Comparative study of Different Search Engines

[1]Shabir Ahmad Rather, [2]Farooq Ahmad Pandith, [3]Ashaq Hussain Shah [4]Rafiq Ahmad Rather

[1,2,3] Department of Computer Applications, [4]Librarian
[1, 4] GDC Anantnag, India, [2]GWC Nawakadal,India, [3]GDC Bijbehara, India

Abstract: This study provides a sophisticated comparative analysis of five search engines—Altavista, Google, HotBot, Scirus, and Bioweb—evaluating their efficiency in retrieving scholarly biotechnology information. Utilizing general search engines as methodological controls, the research assesses retrieval performance through the core metrics of micro-averaged absolute precision, relative recall, and database overlap. Furthermore, the investigation incorporates a longitudinal assessment of the persistence and functional stability of bioinformatics web services retrieved by these tools. Our findings indicate that Scirus exhibits superior retrieval metrics for scholarly content due to the integration of formal journal citations, while the niche engine Bioweb provides entirely unique result sets at the cost of significantly lower precision. Crucially, the utility of high-precision retrieval is found to be fundamentally undermined by the longitudinal decay of URL persistence and a systemic "maintenance crisis" within the bioinformatics community, where a substantial portion of services become unreachable or non-functional within three years of publication.

Index Terms **– Search engines, Scholarly, WWW.**

I. INTRODUCTION

The transition of the World Wide Web from an esoteric research network into the de-facto global information repository has resulted in a massive "information explosion" within the life sciences. In the specialized field of biotechnology, this evolution presents a primary challenge for information scientists: the isolation of high-value "scholarly information"—encompassing peer-reviewed research, patents, and technical databases—from the background noise of marketing materials, media content, and general interest data. General search engines are designed for rapid, high-volume indexing but frequently lack the algorithmic filters necessary to exclude duplicate or non-scholarly results. To mitigate these deficiencies, subject-specific information retrieval (IR) tools have emerged. This research examines five engines categorized by their indexing breadth:

- **General Search Engines:** AltaVista, Google, and HotBot.
- **Science-Specific Engine:** Scirus, covering broad science and technology domains.
- **Biotechnology-Specific Engine:** Bioweb, targeting niche biological research data.

II. RESEARCH METHODOLOGY

Test Environment

Data collection was performed using a standardized set of 20 queries derived from the *Library of Congress Subject Headings (LCSH, 2003)* . To evaluate the sensitivity of search algorithms to Boolean logic and phrase-matching syntax, queries were stratified into three structural categories: simple (single-term), compound (phrased), and complex (Boolean-operated). All searches were conducted within a singular 24-hour window to eliminate variations stemming from dynamic index updates.

Precision Evaluation (Four-Point Scale)

The retrieval efficacy was measured by evaluating the first ten results for each query against a rigorous four-point scale:

- **Score 3:** Full-text research papers, seminar/conference proceedings, or patents.
- **Score 2:** Research abstracts (papers or patents).
- **Score 1:** Scholarly books or specialized technical databases.
- **Score 0:** Irrelevant content (marketing/media), duplicate records under variant URLs, or non-responsive/dead servers.

Mathematical Definitions and IR Metrics

The study utilized micro-averaging as the foundational method for calculating mean values across the query set, ensuring that the total performance was accurately represented across all query types.

- **Precision:** Defined as the sum of the scores of scholarly documents retrieved divided by the total number of evaluated results.
- **Relative Recall:** Calculated as the total number of unique scholarly documents retrieved by an engine divided by the aggregate sum of unique scholarly documents retrieved by all five search engines.
- **Overlap Measurement:** Evaluated using the intersection (n) of result sets. For instance, "G n A" represents the set of identical URLs retrieved by both Google and Altavista.

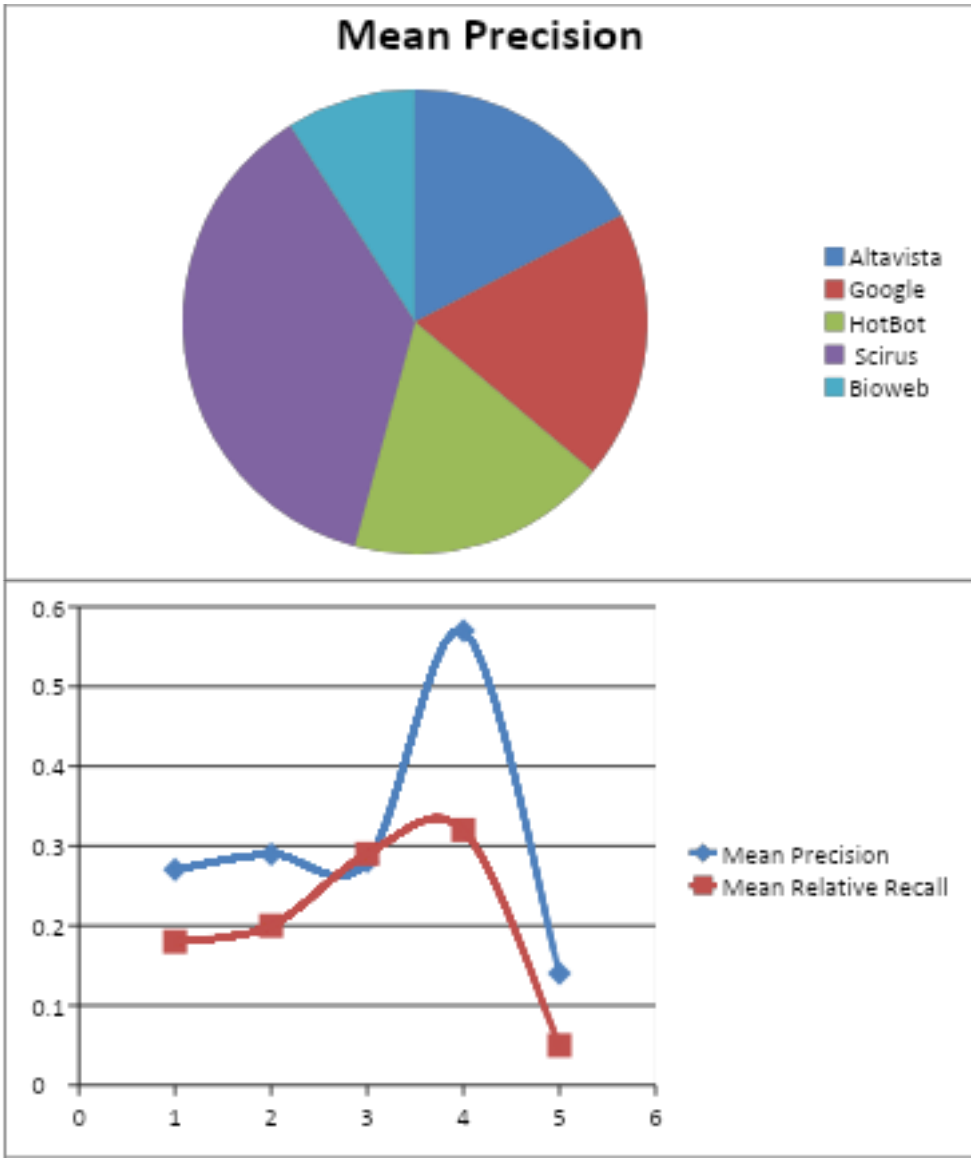
III. COMPARATIVE PERFORMANCE: PRECISION AND RECALL METRICS

The performance metrics highlight significant variance in the ability of these engines to isolate scholarly biotechnology content. Scirus emerged as the most effective tool across all categories.

Table 1: Mean Precision and Relative Recall of Search Engines

Search Engine	Mean Precision	Mean Relative Recall
Altavista	0.27	0.18
Google	0.29	0.20
HotBot	0.28	0.29
Scirus	0.57	0.32
Bioweb	0.14	0.05

The data mentioned in the table is also shown in the pie chart and graph below:



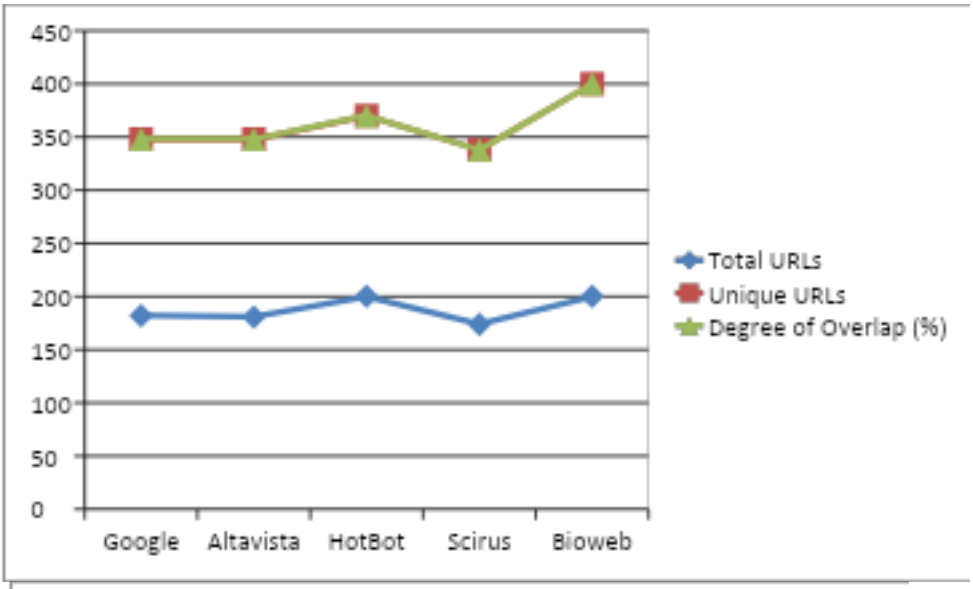
IV. ANALYSIS OF RESULT

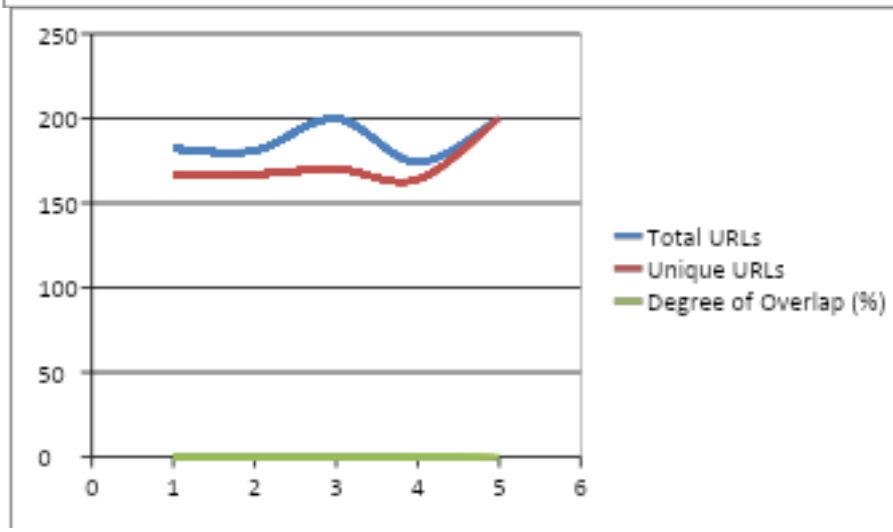
Analyzing the degree of overlap provides insight into the indexing strategies and database breadth of each engine.

Table 2: Degree of Overlap and Result Uniqueness:

Search Engine	Total URLs	Unique URLs	Degree of Overlap (%)
Google	182	166	8.79%
Altavista	181	167	7.74%
HotBot	200	170	15.0%
Scirus	174	164	5.75%
Bioweb	200	200	0.0%

The finding that 92.53% of all URLs across the study were retrieved by only a single engine underscores the fragmentation of biotechnology data. Bioweb's 100% uniqueness confirms its status as a niche, non-crawled repository. Conversely, HotBot's 15% overlap—the studies highest—suggests a reliance on larger, shared databases.





V. PERSISTENCE AND FUNCTIONALITY OF BIOINFORMATICS WEB SERVICES

The longitudinal stability of retrieved information is as critical as the retrieval itself. The utility of high-precision retrieval—exemplified by Scirus—is fundamentally undermined if the underlying bioinformatics web services lack longitudinal stability. A "Score 3" link that is dead three years later effectively reverts to a "Score 0" in practical scientific terms.

State of Web Service Addresses

Longitudinal data indicates that while 72% of services remain available at their originally published addresses, there is a clear decay in URL persistence over time.

- **Transparent Redirects (Light Blue):** 13% of services are reachable via transparent browser redirects or static links provided by the host.
- **Search Engine Discovery (Dark Blue):** 7% of functional services could only be located by actively searching for the service name after the published link failed.
- **Unreachable:** 9% of services are completely unavailable due to server non-responsiveness.

Functionality Verification

Beyond the reachability of the URL, we verified whether the software services were operational:

- **Verified Functional:** 45% (delivered results on example data).
- **Truly Not working:** 13% (site loads, but the underlying application fails).
- **Unable to Test:** 33% (lack of example data or required parameters prevented verification).

The Maintenance Crisis and LT-Score Synthesis

The "Maintenance Crisis" in the field is acute: 78% of services are developed by non-permanent staff (students or post-docs), and 24% have no plan for maintenance post-departure. Our Long-Term Score (LT-Score) analysis demonstrates that persistence is directly correlated with specific variables: services that provide version information, updated contact data, and "fair testing" possibilities receive 1.8 times more citations and exhibit significantly lower longitudinal decay. The absence of these LT-Score variables directly triggers the failure rate of the scientific record.

VI. Conclusion

Scirus is established as the superior tool for scholarly biotechnology retrieval, outperforming all engines through its robust handling of structured queries and integration of journal metadata. While Bioweb provides a unique data niche, its retrieval precision is insufficient for exhaustive research. To preserve the integrity of the scientific record, the community must address the instability of web-based services. We recommend the mandatory adoption of persistent identifiers (DOIs) for web services and the deposition of source code in public repositories to ensure long-term reproducibility and mitigate the risks inherent in the current maintenance crisis.

VII. References

1. Bar-Ilan, J. (1998). On the overlap, the precision and estimated recall of search engines: A case study of the query "Erdos". *Scientometrics* , *42* (2), 207-208.
2. Benson, G. (2010). Submitting to the web server issue. *Nucleic Acids Research* , *38* , W1.
3. Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., et al. (2010). BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Research* , *38* , W689–W694.
4. Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems* , *30* , 379-388.
5. Brazas, M. D., Yim, D. S., Yamada, J. T., & Ouellette, B. F. (2011). The 2011 bioinformatics links directory update. *Nucleic Acids Research* , *39* , 3–7.
6. Caporale, L. H. (1999). Chance favors the prepared genome. *Annals of the New York Academy of Sciences* , *870* , 1-21.
7. Chignell, M. H., Gwizdka, J., & Bodner, R. C. (1999). Discriminating meta-search: A framework for evaluation. *Information Processing and Management*, *35* , 337-362.
8. Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: a comparative study. *Proceedings of the ASIS Annual Conference* , *33* , 127-135.
9. Clarke, S., & Willett, P. (1997). Estimating the recall performance of search engines. *Aslib Proceedings* , *49* (7), 184-189.
10. Ding, W., & Marchionini, G. (1996). A comparative study of the Web search service performance. *Proceedings of the ASIS Annual Conference* , *33* , 136-142.
11. Ding, W., & Marchionini, G. (1998). A comparative study of Web search service performance. *Proceedings of the ASIS Annual Conference* , *35* , 136-142.
12. Ferrara, J., da Silva, A. R., & Delgado, J. (2004). Does overlap mean relevance? *Proceedings of WWW/Internet 2004 (LADIS)* .
13. Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: comprehensive approach for accessible research. *Genome Biology* , *11* , r86.
14. Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web. *Information Processing and Management* , *35* , 141-180.
15. Hall, G. W., & Stokes, H. W. (1993). Integrons: Novel DNA elements which capture genes by site-specific recombination. *Genetica* , *90* , 219-232.
16. Hood, W. W., & Wilson, C. S. (2001). Overlap in bibliographic databases. *Journal of the American Society for Information Science and Technology* , *54* (12), 1091-1103.
17. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., et al. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Research* , *34* , 729–732.
18. Leighton, H. (1996). Performance of four WWW index services. *Winona State University Report* .
19. Leighton, H., & Srivastava, J. (1997). Precision among WWW search services. *Winona State University Report* .
20. Levesque, C., Piche, L., Larson, C., & Roy, P. H. (1994). Characterization of the class 1 integron gene code. *Gene* , *142* (1), 49-54.
21. Library of Congress. (2003). *Library of Congress Subject Headings (Volumes 1-5)* . Washington, DC: Library of Congress.
22. McGinnis, S., & Madden, T. L. (2004). Blast: core of sequence analysis tools. *Nucleic Acids Research* , *32* , 20–25.
23. McWilliam, H., Valentin, F., Goujon, M., Li, W., Narayanasamy, M., et al. (2009). Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Research* , *37* , 6–10.
24. Modi, G. (1996). Searching the Web for gigabucks. *New Scientist* , *150* (2024), 36-40.
25. Naas, T., Mikami, Y., Ibe, T., & Nordmann, P. (2001). Characterization of a multiresistance integron. *Antimicrobial Agents and Chemotherapy* , *45* , 1115-1121.
26. Nicholson, S. (2000). Raising reliability of Web search tool research through replication. *Journal of the American Society for Information Science* , *51* (8), 724-729.
27. Ochman, H., Lawrence, J. G., & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* , *405* , 299-304.
28. Oppenheim, C., Morris, A., McKnight, C., & Lowley, S. (2000). The evaluation of WWW search engines. *Journal of Documentation* , *56* (2), 190-211.
29. Recchia, G. D., & Hall, R. M. (1995). Gene cassettes: a new class of mobile element. *Microbiology* , *141* , 3015-3027.
30. Roberts, R. J. (2003). Editorial. *Nucleic Acids Research* , *31* , 3289.
31. Rowe-Magnus, D. A., & Mazel, D. (1999). Resistance gene capture. *Current Opinion in Microbiology* , *2* (5), 483-488.

32. Rowe-Magnus, D. A., & Mazel, D. (2001). The role of integrons in antibiotic resistance gene dissemination. *Research in Microbiology* , *152* , 737-746.
33. Rowe-Magnus, D. A., Guerout, A. M., Biskri, L., Bouige, P., & Mazel, D. (2002a). Identification of a second family of integron gene cassettes in Vibrio cholerae. *Molecular Microbiology* , *43* , 357-369.
34. Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., et al. (2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* , *38* , 5–16.
35. Schlichting, C., & Nilsen, E. (1996). Signal detection analysis of WWW search engines. *Microsoft Usability Web Conference* .
36. Schultheiss, S. J., Münch, M. C., Andreeva, G. D., & Rätsch, G. (2011). Persistence and availability of web services in computational biology. *PLoS ONE* , *6* (9), e24914.
37. Scoville, R. (1996). Find it on the Net. *PC World* , *14* (1), 125-130.
38. Shafi, S. M., & Rather, R. A. (2005). Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. *Webology* , *2* (2), Article 12.
39. Spink, A., Jansen, B. J., Kathuria, V., & Koshman, S. (2006). Overlap among major web search engines. *Internet Research* , *16* (4), 419-426.
40. Stokes, H. W., O'Gorman, D. B., Recchia, G. D., Parsekhian, M., & Hall, R. M. (1997). Structure and function of 59-base element recombination sites associated with mobile gene cassettes. *Molecular Microbiology* , *26* (4), 731-745.
41. Sundstrom, L. (1998). The promoters of integrons. *Frontiers in Bioscience* , *3* , 485-495.
42. Tague, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management* , *28* (4), 467-490.
43. Thornton, J. (2009). Data curation in biology – past, present and future. *Nature Precedings* , 3225.
44. Thornton, J., & BioSapiens Network. (2009). Annotations for all by all - the biosapiens network. *Genome Biology* , *10* , 401.
45. Veretnik, S., Fink, J. L., & Bourne, P. E. (2008). Computational biology resources lack persistence and usability. *PLoS Computational Biology* , *4* , e1000136.