IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Precision and Recall of Google and Baidu for Retrieval of Scholarly Information in the field of Computer Applications

Aasim Bashir[1], Peerzada Mohd Iqbal[2]

Lovely Professional University, Phagwara, Punjab, India[1]

Senior Professional Assistant, SKUART-K[2]

Abstract: Google and Baidu, both search engines give different kinds of results for scholarly information using Computer Application related search terms. The modus operandi for this research follows a discrete pattern for search engines which are evaluated by taking the first ten results pertaining to "Scholarly Information in the field of Computer Applications" for estimation of precision and recall. 20 search terms were selected which were divided into three groups Simple, Compound and Complex. The search terms were selected by using a tool from the field of Information Science (Sears List of Subject Headings). The result will cover major fields like Comprehensiveness, Duplication, Currency, Dead Links, Fluctuations and Search Capabilities.

Index Terms: Baidu and queries, Computer Application, Google, Precision, Recall, Scholarly information.

1. INTRODUCTION

World Wide Web has become a part in our life, as the internet has become an important tool for searching and collecting information and also provides information from different WebPages over the internet. The ultimate goal in planning, designing and publishing a webpage is to share information. However the number of WebPages added to the internet or World Wide Web on a daily basis has made web a hub of all kind of data and information available, and provides a challenge for information retrieval. The amounts of information and data which is available on the internet and the domain name and hosts are growing rapidly. To overthrow these retrieval problems, more than twenty companies and institutions have developed tools such as Yahoo, Google, Baidu, Bing and many others [1].

The web is expanding exponentially. In September 2013, The Indexed Web contains nearly about 3.84 million web pages; The Dutch Indexed Web contains at least 238.92 million pages this expansion has led to reliance on the search tools or search engines for web results. This in turn casts responsibility on the search engines to meet the needs and expectations of the scholarly community. Using more than one search engine is impractical if overlapping of results is frequent and substantial. Overlapping is genuine if the results are more relevant to the user's queries. Use of different search engines simultaneously reduces searching time and increases efficiency [4].

Search engine is software whose work is to give the best search results on the search terms provided by the user from the www World Wide Web) or in other words we can say that it is a program for the retrieval of data, files, or document from a database or network, especially the internet. The information may be a specialist in web pages, images, information, pdf files, document files, presentation files and other types of files. This search result depends on the content of pages which are on the web and return the result on the keywords. When we do a search actually we are searching the web, we are searching the index of the web [3].

The search engines contains millions and sometimes billions of pages, many search engines not only just search the pages but also display the results depending upon their importance. This importance is commonly determined by using various algorithms.

It was found in 1998; Google's rapid growth made itself an unbeatable giant in the search engine battlefield. However, its dominance has yet to become an international absolute. Google Inc. struggles arduously on the digital battlefield in China's Internet search engine market. In China, Baidu.com has been described as China's Google for years and challenged Google's expansion. "The search engine marketing strategies adopted in China and the Western countries through these two illustrative cases, namely, Google and Baidu use search engine optimization (SEO) to rank their sites higher for queries. Baidu, however, offers paid search placement, or the selling of engine results for particular keywords to the higher bidders. Whereas Google has been providing innovative services ranging from Google Map to Google Blog, Baidu remains focused on search services – the one that it does best". The main emphasis lies on the fact whether Google is best or Baidu for retrieving scholarly information using Computer Application related search terms [2].

Proposed methodology is given to identify the search engines for retrieval of scholarly information in the field of Computer Application, analysis of precision and recall Evaluation of the selected engines and for understanding the effect of nature and types of queries on recall and precision of the selected engine based on simple, compound and complex terms.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

2. RELATED WORK

Search engine is a website dedicated to find other websites or information. In today's world Search engine is a key tool in finding information. The matter of concern is whether the information retrieved is the same we are looking for or mere relevant. For this a number of tools were developed in order to shorten this gap. Google is the most popular search engine in the World mostly used and widely accepted. Whenever we say search engine Google is the first web site which comes in our mind. But in some countries the local search engines perform better like Baidu in china and Japan. Still both search engines do the same task. The overall quality depends on perfectness of the material retrieved. Precision and recall are two major tools to check the relevancy of the material we are in search for [8].

Boyun chiou's research concludes Baidu to be Google of china as china is world's number one in populations so baidu is widely used in china. The researchers suggest Baidu to be number one search engine in china and taking the market of Google in china. His research on Yendex purely thrusts on baidu to be more precise than Google [2]. Shafi and Rather had presented a research and analysis on five search engines for retrieving scholarly information using biotechnology related search terms. The search engines are evaluated taking the first ten results pertaining to scholarly information for estimation of precision and recall [7].

According to Linda, Sau-ling LAI her research a comparative analysis of Google and Baidu has concluded Google to be best in providing effective services and Baidu in providing search excellence and pin pointed results [5].

3. PROPOSED METHODOLOGY

Terms we have to select are not directly possible in developed and multidimensional field like Computer Science. Therefore, experts were consulted for purpose of term related queries from the field of Information Science; they consulted classification schemes like Universal Decimal Classification and Dewey decimal classification to understand broad/ narrow structure of Computer Science. It helped to get five terms/ fields i.e., Computer Engineering, Information technology, Electronics and Communication, Network and Internet, Artificial Intelligence & Cryptography. These terms were browsed in LC list of subject headings which provided many other related terms like Related Term (RT) and Narrow Term (NT). Further NT and RT attached to each preferred or standard term were also browsed which retrieved a large number of computer terms. At first instance 190 Computer Science related terms were identified. Some terms occurred more than once and duplication removed. It reduced the number to 140. Later terms were divided into three broad groups under utility, Process and test. Utility denotes applications of Computer Science in various fields and about 70 terms came under this group. Process refers to a method of developing or processing some data, information & ideas and 40 terms fall under this group. Test means transformation of information to a desired level and 30 terms came under this group. Further under each category, these were further sub divided into 11 groups. Utility divided into 7 groups, process subdivided into two groups & test into three groups. The terms in each group were arranged alphabetically and each term was given a tag. Later 15% of the terms were selected from each group using "Systematic sampling" (i.e., first item selected randomly and next items after specific intervals). It further reduced the number to 20. Finally the selected terms were classified in three groups under simple, compound and complex terms (Table: 1). this will be done in order to investigate how search engines will control and handle simple and phrased terms. Simple terms containing a single word will submit to the search engines in the natural form i.e., without punctuation marks. Compound terms consisting of two words will be submitted to the search engines in the form of phrases as suggested by respective search engines. Complex terms composed of more than two words of phrases, will sent to the search engines with suitable Boolean operators "AND" and "OR" between the terms to perform special searches.

S. No	Simple Terms	Compound Terms	Complex Terms
1	Antivirus	Integrated Circuits	Database Management System
2	Coding	Mainframe Computers	Object oriented programming
3	Software	Programming Structures	Command line interface
4	Data	Arithmetic calculations	Asynchronous transfer mode
5	Networking	Hypertext Pre-Processor	New Technology File System
6	Windows	Proxy Server	Structured Query Language
7	----------	Screen Saver	Rapid Application Development

TABLE 1: SIMPLE, COMPOUND AND COMPLEX SEARCH TERMS

The process will be carried out in three stages. In the first stage, related material will be collected in print and electronic format for the study. In the second stage, search engines will be selected and search terms will be drawn subsequently. In the third stage, the data will be analyzed for results.

Search Engines for the Study

The search engines investigated will be:

- Google (Western Countries)
- Baidu (China and Japan)

DOI 10.17148/IJARCCE.2016.5224

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

4. DATA ANALYSIS

Nature of the term	Search term	Google		Baidu	
		Total no of url's retrieved	Scholarly url's retrieved	Total no of url's retrieved	Scholarly url's retrieved
SIMPLE TERMS	Antivirus	56,600,000	10,100,000	14,700,000	90,200
	Coding	81,900,000	61,900,000	26,100,000	827,000
	Software	863,000,000	104,000,000	100,000,000	5,020,000
	Data	1,520,000,000	189,000,000	100,000,000	13,400,000
	Networking	176,000,000	91,800,000	61,500,000	934,000
	Windows	797,000,000	42,200,000	100,000,000	4,070,000
Averages		**582,416,666**	**83,166,666**	**67,050,000**	**4,056,866**
COMPOUND TERMS	Integrated Circuits	89,50,000	65,50,000	3,960,000	582,000
	Mainframe computers	3,050,000	190,000	584,000	40,200
	Programming Structures	5,56,00,000	2,79,00,000	3,560,000	445,000
	Arithmetic Calculation	1,14,00,000	96,00,000	715,000	102,000
	Hypertext Processor	13,80,000	49,700	166,000	683
	Proxy Server	3,34,00,000	1,09,00,000	12,900,000	220,000
	Screen Saver	3,25,00,000	17,40,000	2,130,000	62,300
Averages		**20,897,142**	**8,132,814**	**3,430,714**	**207,454**

TABLE I: COMPREHENSIVNESS

Search Engine	Total Results	Scholarly publication	Percentage
Google	301656904	45649740	15.14
Baidu	35240357	2132160	6.06

TABLE II: MEAN SCHOLARLY COVERAGE

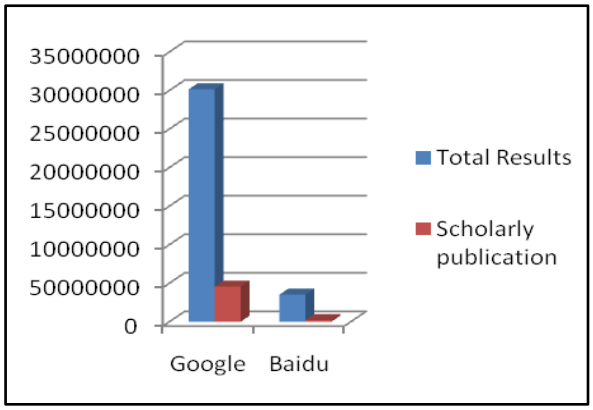

TABLE IV: MEAN SCHOLARLY COVERAGE OF GOOGLE AND BAIDU

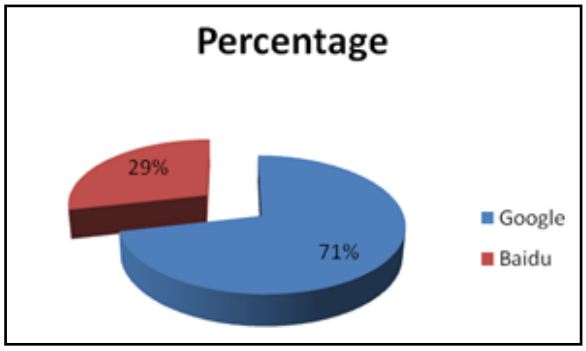

TABLE V: COMPARISON OF SCHOLARLY PUBLICATION

Search Engine	Total Hits	No of duplicate hits	Percentage of duplicate hits
GOOGLE	38	19	50%
BAIDU	23	12	52.17%

TABLE VI: DUPLICATION

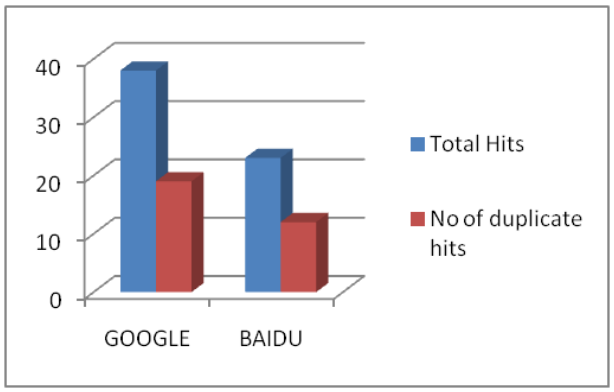

TABLE VII: MEAN DUPLICATION OF GOOGLE AND BAIDU

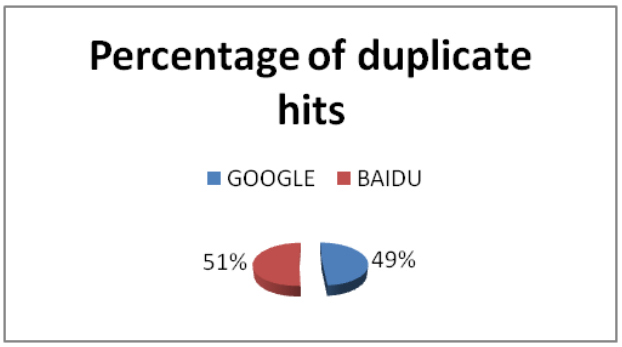

TABLE VIII: COMPARISION OF DUPLICATE HITS

Search Engine	Total results Evaluated	No of dead links	Percentage of dead links
GOOGLE	38	2	5.26%
BAIDU	23	3	13.04%

TABLE IX: DEADLINKS

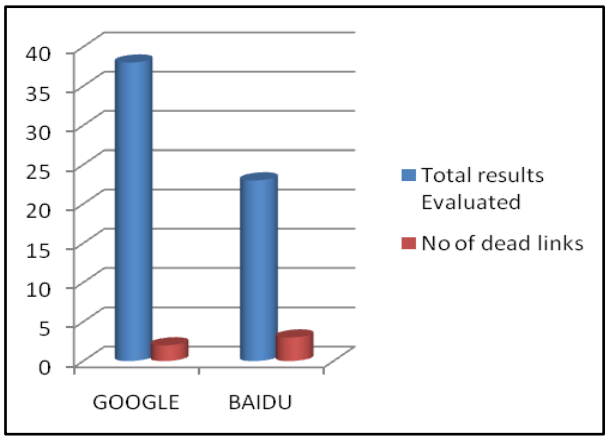

TABLE X: MEAN DUPLICATION OF GOOGLE AND BAIDU

DOI 10.17148/IJARCCE.2016.5224

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

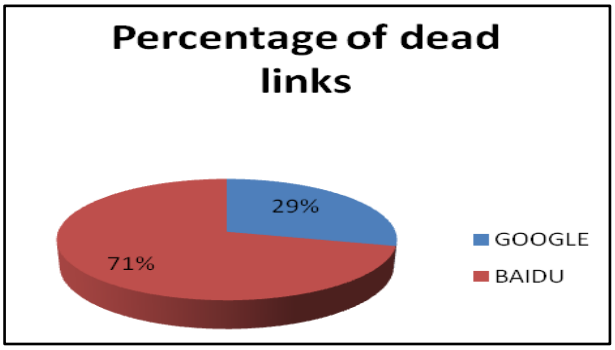

TABLE XI: PERCENTAGE OF DUPLICATION.

Nature of the term	Search term	Precision for first Ten Results										Sum of Precision ∑ P	Average Precision P= ∑ P/n
		I	II	III	IV	V	VI	VII	VIII	IX	X		
SIMPLE	Antivirus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Coding	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Software	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Data	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Networking	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Windows	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMPOUND	Integrated Circuits	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.10
	Mainframe computers	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Programming Structures	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Arithmetic Calculation	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Hypertext Processor	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.10
	Proxy Server	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Screen Saver	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.10

Mean Precision= 0.23
TABLE X11: PRECESSION OF GOOGLE

Nature of the term	Search term	Relative Recall of Search Engine BAIDU										Sum of Precision	Average Precision
		I	II	III	IV	V	VI	VII	VIII	IX	X	∑ P	P= ∑ P/n
SIMPLE	Antivirus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Coding	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Software	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Data	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Networking	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Windows	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMPOUND	Integrated Circuits	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.10
	Mainframe computers	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Programming Structures	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.10
	Arithmetic Calculation	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	2.00	0.20
	Hypertext Processor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Proxy Server	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Screen Saver	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Mean Precision= 0.30
TABLE XIII: PRECISION OF BAIDU

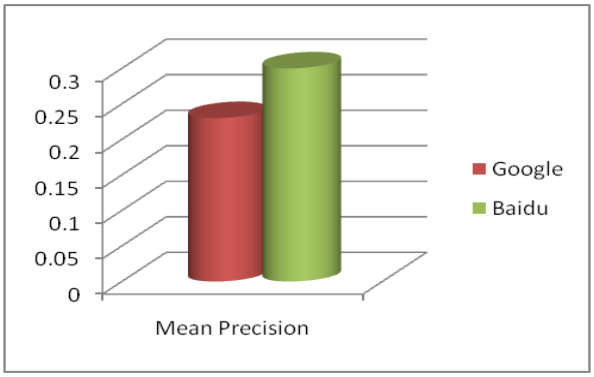

TABLE XV: MEAN PRECISION OF GOOGLE AND BAIDU

Search Engine	Mean Precision
Google	0.23
Baidu	0.30

TABLE XIV: PRECISION OF GOOGLE AND BAIDU

REFERENCES

[1] Anna Paananen (2012). Comparative Analysis of Yandex and Google Search Engines. Helsinki Metropolia: University of Applied Sciences Retrieved 7th September, 2013
[2] Chiou, Bo-Yun. (2009). Google takes on China : a cross-cultural analysis of internet service design. University Ave., Muncie. Retrieved 6th September, 2013
[3] Clarke, S., & Willett, P. (1997). Estimating the recall performance of search engines. ASLIB Proceedings, 49 (7), 184-189.
[4] Maurice de Kunder (2013). Worldwidewebsize.com (2013). Retrieved 6th September, 2013
[5] Linda, Sau-ling (2011). In Search of Excellence – Google vs Baidu. World Academy of Science, Engineering and Technology. 60:2011. Retrieved 7th September, 2013
[6] Information Services (Cardiff University). UK. Retrieved 26 September 2013
[7] 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.
[8] University Libraries (2013). University Libraries. New York.

BIOGRAPHIES

Aasim Bashir Masters In Technology (Computer Science) from Lovely Professional University, Phagwara, India. Masters in Computer Application from Lovely Professional University. He has received his bachelor's degree from Kashmir University. Currently working on International projects of SEO (Search Engine Optimization) as well as SMO (Social Media Optimization). His area of research is data mining and knowledge management process, internet technologies, Search Engine Optimization.

Peerzada Mohd Iqbal Masters In Library Science, Masters in Information Technology, Double M.Phil in Library Science and B.ED from Kashmir University. His area of research is Content Management, Digital Libraries, and System Analysis and Designing, Knowledge management process, Data Mining and Internet Research.