On Low Overlap among Search Results of Academic Search Engines

Anasua Mitra
Indian Institute of Technology
Guwahati, Assam 781039, India
anasua.mitra@iitg.ernet.in

Amit Awekar
Indian Institute of Technology
Guwahati, Assam 781039, India
awekar@iitg.ernet.in

ABSTRACT

Number of published scholarly articles is growing exponentially. To tackle this information overload, researchers are increasingly depending on niche academic search engines. Recent works have shown that two major general web search engines: Google and Bing, have high level of agreement in their top search results. In contrast, we show that various academic search engines have low degree of agreement among themselves. We performed experiments using 2500 queries over four academic search engines. We observe that overlap in search result sets of any combination of academic search engines is significantly low and in most of the cases the search result sets are mutually exclusive. We also discuss implications of this low overlap.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

Keywords

Web mining; Scholarly data; Search engine comparison

1. INTRODUCTION

Number of published research papers approximately doubles after nine years[3]. This robust trend which is consistently observed after the second world war can be attributed to many factors. Sheer number of researchers is increasing with improvements in academic and research infrastructure in countries such as China and India. Researchers increasingly face publish-or-perish paradigm in research universities and laboratories. Web based systems for paper submission, reviewing and publication have significantly brought down the timespan and cost of publishing a paper. As a result, large number of low quality and even predatory research conferences and journals have emerged.

It is impossible to manually keep track of all relevant literature for any research topic. For example, in 2016 more than 2000 papers were published with words "deep learning" mentioned in the title[1]. Actual number of papers published in 2016 on deep learning will be far more than that. In response to this information overload, many academic search engines (ASEs) have been developed to find appropriate research articles. These ASEs differ in multiple aspects: broad vs. specific topic coverage, web crawl vs. curated databases as source of data, commercial vs. non profit. ASEs play significant role in deciding which research papers are read by researchers. Therefore it is necessary to study ASEs separately from general search engines.

2. RELATED WORK

Overlap in the coverage and search results of web search engines is well studied. During mid 90s, multiple search engines such as AltaVista, Excite, and Lycos were competing with each other. These search engines covered only about 3 to 4% of the web[5] and overlap in their search results was as low as 1.4%[2]. By the year 2005, major search engines were indexing more than 60% of the web, but overlap in their results was still only about 11%. Currently, there are only two major web search engines: Google and Bing. Recent work by Agrwal et. al. has shown that both these web search engines have high level of agreement between them[1]. This convergence can be attributed to multiple factors such as high coverage of both search engines, maturity in ranking algorithms, and possibility of copying results from each other. Bibliographic datasets have been shown to have low overlap in their coverage[4]. Various works have tried to predict coverage of ASEs[6]. However to the best of our knowledge, there is no existing work that systematically studies overlap in the search results of ASEs.

3. OUR WORK

Motivation for systematically studying agreement among multiple ASEs primarily came from our own experience of using ASEs for research literature review. Almost mutually exclusive results obtained from multiple ASEs, prompted us to verify whether it was just a coincidence. This observation was in sharp contrast with recent work by Agrawal et. al.[1] where they observed strong agreement among web search engines.

We queried four ASEs: Google Scholar (GS), Semantic Scholar (SS), Microsoft Academic (MA), and Scopus (SC). Main reason for choosing these ASEs was that they are popular in computer science and engineering domain. We
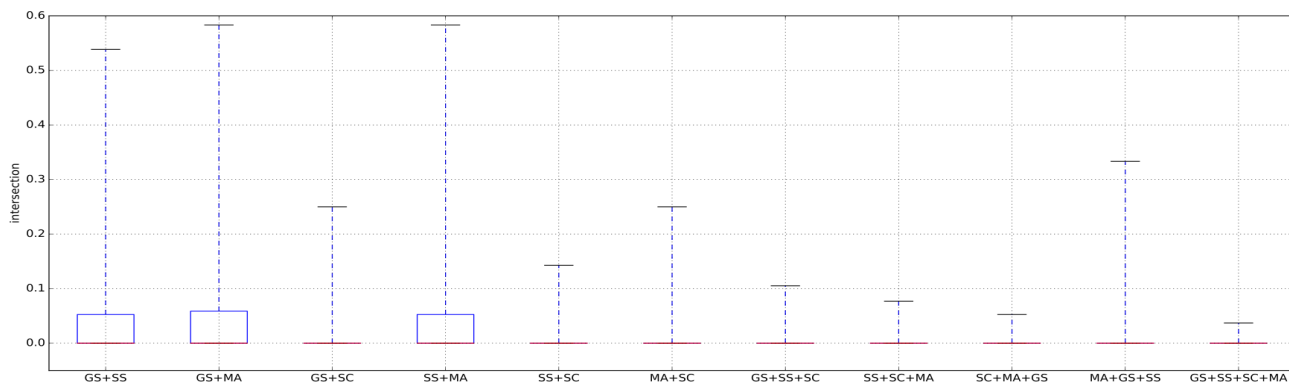
---

[1]Source: Google Scholar

Figure 1: Overlap in Search Result Sets of Academic Search Engines

collected approximately 2300 query terms from 2012 ACM Computing Classification System[2]. This system arranges various computer science topics including technology, product, organization names, eminent researchers in computing into a poly-hierarchy ontology. It sufficiently covers broad spectrum of topics in computer science from coarse to fine granularity. We also manually collected over 200 keywords from papers published in ACM SIGKDD 2016 conference.

We sent these 2500 queries to all four selected ASEs. We looked at maximum top eight results as each ASE returns at least eight results on the first page. We ignored the order in results and treated them as sets. If we consider order of results, then it will further drive down the similarity scores. We computed similarity using the Jaccard similarity, that is ratio of size of intersection and union.

Please refer to Figure 1. X axis represents all possible combination of chosen ASEs. There are total eleven such combinations as we considered four ASEs. Y axis represents Jaccard similarity for each combination using boxplots. We obtained 2500 similarity scores for each ASE combination. Out of these scores, the figure depicts maximum (top black line), minimum (bottom black line), median (red line), 25 percentile, and 75 percentile (blue box) scores for each combination of ASEs.

For all combinations, minimum score is always zero. It means that we have at least one query per combination such that their intersection set is empty. For all combinations, median score is also zero indicating that for most of the queries search result sets of ASEs are mutually exclusive. For each query, very few research articles appear in the top results list of all four ASEs. This shows strong disagreement among ASEs. Our queries covered topics of coarse as well as fine granularity and some of the popular topics from recent top conference site. However, we did not see any correlation between topic granularity and overlap in search results.

GS and MA have comparable coverage of research literature (160 million and 150 million documents respectively). These two ASEs cover multiple topics including computer science. While SS covers only about 10 million documents, mostly from computer science. Compared to these three ASEs, Scopus has far small share of computer science documents. All our queries are from computer science domain. Therefore blue box comprising of 25 percentile and 75 per-

centile similarity scores is wide for combinations consisting of GS, MA, and SS. Where as combinations involving SC, the blue box almost flattens to line overlapping with minimum value of zero.

4. CONCLUSION AND FUTURE WORK

Unlike web search engines, ASEs do not have one or two dominant players. This has led to multiple systems that differ in coverage of research literature and ranking algorithms. Therefore overlap among search results of ASEs is significantly low. As a result, users of ASEs have to look across multiple ASEs to find relevant research literature. We are working on extending this study in three ways. First, we are including more ASEs in the comparison. Second, we are using more diverse queries related to other subjects apart from just computer science. Third, we want to compare ASEs based on quality of search results.

5. REFERENCES

[1] R. Agrawal, B. Golshan, and E. Papalexakis. A study of distinctiveness in web results of two search engines. In *International Conference on World Wide Web*, pages 267–273, 2015.

[2] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems*, 30(1):379–388, 1998.

[3] L. Bornmann and R. Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015.

[4] W. W. Hood and C. S. Wilson. Overlap in bibliographic databases. *Journal of the American Society for Information Science and Technology*, 54(12):1091–1103, 2003.

[5] S. Lawrence and C. L. Giles. Searching the world wide web. *Science*, 280(5360):98–100, 1998.

[6] E. Orduna-Malea, J. M. Ayllón, A. Martín-Martín, and E. D. López-Cózar. Methods for estimating the size of google scholar. *Scientometrics*, 104(3):931–949, 2015.

---

[2] https://www.acm.org/publications/class-2012