On Low Overlap among Search Results of Academic Search Engines

An attempt to quantify amount of disagreements among ASEs

Anasua Mitra & Amit Awekar
Indian Institute of Technology Guwahati, India
anasua.mitra@iitg.ernet.in, 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.

Motivation

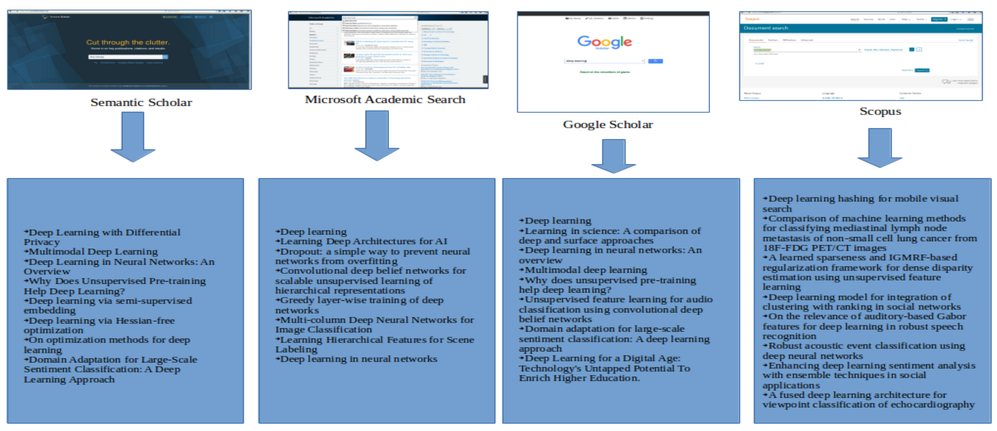

Figure 1: The result-set on querying Google Scholar, Microsoft Academic Search, Semantic Scholar, Scopus for the keyword "Deep Learning".

Framework : Input

1. We collected approximately 2300 query terms from 2012 ACM Computing Classification System. 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.
2. 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.

Search Engines	Estimated Size (in millions)	Year	Broad Topics
Google Scholar	160	2004	Multidisciplinary
Microsoft Academic Search	150	2011	Multidisciplinary
Semantic Scholar	10	2015	Computer Science, Neuro-Science
Scopus	40	2004	Multidisciplinary

Table 1: Comparison of Academic Search Engines

Framework : Architecture

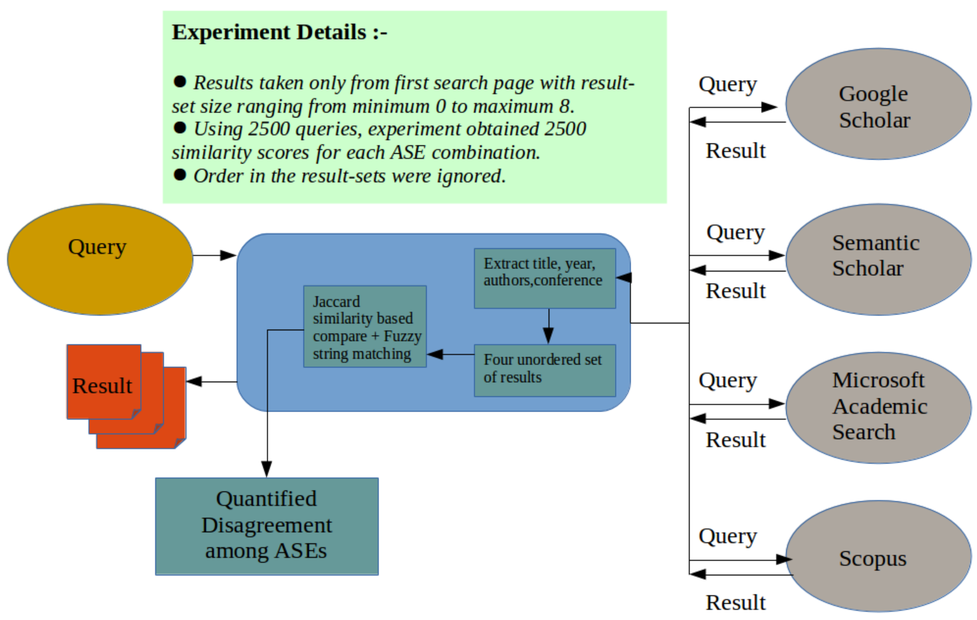

Figure 2: Framework for experiment

Results

Please refer to Figures, X axis represents all possible combination of chosen ASEs. Y axis represents Jaccard similarity for each combination using boxplots. There are total eleven such combinations as we considered four ASEs. 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 (except for first two figures) 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.

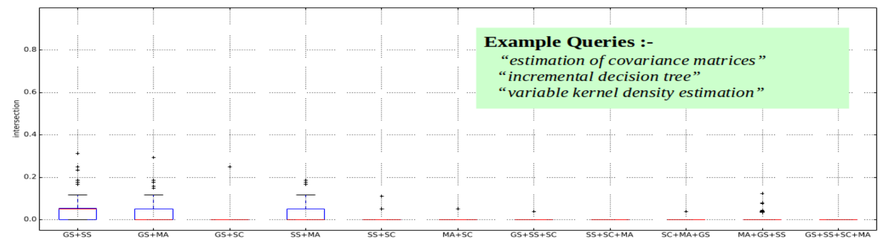

Figure 3: Degree of agreement among ASEs on keywords collected from papers published in SIGKDD 2016

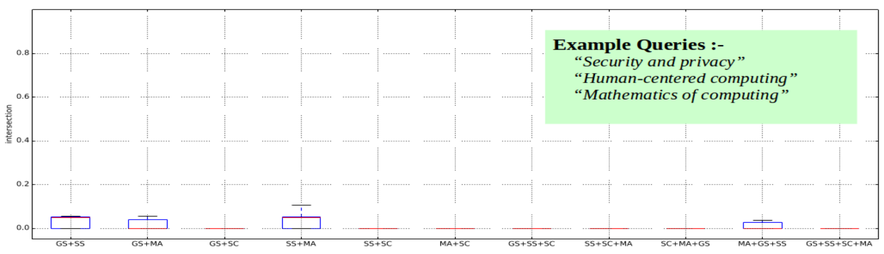

Figure 4: Degree of agreement among ASEs on top level keywords of ACM Computing Classification System

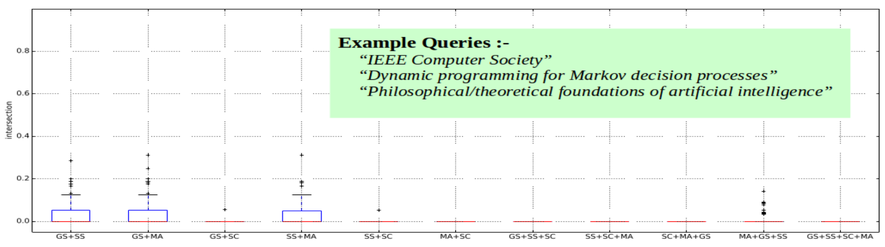

Figure 5: Degree of agreement among ASEs on second level keywords of ACM Computing Classification System

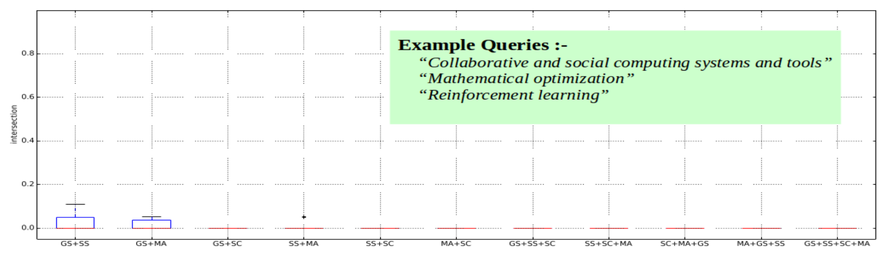

Figure 6: Degree of agreement among ASEs on third level keywords of ACM Computing Classification System

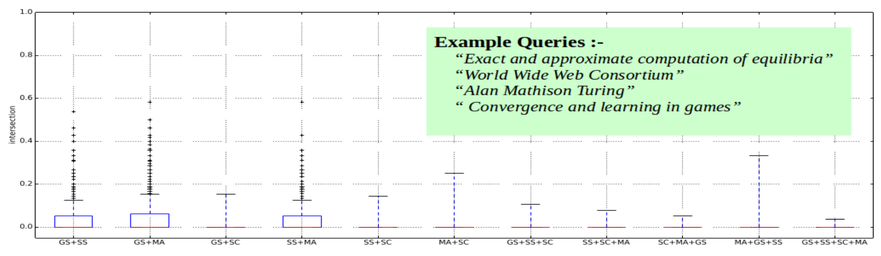

Figure 7: Degree of agreement among ASEs on leaf level keywords of ACM Computing Classification System

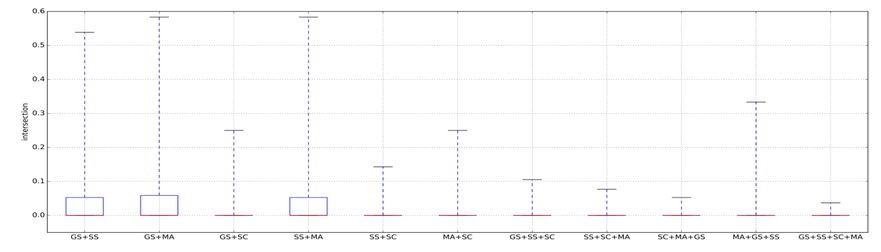

Figure 8: Degree of agreement among ASEs on all keywords of ACM Computing Classification System

Conclusions

- Overlap among search results of ASEs is significantly low for queries pertaining to computer science. So users of ASEs have to look across multiple ASEs to find relevant research literature.

Forthcoming Research

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.

References

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