7-2021

Do Search Engines Display All Search Hits?

Mufazil Ali
University of Kashmir, India, mufazilbhat@yahoo.com

Fayaz Ahmad Loan
University of Kashmir, Srinagar, fayazlib@yahoo.co.in

Do Search Engines Display All Search Hits?

Mufazil Ali, Ph.D. Scholar, Centre of Central Asian Studies, University of Kashmir, Srinagar, Jammu and Kashmir, (India). E-mail: mufazilbhat@yahoo.com

Fayaz Ahmad Loan, Centre of Central Asian Studies, University of Kashmir, Srinagar, Jammu and Kashmir, (India). E-mail: fayazlib@yahoo.co.in

ABSTRACT

Purpose: The web is used to gather information through Search tools globally. These search tools display results as HITS (Hyper-Text Induced Theme Search). This study aims to explore how accurate search tools are when search hits are counted and displayed.

Methodology: The paper began with search tools and search terms recognition. Academic search tools Google Scholar, BASE (Bielefeld Academic Search Engine**),** CORE (Connecting Repositories) were identified. Using the Dewey Decimal Classification (DDC), the subject areas were selected from the fields of Economics and Political Science. Search terms were selected from the Sears List of Subject Headings (SLSH). The searches were conducted in the simple search mode of the search engines. The total number of hits shown by search tools was recorded, and the hits were then manually counted while navigating from one page to the next to identify the true number of search hits.

Findings: The findings reveal that there is a large difference in the number of hits claimed by the BASE**,** CORE, and Google Scholar and actual hits displayed. However, the actual hits don't vary significantly between and among search engines.

Research Limitations: The study is limited to only three academic search tools.

Originality/value: Few studies have been conducted to estimate the search results of search engines. However, no study has covered these three academic search engines so far.

Keywords: Search Engines, Google Scholar, CORE, BASE, Search Results, Search Hits, Count Estimation

Paper Type: Research Paper

INTRODUCTION

The tremendous development of the Web has converted the Internet into a vast knowledge resource with a range of content, often improperly arranged. The world wide web is expanding, with new websites being launched every second. The amount of data we produce every day is truly mind-boggling. In 2018, it was estimated that there are 2.5 quintillion bytes of data created each day and over the last two years alone 90 percent of the data in the world was generated **(Marr, 2018)**. Specific techniques and tools are being used to search the information available on the web **(Hölscher and Strube, 2000)**. The search engines are now the most important resources for finding information on the Internet **(Bar-Ilan, Levene, and Mat-Hassan, 2006)**. Web search engines are used to expand their access to electronic content. Web search systems like Google, Bing, and Yahoo are now tools to discover things that users use every day **(Bernard and Amanda, 2004)**. On average, Google now processes more than 40,000 searches every second and 3.5 billion searches per day. While 77% of searches are conducted on Google, it would be remiss not to remember other search engines are also contributing to our daily data generation. Worldwide, 5 billion searches are conducted on search engines daily **(Marr, 2018)**. Even though users use the same search phrase to search information, each search tool provides different outcomes. More than one search tool and search phrases have also been used where users cannot locate useful information according to information requirements **(Kaushik, 2012)**. The various search engines have varying characteristics, efficiencies, and databases. Therefore, to identify the most appropriate search platform for information seekers, the popular characteristics of different search engines must be evaluated. Count estimates provided by the Web search engine are considered as a yardstick to measure the size of the database of a search engine. The present study is performed to evaluate the accuracy of search tools in terms of count estimation.

LITERATURE REVIEW

Web Search Engine' is a database-composed recovery tool and that all kinds of search engines are evaluated using their basic functionality (database information, recovery tools, and search interface) **(Poulter, 1997)**. **Xie (2004)** highlighted three essential mechanisms for online information retrieval system validation viz the architecture of interfaces, system performance, and coverage of materials. According to **Dudek, Mastora & Landoni (2007),** "Search engines

are directly interlinked with their usability and performance". Also, users prefer search engines because of the ease and easiness of the search engine. The overall efficiency of the information retrieval system is based on different aspects of the system and the actions of its users in the quest of information, for instance, system speed, user interface, query language, and engine functionality **(Carterette, Kanoulas, & Yilmaz, 2012)**. **Loan and Rufai (2014)** tested the accuracy of Google search engines in counting search hits, and it was discovered that the number of results shown and the real number of results varies greatly. **Malla & Loan, (2018)** conducted a study on five search tools. The actual number of Hits registered, the number of HITS claimed by Google, Yahoo, and Google Scholar varies greatly. However, the number of HITS shown by the CORE and SSRN is identical to the actual number of Hits counted. The present study will also analyze the accuracy of search tools in terms of count estimation. However, this study is a bit different because it will analyze the three major academic search tools- Google Scholar, CORE, and Base.

RESEARCH DEIGN

a) Purpose of the study

Throughout the world, search engines are used to find content. These tools provide results as HITS (Hyper-Text Induced Topic Search) for any subject area. This paper aims to assess the performance with which search tools count search results.

b) Methodology

The research began with search tools and search terms recognition. Academic search tools Google Scholar, BASE (Bielefeld Academic Search Engine**),** CORE (Connecting Repositories) were identified. Using the Dewey Decimal Classification (DDC), the subject areas were selected from the fields of Economics and Political Science. Search terms were selected from the Sears List of Subject Headings (SLSH). Besides, to retrieve area-specific information, Central Asia was added to each term. The searches were conducted in the simple search mode of the search engines. The total number of hits shown by search tools was recorded, and the hits were then manually counted when navigating from one page to the next to determine the true number of search hits.

c) Hypotheses

The following hypotheses were formulated:

I. H0: (Null Hypothesis): "Actual HITS of search engines in Political Science does not differ significantly"

II. H0: (Null Hypothesis): "Actual HITS of search engines in Economics does not differ significantly"

DATA ANALYSIS

a) Count Estimation of Google Scholar

Google Scholar results show that the hits claimed and the actual number of hits displayed mismatch. Google Scholar claims thousands to millions of hits for any combination of keywords; however, the findings reveal that Google scholar retrieves or displayed 1000 hits at the most or less, as few terms "*Politics in Central Asia*" retrieves 989 results and "*Political Rights in Central Asia*" displays 990 hits only. The variation between hits shown and actual hits displayed is indeed quite massive (Table 1).

Table 1: Google Scholar Hits in Political Science

No.	Search Term	Listed	Displayed	Difference
1.	Politics **AND** Central Asia	2,270,000	989	2,269,011
2.	Governments **AND** Central Asia	2,960,000	1000	2,959,000
3.	International relations **AND** Central Asia	3,540,000	1000	3,539,000
4.	Political rights **AND** Central Asia	2,340,000	990	2,339,010
5.	Elections **AND** Central Asia	524,000	1000	523,000
6.	Colonization **AND** Central Asia	268,000	1000	267,000
7.	Legislation **AND** Central Asia	717,000	1000	716,000
8.	Political parties **AND** Central Asia	2,170,000	1000	2,169,000
9.	Slavery **AND** Central Asia	293,000	1000	2,922,000
10.	Geopolitics **AND** Central Asia	854,000	1000	853,000

The findings show a large variation in the outcomes found by Google Scholar in all key terms in Economics as well. Since Google Scholar retrieves thousands to millions of hits for every

keyword, but Google Scholar doesn't provide access to more than 1000 results. There's an enormous difference between the result shown and actual hits (Table 2).

No.	Table 2: Google Scholar Hits in Economics			
	Search Term	**Claimed**	**Displayed**	**Difference**
1.	Economic theories **AND** Central Asia	1,220,000	1000	1,219,000
2.	Labor economics **AND** Central Asia	1,790,000	1000	1,789,000
3.	Financial economics **AND** Central Asia	2,830,000	1000	2,829,000
4.	Land economics **AND** Central Asia	1,830,000	1000	1,829,000
5.	Cooperatives **AND** Central Asia	118,000	1000	117,000
6.	Marxism **AND** Central Asia	137,000	1000	136,000
7.	Public finance **AND** Central Asia	1,830,000	1000	1,829,000
8.	International economics **AND** Central Asia	2,560,000	1000	2,559,000
9.	Agriculture production **AND** Central Asia	2,520,000	1000	2,519,000
10.	Macroeconomics **AND** Central Asia	87,100	1000	86,100

b) Count Estimation of BASE

The BASE search engine evaluation demonstrates that results are obtained in thousands in most cases, but results do not reach above one thousand in any case. For eight search terms, 1000 results are displayed whereas the term "*Geopolitics in Central Asia"* fetches 744 hits and "*Slavery in Central Asia*" 165 hits only. However, the distinction between hits claimed and hits displayed is indeed very less in most of the cases and there is no difference in search hits retrieving documents below one thousand (Table 3).

No.	Table 3: BASE Search Hits in Political Science			
	Search Term	**Claimed**	**Displayed**	**Difference**
1.	Politics **AND** Central Asia	11,457	1000	10,457
2.	Governments **AND** Central Asia	12,548	1000	11,548
3.	International relations **AND** Central Asia	8,895	1000	7,895
4.	Political rights **AND** Central Asia	2,340	1000	1,340
5.	Elections **AND** Central Asia	1,240	1000	240
6.	Colonization **AND** Central Asia	1,597	1000	597
7.	Legislation **AND** Central Asia	1,880	1000	880

8.	Political parties **AND** Central Asia	1,448	1000	448
9.	Slavery **AND** Central Asia	165	165	0
10.	Geopolitics **AND** Central Asia	744	744	0

The BASE search engine gives output results in thousands for most of the search terms but the terms "*Marxism in Central Asia*" displayed just 69 results and "*Macroeconomics in Central Asia*" 691 results only. Matter of fact, the gap between reported hits and visible hits was not too high (Table 4).

Table 4: BASE Hits in Economics

No.	Search Term	Claimed	Displayed	Difference
1.	Economic theories **AND** Central Asia	2,086	1000	1,086
2.	Labor economics **AND** Central Asia	2,917	1000	917
3.	Financial economics **AND** Central Asia	5,181	1000	4,181
4.	Land economics **AND** Central Asia	3,147	1000	2,147
5.	Cooperatives **AND** Central Asia	6,506	1000	5,506
6.	Marxism **AND** Central Asia	69	69	0
7.	Public finance **AND** Central Asia	4,199	1000	3,199
8.	International economics **AND** Central Asia	9,492	1000	8,492
9.	Agriculture production **AND** Central Asia	3,731	1000	2,731
10.	Macroeconomics **AND** Central Asia	691	691	0

c) Count Estimation of CORE

CORE listed millions of hits for each search term of Political Science, but the total number of hits displayed for each term is 1000 only. There is a sea of difference between the claimed and actual hits (Table 5).

Table 5: CORE Hits in Political Science

No.	Search Term	Claimed	Displayed	Difference
1.	Politics **AND** Central Asia	49,760,965	1000	49,759,965
2.	Governments **AND** Central Asia	50,547,452	1000	50,546,452
3.	International relations **AND** Central Asia	51,618,725	1000	51,617,725
4.	Political rights **AND** Central Asia	50,132,607	1000	50,131,607

5.	Elections **AND** Central Asia	49,935,448	1000	49,934,448
6.	Colonization **AND** Central Asia	49,937,296	1000	49,936,296
7.	Legislation **AND** Central Asia	49,956,365	1000	49,955,365
8.	Political parties **AND** Central Asia	50,072,341	1000	50,071,341
9.	Slavery **AND** Central Asia	56,595,251	1000	56,594,251
10.	Geopolitics **AND** Central Asia	49,933,059	1000	49,932,059

CORE listed millions of hits for each search term in Economics as well, but the total number of hits displayed doesn't exceed more than 1000 thresholds. Indeed, the difference between seen hits and real hits is too huge (Table 6).

Table 6: CORE Hits in Economics

No.	Search Term	Claimed	Displayed	Difference
1.	Economic theories **AND** Central Asia	50,170,625	1000	50,169,625
2.	Labor economics **AND** Central Asia	50,094,480	1000	50,093,480
3.	Financial economics **AND** Central Asia	50,092,405	1000	50,091,405
4.	Land economics **AND** Central Asia	50,138,985	1000	50,137,985
5.	Cooperatives **AND** Central Asia	49,979,403	1000	49,978,403
6.	Marxism **AND** Central Asia	49,978,134	1000	49,977,134
7.	Public finance **AND** Central Asia	50,559,726	1000	50,558,726
8.	International economics **AND** Central Asia	51,741,781	1000	51,740,781
9.	Agriculture production **AND** Central Asia	50,493,573	1000	50,492,573
10.	Macroeconomics **AND** Central Asia	50,255,180	1000	50,254,180

d) Testing and Verification of Hypothesis

To verify the null hypothesis that, "The actual hits of all three search engines do not differ significantly" the one-way ANOVA, is being carried out. This test is used to compare means and variance between search tools. In the present case, we apply the test on the actual hits (hits displayed) for random samples of three search tools to verify "Is there sufficient evidence at the 0.05 level of significance to conclude that there is a difference in the mean of actual HITS among search engines?"

The table displays the results of the one-way ANOVA analysis.

The BASE search engine had a Mean of 890.9 and SD=267.4; the CORE had a Mean of 1000 and SD = 0; Google Scholar had a Mean of 997.9 and SD 4.43. The analysis of variance showed that the difference between the mean precision of BASE, CORE, and Google Scholar was not statistically significant, F (2, 27) = 1.6, p =0.21). The ANOVA results revealed that there are no differences among search engines in terms of actual HITS (or hits displayed) in the subject of Political Science. The test confirms that the differences among the search engines are not statistically significant and hence, we accept the null hypothesis.

SUMMARY					
Groups	*Count*	*Sum*	*Average*	*Variance*	*SD*
BASE	10	8909	890.9	71525.88	267.4
CORE	10	10000	1000	0	0
Google Scholar	10	9979	997.9	19.65556	4.43

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	77854.07	2	38927.03	1.632263	0.214192	3.354131
Within Groups	643909.8	27	23848.51			
Total	721763.9	29				

The BASE search engine had a Mean of 876, SD=299.7; the CORE had a Mean of 1000 and SD = 0; Google Scholar had a Mean of 1000 and SD 0. The analysis of variance showed that the difference between the Mean precision of BASE, CORE, and Google Scholar was not statistically significant, F (2, 27) = 1.7, p =0.19). The ANOVA results revealed that there are no differences among search engines in terms of actual HITS in the subject of Economics. The test confirms that the differences among the search engines are not statistically significant. Since the null hypothesis is accepted.

SUMMARY					
Groups	*Count*	*Sum*	*Average*	*Variance*	*SD*
BASE	10	8760	876	89831.33	299.7188
CORE	10	10000	1000	0	0
Google Scholar	10	10000	1000	0	0

ANOVA						
Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	102506.7	2	51253.33	1.711652	0.199584	3.35413083
Within Groups	808482	27	29943.78			
Total	910988.7	29				

CONCLUSION

Google Scholar results show that the hits claimed and the actual number of hits displayed mismatch in all search terms in both disciplines - *Political Science* and *Economics*. The variation between hits claimed and actual hits displayed is indeed quite massive. Google Scholar claims thousands to millions of hits for any combination of keywords; however, the findings reveal that Google scholar retrieves or displays 1000 hits at the most or less in few cases. In the case of the BASE search engine, the distinction between hits claimed and hits displayed is indeed very less in most of the cases and there is no difference in search hits retrieving documents below one thousand. The matter of fact is that the gap between reported hits and visible hits was either accurate or not too high in the case of the BASE in both disciplines. CORE also plays with the number game like Google Scholar. CORE listed millions of hits for each search term in both subjects but the total number of hits displayed doesn't exceed more than 1000 thresholds.

The common among all the three search engines is that the maximum number of hits displayed doesn't cross the 1000 mark in any case. We can assume that it is a policy of the search engines to display only up to 1000 hits, but in few cases, Google Scholar claimed to have hits in millions but failed to display even 1000 hits (see Table 1). It gives us an indication that there is a possibility that search engines like Google Scholar and CORE play number games and

manipulate the data to show their supremacy and size of databases. However, the testing and verification of the hypotheses proved that there is no significant difference between and among search engines in the actual number of results and all search engines almost display the same number of hits

REFERENCES

Bar-Ilan, J., Levene, M., & Mat-Hassan, M. (2006). Methods for evaluating dynamic changes in search engine rankings: a case study. *Journal of Documentation*, *62*(6), 708-729.

Spink, A., & Jansen, B. J. (2004). A study of web search trends. *Webology*, *1*(2), 4. https://www.webology.org/data-cms/articles/20200515033219pma4.pdf

Carterette, B., Kanoulas, E., & Yilmaz, E. (2012). Evaluating Web retrieval effectiveness. In *Web search engine research*. Emerald Group Publishing Limited.

Dudek, D., Mastora, A., & Landoni, M. (2007). Is Google the answer? A study into usability of search engines. *Library Review, 56*, 3, 224-233.

Hölscher, C., & Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer networks*, *33*(1-6), 337-346. doi:10.1016/s1389-1286(00)00031-1

Jansen, B. J., Goodrum, A., & Spink, A. (2000). Searching for multimedia: analysis of audio, video and image web queries. *World Wide Web*, *3*(4), 249-254.

Kaushik, A. (2012). Judging the Capability of Search Engines and Search Terms. *International Journal of Information Dissemination and Technology*, 2(1), 6-11.

Loan, F. A. & Rufai, S.R. (2014). Myth and reality of the Google search results in terms of count estimation. *International Journal of Information Dissemination and Technology*, 4(2), 101-105.

Malla, M. A., & Loan, F. A. (2018). The Accuracy of Search Tools in Counting the Search Hits-A Case Study of the Five Search Tools. In *Building Smart Libraries Changes Challenges Issues & Strategies* (pp. 385-390). New Delhi: Asian Library Association.

Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. *Forbes. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=6ae48cd460ba*

Poulter, A. (1997). The design of World Wide Web search engines: a critical review. *Program: electronic library & information systems*, *31*(2), 131-145.

Xie, H. (2004). Online IR system evaluation: online databases versus Web search engines. *Online information review*, *28*(3), 211-219.