10 Point Scale for Evaluating Academic Search Engines

Kanishka Parankusham
dept. of Computer Science
University of South Dakota
Vermillion, USA
Kanishka.Parankusham@coyotes.usd.edu

***Abstract*—This work explores the challenges of evaluating academic search engines in a landscape where results increasingly homogenise. To address this, we propose a ten point scoring system. One of the key metrics is time based mean reciprocal rank. The evaluation process delves deeper than mere relevance and precision, incorporating factors such as user experience, advanced search functionalities, multilingual support, and open-access availability. This holistic approach provides a nuanced understanding of each search engine's performance. We demonstrate the effectiveness of our framework by evaluating three prominent academic search engines. The results reveal key strengths and weaknesses of each platform, highlighting the importance of diverse evaluation metrics. Our work offers a valuable tool for researchers to make informed choices about the academic search engines they utilize. Moreover, it paves the way for further research to refine and expand the evaluation framework, ultimately contributing to the advancement of the academic search landscape.**

***Index Terms*—Academic Search Engine, 10 point scale, Mean Reciprocal Rank**

I. Introduction

In recent years, search results for search engines have become increasingly similar, with top players consistently delivering relevant results. However, in today's saturated landscape [11], the need for differentiation extends beyond mere accuracy. To truly stand out, academic search engines must offer a compelling user experience, featuring intuitive interfaces, reduced clutter, and personalization options. This paper proposes a points-based model designed to comprehensively evaluate and rank academic search engines, moving beyond simple relevance to encompass these critical user-centric factors.

Academic search engines, despite facing increasing competition, remain essential for researchers and students. Their unique features, rooted in their historical significance [12] and addressing the enduring demand for specialised scholarly information, ensure their continued relevance. Evolving from early roots, they offer advanced functionalities absent in general search engines, including robust filtering, citation tracker, and etc, proving crucial for academic research. These unique capabilities, combined with their focus on in-depth, specialised functionalities, guarantee their continued relevance and unwavering support for research and learning across academic disciplines. We might have seen many articles, research papers on how one search engine(even academic) compares to the other but they have never formally interpreted it in a points based system. Points based systems have many use cases in research or in sports [4]. It generalises how better a system is compared to the other. In recent years such type of scoring helped improve the overall quality of systems.

While relevance and precision remains king, user loyalty towards an academic search engine transcends mere accuracy, embracing a holistic spectrum of features, user experience, and many such features. To address this dynamic landscape, we propose a 10-point model for comprehensive evaluation and ranking. This model empowers user choices, ultimately igniting a transformative spark within the academic search landscape, propelling it towards a future of improvements in user satisfaction and advancement in scholarly knowledge discovery in many directions.

II. Methodology

By focusing on a diverse range of capabilities, our model goes beyond simple relevance and provides a more holistic evaluation of academic search engines. Let's dive into them.

A. Relevance and Precision

In today's competitive digital ecosystem, the differences in search engine results and website rankings have largely homogenised, significantly narrowing the playing field and making it increasingly difficult for websites to stand out. Our model doesn't directly include a relevance factor. Instead, we calculate the Mean Reciprocal Rank (MRR) of each evaluation set where papers are taken from recent conferences to historic journals. We can call it time bound MRR serves as a proxy for relevance by measuring how the model ranks and retrieves the most relevant document. The calculated MRR is then used to determine a weight between 0 and 1 for each evaluation set, reflecting its importance. Rank of the required document plays an important role in determining the score as all the search results in today's academic search engine show relevant results. To calculate the score of the academic search engine, we first select 4 research articles published in different time periods. This check assesses the search engine's temporal coverage, evaluating how far back it can retrieve relevant articles and ensuring the database's recency.

To assess the system's robustness, we conduct three tests:

- Exact Match: We evaluate the MRR using the complete document title, measuring the system's ability to retrieve the exact document when provided with a precise query.
- Partial Match: We utilise partial search terms to gauge the system's ability to handle incomplete queries and retrieve relevant documents despite missing information.
- Ambiguous Match: We employ ambiguous wording from the document itself to analyse the system's performance with complex or unclear queries, evaluating its ability to interpret diverse user inputs.

The average MRR across these three tests serves as the final score for the relevance category, reflecting the system's overall robustness.

B. Comprehensiveness

The size of the database plays a crucial role in user searches for academic articles. A larger database is more likely to contain the desired article compared to a smaller one. To calculate the comprehensiveness score, we assign the largest database a score of 1 and normalise the scores of other databases on a 0-1 scale. While a larger database offers an advantage, its impact is relatively small when considering the nine other evaluation factors.

C. Multilingualism

In today's globalized academic landscape, multilingualism plays a vital role. Researchers publish in various languages, and accessing these diverse resources can be invaluable. While technology facilitates text translation, incorporating multilingual search capabilities directly within academic search engines provides a more seamless experience. Therefore, we assess multilingualism based on two factors:

- Supported Languages: We evaluate the number of languages supported by each search engine. The engine with the most languages receives a score of 0.5, and other scores are normalized between 0 and 0.5 based on this maximum.
- Language Understanding: We assess each search engine's ability to interpret and retrieve relevant results for queries written in different languages. This score ranges from 0 to 0.5, with 0.5 representing the highest level of understanding and retrieval accuracy across various languages.

Combined Score: The final multilingualism score is calculated by averaging the scores from both factors, providing a comprehensive assessment of each search engine's capabilities in this domain.

D. Advanced Search Capability

While most academic search engines offer some form of advanced search functionality, the quality and comprehensiveness of these features vary significantly. A robust and well-designed advanced search capability plays a critical role in helping researchers efficiently locate relevant articles and journals, particularly within vast databases and diverse topic areas. This efficiency translates to time saved and research productivity enhanced.

Therefore, search engines featuring advanced search capabilities, including the ability to refine searches by specific criteria (e.g., author, publication date, methodology), Boolean operators, and specialised filters (e.g., document type, research area), will receive a score of 1. For search engines lacking such features, the score will be 0.

E. Responsiveness

In today's fast-paced academic environment, user satisfaction hinges on a responsive search experience. With large databases and complex queries, search speed becomes crucial for research efficiency. Users expect rapid retrieval of relevant results to maintain engagement and optimise their workflow. Therefore, responsiveness constitutes a critical factor in our evaluation model [13].

To assess responsiveness, we measure the average time it takes for a search engine to deliver results. Any search exceeding a 4-second response time receives a score of 0, reflecting a significant impact on user experience and research productivity.

F. User-Friendly Interface

A user-friendly interface is paramount for any website, especially academic search engines where users navigate complex information landscapes. The interface should be intuitive, clean, and minimise clutter to maximise user satisfaction and research efficiency. Accessing advanced search tools and filtering options should be straightforward and unobtrusive, while information presentation should be clear and concise to avoid overwhelming users. [14]

In our evaluation, search engines exhibiting an optimal interface with these characteristics will receive a score of 1. However, any significant flaws in design, layout, or information architecture that hinder usability will lead to a 0.

G. Promoting Open Access

Promoting open access to academic research is crucial for democratising knowledge and accelerating scholarly progress. Academic search engines play a vital role in this endeavour by providing researchers with readily accessible information.

Therefore, we consider open access support a vital element in our evaluation model. Search engines that clearly indicate whether an article is freely available and, if not, provide links to free access versions (e.g., institutional repositories, preprint servers) will receive a score of 1. This functionality empowers users to easily access research without encountering paywalls or subscription barriers. Conversely, search engines lacking open access support will receive a score of 0, reflecting their limited contribution to knowledge dissemination.

H. Snippets and Meta Descriptions

In today's information-rich environment, users rely heavily on snippets and meta descriptions to make informed decisions about which research articles to explore further. These concise summaries provide users with a valuable preview of the content, allowing them to assess its relevance and potential

value without delving into the full text. This saves valuable time and effort, especially for users navigating large search results pages.

Therefore, our evaluation model assigns significant weight to the presence and quality of snippets and meta descriptions within academic search engines. Search engines providing informative, accurate, and engaging snippets that summarize key aspects of the article, including relevant keywords and research findings, receive a score of 1. Additionally, the ability to access abstracts or brief summaries directly within the search results further enhances user experience and earns a score of 1.

Conversely, search engines lacking snippets or providing misleading or irrelevant information receive a score of 0. This reflects the negative impact such limitations have on user efficiency and research productivity.

I. Metadata Accuracy

Accurate and reliable metadata is critical for researchers navigating the ever-expanding landscape of scholarly information. Precise details such as authors, publication date, abstract information, and publication venue (e.g., conference, journal, university) allow users to quickly assess the relevance and credibility of research articles. Inaccurate or incomplete metadata can lead to missed opportunities, wasted time, and hampered research progress. As there are many versions of the paper the search engine has to use trusted sources to display the metadata.

Therefore, our evaluation model heavily emphasizes metadata accuracy. Search engines that consistently display accurate and complete metadata across all four key parameters mentioned (year, conference/source, abstract glimpse, and author names) receive a score of 1. This reflects their commitment to providing users with reliable information and facilitating efficient research workflow.

Conversely, search engines exhibiting frequent inaccuracies or inconsistencies in their metadata receive a lower score. This reflects the negative impact such errors have on user experience, research accuracy, and overall trust in the platform. Additionally, search engines must prioritise showcasing trusted and authentic sources whenever possible to ensure the integrity of their results.

J. Additional Features

While the previously discussed factors focus on core search functionalities, additional features can significantly enhance the user experience and research workflow for academics. These features go beyond basic search functionality and provide workspace efficiency and convenience for regular users.

Therefore, we evaluate the presence and effectiveness of the following features:

- Bookmarking: The ability to store and easily access relevant articles for future reference is crucial for researchers. Search engines with robust bookmarking features receive a higher score.
- Author/Institute Information and Work Discovery: Providing easy access to author and institutional information directly from the search results, along with the ability to explore an author's or institute's previous work with a single click, allows researchers to quickly understand the context and credibility of the research and facilitate deeper research and exploration of related areas of study.
- Keyword Alerts: The ability to set up alerts for new articles published with specific keywords keeps researchers informed about the latest developments in their specific research fields.

Search engines offering all four of these features efficiently and effectively receive a score of 1, demonstrating their commitment to enhancing user experience and research productivity. Conversely, the absence of these features, or their implementation in a cumbersome or ineffective manner, results in a lower score.

III. Evaluation and Results

We evaluated three popular academic search engines on December 9th, 2023, based on their popularity and comprehensiveness. We excluded search engines based on another engine or requiring login procedures to ensure a focused analysis. The evaluated search engines were Google Scholar, BASE (Bielefeld Academic Search Engine), and The Lens. We assessed each search engine based on all specific parameters, followed by observations for each.

We will be using four academic papers for the evaluation, focusing on their coverage of both recent and historical publications.

- Paper 1: "Unraveling the 'Anomaly' in Time Series Anomaly Detection: A Self-supervised Tri-domain Solution" (November 19, 2023), will be used to assess whether the search engines retrieve the latest academic papers effectively [6].
- Paper 2: "The Probabilistic Relevance Framework: BM25 and Beyond" (2009), will further evaluate the search engines' ability to retrieve relevant results from a broader timeframe [7].
- Paper 3: "The Anatomy of a Large-Scale Hypertextual Web Search Engine" (1998), will provide insight into how well the search engines handle older academic literature [8].
- Paper 4: "Relevance weighting of search terms" (1976), will test the search engines' ability to retrieve relevant historical documents and assess the accuracy of their metadata definitions [9].

While we could delve further back to 19th-century research articles published by Nature and other esteemed journals, this selection provides a robust starting point for our comprehensive testing.

A. Relevance and precision

1) Test 1: Full wording
   Observations:

TABLE I

	Paper 1	Paper 2	Paper 3	Paper 4	MRR
Google scholar	1	1	1	1	1
Base Search	1	0	1	1	0.75
The Lens	0	0	0.5	0.5	0.12

- Google Scholar: Retrieved all four documents, with the most recent article ranked first.
- BASE: Retrieved the latest article ("Unraveling the 'Anomaly'...") but failed to retrieve "The Probabilistic Relevance Framework: BM25 and Beyond." This is likely due to BASE not being a web crawler and only indexing academic databases it has partnerships with.
- The Lens: Failed to retrieve the two most recent articles. It found the third article ("The Anatomy of a Large-Scale Hypertextual Web Search Engine") in second place, but the first ranked result lacked proper metadata, making the second result the relevant one. The same issue occurred with the fourth article.

2) Test 2: Partial word test
To further evaluate the search engines' effectiveness, we conducted a test where we removed a part of the title of each research paper and assessed whether they still yielded relevant results. This allowed us to analyze their ability to handle incomplete or partially missing information.

TABLE II

	Paper 1	Paper 2	Paper 3	Paper 4	MRR
Google scholar	1	1	1	1	1
Base Search	1	0	1	1	0.75
The Lens	0	1	1	0.33	0.58

Observations:

- Google Scholar: Maintained its consistent performance by retrieving all documents in the first result.
- BASE: Continued its inability to retrieve the second paper but successfully ranked the remaining papers in top positions. This reinforces the hypothesis that BASE focuses on partnered databases and might lack coverage for certain publications.
- The Lens: Interestingly, The Lens improved in the second evaluation by retrieving the second and third papers. It even ranked them in the first position, showcasing a potential learning mechanism. However, its inability to recognize the colon character in the second paper's title raises questions about its search algorithm's robustness.

3) Test 3: Ambiguous wording We have removed most of the terms from the research papers and only included rich words can be identified.
Observations:

TABLE III

	Paper 1	Paper 2	Paper 3	Paper 4	MRR
Google scholar	1	1	1	1	1
Base Search	1	0	1	1	0.75
The Lens	0	0.33	0.5	0.14	0.24

Google Scholar and BASE: Both search engines surprisingly retrieved relevant results despite the significantly reduced title information. This suggests their algorithms are adept at handling incomplete or ambiguous queries and leveraging remaining keywords effectively. BASE's performance particularly highlights its potential to rival Google Scholar if it improves its search capabilities further.
The Lens: While The Lens also retrieved the documents, their ranking significantly dropped compared to the tests with full titles.
Overall: The tests with reduced titles reveal interesting insights into each search engine's strengths and weaknesses. Google Scholar and BASE demonstrate commendable resilience against incomplete information, while The Lens requires improvement in handling ambiguous queries and prioritising relevant results.
As we calculated MRR for each test, we need to average it to get a normalised score.

- Google scholar : 1
- Base search : 0.75
- The Lens: 0.31

B. Comprehensiveness

- Google scholar: 389,000,000 documents. Score: 1 [1].
- Base Search: 278,600,000 documents. Score: 0.71 [2].
- The Lens: 266,403,420 documents Score: 0.68 [3].

Observation:
It might be unfair for other search engines to have a larger database but this is an important factor to consider if you are not particular about the domain of the works but a general student.

C. Multilingualism

- Google scholar : No of languages supported: 40. Score: 0.5. Documents in different languages: 0.5: Score 1.
- Base search: No of languages supported: 8. Score: 0.1. Documents in different languages: 0.5 Score: 0.6.
- The Lens: No of languages supported: 8. Score: 0.15. Documents in different languages: 0.0 Score: 0.15.

Observation:
Google Scholar's multilingual support (32 additional languages) significantly distanced it from the second position. Notably, both Google Scholar and BASE retrieved documents in various languages, while The Lens failed to do so. This highlights the advantage of incorporating multilingual capabilities during website or search engine development, as it simplifies the implementation process compared to retrofitting it later.

D. Responsiveness

- Google scholar: 1
- Base search: 1
- The Lens: 1

Observation:

All evaluated search engines provided responsive performance. It's noteworthy, however, that BASE was the slowest, barely passing the 4-second mark.

E. User-friendly interface

- Google scholar: 1
- Base search: 1
- The Lens: 1

Observation:

All three evaluated websites maintained user-friendly interfaces with minimal clutter. This stands in contrast to the growing trend of requiring user accounts for academic search, a practice detrimental to accessibility. While The Lens presented a more information-dense interface, regular users can likely adapt and navigate it effectively.

F. Open access support

- Google scholar: 1
- Base search: 1
- The Lens: 1

Observation:

All three evaluated search engines commendable showcase open access support for documents, a crucial feature for academic research. Notably, they prioritize displaying free versions of articles over paid ones, demonstrating a commitment to accessibility and scholarly dissemination.

G. Snippets or Meta Description

- Google scholar: 1
- Base search: 1
- The Lens: 1

Observation:

All three evaluated academic search engines provide snippet previews of document abstracts, each with their own unique presentation style. This feature proves highly valuable when browsing through extensive search results, allowing for faster assessment and identification of relevant articles.

H. Metadata accuracy

Google scholar: Score: 1

1) Year: 0.25
2) Conference or source: 0.25
3) Abstract glimpse: 0.25
4) Authors: 0.25

Base search: Score: 1

1) Year: 0.25
2) Conference or source: 0.25
3) Abstract glimpse: 0.25
4) Authors: 0.25

The Lens: Score: 0.75

1) Year: 0.25
2) Conference or source: 0.25
3) Abstract glimpse: 0.25
4) Authors: 0

Observation:

All three academic search engines accurately displayed the publication year of each paper. Google Scholar further enhanced the user experience by automatically generating author profiles showcasing their entire body of work on a single page when clicked on. While BASE offers a similar feature, it requires users to register and claim each individual work, creating an unnecessary hurdle. The Lens also displays author profiles and even institute profiles, but the author profiles were unfortunately not functional during testing.

I. Additional features

1) Ability to save a page
2) Author or institute based search
3) Keyword alert

- Google scholar: Score: 1
- Base search: Score: 1
- The Lens: Score: 1

Observation:

While a dedicated evaluation of additional features is warranted, several key functionalities are expected of an academic search engine. All three evaluated search engines offer basic features such as saving papers, performing author or institute-based searches, and setting keyword alerts. These features are essential for streamlined research workflows and enhance user experience.

J. Final results

- Google scholar: Score: 10
- Base search: Score: 9.06
- The Lens: Score: 7.89

Observation:

Google Scholar achieved a perfect score of 10, highlighting its exceptional relevance and precision in retrieving documents. While further research may explore the impact of prioritizing these factors, Google Scholar's current performance sets a high benchmark. BASE, with a score of 9.06, demonstrates impressive capabilities despite competing against a tech giant. This commendable performance indicates its potential as a strong alternative.

The Lens, however, presents a mixed bag of results. While its user interface and graphical representations are appealing, its core search functionality suffers from inconsistency and unreliability. This lack of precision and accuracy undermines its overall effectiveness as an academic search engine.

Further analysis is required to assess the optimal balance between relevance, precision, and user experience for academic search engines.

IV. Conclusion

This analysis has delved into the evaluation of various academic search engines, highlighting the critical role of robust metrics, including the time-bound mean reciprocal rank (TMRR), in assessing their effectiveness. While the specific performance of individual search engines like Google Scholar, BASE, and The Lens was discussed, the focus here shifts to the broader significance of evaluation metrics in guiding researchers and developers.

The evaluation process revealed the importance of considering various metrics beyond simple relevance and precision. User-friendliness, advanced search features, multilingual support, open access availability, and TMRR, which measures the speed and accuracy of retrieving relevant results, all emerged as crucial factors impacting the overall user experience and research efficiency.

The lack of a standardised and comprehensive evaluation framework remains a challenge in the academic search landscape. This poses difficulties in accurately comparing different platforms.

This paper contributes to bridging this gap by introducing and demonstrating the effectiveness of TMRR as a valuable metric for evaluating academic search engines. Its ability to account for time sensitivity provides a more realistic and nuanced assessment of search performance compared to traditional metrics.

Further research is crucial to refine and establish a comprehensive set of evaluation metrics that effectively capture the multifaceted nature of academic search. This includes metrics that assess relevance, accuracy, user experience, advanced functionalities, accessibility, and, importantly, TMRR to ensure efficient and timely access to critical research resources.

By prioritising the development and implementation of robust evaluation metrics, we can empower researchers to make informed choices about the search engines they utilise. This, in turn, will drive the development of more sophisticated and user-centred academic search tools, ultimately enhancing the research experience for all. Ultimately, the future of academic search lies in the meticulous development and application of evaluation metrics.

References

[1] Gusenbauer, Michael. "Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases." Scientometrics 118.1 (2019): 177-214.

[2] Bielefeld University. "BASE Statistics for 2021." Bielefeld University, 31 Dec. 2021.

[3] The Lens. "Results the Lens - Free & Open Patent and Scholarly Search." The Lens - Free & Open Patent and Scholarly Search, Cambia, 12 Dec. 2023, www.lens.org/lens/search/scholar/structured. Accessed 13 Dec. 2023.

[4] Coelho, Pedro S., and Susana P. Esteves. "The choice between a fivepoint and a ten-point scale in the framework of customer satisfaction measurement." International Journal of Market Research 49.3 (2007): 313-339.

[5] Pieper, Dirk, and Friedrich Summann. "Bielefeld Academic Search Engine (BASE) An end-user oriented institutional repository search service." Library Hi Tech 24.4 (2006): 614-619..

[6] Sun, Yuting, et al. "Unraveling theAnomaly'in Time Series Anomaly Detection: A Self-supervised Tri-domain Solution." arXiv preprint arXiv:2311.11235 (2023).

[7] Robertson, Stephen, and Hugo Zaragoza. "The probabilistic relevance framework: BM25 and beyond." Foundations and Trends® in Information Retrieval 3.4 (2009): 333-389.

[8] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." Computer networks and ISDN systems 30.1-7 (1998): 107-117.

[9] Robertson, Stephen E., and K. Sparck Jones. "Relevance weighting of search terms." Journal of the American Society for Information science 27.3 (1976): 129-146.

[10] Ortega, José Luis. Academic search engines: A quantitative outlook. Elsevier, 2014.

[11] Gusenbauer, Michael, and Neal R. Haddaway. "Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources." Research synthesis methods 11.2 (2020): 181-217.

[12] Walker, Stephen. "OKAPI: Evaluating and enhancing an experimental online catalog." (1987).

[13] Google. "About PageSpeed Insights." Google Developers, Alphabet Inc., 23 May 2023, developers.google.com/speed/docs/insights/v5/about.

[14] Ping Zhang, R. V. Small, G. M. von Dran and S. Barcellos, "Websites that satisfy users: a theoretical framework for Web user interface design and evaluation," Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers, Maui, HI, USA, 1999, pp. 8 pp.-, doi: 10.1109/HICSS.1999.772668.