Received: 21 Mar. 2022 | Revised: 29 Mar. 2023 | Accepted: 16 Apr. 2023 | Published online: 23 Apr. 2023

RESEARCH ARTICLE

Journal of Global Humanities and Social Sciences
2023,Vol. 4(2)52-58
DOI: 10.47852/bonviewGHSS23208570202

Research Progress and Hotspot of Information Retrieval Correlation based on CiteSpace

BON VIEW PUBLISHING

Jin Chen*[1]

[1]*Nanjing University,China*

Abstract: Taking 1760 journal papers of information retrieval relevance in the core database of Web of Science as the research object, using literature co-citation network and keyword co-occurrence network analysis, with information visualization as the means, this paper summarizes the research on the relevance of information retrieval. It is found that the current information retrieval correlation research network is concentrated, which is mainly divided into two main knowledge groups, retrieval algorithm and correlation cognition, with few frontier branches. And the knowledge fusion between the two groups needs to be strengthened. Since the emergence of natural language processing, information retrieval relevance has been improved along the path of natural language processing-machine learning-deep learning.

Keywords: information retrieval; Correlation; Citespace

1. Introduction

Relevance has become the core concept in the field of information science since the International Scientific Information Conference was held in 1958(Mizzaro, 1997). In information science, relevance is considered to be the relationship between information or information objects (X_s) and context(Borlund, 2003). And the basis of realizing the relationship between the two is some attributes of the expected expression form reflecting relevance. Saracevic further summarizes eight attributes of correlation: Relationships, intentions, contexts (both internal and external), inferences, choices, interactions, and measures. They are all interconnected, and conceptualize the general understanding of relevance(Cosijn & Ingwersen, 2000). However, these general understandings cannot be smoothly translated into theories and models. How to obtain, measure and improve relevance in information retrieval systems is a perennial research topic in the field of information science(Allan et al., 2012).

In recent years, with the strengthening of relevance theory and the improvement of information retrieval research systems, the field of information retrieval relevance research has formed a variety of research focus and frontier branches. Therefore, identifying the origin and evolution of information retrieval correlation research, and identifying the classical theories and knowledge clustering of information retrieval correlation research is conducive to building a theoretical database of information retrieval correlation research, which provides references for scholars in this field to analyze the research status of information retrieval correlation and expand new research frontiers.

2. Data sources and processing methods

In this paper, the relevant literature of information retrieval relevance is taken as the research object, and Citespace is used for quantitative analysis of cited literature and citations.

We extract the basic knowledge in the field of information retrieval relevance, and grasp the latest progress and evolution path of information retrieval relevance research, so as to provide a basis for the research in this direction.

In order to ensure the accuracy,

Corresponding Author: Jin Chen
Nanjing University,China
Email:Chenjin_98@yeah.net

comprehensiveness and high interpretability of the original data, this paper takes the three core databases of Web of Science as data sources. Finally, it obtains 1760 pieces of literature and 32,017 self-cited literature without self-cited references.

Through preliminary analysis, the annual distribution (figure 1) and subject classification (table 1) of the literature are obtained, forming a general cognition in the field of information retrieval relevance research.

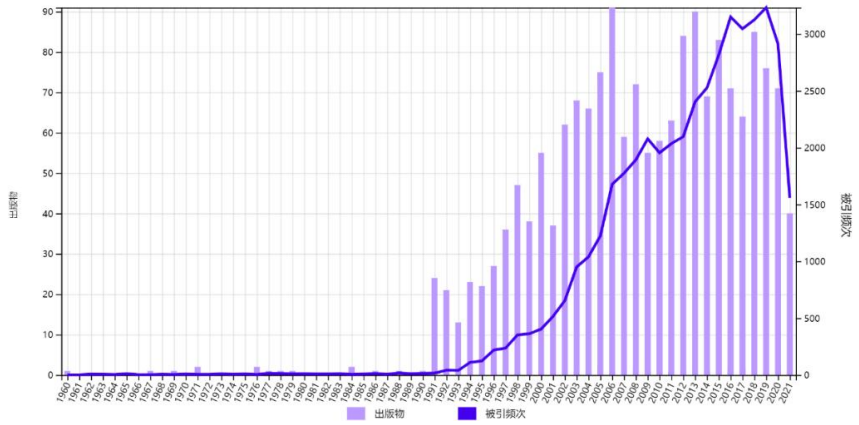

Figure 1. Citation frequency and publication distribution by year

In the 30 years from 1960, only 15 relevant pieces of literature were published in this field. It can be seen that in the embryonic stage (1960-1990), the correlation did not attract wide attention in the field of information retrieval. Since information retrieval entered the online retrieval stage in 1991, this kind of research has seen a blowout growth, and the research popularity and importance of the correlation problem of information retrieval has been on the rise, maintaining the explosive growth period of 16 years in total (1991-2006). From 2007 to 2018, the number of literature publications fluctuated slightly, while the number of citations basically kept increasing, and this kind of research entered a plateau. Until 2019, the number of publications and citations of relevant studies on information retrieval showed a declining trend, and the research entered the bottleneck stage.

Disciplines	Literature Quantity	Literature Proportion
Computer Science	1,410	80.11%
Information Science Library Science	698	39.66%
Engineering	213	12.10%
Medical Informatics	67	3.81%
Health Care Sciences Services	47	2.67%
Operations Research Management Science	40	2.27%
Telecommunications	39	2.22%
Mathematical Computational Biology	34	1.93%
Psychology	34	1.93%
Business Economics	25	1.42%
Biochemistry Molecular Biology	24	1.36%
Mathematics	17	0.97%
Physical Geography	17	0.97%
Biotechnology Applied Microbiology	16	0.91%
Imaging Science Photographic Technology	14	0.80%

Table 1. Number of literature in the top 15 disciplines

From the perspective of the distribution of research disciplines, it is mainly concentrated in computer science (80.11%), library and information science (39.66%), and engineering (12.10%). In addition, some scholars in health care, operations research, psychology, management economics, mathematics, biochemistry and molecular biology, and physical geography are also engaged in related research. The relevance of information retrieval has become a research topic of common concern in multiple disciplines. However, the authoritative literature on the relevance of information retrieval mainly focuses on the research of relevance model and information retrieval system construction, and the application research in other disciplines has not been fully developed.

3. Bibliometric analysis of information retrieval correlation research

3.1 Knowledge group identification

The research domain can be conceptualized as a time mapping from the research frontier to the knowledge base. The cited literature in the original data form the knowledge base of this field, and the corresponding citations form the research frontier(Chen, 2004). Therefore, the clustering and evolution research of cited literature is the basis of exploring research frontiers, which can reveal the critical knowledge turning points of the evolution of research frontiers and show the connections between research frontiers(Meng et al., 2020). Using the data samples mentioned above, this paper draws the knowledge map of the research field of "information retrieval relevance" based on the co-citation network, and analyzes the theoretical structure of the research field of "information retrieval relevance".

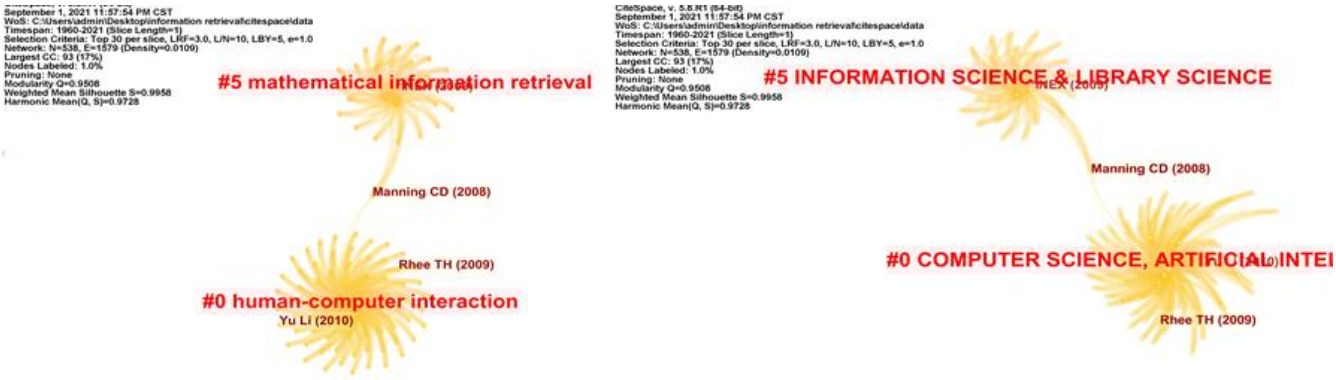

Figure 2. Literature co-citation network

Author	Time	Title	Journal
#0			
Anthony, L., Yang, J., Koedinger, K.R.	2007	Adapting handwriting recognition for applications in algebra learning	Proceedings of the ACM Workshop on Educational Multimedia and Multimedia Education
Adeel M, Cheung H S, Khiyal S H.	2008	Math go! prototype of a content based mathematical formula search engine	Journal of Theoretical & Applied Information Technology
Manning C.D., Raghavan P., Schütze H.	2008	Introduction to Information Retrieval	Cambridge University Press

Table 2. Information of node documents in literature co-citation network

Aly W, Uchida S, Suzuki M.	2008	Identifying subscripts and superscripts in mathematical documents	Mathematics in Computer Science
#5			
Fuhr N, Lalmas M, Malik S, et al.	2004	Initiative for the evaluation of XML retrieval	Proceedings of the Third Workshop (INEX 2004)
Balatsoukas, P., Morris, A., & O'Brien, A.	2009	An evaluation framework of user interaction with metadata surrogates	Journal of Information Science,
Betsi, S., Lalmas, M., Tombros, A., & Tsikrika, T.	2006	User expectations from XML Element retrieval	Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval

Table 2. Information of node documents in literature co-citation network

From the map of the literature co-cited network, the correlation research of information retrieval presents a typical initial state of research: the cited network is strong in concentration, the network overlap is high, and the research branches of the system are few. Node documents in the atlas show a strong correlation degree and have a strong interpretation of each other. In addition, some key node documents are located at the junction of knowledge groups and play a connecting role between groups, providing theoretical support and direction for subsequent research(Mou et al., 2019). Combined with the literature co-citation network map (Figure 2) and Information on key node literatures (Table 2), this paper divides the research field of information retrieval correlation into two knowledge groups: retrieval algorithm and correlation cognition.

#0 cluster is the largest cluster group in the field of information retrieval correlation research -- "retrieval algorithm" group. This group has a large amount of literature, a high degree of centrality, and high internal connection intensity. It focuses on computer science and artificial intelligence.Through the study of the original literature, it is found that the research of this group focuses on the two aspects of "text and multimedia information retrieval query recognition" and "multimedia information retrieval model construction."

#5 clustering -- the "correlation cognition" group was first cited before group #0, and the map showed that the distribution of node centrality was relatively balanced, forming a polymorphism research center and producing rich results and conclusions, which played a connecting role in the whole research field. The #5 knowledge group explores and demonstrates the theories and methods of information retrieval system research from the perspective of "cognition of user relevance", which plays a linking role in the entire research field. Its main research scope includes three aspects: user information retrieval relevance judgment basis, relevance discussion based on relevance theory, semantic information indexing and semantic information query extension.

As can be seen from the map of the literature co-citation network, there is a strong independence between group #0 and group #5. The research on "correlation cognition" in the field of library and information has not been integrated with the construction of "retrieval algorithm", and there is still a certain gap between disciplines.

3.2 Research topic identification

The distribution and evolution of research topics can intuitively reflect the changes of hot fields, analysis perspectives and research methods in

different time sequences(Zhang et al., 2021). As a concise expression of the research topic of academic papers, the relevance of keywords can reveal the internal relationship of knowledge in the subject field to a certain extent. In view of this, this paper identifies the main research directions and hot spots of information retrieval correlation research through keyword co-occurrence analysis, and makes a judgment on the development and change of the topic structure of this research field. This paper prunes and merges the keyword co-occurrence network through the PathFinder algorithm, and obtains the key path of keyword co-occurrence (Figure3，Figure4).

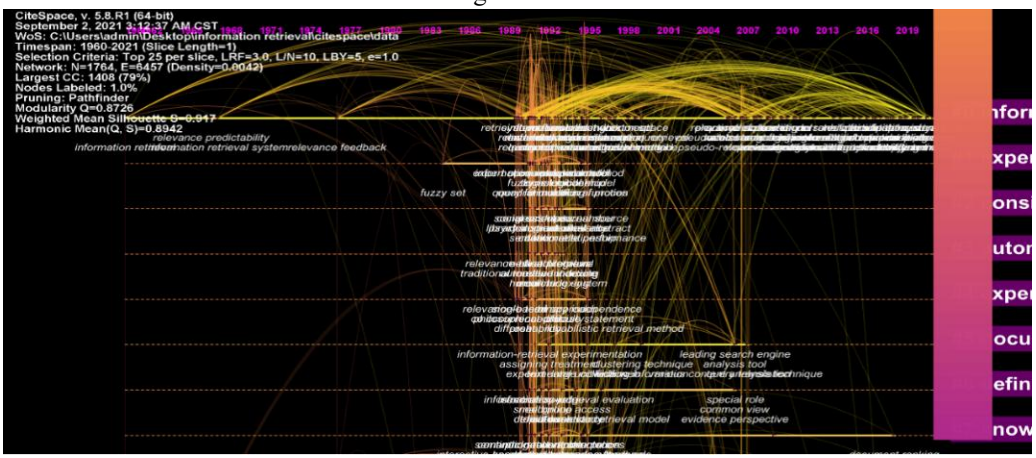

Figure 3. Topic evolution path Diagram (Part 1)

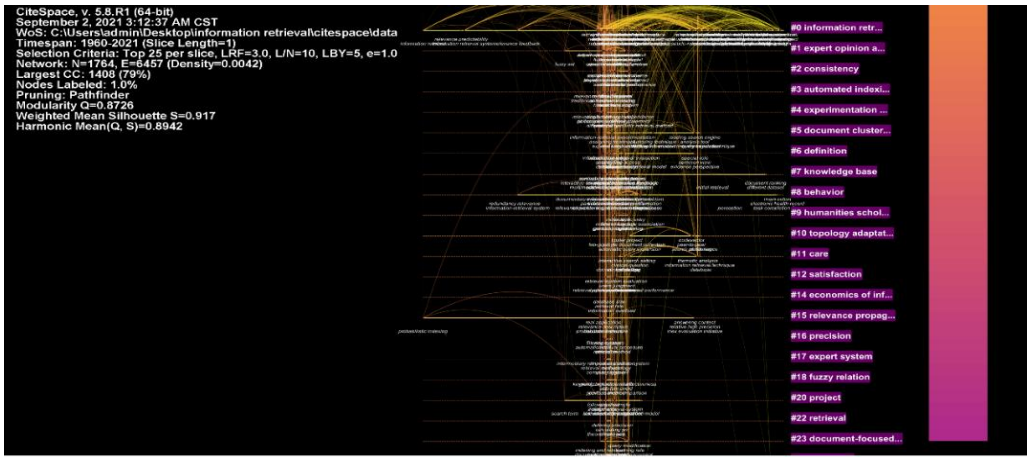

Figure 4 Topic evolution Path (global)

The correlation research of information retrieval was in its infancy from 1960 to 1990. At this stage, the research topic focuses on the introduction of relevance in information retrieval. Keywords such as information retrieval, relevance feedback and relevance feedback show strong centrality and are the core nodes of the whole research field. At this stage, scholars represented by Saracevic and Mizzaro systematically established the theoretical framework of correlation. Saracevic summed up his predecessors' work, The correlation model is divided into five types: system correlation, communication correlation, situation correlation, psychological correlation and interactive correlation(Saracevic, 1996). It lays a solid theoretical foundation for the follow-up study.

Information retrieval correlation research has developed rapidly from 1990 to 2006, and has entered a prosperous period, with diversified research topics and complex and divergent research networks. According to the analysis of co-occurrence word network, more keywords appeared in this period and had a relatively balanced centrality. The 1990s marked an important turning point in the development of online retrieval. With the rapid development of the Internet and the emergence of hypertext technology, the development of retrieval software based on client/server has realized the

transfer of the original host system to the server, so that the client/server online retrieval mode began to replace the previous terminal/host structure. At that time, the logical model began to be transformed into a retrieval model, and basic concepts of information retrieval systems such as algorithms, Boolean retrieval, and recall rates emerged one after another from the topic evolution path diagram (Figure 4).

From 2007 to 2010, the research on correlation of information retrieval had a brief period of stagnation and entered the gestation period. Although natural language processing was introduced into this field in 2009, there was still no improvement. Until 2011, machine learning injected new vitality into the correlation research of information retrieval. In 2019, deep learning further promoted the development of this field, ushering in the second spring after the outbreak of relevant literature. The information retrieval model carries the multi-wave of natural language processing, which also ushered in great progress. Relevance ranking has become a research hotspot in academia (Figure 5). On the other hand, scholars in the field of graphics take a new approach, focusing on knowledge-based and Behavior (Figure 6). At the same time, due to the birth of the computerized medical record system and the development of medical informatics in the 1990s, medical information retrieval has also become a hot research branch in this field (Figure 5).

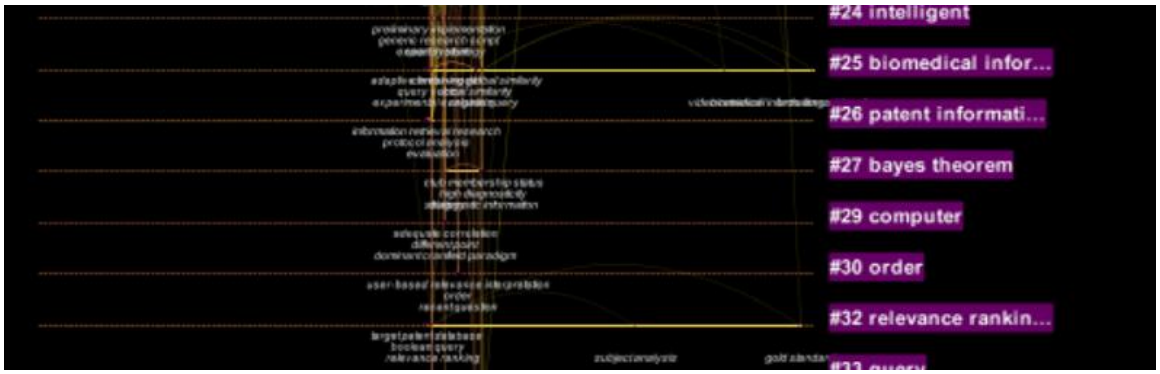

Figure 5 Topic evolution path diagram (Part 2)

Figure 6 Topic evolution path diagram (Part 3)

To sum up, the research on relevance of information retrieval started in 1960, developed by leaps and leaps after the emergence of online information retrieval in 1991, completed the construction of knowledge base in 5 years (1991-1995), and accumulated a large number of diversified researches in 15 years. Since the emergence of natural language processing, information retrieval relevance has been improved along the path of natural language processing-machine learning-deep learning. At the same time, knowledge representation, user behavior and the interdisciplinary branch of medical information retrieval have also developed.

4.Conclusion

Through the previous analysis, we found that the current research on the relevance of information retrieval reflects the following characteristics:

(1) Subsequent research was lackluster. After more than 30 years of development, the relevance of information retrieval has been further studied, and

the research on the relevance of information retrieval in multiple disciplines has generally shown a growing trend, but it has entered a bottleneck period in recent years.

(2) Knowledge presents the phenomenon of group aggregation. At present, the correlation research network of information retrieval is concentrated, which is mainly divided into two main knowledge groups — retrieval algorithm and correlation cognition. And the knowledge fusion between the two groups needs to be strengthened.

(3) There is a lack of diversity in research directions. Since around 2010, the research on the relevance of information retrieval has only focused on the relevance ranking algorithm, lacking multi-disciplinary and multi-view research branches.

In addition, this study has certain limitations. First of all, the type of data that the analysis software can handle is limited. This study only selects the literature of the Web of Science (WoS) database as the data source. Future research can further enrich the data source. Secondly, this study only discusses the international literature, and it can also be compared with the research of domestic scholars in the future.

Conflict of Interest

The authors declare that they have no conflicts of interest to this work.

References

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, *48*(9), 810–832.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, *54*(10), 913–925.

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, *36*(4), 533–550.

Allan, J., Croft, B., Moffat, A., & Sanderson, M. (2012). Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in lorne. *In Acm Sigir Forum New York, NY, USA: ACM.*, *46*(1), 2–32.

Chen, C. (2004). Searching for intellectual turning points. *Progressive Knowledge Domain Visualization. Proceedings of the National Academy of Sciences*, *101*(suppl_1), 5303–5310.

Meng, L., Wen, K. H., Brewin, R., & Wu, Q. (2020). Knowledge atlas on the relationship between urban street space and residents' health—a bibliometric analysis based on VOSviewer and CiteSpace . *Sustainability*, *12*(6), 2384.

Mou, J., Cui, Y., & Kurcz, K. (2019). Bibliometric and visualized analysis of research on major e-commerce journals using CiteSpace. *Journal of Electronic Commerce Research*, *20*(4), 219–237.

Zhang, X., Zhang, Y., Wang, Y., & Fath, B. D. (2021). Research progress and hotspot analysis for reactive nitrogen flows in macroscopic systems based on a CiteSpace analysis. *Ecological Modelling*, *4*(43), 109456.

Saracevic, T. (1996). Relevance reconsidered. *In Proceedings of the Second Conference on Conceptions of Library and Information Science* , *10*(CoLIS 2), 201–218.

How to Cite: Chen, J. (2023). Research Progress and Hotspot of Information Retrieval Correlation based on CiteSpace. *Journal of Global Humanities and Social Sciences*, 04(04),52–58.
https://doi.org/10.47852/bonviewGHSS23208570202