JDIS | Journal of Data and Information Science
Research Paper

Understanding teams and productivity in information retrieval research: Academia, industry, and cross-community collaborations

Jiaqi Lei[1,2*], Liang Hu[2*], Yi Bu[2†], Jiqun Liu[3†]

[1]Institute of Education, Tsinghua University, Beijing 100084, China
[2]Department of Information Management, Peking University, Beijing 100871, China
[3]School of Library and Information Studies, The University of Oklahoma, Norman, Oklahoma 73019, U.S.A.

Citation: Lei, J. Q., Hu, L., Bu, Y., & Liu, J. Q. (2026). Understanding teams and productivity in information retrieval research: Academia, industry, and cross-community collaborations. *Journal of Data and Information Science*, 11(1), 132-154. https://doi.org/10.2478/jdis-2025-0051

DOI: 10.2478/jdis-2025-0051 **CSTR:** 32295.14.jdis-2025-0051

Received: Jun. 6, 2025; Revised: Aug. 7, 2025; Sep. 17, 2025; Accepted: Sep. 19, 2025

ABSTRACT

Purpose: Prior Information Retrieval (IR) research synthesizes progress from individual studies, yet academia-industry collaboration dynamics remain unexplored. This study investigates: (1) productivity patterns and venues, (2) citations-downloads relationships, (3) topic evolution, and (4) collaboration trends.

Design/methodology/approach: We perform an analysis of 53,471 ACM IR papers (2000–2018) using bibliometrics and DistilBERT topic modeling.

Findings: We find that industry-involved papers preferred WWW/CIKM venues; collaborations dominated RecSys/CSCW. We see that academia-industry collaborations achieved the highest download-to-citation conversion rates. Academia focused on algorithms; industry on applications; collaborations bridged both with rising human-centered themes.

Research implications: This is a pioneering large-scale bibliometrics revealing collaboration's impact on IR knowledge evolution and provides a methodological framework for cross-sector analysis.

Practical implications: The paper identifies optimal venues (RecSys/CSCW) for partnerships and guides joint initiatives (shared datasets, grants) to bridge academia-industry divides and enhance research translation.

* These authors contributed to this paper equally.
† Corresponding authors: Yi Bu (Email: buyi@pku.edu.cn); Jiqun Liu (Email: jiqunliu@ou.edu).

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/343002

Originality/value: This is the first large-scale bibliometric analysis of IR academia-industry collaboration. The paper finds many novel insights, including the fact that collaboration boosts citation efficiency, enables complementary specialization, and drives topic convergence.

Keywords: Information Retrieval; Bibliometrics; Productivity; Research topic; Scientific collaboration; Academic-industry collaboration.

1 Introduction

Information retrieval (IR) research seeks to characterize, support, and improve the process of retrieving relevant information and documents that satisfy users' information needs (Kobayashi & Takeda, 2000). It is an interdisciplinary research field that emphasizes the value of evaluation and brings together knowledge and methods from computer science, library and information studies, human-computer interaction, and other related areas. IR attracts attention and research efforts from both academia and industry, fostering interdisciplinary collaborations. In the past two decades, as search systems become increasingly ubiquitous in different modalities of human-information interactions (such as desktop search, mobile search, conversational or spoken information seeking, and chat-based search), both researchers and system developers from academia and industry have made significant contributions to IR algorithms, interactive search systems, user models, as well as evaluation techniques. Furthermore, the area has witnessed a resurgence of interest in natural language processing and deep learning. Artificial Intelligence (AI) techniques mark a series of unique contributions from industry researchers to the area for developing and evaluating modern intelligent search systems (Culpepper et al., 2018; Li & Lu, 2016; Yates et al., 2021).

While previous surveys and workshops have focused on summarizing progress and synthesizing knowledge and techniques from individual studies and data-driven experiments (e.g., Liu, 2021), the extent of contributions and collaborations between researchers from different communities (e.g., academia and industry) in advancing IR knowledge remains unclear. To address this gap, this study aims to answer the following four research questions. This paper begins by focusing on a straightforward question of whether scientists from academia and industry have preferences and differences in their choice of academic conferences for publishing their papers. Thus, the first research question is formulated as follows:

RQ1: What are the patterns of productivity and preferred venues characterize IR studies by Academia, Industry, and Academia-Industry Collaboration?

The number of full-text paper downloads as an important element of alternative metrics of traditional citation-based measures has now received attention from scholars worldwide. Garfield (1996) proposed the use of web downloads instead of citations of scientific publications to resolve the problem that there are temporal lags in the evaluation of scientific publications using citation analysis. Due to the timeliness of downloads and their complementary effect on citations, previous studies have explored the relationship between download and citation counts (Hu et al., 2021).

This paper specifically investigates the relationship between citation and download counts between academia and industry in the field of IR. Consequently, the second research question is as follows:

RQ2: What is the relationship between citation and downloads counts in Academia, Industry, and Academia-Industry Collaboration?

Given the distinct working culture and various orientations of academia and industry, researchers from academia may tend to lean towards theoretical research while those from industry may focus more on practical applications and systems. That being said, we may expect to observe significant differences in the topics covered by these researchers. When researchers from academia and industry collaborate, one of the issues this paper will explore is the changing focus of research attention and emerging research topics. Hence, the third research question is proposed:

RQ3: How do the research topics change over time in the three types of papers: Academia (all co-authors are from academia), Industry (all co-authors are from industry), and Academia-Industry Collaborations (some co-authors are from academic while others industry)?

As one of the focuses of this paper is on the collaborative outputs from academia and industry, we are eager to know what types of papers are more likely to involve large-team collaborations and what the characteristics of author team sizes in an Academia-Industry collaborations are. Through statistical analysis on author teams, this paper aims to explore these and provide reasonable explanations. Additionally, understanding the changing trends in collaboration between academia and industry is an important aspect of this research. Thus, the final research question is as follows:

RQ4: What is the preferred size of collaboration teams in the three categories of papers, and how does collaboration between authors from academia and industry evolve over time?

Through empirical analysis of the above four questions, this study mainly draws the following contributions:

- Findings from the current study may offer a new perspective for analyzing the advance and emerging trends in IR research and helps demonstrate the cross-community collaborations and scientific contributions of academia and industry.
- Findings from this study can help clarify the roles of academic and industry collaborators in IR research projects, partially break the boundaries that hinders productive collaborations.
- Findings from this study may inspire more joint grant proposals, collaborative evaluation experiments, and cross-community initiatives that will both advance the knowledge and enhance commercial search systems.

2 Related work

2.1 Academia-industry collaborations

Previous studies have noted discrepancy between researches from academia and industry. Academia typically tends to focuses on basic research and scientific exploration while industry is

driven by needs in product development and commercial purposes (Ahmed et al., 2023). Academia places more emphasis on scientific breakthroughs and advances in knowledge, while industry links research and development activities to market needs and commercial interests (Spicer et al., 2022).

In recent years, collaborative behavior between academia and industry has become increasingly common (Wuchty et al., 2007; Zhang et al., 2018). However, cultural differences between academia and industry may make this collaborative system difficult and challenging. Jasny et al. (2017) explore such collaborative systems where incomplete communication and sharing of technology, data, or materials interfere with future research, and advocate for leadership, and support from funding agencies, journals, and other stakeholders. Marijan and Gotlieb (2021) address the challenges in establishing effective scientific collaboration between academia and industry. The Certus model was proposed to facilitate participatory knowledge creation when solving problems. Furthermore, recent studies indicate that the gap in research collaboration between academia and industry is progressively narrowing (Etzkowitz & Leydesdorff, 2000; Rhoten & Powell, 2007). This trend is driven by a growing recognition of the value of interdisciplinary approaches and techniques, increased support from funding agencies and research institutions, and advancements in technology that enable seamless communication, varying modalities of scholarly communication, and active knowledge sharing across platforms.

Despite the difficulties and challenges, the academic-industrial collaboration model has great merit, especially for scientific areas that have deep roots in applications and practical evaluations. Collaboration between academia and industry can "translate" scientific discoveries into tangible products and industrial impact, commercializing researches that would otherwise go undiscovered. In addition, collaboration between academia and industry can promote knowledge sharing and technology transfer, and improve the application of researches (Noyons et al., 1994; Perkmann & Walsh, 2009). The industry community gain new business opportunities and competitive advantages from the research results of academia, and academia community obtain more research resources and financial supports (Owen-Smith, 2003; Van Looy et al., 2006). In short, cooperation between academia and industry can maximize the value of researches and promote the development of IR researches. Recent collaborations between academia and industry have led to significant advancements in AI, Natural Language Processing (NLP), and machine learning, bridging theoretical research with practical applications. In the domain of AI-driven customer service, AI-powered systems such as chatbots and voice assistants have gained traction, automating responses and enhancing customer experience through NLP and machine learning (Rani et al., 2024). In the field of bioinformatics, Serajian et al. (2023) developed MTB++, a machine learning classifier for predicting antibiotic resistance in Mycobacterium tuberculosis. The study provided valuable insights into sequence similarities with antibiotic resistance genes, enhancing the understanding of resistance mechanisms in MTB. In social media analysis, the application of BERT for gender polarity detection has been explored, with particular attention to how emojis and emoticons influence sentiment classification in short texts, as discussed by Jazi et al. (2024). Moreover,

Shahin et al. (2024) introduced an innovative approach for extracting the voice of customer (VoC) data by leveraging GPT-3.5 Turbo model, marking a significant advancement over traditional methods. Additionally, the integration of this NLP technology with Lean Six Sigma 4.0 principles promises to enhance customer-centric strategies in the context of Industry 4.0, offering more comprehensive, real-time insights for decision-making in product development and process improvement.

2.2 Emerging topics and collaborations in Information Retrieval research

Previous research in the field of IR has encompassed a variety of approaches, such as user studies, simulation-based experiments, and naturalist studies, which have been employed to address diverse unresolved challenges and emerging problems both within and beyond technical, system-oriented and evaluation aspects of IR. Keyvan and Huang (2023) conducted a survey focusing on techniques, tools, and methods used to comprehend ambiguous queries in Conversational Search Systems (CSS) deployed in everyday-life and workplace settings, such as chatbots, Apple's Siri, Amazon Alexa, and Google Assistant. Ambiguous query clarification and search result re-ranking, among other open questions, have been extensively explored and discussed in publications from academia, industry, and collaborative projects involving both sides (e.g., Gao et al., 2020; Thomas et al., 2021; Zamani et al., 2020). In the recent three years, the growing interests and cross-domain collaborations in CSS have also been boosted by the innovation and application of large language models (LLMs) and AI-enabled chatbots, which open new opportunities for joint scientific projects and rapid research translation from research-lab-curated models and evaluations to industry implementations.

Besides, the emergence of algorithmic fairness, accountability, transparency, and ethics (FATE) as a notable research topic has attracted attention from both academic and industry scholars in the IR community. This line of research has resulted in a series of publications, industry sessions, collaborative workshops, tutorials, and funding projects (Castillo, 2019; Ekstrand et al., 2019; Gao & Shah, 2021). The FATE-IR research has brought together a diverse group of researchers and practitioners who contribute to both the conceptualization and technical aspects of responsible IR research agenda (Olteanu et al., 2019). Similar concerns have been increasingly discussed in the broader AI and HCI literature, such as Mitchell et al. (2019) on AI accountability.

Currently, there are few articles examining the collaboration between academia and industry in the field of IR. Zaharia and Kaburakis (2016) explore trends in collaboration barriers among various research involvement levels of U.S. sport firms with sport management academia. Ahmed et al. (2023) examine the current state of research in AI from industry and academia working together to tip the scales in favor of industry. A previous research-in-progress paper of the current one examines the features and differences regarding productivity, authorship, and impact of the three types of studies and also pay special attention to the research problems and topics that attract and foster academia-industry collaborations in the recent two decades of IR studies (Lei et al., 2023). Built on the preliminary analysis of the collaboration between academia and industry in IR

in terms of productivity, authorship, impact, and topic distribution, the current paper will leverage an extended dataset, aiming to answer the four research questions (RQs) proposed above.

3 Data preparation

The empirical data employed in our analysis mainly comes from Association for Computing Machinery (ACM) Digital Library[①], which is a comprehensive repository of articles in the field of computer science and related areas. The original dataset we utilized comprised a total of 295,561 articles published in ACM from 1951 to 2018. The dataset includes information such as the publication date, title, abstract, keywords, author, author IDs and author institution IDs and names, citation count, and download count for each paper. To filter out articles in the field of IR, we identified a set of 200+ keywords/phrases (hereafter "keywords"; detailed results are provided in Table A1) from the whole keywords set. These keywords were selected by one of the collaborators of this study, who has research experience in the field of IR, and are considered to be representative of the field. Then the keywords were used for matching in the article keyword field in the original dataset, which ultimately matched 53,471 articles published between 2000 and 2018, referring to this dataset as the ACM dataset.

The Research Organization Registry (ROR)[②] is an inclusive and community-driven global registry that maintains open persistent identifiers for research organizations. The ROR dataset provides categories to obtain the types of author institutions, which allows us to classify authors into three categories: academia, industry, and others (e.g., facilities, health, governments, etc.). This classification process was facilitated by extracting 137,843 (author, institution) data pairs from the ACM dataset and matching them with the corresponding entries in ROR dataset. Our matching efforts successfully identified 125,668 data pairs, accounting for 91.17% of the total data. However, there remained 8.83% of the data pairs that did not find a direct match. To ensure that no significant institutions were overlooked, we took a proactive measure. Specifically, we manually supplemented the type labels for institutions that appeared 10 times or more in the dataset. This supplementation involved conducting ROR searches to assign appropriate type labels. As a result, only 3.69% of the author institutions did not have a corresponding type match.

We conducted further classification of scientific publications into four distinct types: publications authored/co-authored exclusively by individuals from academia (Academia), publications authored/co-authored exclusively by individuals from industry (Industry), publications co-authored by individuals from both academia and industry (Academia-Industry Collaboration), and others. A single publication may correspond to multiple authors, and a single author may be affiliated with multiple institutions. This results in a mapping relationship between the publication and multiple institutions. Then, the type of the publication is determined based on the types of these institutions. For example, if all the author-affiliated institutions corresponding to a publication

① https://dl.acm.org
② https://ror.org

are academic, then the publication is classified as an academic type. If the author-affiliated institutions corresponding to a publication are both academic and industrial, then the publication is classified as a collaboration type. Our analysis revealed that within the ACM dataset, there were 37,034 papers classified as Academia, 4,941 papers classified as Industry, 1,986 papers classified as Academia-Industry Collaboration, and 2,604 papers classified as other types. These numbers indicate that the majority of papers in the field of IR are authored/co-authored by individuals within the academic community, followed by the number of papers published collaboratively between academia and industry, and the type with the fewest number of papers published entirely by researchers from the industrial community. This paper primarily focuses on the first three types, which collectively comprise a total of 46,565 papers. The specific data processing steps are illustrated in Figure 1. The distribution of the number of papers in three types over the years is shown in Figure 2. It can be observed that as the years progress, the number of papers in the field of IR in the ACM dataset continues to increase, with academic papers consistently being the predominant type.

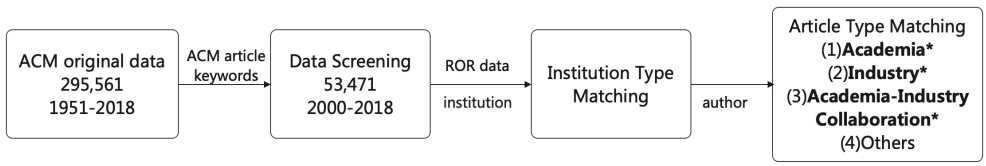


Figure 1. Flow chart of data processing. * indicates the focus of this current paper.

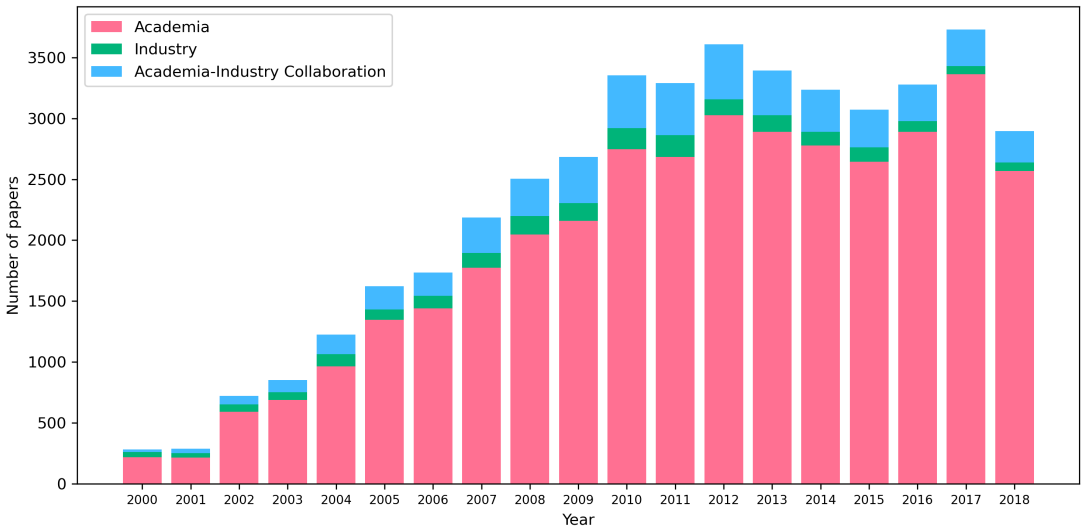


Figure 2. Distribution of the number of papers in three types over the years.

Due to the incomplete nature of the ACM dataset, some IR reviewed conferences may not be

fully included in our dataset. Papers from some of the refereed conferences may represent high-quality research and recent advances in the field of IR. The exclusion of these papers may result in a dataset that lacks the necessary standardization and consistency in evaluating and comparing IR technologies. However, the ACM Digital Library includes conference papers from as comprehensive a range of IR fields as possible, and most papers in the IR field are presented at conferences. Even if all papers published in refereed conferences are excluded, conference papers included in the dataset may come from more than one conference, and this diversity helps researchers gain a broad understanding of different conferences, different research communities, and different research topics.

This paper categorizes authors according to the type of institution they belong to. The names of the authors and their institutions at the time of publishing the corresponding paper are given in the original dataset. Therefore, while the authors' institutional affiliations and their categories may have changed over their academic careers, they will change accordingly in our dataset. To ensure the accuracy of the data, we also extracted a portion of the data for manual inspection. Specifically, we randomly selected approximately 50 authors along with their corresponding affiliations, manually retrieved the types of these affiliations, and verified their consistency with the type labels in the dataset. The results indicated that the error rate is quite low (<5%).

4 Methods

In this section, we clarify the indicators and methods used for each research question. In the *productivity patterns and preferred venues* part, visualization is used to compare the conferences that published the largest number of the three types of articles. In the *citations and downloads* part, we explore the correlation between citations and downloads for the three types of articles using heat maps of the correlation coefficient matrix. To further explore the relationship between downloads and citations, we introduced a simple metric, namely conversion rate, which is calculated as:

$$CR_i = citations_i / downloads_i \quad (1)$$

In Equation (1), CR_i stands for conversion rate and i refers to a scientific publication. $citations_i$ and $downloads_i$ indicate the number of citations and downloads of publication i. This metric is used to measure how many cumulative downloads of each article will be converted into real citations.

In the *research topic analysis* part, a pre-trained BERT is employed to extract keywords and get the five most important words for each year and each paper type to investigate the potential changes in research topics over time. To investigate the potential changes in research topics among the three types of articles over time, particularly for papers resulting from collaborations between academia and industry, we employ a large-scale pre-trained language model called DistilBERT for keyword extraction. DistilBERT is distilled from BERT using Knowledge Distillation techniques. Compared to BERT, it has a smaller model size and faster inference speed, making it more efficient and flexible (Sanh et al., 2020). We first divide articles by their types and publication years and combine the title and abstract of an article into one document. Next, we use CountVectorizer to

extract n-grams (phrases) from the text as candidate keywords. For the length of the keywords, we tested n-grams ranging from two to five. Based on the results, n-grams with a length of two yielded the most semantically complete keywords. After obtaining the document and candidate keywords, we load the pre-trained SentenceTransformer (Sanh et al., 2020) model and compute the vector representations for both the text and the candidate keywords. We then calculate the cosine similarity between the document vector and the candidate keyword vectors, selecting the top five keywords with the highest similarity to the document as the keywords for that paper type and year. In this way, the top five keywords that best represent the research topics of corresponding year and paper type are obtained, as shown in Table A2. In the *scientific collaborations* part, we do some team size-wise explorations on different types of publications as a supplement.

5 Analysis and results

This section focuses on presenting the findings obtained from the analysis of the pre-processed ACM publication dataset discussed earlier. The research questions are addressed across four main areas, namely productivity patterns and preferred venues for three types of articles, the relationship between the number of citations and downloads of the three types of papers and the derived questions, changes in dissertation research topics over time, changes in the number of partners and variations in the number of partnerships and changes in partnerships.

5.1 Productivity patterns and preferred venues

There are a total of 2,001 different conferences in the dataset. Figure 3 shows the top 50 conferences in the dataset ranked by the number of publications, along with their respective publication counts. It can be observed that in the field of IR, the conference with the largest cumulative number of submissions is CHI (Conference on Human Factors in Computing Systems), followed by WWW (International World Wide Web Conference), CIKM (Conference on Information and Knowledge Management), and SIGIR (Special Interest Group on Information Retrieval). Figure 4 presents visual representations of the top ten published conferences within each of the three categories: Academia, Industry, and Academia-Industry Collaboration. The pie charts provide a clear overview of the distribution of publications among these conferences. The percentages in the figure represent the proportion of articles of the corresponding conference within that category. The top left figure represents Academia, the top right figure represents Industry, and the bottom figure represents the Academia-Industry Collaboration. Conferences are shown in the legend from top to bottom in descending order of representation. In the three categories, CHI, WWW and CIKM hold the top three positions. However, their rankings vary within each category. Interestingly, WWW emerges as the top conference in both the Industry and Academia-Industry Collaboration categories. Furthermore, the top five conference rankings in these two categories remain consistent. This observation suggests that when a co-author from industry is involved, the paper's content is more likely to have an industrial focus, leading to a preference for conferences aligned with purely industrial papers.

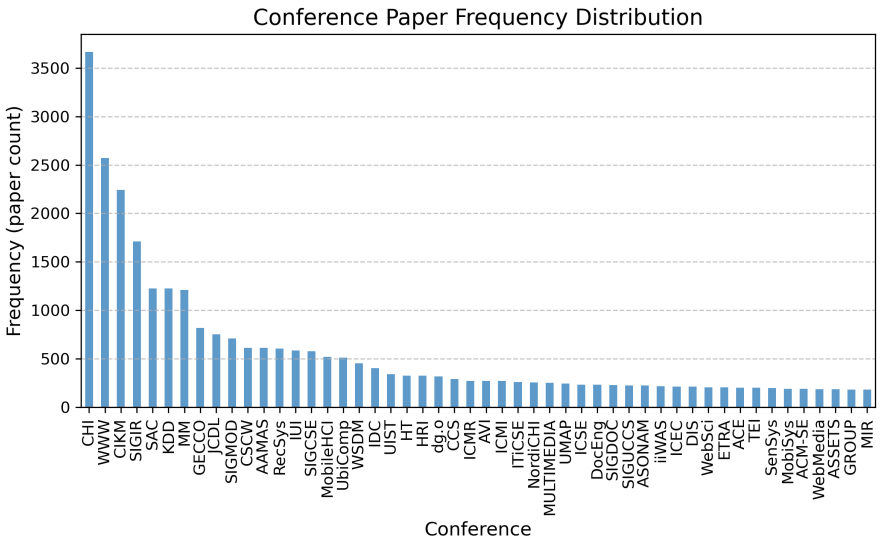


Figure 3. The distribution of conference frequency.

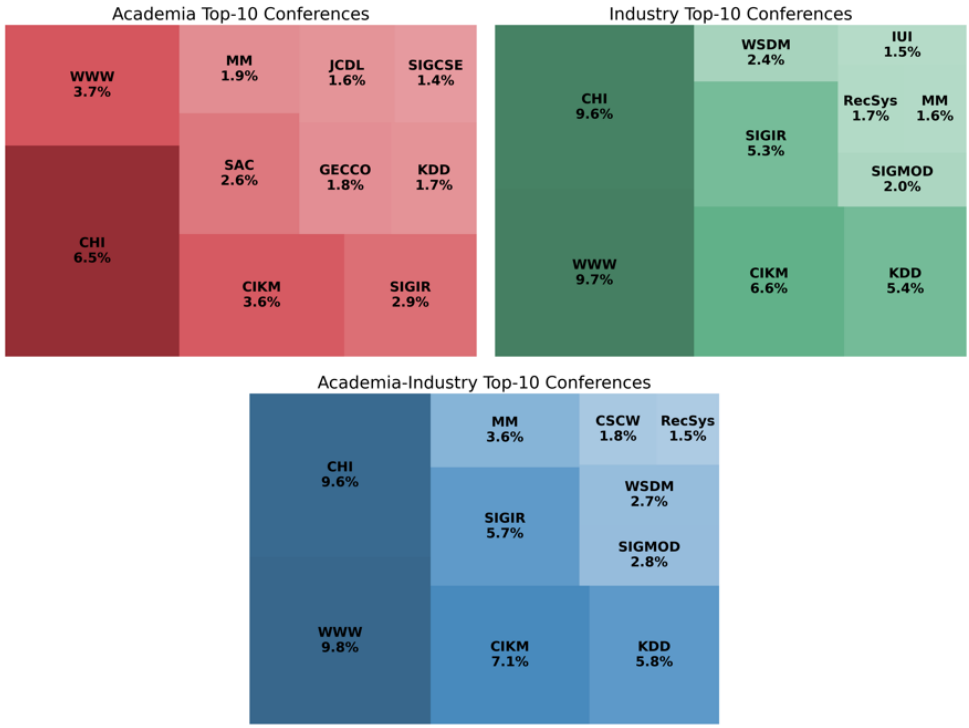


Figure 4. The top-10 published conferences in the three categories.

Notably, two conferences, RecSys (The ACM Conference Series on Recommender Systems) and CSCW (Conference on Computer supported cooperative work), only appear in the top ten list of conferences for the Academia-Industry Collaboration category. This finding indicates that these conferences are more prevalent in collaborative research efforts between academia and industry. Overall, these insights shed light on the conference preferences and content orientations within each category, highlighting the influence of industry collaboration on the publication choices and directions of research papers.

5.2 Citations and downloads

With the development of the Internet, the electronification of academic papers is becoming more and more popular, and almost all journal papers are able to be accessed through online databases. Before a paper is cited, there will be browsing, downloading, reading and other use behaviors, of which downloading behavior is more easily recorded and targeted. Therefore, downloads have gradually become one of the mainstream and important alternative measures (Schloegl & Gorraiz, 2011). Supplementing the analysis of downloads and citations enables a more comprehensive assessment of the impact of the focal paper, and also helps to understand the dynamics of the paper's dissemination and acceptance process, thus facilitating more effective academic communication and knowledge dissemination. To initially explore the relationship between the number of citations and downloads, we calculated the Pearson's correlation coefficients between the three types of data and visualized them in heat maps, as shown in Figure 5. In the figure, citation_count, downloads_cu, downloads_12month, and downloads_6week represent the total number of citations, the total number of downloads, the number of downloads in the last year, and the number of downloads in the last six weeks (during the time frame from which the dataset is acquired). The results of the Academia, Industry, and Academia-Industry Collaboration are shown from the top to the bottom panels of Figure 5. It is evident that Academia and Academia-Industry Collaboration exhibit similar patterns and regularities while Industry shows distinct differences. Additionally, the figure reveals a strong correlation between the number of citations and the cumulative number of downloads for all three types of papers, with correlation coefficients exceeding 0.60. This suggests that the number of citations and cumulative downloads can serve as predictors for each other. In particular, the correlation coefficients between downloads within six weeks and downloads within one year are particularly high for all the three types of papers (greater than 0.7). In the case of Industry, this coefficient even reaches 0.94. However, when examining the correlation between the total number of downloads and the number of downloads within six weeks or within one year, the coefficients are relatively small. This indicates that the significant variation in downloads over time is not adequately captured by the total download count. Moreover, the correlation between citations and downloads strengthens when a team include scientists from academia. Conversely, in the case of Industry, the correlation between citations and downloads within six weeks surpasses the correlation between downloads within one year.

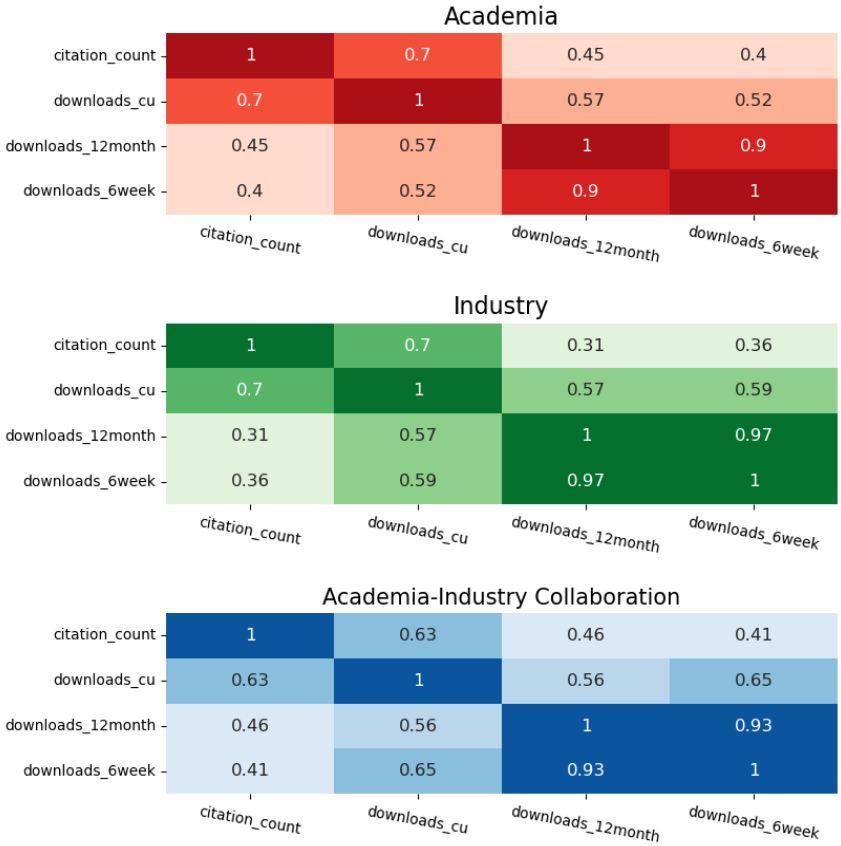


Figure 5. Heat map of the correlation coefficient matrix for the three types of articles.

To present the findings more effectively, raw data was grouped into 15 bins according to their number of downloads. Figure 6 clearly demonstrates that scientific publications originating from Academia-Industry Collaboration consistently exhibit higher conversion rates compared to articles purely from Academia and purely from Industry, regardless of the number of downloads. Moreover, a One-Way ANOVA on the conversion rates of the three types of publications and found a significant difference in their mean conversion rates (F-statistic: 364.2675, P-value: 0.0000). These suggests that, for an equivalent number of downloads, articles by Academia-Industry Collaborations are more likely to receive citations. These results provide invaluable insights into the impact of collaborative efforts between academia and industry on citation rates. A greater conversion rate observed in the Academia-Industry Collaboration category may imply the added value and influence of this collaborative research approach.

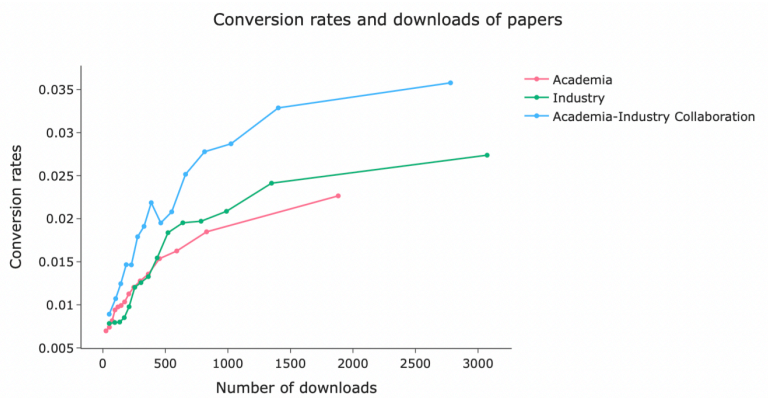


Figure 6. Conversion rate for the three types of articles.

5.3 Research topic analysis

A pre-trained BERT is employed to extract keywords and get the five most important words for each year and each paper type to investigate the potential changes in research topics over time. The top 5 keywords that best represent the research topics of corresponding year and paper type are obtained, as shown in Table A2. Overall, we found that academic research tends to involve topics of diverse types, whereas industry research usually focuses on specific platforms, data formats, and commercial applications. In addition, compared to academic research that main focuses on digital libraries and algorithmic improvement, industry research has studied a series of practical challenges related to multimedia content processing (e.g., video, music), Web search, and social media platforms. With respect to the collaborations between two communities, the topics involve both algorithmic studies and tool- or dataset-specific experiments, which are likely to bring together the scientific research strength of academia and industry's advantages in tools, platforms, and datasets generated by large user pools. In recent years, academia-industry collaborative research also pays increasing attention to human-centered topics and methods, such as online and social media bullying, chatbot developers, and crowdsourcing analytics. This phenomenon is aligned with recent growing trends in human-centered computing research that emphasizes not only algorithmic and system-oriented effectiveness, but also individual users' experiences, fairness and ethics, as well as broader cultural and societal impacts.

We then pay particular interests in the semantic similarity (measured by the cosine distance between two vectors representing the publications) among the three types of publications, demonstrated in Figure 7. While no obvious increasing or decreasing trends are observed in any of the categories, the similarity values between the categories range predominantly between 0.5 and 0.8. These findings suggest that research conducted within academia and industry may either align or diverge in preferences from year to year. However, collaborative papers between academia and industry contribute to "shifting" these dynamics, potentially resulting in either increased or decreased similarities between the two.

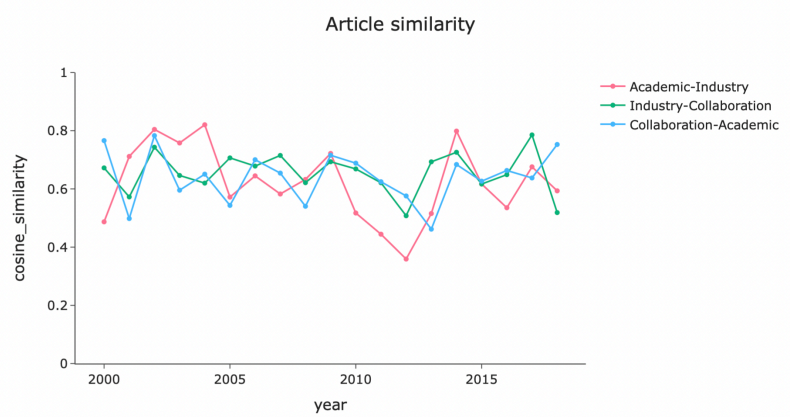


Figure 7. Variation of cosine similarity with year for three types of articles. "Academic-industry" indicates the similarity between publications by authors purely from academia and publications by authors purely from industry. "Industry-collaboration" indicates the similarity between publications by authors purely from industry and publications co-authored by scientists from academia and industry. "Collaboration-academic" refers to the similarity between publications co-authored by scientists from academia and industry and publications by authors purely from academia.

5.4 Scientific collaborations

Figure 8 shows the percentage of the number of co-authors for the three types of papers in the form of a line graph. The proportion of single-author articles published by industry is 27.29%, significantly greater than that by academia (12.12%). This observation implies that researchers from industry have a stronger inclination towards conducting independent research. On the other hand, when the number of co-authors exceeds four, the percentage of Academia-Industry Collaboration consistently surpasses that of both Industry and Academia, which suggests that when academia-industry collaborations mostly occur in large teams.

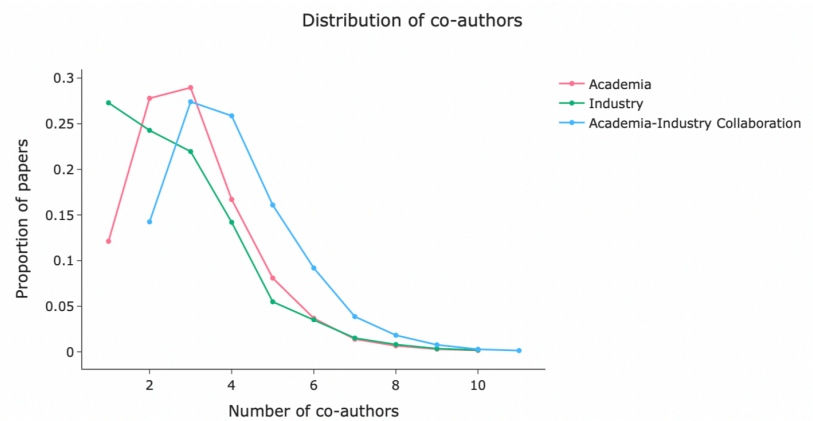


Figure 8. Distribution of the number of co-authors for each type of publications. Since the Academia-Industry Collaboration is defined as having at least one author from academia and one author from industry, the blue curve representing Academia-Industry Collaboration starts with a number of co-authors of two.

6 Discussion

In this paper, we explore four aspects of IR research: productivity patterns and preferred venues, the relationship between citations and downloads, changes in research topics, and changes in patterns of scientific collaboration, by analyzing and comparing publication pairs from both industrial and academic researchers in the field of IR. In the productivity patterns and preferred venues part (RQ1), we found that the inclusion of authors from industry makes the research content more likely to have an industrial focus, leading to a preference for conferences aligned with purely industrial papers. In the citations and downloads part (RQ2), we found that the relationship between citations and downloads is similar for Academia and Academia-Industry Collaboration, but differs more significantly for Industry, with Academia-Industry Collaboration more likely to achieve higher download conversion rates, suggesting that collaboration can increase the impact of research. In the research topic analysis part (RQ3), we found that Academic research covers diverse topics, while industry research focuses on specific platforms, data formats, and commercial applications; Collaborations between academia and industry involve both algorithmic studies and tool- or dataset-specific experiments; Also, recent academia-industry collaborative research pays increasing attention to human-centered challenges, research topics and methods, such as cyberbullying, chatbot development, and crowdsourcing analytics. In the scientific collaborations part (RQ4), we found that, among the collaboration models, Academia-Industry Collaboration is more oriented towards large teamwork. These conclusions help the Industry and Academia understand each other's research characteristics, thereby better promoting practical cooperation between them.

Moreover, our research holds significant theoretical implications. We employed a combination of strategies, such as utilizing ACM and matching multiple datasets, to obtain the most comprehensive and accurate dataset possible, and serves as an initial step toward understanding how collaborations, research topics, and productivity evolve over time in IR community, a key interdisciplinary field in computing research. This study provides a more comprehensive exploration of academic-industry collaboration in IR in terms of content, citations, and modes of collaboration. For researchers in the field of IR, this paper reveals the impact and benefits of collaboration between academia and industry, encourages active collaboration between researchers in both fields and advances science in IR; for researchers studying the patterns of collaboration between academia and industry, this paper differs from other articles that start with industrial topics, and defines the research area in a field that is "binational" in nature - information retrieval - providing new research ideas, directions, and testbeds.

Undoubtedly, this study has certain limitations that should be addressed and could inspire further explorations. Firstly, our dataset only covers information up to 2018, and considering that IR is constantly evolving, there may have been significant developments among certain sub-areas, such as conversational IR, neural IR, explainability, and LLM-enabled chat search, in recent years. Acquiring more up-to-date data would enhance the robustness of our conclusions and strengthen

their validity in capturing recent trends. Secondly, similar to other computing fields, IR community consider conferences have been a driving force in publishing latest breakthroughs and setting the research agenda in IR. However, since few co-authored papers on planning are included in our chosen dataset, we may not be able to get an accurate and complete understanding of the history of IR when exploring the evolution of IR field and the change of research topics. Additionally, since we did not have access to full-text data, our analysis of research topics was confined to title and abstract pairs. The dataset we obtained does not provide time-windowed citation data, thus we can only use cumulative citation counts. Having complete data and more fine-grained features would have allowed for a more comprehensive exploration of the changes in research topics and team structures over time across the three types of papers. More data can also help supplement control variables in the regression model, such as the impact factor of the journal where the paper is published, the average citation number of all the authors, etc., to enhance the robustness of the regression. In addition, ACM dataset do not contain some of the journal papers in the IR field. However, we initially used the MAG dataset to find areas related to the field of IR, which involves a smaller content of papers, and most of the papers studying IR are published as conference papers. Therefore, ACM was chosen as a more complete reflection of the history of the field of IR than MAG. While this study provides insights into the patterns of knowledge production in the field of IR, it does not delve deeply into the underlying reasons behind the different patterns observed for the three types of knowledge production (academic, industrial, and collaborative). Additionally, the exploration of the evolution of collaboration models and research topics is limited to descriptive analysis. Future research could build upon this by, for example, exploring the impact of different collaboration models on the influence of the research itself and their contributions to the field of IR through causal inference methods.

To further enhance the study, future research should aim to acquire more recent data, access complete article texts, and employ advanced techniques for a more nuanced analysis of collaboration patterns, research topics, and target research problems that potentially connect academic studies and industry applications. Because of the limited content of datasets we had access to, our definition of collaboration was limited to cases where scientists co-authored papers, which may missed out other possible forms of collaboration that might not necessarily result in peer-reviewed publications, such as graduate students doing internships in companies, professors acting as consultants to companies and producing informal reports, and academia-industry collaborative grant proposal development and system design. In future research, one could explore richer collaborations that might raise more interesting questions and provide some hands-on experience, and also investigate the available resources and infrastructure, policies in industry and research institutions, and regulations that may affect the form and scope of collaborations. In addition, with richer and up-to-date data on team collaborations, researchers could investigate academia-industry collaboration patterns in a wider range of information interactions that evolve rapidly, such as conversational search, retrieval-augmented generation (RAG), and LLM-enabled chat interactions.

Acknowledgements

An early version of this paper was presented at iConference 2023. The authors are grateful to the two anonymous reviewers for their insightful comments.

Funding information

Yi Bu's participation in this work was in part supported by the National Science Foundation of China (#24&ZD072).

Conflict of interests

Yi Bu is an editorial board member of the *Journal of Data and Information Science*.

Author contributions

Jiaqi Lei (radium@stu.pku.edu.cn): Investigation, Methodology, Visualization, Writing - original draft;

Liang Hu (huliang@stu.pku.edu.cn): Investigation, Methodology, Visualization, Writing - original draft;

Yi Bu (buyi@pku.edu.cn): Funding acquisition, Supervision, Writing - review & editing;

Jiqun Liu (jiqunliu@ou.edu): Project administration, Supervision, Validation, Writing - review & editing.

References

Ahmed, N., Wahed, M., & Thompson, N. C. (2023). The growing influence of industry in AI research. *Science*, *379*(6635), 884–886. https://doi.org/10.1126/science.ade2420

Castillo, C. (2019). Fairness and Transparency in Ranking. *ACM SIGIR Forum*, *52*(2), 64–71. https://doi.org/10.1145/3308774.3308783

Culpepper, J. S., Diaz, F., & Smucker, M. D. (2018). Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *ACM SIGIR Forum*, *52*(1), 34–90. https://doi.org/10.1145/3274784.3274788

Ekstrand, M. D., Burke, R., & Diaz, F. (2019). Fairness and Discrimination in Retrieval and Recommendation. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1403–1404. https://doi.org/10.1145/3331184.3331380

Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: From National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research Policy*, *29*(2), 109–123. https://doi.org/10.1016/S0048-7333(99)00055-4

Gao, J., Xiong, C., & Bennett, P. (2020). Recent Advances in Conversational Information Retrieval. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2421–2424. https://doi.org/10.1145/3397271.3401418

Gao, R., & Shah, C. (2021). Addressing Bias and Fairness in Search Systems. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2643–2646. https://doi.org/10.1145/3404835.3462807

Garfield, E. (1996). Fortnightly Review: How can impact factors be improved? *BMJ*, *313*(7054), 411–413. https://doi.org/10.1136/bmj.313.7054.411

Hu, B., Ding, Y., Dong, X., Bu, Y., & Ding, Y. (2021). On the relationship between download and citation counts: An introduction of Granger-causality inference. *Journal of Informetrics*, *15*(2), 101125. https://doi.org/10.1016/j.joi.2020.101125

Jasny, B. R., Wigginton, N., McNutt, M., Bubela, T., Buck, S., Cook-Deegan, R., Gardner, T., Hanson, B., Hustad, C., Kiermer, V., Lazer, D., Lupia, A., Manrai, A., McConnell, L., Noonan, K., Phimister, E., Simon, B., Strandburg, K., Summers, Z., & Watts, D. (2017). Fostering reproducibility in industry-academia research. *Science*, *357*(6353), 759–761. https://doi.org/10.1126/science.aan4906

Jazi, S. Y., Mirzaeinia, A., & Jazi, S. Y. (2024). *Analyzing Gender Polarity in Short Social Media Texts with BERT: The Role of Emojis and Emoticons*. https://doi.org/10.13140/RG.2.2.15772.50568

Keyvan, K., & Huang, J. X. (2023). How to Approach Ambiguous Queries in Conversational Search: A Survey of Techniques, Approaches, Tools, and Challenges. *ACM Computing Surveys*, *55*(6), 1–40. https://doi.org/10.1145/3534965

Kobayashi, M., & Takeda, K. (2000). Information retrieval on the web. *ACM Computing Surveys*, *32*(2), 144–173. https://doi.org/10.1145/358923.358934

Lei, J., Bu, Y., & Liu, J. (2023). Information Retrieval Research in Academia and Industry: A Preliminary Analysis of Productivity, Authorship, Impact, and Topic Distribution. In I. Sserwanga, A. Goulding, H. Moulaison-Sandy, J. T. Du, A. L. Soares, V. Hessami, & R. D. Frank (Eds.), *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity* (Vol. 13972, pp. 360–370). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28032-0_29

Li, H., & Lu, Z. (2016). Deep Learning for Information Retrieval. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1203–1206. https://doi.org/10.1145/2911451.2914800

Liu, J. (2021). Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management*, 58(3), 102522. 10.1016/j.ipm.2021.102522

Marijan, D., & Gotlieb, A. (2021). Industry-Academia research collaboration in software engineering: The Certus model. *Information and Software Technology*, *132*, 106473. https://doi.org/10.1016/j.infsof.2020.106473

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). 10.1145/3287560.3287596

Noyons, E. C. M., Van Raan, A. F. J., Grupp, H., & Schmoch, U. (1994). Exploring the science and technology interface: Inventor-author relations in laser medicine research. *Research Policy*, *23*(4), 443–457. https://doi.org/10.1016/0048-7333(94)90007-8

Olteanu, A., Garcia-Gathright, J., De Rijke, M., Ekstrand, M. D., Roegiest, A., Lipani, A., Beutel, A., Olteanu, A., Lucic, A., Stoica, A.-A., Das, A., Biega, A., Voorn, B., Hauff, C., Spina, D., Lewis, D., Oard, D. W., Yilmaz, E., Hasibi, F., … Kamishima, T. (2019). FACTS-IR: Fairness, accountability, confidentiality, transparency, and safety in information retrieval. *ACM SIGIR Forum*, *53*(2), 20–43. https://doi.org/10.1145/3458553.3458556

Owen-Smith, J. (2003). From separate systems to a hybrid order: Accumulative advantage across public and private science at Research One universities. *Research Policy*, *32*(6), 1081–1104. https://doi.org/10.1016/S0048-7333(02)00111-7

Perkmann, M., & Walsh, K. (2009). The two faces of collaboration: Impacts of university-industry relations on public research. *Industrial and Corporate Change*, *18*(6), 1033–1065. https://doi.org/10.1093/icc/dtp015

Rani, Y. A., Balaram, A., Sirisha, M. R., Nabi, S. A., Renuka, P., & Kiran, A. (2024). AI Enhanced Customer Service Chatbot. *2024 International Conference on Science Technology Engineering and Management (ICSTEM)*, 1–5. https://doi.org/10.1109/ICSTEM61137.2024.10561155

Rhoten, D., & Powell, W. W. (2007). The Frontiers of Intellectual Property: Expanded Protection versus New Models of Open Science. *Annual Review of Law and Social Science*, *3*(1), 345–373. https://doi.org/10.1146/annurev.lawsocsci.3.081806.112900

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (No. arXiv:1910.01108). arXiv. http://arxiv.org/abs/1910.01108

Schloegl, C., & Gorraiz, J. (2011). Global usage versus global citation metrics: The case of pharmacology journals. *Journal of the American Society for Information Science and Technology*, *62*(1), 161–170. https://doi.org/10.1002/asi.21420

Serajian, M., Marini, S., Alanko, J. N., Noyes, N. R., Prosperi, M., & Boucher, C. (2023). *Scalable De Novo Classification of Antibiotic Resistance of Mycobacterium Tuberculosis*. Bioinformatics. https://doi.org/10.1101/2023.11.16.567394

Shahin, M., Chen, F. F., Hosseinzadeh, A., Maghanaki, M., & Eghbalian, A. (2024). A novel approach to voice of customer extraction using GPT-3.5 Turbo: Linking advanced NLP and Lean Six Sigma 4.0. *The International Journal of Advanced Manufacturing Technology*, *131*(7–8), 3615–3630. https://doi.org/10.1007/s00170-024-13167-w

Spicer, A. J., Colcomb, P.-A., & Kraft, A. (2022). Mind the gap: Closing the growing chasm between academia and industry. *Nature Biotechnology*, *40*(11), 1693–1696. https://doi.org/10.1038/s41587-022-01543-4

Thomas, P., Czerwinksi, M., Mcduff, D., & Craswell, N. (2021). Theories of Conversation for Conversational IR. *ACM Transactions on Information Systems*, *39*(4), 1–23. https://doi.org/10.1145/3439869

Van Looy, B., Callaert, J., & Debackere, K. (2006). Publication and patent behavior of academic researchers: Conflicting, reinforcing or merely co-existing? *Research Policy*, *35*(4), 596–608. https://doi.org/10.1016/j.respol.2006.02.003

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science*, *316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099

Yates, A., Nogueira, R., & Lin, J. (2021). Pretrained Transformers for Text Ranking: BERT and Beyond. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 1154–1156. https://doi.org/10.1145/3437963.3441667

Zaharia, N., & Kaburakis, A. (2016). Bridging the Gap: U.S. Sport Managers on Barriers to Industry–Academia Research Collaboration. *Journal of Sport Management*, *30*(3), 248–264. https://doi.org/10.1123/jsm.2015-0010

Zamani, H., Dumais, S., Craswell, N., Bennett, P., & Lueck, G. (2020). Generating Clarifying Questions for Information Retrieval. *Proceedings of The Web Conference 2020*, 418–428. https://doi.org/10.1145/3366423.3380126

Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2018). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology*, *69*(1), 72–86. https://doi.org/10.1002/asi.23916



Copyright: © 2026 Jiaqi Lei, Liang Hu, Yi Bu, Jiqun Liu. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Appendix

Table A1. ACM article keywords/phrases (ranked in an alphabetical order by column; all keywords/phrases lowercased).

accountability information retrieval	explainability information retrieval	navigation	sentiment analysis
active learning	exploratory search	neural network	similarity
adaptation	eye tracking	neural networks	similarity measure
annotation	faceted search	novelty	similarity search
annotations	fairness information retrieval	online social networks	social media
audio	feature extraction	ontologies	social network
augmented reality	feature selection	ontology	social network analysis
benchmark	federated search	open data	social networks
big data	filtering	opinion mining	social search
blog	flickr	optimization	social tagging
browsing	folksonomy	P2P	spam
caching	geographic information retrieval	pagerank	spoken search system
CBIR	graph mining	passage retrieval	sponsored search
children	group recommendation	peer-to-peer	summarization
classification	hashing	performance	supervised learning
cloud computing	human-computer interaction	performance evaluation	SVM
clustering	image annotation	personal information management	tagging
collaboration	image classification	personalization	tags
collaborative filtering	image retrieval	personalized search	test collection
collaborative tagging	image search	privacy	test collections
community detection	implicit feedback	pseudo relevance feedback	text categorization
complex event processing	index	pseudo-relevance feedback	text classification
content analysis	indexing	query	text mining
content-based filtering	information extraction	query classification	time series
content-based image retrieval	information filtering	query expansion	topic model
content-based retrieval	information retrieval	query formulation	topic modeling
context	information seeking	query intent	topic models
context-awareness	information visualization	query log analysis	transfer learning
conversational information retrieval	interaction	query logs	transparency information retrieval
convolutional neural networks	interactive information retrieval	query performance prediction	trust
correlation	interoperability	query processing	twitter
credibility	inverted index	query reformulation	unsupervised learning
cross-language information retrieval	kernel methods	query suggestion	usability
cross-modal retrieval	keyword search	question answering	user behavior
crowdsourcing	knowledge base	random walk	user interaction
data integration	knowledge management	ranking	user interface

accountability information retrieval	explainability information retrieval	navigation	sentiment analysis
data mining	language model	RDF	user interfaces
database	language modeling	recommendation	user modeling
deep learning	language models	recommendation system	user profile
digital humanities	learning	recommendation systems	user profiling
digital libraries	learning to rank	recommender system	user studies
digital library	lifelogging	recommender systems	user study
digital preservation	link analysis	relation extraction	video
dimensionality reduction	linked data	relevance	video analysis
distributed information retrieval	locality sensitive hashing	relevance feedback	video annotation
diversification	location-based services	re-ranking	video retrieval
diversity	log analysis	responsible information retrieval	video search
document clustering	machine learning	retrieval	video summarization
document representation	machine translation	retrieval models	visualization
document retrieval	MapReduce	sampling	web
e-commerce	matrix factorization	scalability	web 2.0
education	measurement	search	web mining
efficiency	metadata	search behavior	web search
e-government	mobile	search engine	web search engine
emotion	mobile computing	search engines	web service
enterprise search	mobile devices	semantic relatedness	web services
entity linking	multimedia	semantic search	wiki
ethnics information retrieval	multimedia retrieval	semantic similarity	wikipedia
evaluation	music	semantic web	word embeddings
event detection	music information retrieval	semantics	world wide web
events	music recommendation	semi-supervised learning	XML
experimentation	named entity recognition	sensor networks	XML retrieval
expert finding	natural language processing		

Table A2. Research topics over time.

	Academia	Industry	Academia-Industry Collaboration
2000	intelligent libraries library classes library technologies novel browser internet classrooms	video classroom video watermarking video performance video technical video recording	learning algorithms analysis hashing proliferation internet classifies algorithms learning algorithm
2001	mouse popular mouse 3d popular multiplayer 3d computing powerful 3d	advanced algorithms computationally feasible researchers improve modeling useful interface powerful	software engineers designer needs designing ontology xml software documentation engineers

	Academia	Industry	Academia-Industry Collaboration
2002	libraries tutorial databases attractive library technology efficient indexing databases tutoring	designing web auctions improving expanding rehearsal bioinformatics emerging search engines	offering algorithms tackling algorithms semantic web algorithms software web query
2003	optimization cancer audition algorithms partitioning algorithms algorithms comparison algorithms haplotype	music photo2video retrieves songs music concert music database new songs	simplifying web internet experiments popular web online semantic new spam
2004	algorithm updating novel algorithms algorithms lessons developing algorithms algorithms learning	toolkit debugging browser optimizations search algorithms web verification algorithms methodology	simplifying web internet experiments popular web online semantic new spam
2005	valuable indexing algorithms improving efficient ontology interesting research important crosscutting	algorithms methodology magic instructional data webgazeanalyzer webgazeanalyzer brings laser pointer	holistic algorithms novel internetworking smart phones algorithms scalability wireless broadband
2006	valuable indexing algorithms improving efficient ontology interesting research important crosscutting	research challenging executed twice algorithms counter bioinformatics motivation steep learning	challenge traffic traffic engineering algorithms damping algorithms research engineering algorithms
2007	servers cheating privacy vulnerabilities online personalization leakage internet online anonymity	huge database hashing billions largest commerce amazon highly huge collections	major browsers algorithms large algorithms widely extensive programming fastdash developer
2008	wikipedia huge browsers popular huge databases favourite websites winning podcasts	study robots survey robots querybuilder query querylogs bundling botnet detection	algorithm recommender study novel research recent tagging podcasts researchers paper
2009	rfid popular patents important rfid algorithms patents essential tagging expert	browsers actively growth online online advertising online surveys pushing browsers	internet researchers importance researchers benchmarking browsers research researchers data researchers
2010	new algorithms hashtag innovation evolving wiki new apps discovery bioinformatics	research revolutionize expensive simulators database huge budget challenging consumption skyrocketed	wikiprojects increased pagerank algorithm playlists photoselect brainstorming stylus retagging online
2011	design tutorial bioinformaticians designing designing sparql clustering bioinformatics tutorials technologies	driver safety automobiles tutorial refocus driving driver infotainment opportunistic driver	attractive websites understanding internet moderating online good websites modeling internet

	Academia	Industry	Academia-Industry Collaboration
2012	comics techcommix cancer increased cyberinfrastructure scientist algorithms physicists scientists seeking	improving tweet threefold allows reducing aliasing algorithm reducing strategies threefold	avatar conferencing rearranging videos video hashing video tutorials media tutorials
2013	detectives solving algorithms greedy played detectives crime notepad detecting cyberbullying	latest poll proposed algorithm tweeted wedding motivation graphbuilder research innovations	project openstreetmap tutorial overview openstreetmap editors pattern openstreetmap modeling tutorial
2014	discriminating online misinformation crowdturfing instagram traffic kickstarter interview internet challenging	improving online improvements online google overtaking web analytics wikipedia benefitted	pagerank algorithm web searchers search engines search engine online surveys
2015	apps revolutionize apps changing simplifying mobile investigating smartphone developing smartwatch	improve genetic experiments expensive expensive experiments novel algorithms quantitative genetic	youtube tutorials movielens netflix netflix datasets netflix dataset youtube flickr
2016	learning analytics analytics learning agile analytics learning smartwatches learning evolution	google facebook facebook microsoft facebook conducting traffic staggering economics rigorously	chatbots developers needed mobile designing android prototype chatbot apps study
2017	videos rebooting videoconferencing application video tutorials video study videoconferencing	google facebook facebook microsoft facebook conducting traffic staggering economics rigorously	online bullying facebook misleading bullying twitter online distractions reducingcontroversy homepage
2018	videos rebooting videoconferencing application video tutorials video study videoconferencing	important research seismic interpreters cryptography needed noisy training research challenges	crowdsourcing analytics reuse networking cache management cache reuse web transformational