbin, Lu Wei, Zhang Xiaojuan, et al. Query specificity feature analysis and automatic recognition[J]. New Technology of Library and Information Service, 2015, 31(2): 15-23. [6] JANSEN B J, SPINK A, SARACEVIC T. Real life, real users, and real needs: a study and analysis of user queries on the web[J]. Information processing & management, 2000, 36(2): 207-227. [7] Yu Huijia, Liu Yiqun, Zhang Min, et al. Search engine user behavior analysis based on large-scale log analysis[J]. Journal of Chinese Information Processing, 2007, 21(1): 109-114. [8] Tong Guoping, Sun Jianjun. User behavior analysis based on search logs[J]. New Technology of Library and Information Service, 2015, 31(7): 80-88. [9] LI X, SCHIJVENAARS B A, RIJKE M. Investigating queries and search failures in academic search[J]. Information processing & management, 2017, 53(3): 666-683. [10] LIN T, PANTEL P, GAMON M, et al. Active objects: actions for entity-centric search[C]//Proceedings of the 21st international conference on World Wide Web. France: ACM, 2012: 589-598. [11] Jiang Tingting, Wang Miao, Gao Huiqin. OPAC system user search behavior log analysis: a case study of Wuhan University Library[J]. Library and Information Service, 2015(5): 46-54. [12] CHAPMAN S, DESAI S, HAGEDORN K, et al. Manually classifying user search queries on an academic library Website[J]. Journal of web librarianship, 2013, 7(4): 401-421. [13] Lu Wei, Zhou Hongxia, Zhang Xiaojuan. A review of query intent research[J]. Journal of Library Science in China, 2013, 39(1): 100-111. [14] ROSE D E, LEVINSON D. Understanding user goals in Web search[C]//Proceedings of the 13th international conference on World Wide Web. ACM, 2004: 13-19. [15] He Guoxiu, Zhang Xiaojuan. Discussion on improved methods for automatic query intent classification[J]. Digital Library Forum, 2018(1): 53-60. [16] Zhang Xiaojuan, Lu Wei, Lei Shengwei. Automatic recognition of news intent based on

query feature analysis[J]. Library and Information Service, 2014, 58(20): 82-90. [17] KHABSA M, WU Z, GILES C L. Towards better understanding of academic search[C]//2016 IEEE/ACM joint conference on digital libraries. IEEE, 2016: 111-114. [18] Sun Zhen, Wang Huilin. A review of named entity recognition research progress[J]. Data Analysis and Knowledge Discovery, 2010, 26(6): 42-47. [19] Wu Dan, Yan Ting, Jin Guodong. Comparison and evaluation of online Q&A communities and collaborative reference consultation[J]. Journal of Library Science in China, 2011, 37(4): 94-105. [20] Liu Gaoyong, Deng Shengli. Research on the evolution and development of social Q&A services[J]. Library Tribune, 2013, 33(1): 17-21. [21] Deng Shengli. Comparative study of domestic and foreign interactive Q&A platforms and countermeasures[J]. Information Studies: Theory & Application, 2009(3): 50-55.

Author Contributions: Wang Ruixue: experimental design, data cleaning, and paper drafting; Fang Jing: data cleaning, experimental operation, and paper drafting; Li Xin: data cleaning and paper revision; Lu Wei: experimental design and paper revision; Zhang Xian: data cleaning and experimental design.

The promotional content about "中国发明与专利" magazine has been omitted as it is not part of the academic paper.

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