Technical Journal, University of Engineering and Technology (UET) Taxila, Pakistan *Vol. 28 No. 3-2023*
ISSN:1813-1786 (Print) 2313-7770 (Online)

Performance Comparative Analysis of Web Search Engines for Retrieving Computer Science Research Articles Using Information Retrieval Approaches

S. A. A. Shah[1], S. Ali[2], S. F. Solehria[3]

[1] *Department of Library and Information Sciences, Sarhad University of Information Technology, Pakistan*
[2] *Department of Computer Science, University of Peshawar, Pakistan*
[3] *Department of Computer Science, Sarhad University of Information Technology, Pakistan*

[2] shoonikhan@uop.edu.pk

Abstract- More than 200 computer science research journals publish thousands of research articles yearly. The publishers maintain explicit repositories of research articles and the bibliographic databases are locked behind paywalls. The exponential growth in the number of articles makes it troublesome for computer science research scholars to search and retrieve relevant research articles quickly and precisely. The web search engines (WSEs) are used by scholars to ease their searching and retrieving operations. However, the enlarged list of web search engines and lack of availability of their performances information makes WSE selection difficult for the scholars. This paper presents an empirical performance comparative analysis of the top five WSEs (i.e., Google, Bing, Yahoo, Baidu and Yandex) to retrieve computer science research articles in response to domain-specific queries varying in complexity. The comparison and evaluation methodology, performance analysis using information retrieval metrics (i.e., response time, recall, and precision), and systematic comparison of results are presented to highlight strengths and weakness of the WSEs. The results have shown that Google is performance-wise better WSE for searching and retrieving computer science research articles. The findings could be helpful for computer science research scholars in using an appropriate WSE in their searching and retrieving processes.

Keywords- Web Search Engine, Information Retrieval, Computer Science, Response Time, Recall and Precision, Research Articles, Crawler, World Wide Web

I. INTRODUCTION

The World Wide Web (WWW) has been proven as huge information source and fastest communication medium. The increasing advancements in WWW and digital storage technologies, and easy sharing of information globally have enabled developers to develop websites of numerous types pertaining information on the different topics [1-3]. To achieve the goal of information sharing, the rapid growth of the Web (i.e. 1.88 billion websites with more than 1 trillion web pages by 2021) and the number of domain names registered have turned WWW into a de-facto standard for retrieving information by billions of people around the globe in just a few clicks instead of being an esoteric system restricted to a specific class of individuals [4]. However, the addition of large amount of web resources has turned the Web into an ocean of different types of information, eventually has increased the challenges of information retrieval (IR) [5]. To facilitate the retrieving processes of users' for finding specific information in the haystack of information, the first Web Search Engine (WSE) was developed in 1994 namely World Wide Web Worm (WWWW), which could reference 110,000 web pages and answer 1500 search queries daily [6-7]. The WSEs are tools developed to facilitate users and reduce the time required for finding and retrieving information on the Web [8]. The WSEs are ranked at top of the highly accesses websites [9]. In the past several years, a number of WSEs with different ranking/indexing methods, scope of coverage, features, and interfaces are introduced by the researchers, academia and organizations [10]. For example, Google has indexed more than 64 billion web pages and received 8.5 billion searches per day in 2022 [11]. The WSEs builds and keep an index of the terms within the Web documents and provides search result in the form of ranked list of relevant documents. However, some of the results are usually relevant, whereas, majority is irrelevant to the users [9]. In addition, not all of the WSEs exhibits best performance for all types of search queries [10].

Therefore, selection of an appropriate WSE for a query is an increasingly difficult task for users [10].

The computer science is one of the most prominent fields of research with more than 200 computer science specific journals publishing thousands of research papers yearly. Among the biggest users of the WSEs are the computer science research scholars, who mainly use this amazing technology for retrieving relevant and precise research material of interest. However, the WSEs varies and differs in algorithms and functionalities from each other. Therefore, the WSEs due to their inherent shortcomings almost disappoint the research scholars and encounter them with difficulties during their retrieving activities. The WSEs are believed to return the relevant results but there are counter arguments to this belief and the WSEs could return non-relevant even biased results such as Mr. Donald Trump accusation on Google for displaying negative results when his name is searched on the Google [12]. The same applies to retrieving non-relevant scholarly articles in response to scholar's search queries. Although, some of the WSEs have a leap forward by introducing separate applications (known as bibliographic databases) for retrieving scholarly information such as Google's Google Scholar and Baidu Scholar. However, not all WSEs provide such separate applications. Therefore, the WSEs remain the primary tool for the scholars to search their required scholarly articles. Thus, the scope of this empirical performance comparative study is restricted to WSEs to provide a fair comparison, analysis and evaluations.

Like Information Retrieval (IR) systems, the performance evaluation of WSEs for retrieving computer science research articles is essential for their success, viability, reliability, and effectiveness. However, the lack of retrieval performance information limits research scholars to one or two of the WSEs which they feel comfortable and sometimes feel frustration for not easily retrieving required research articles. Therefore, identification of a best WSE which could actually satisfy the scholars' needs is essential. In addition, the increasing number of the WSEs have opened an avenue of research for examining their performance for retrieving computer science research articles [13-14].

This paper employs an empirical approach for the domain-focused performance comparison of the top five WSEs (i.e., Google, Bing, Yahoo, Baidu and Yandex) for retrieving computer science research articles using IR methods. Only five WSEs are selected from a big list of WSEs due to limitation of time and resources; however, the methodology and results could provide a benchmark for comparative analysis of other WSEs in other domains of research. Different measures are used for the WSEs performance evaluations including precision, coverage, response time, recall, and interface [15-16]. For the successful completion of this study, the WSEs performance comparative analysis is defined in terms of IR metrics: response time, recall, and precision. The performance metrics are of importance due to: (1) the recall and precision indictors would provide valuable information to determine effectiveness of ranking and indexing algorithms used by the WSEs; (2) the response time indicator would provide valuable information to determine efficiency of the searching and retrieving algorithms used by the WSEs. A WSE with low response time, high recall, and high precision will stand top in this performance comparative analysis study. A systematic methodology is defined for the selection of search queries and WSEs, selectin of participants for execution of search queries in a specialized test environment, collection and analysis of tests data and derivation of results. The results of the empirical study have depicted that Google has better performance than other WSEs in searching and retrieving computer science research articles and could be the best choice for the scholars having access to various online journals or databases like IEEE Explore, and ACM etc. The results are expected to increase knowledge of the scholars regarding WSEs capabilities and enable to make inferences about WSEs, and provide avenues for future research.

II. IR AND WSES OVERVIEW

The WSEs, by virtue, are within the field of IR [3]. In addition to providing features for representing, storing, access and searching, and manipulating huge collections of electronic documents [17], advanced web IR systems includes modeling, documents classification and categorization, systems architecture, user interfaces, data visualization, filtering, and languages [3]. However, IR is different than data retrieval. The data retrieval deals with structured and semantic data (e.g., data stored in relational databases), and a search query result at any instant of time will be accurate by returning the same data until changed in a database. The IR deals with unstructured and semantically ambiguous natural language text, and a search query result could be inaccurate as long as the error is significant. Therefore, an IR system implements mechanism(s) to interpret the information item by extracting its synthetic and semantic information and rank them according to relevancy to a user query to retrieve more relevant and few non-relevant results as much as possible.

A WSE is a large database of web accessible resources including web pages, documents, images, software programs, newsgroups, etc., enabling users to easily locate their required information on the WWW. The

Web search tools can be either directories or WSEs [9, 18]. The directories (e.g., DMOZ open directory, Yahoo directory, etc.) are collection of human-reviewed web pages and arranged into topical categories. The WSEs creates database by software known as crawlers, spiders or robots. The directors are based on manual browsing and the WSEs are based on automatic searching by taking keywords quires and provide a comprehensive list of relevant results using mathematical formula(s). The WSEs can be classified into different groups using features, function, geographical scope, and applications [18]. The crawler-based WSEs automatically and constantly crawl the Web and automatically updates the database with new or altered information about the Web resources. The hybrid WSEs combines crawler-based WSEs results and directory results and most of the WSEs are moving to hybrid-based model (e.g., Google, Yahoo, etc.). The metaWSEs takes results from other WSEs and combines them into a single large listing (e.g., Metacrawler, Dogpile, Searx, etc.). The domain specific WSEs are developed to satisfy user requirements in a specific domain [19]. For example, Froogle, BizRate, PriceSpy, PriceGrabber, etc., are WSEs specially developed for e-commerce and shopping. The WSEs indexes and searches web resources using keywords and phrases, and are mostly specialized for English languages. However, the WSEs specialized in other languages (e.g., Chinese, Russian, Korean, Japanese, etc.) are also available [3].

The WSEs are architecturally alike IR systems rendering almost the same features and functions. A WSE takes users' search queries consisting of keywords and phrases to convey semantics and information needs, and returns references to the relevant information items stored on the Web. A WSE architecture must satisfy two main criteria: (1) effectiveness in satisfying relevancy criterion to improve quality; and (2) efficiency in speed to have low response time and high throughput. The WSEs architecturally vary from each other; however, they constitutes four basic components: interface, crawler or spider, indexer, and query module [9]. Typical architecture of a centralized WSEs is shown in Fig. 1. The WSEs work by sending crawlers to search and fetch keywords and phrases information from as many web resources as possible. The indexer reads the information from crawlers, stores in a predefined database/catalogue, and creates index based on the information in the web resources. Every WSE implements a proprietary algorithm (e.g., frequency of keywords, relevancy of information, links, etc.) to create indices for efficient searching of required information and rank web resources to return meaningful and relevant results for search queries. For example, PageRank algorithm is used by Google. This enables WSEs to look for the keywords or phrases in the index of database instead of directly searching in the web resources. The interface module enables users to present keywords or phrases queries to a WSE, perform advanced configurations and settings, and display the web resources returned as result of search queries. The results normally include URL of a web resource title, size of text portion, few sentences, etc. The user examines the results retrieved for the presence of required keywords or phrases in the returned web resources.

The detailed discussion of WSEs' features and mechanisms can be found in [20]. Recognizing the economic benefits, a number of WSEs are designed and developed by the different organizations with varying capabilities, features, and applications. Although, the WSEs are designed to search the mammoth amount of information with impressive performance but not all WSEs exhibits best performance for all types of queries and are usually subjected to enormous critics including retrieving duplicate, and irrelevant information [10, 21]. It could be due to WSEs are creating and maintaining gigantic databases consisting of huge magnitude of information belonging to divers fields including media, marketing, entertainment, advertisement, etc. Thus, the tremendous growth of the Web poses serious challenges for the WSEs: (1) growth of the information is exponential as compared to the available technologies to index and results may become out-of-date; (2) the rapidly updatable web pages require periodic crawling and indexing, which could be missed by the WSEs [9]. This has given rise to research field of WSEs performance comparisons and evaluations, and different methods and measures are proposed by the research community over the time.

III. LITERATURE REVIEW

The WSEs performance comparison and evaluation has been the subject of interest for researchers since years. A detailed survey of WSEs evaluation can be found in [8]. However, most of the available researches have compared precision of WSEs using general subject queries with minimal consistency in terms of methodology and WSEs selection. Leighton [22] has examined Infoseek, Lycos, WebCrawler and WWWWorm, and found Lycos and Infoseek as higher in precision. Ding [23] has formulated five complex queries for investigating precision, repetition, and overlap degree of Infoseek, Lycos and Open Text. Using criteria of first twenty hits for precision calculation has shown Lycos and Open Text as superior with returning best results. Leighton [24] has examined AltaVista, Excite, HotBot, Infoseek and Lycos by executing fifteen complex search queries,

used criteria of first twenty hits for evaluation, and came up with the conclusion of AltaVista, Excite, and Infoseek the top three services in precision and response time evaluation. Chu [15] has investigated AltaVista, Excite and Lycos for their search capabilities and precision. The authors have used ten search queries of varying complexity by evaluating the first ten results for relevance assessment and revealed that AltaVista outperformed Excite and Lycos both in search facilities and retrieval performance. Clarke. [13] has examined AltaVista, Excite and Lycos by using thirty queries of variable nature and examined first twenty hits for evaluation and found Alta Vista as the one with best results in terms of precision, recall and coverage. Bar-Ilan [25] has used a single query "Erdos" and evaluated six WSEs. All of the 6,681 retrieved documents were checked for precision, overlap, and an estimated recall. However the report showed that none of the WSEs has high recall. Edosomwan [9] has examined Google, Yahoo, AlltheWeb, Gigablast, Zworks, AltaVista, and Bing with list of ten queries of varying complexity, evaluated first ten hits, and found Google as the best search engine in terms of precision and recall. Li [10] has examined AltaVista, Google, and Infoseek using automatic performance comparison and found that automatic method can be fast and flexible. However, the results have shown that the manual and automatic performance analysis showed the same results and Google is found best in term of precision.

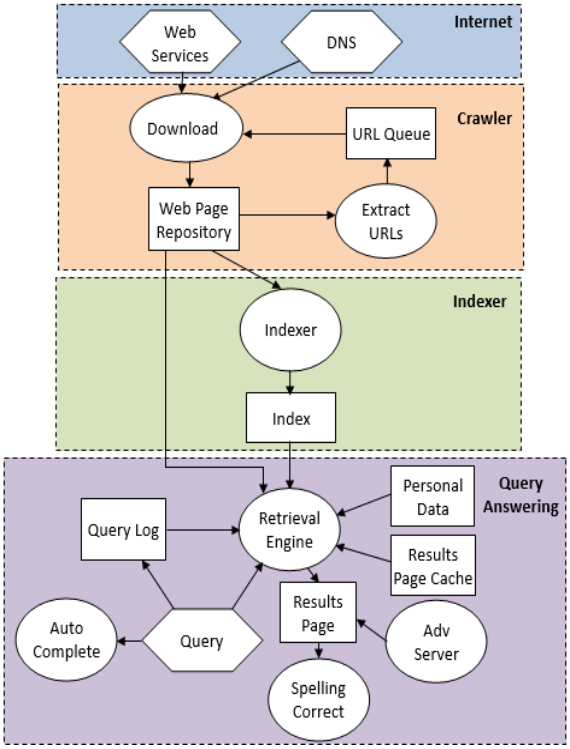

Fig. 1: Archiecture of a centralized WSE.

However, there is no significant work on the WSEs performance comparison and evaluation for domain-specific and specialized queries. Wishard [26] has used three queries (keywords and phrases) from the earth science domain and found that Go2, InfoMine, and Argus Clearinghouse have high precision for catalogue type search engines and Excite, Infoseek and Northern Light have high precision for the keywords type search engines. Lebedev [27] has examined seven search engine using eight different keywords from physics and chemistry, and used the size of the returned list as the primary performance measurement indicator instead of measuring precision of the returned links. Shafi [21] has examined AltaVista, Google, HotBot, Scirus and Bioweb using search queries of varying complexity from biotechnology domain, evaluated the first ten results pertaining the scholarly information, and found Scirus as top search engine for retrieving scholarly information in biotechnology domain. Anuyah [28] has examined Google and Bing WSEs for identification of their strengths and limitations when responding to search queries aligned with task objectives and user capabilities resources related to reading skills and curriculum relevant but not including hate speech, sexual content and non-opinionated. It is found that existing WSEs have limitations and are needed to improve their filtering and ranking algorithms. Gusenbauer [29] has examined the 28 widely used academic search systems including Google Scholar, PubMed and Web of Science for their systematic research search qualities and found that search systems have substantial differences in performance for usability in systematic research search qualities and Google Scholar cannot be regarded as an appropriate principal search system. Martin-Martin [30] has compared six bibliographic databases (i.e., Microsoft Academic, Dimensions, OpenCitations Index of CrossRef open DOI-to-DOI citations (COCI), Web of Science Core Collection (WOS), Scopus and Google Scholar) for retrieving citations across different subject categories and found Google Scholar the most comprehensive source. CheshmehSohrabi [31] has compared four general search engines (i.e., Google, Yahoo and DuckDuckGo) and three specialized search engines (Flicker, PicSearch and GettyImages) for image retrieval and found that general search engines and specialized search engines have average recall 76.32% and 24.51% respectively and average precision 82.08% and 32.21% respectively and existence of competition between general search engines for image retrieval.

Conclusively, the literature review has shown that several of the researchers have examined different WSEs either for general or domain-specific search queries with varying complexities, and have described their performances and features, search strategies,

precisions, response time, and coverage. However, none of them have attempted performance comparison analysis of the modern WSEs for searching and retrieving contents and relevancy of computer science research articles.

IV. MATERIALS AND METHODOLOGY

The methodology used in this empirical performance comparison study is based on the Cranfield approach for IR systems evaluation [32]. The Cranfield approach is based on test collection(s) comprising of re-usable and standardized resources (i.e., keywords/topics, documents and relevance measurements) for the evaluation of IR systems. These in combination with evaluation measures empowers users for operational settings and quantifying effectiveness of an IR system [33]. The evaluation using Cranfield approach typically requires: (1) selection of IR systems to compare, (2) creating a collection of search queries for judging relevance, (3) creating a ranked list of documents for each query, (4) computing the effectiveness of each IR system for each query as a function of relevant documents retrieved, (5) computing the overall effectiveness of a systems by averaging the score over all queries, and (6) relatively ranking the IR systems using the scores . Statistical analysis can be used in support for determining the significance of differences of effectiveness scores for IR systems [33].

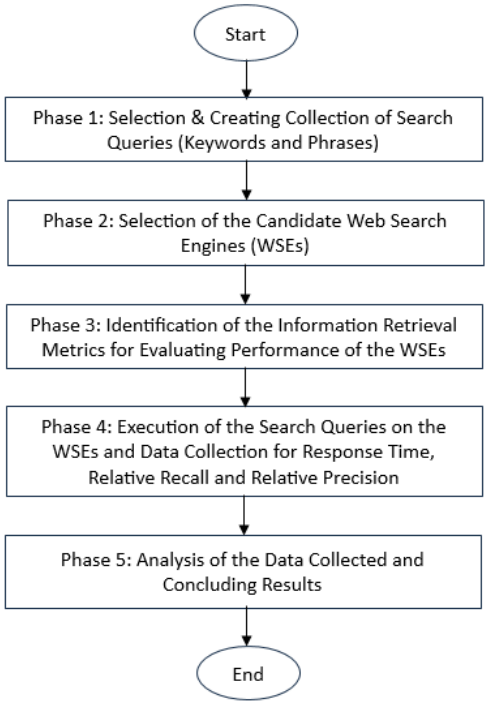

Fig. 2: Schematic diagram of the methodology phases.

Keeping in view of the Cranfield approach guidelines [31], the methodology of this research is composed of five phases (shown in Fig. 2). In the first phase, search queries (i.e., keywords and phrases) for the study are collected from the participants and computer science research literature available in the electronic format. In the second phase, the candidates WSEs for the study are selected. In the third phase, the IR metrics for evaluating performances of the candidate WSEs are identified. In the fourth phase, participants are selected for executing the search queries on the selected WSEs and collection of tests results data (i.e., response time, relative recall, and precise precision). In the last phase, the collected results data is analyzed and results are concluded.

Sample Search Queries

To fairly constitute search queries list comprising of keywords and phrases, volunteer and enthusiastic 20 research students (i.e., MS and PhD) from the different research areas of computer science (e.g., software engineering, web engineering and web semantics, computer networks, wireless sensors networks, databases and data mining, digital image processing, etc.) at the Department of Computer Science, University of Peshawar are interviewed and an extensive list of 150 most frequently used search queries is created. The search queries list is cleaned and scrubbed to improve its quality. Firstly, the search queries in the list are verified by checking in the relevant research literature and the search queries list is refined by removing duplications, incomplete and inaccurate search queries. Secondly, the search queries in the list with high frequencies (i.e., recommended by most of the participants and verified from the search requests from the Google Keyword Planner) and high scale of occurrence in most of the computer science research literature are selected. Therefore, for the search queries' tests, thirty search queries are selected from the total population of 150.

The selected search queries are of varying complexities either comprising of single or multiple words. Therefore, three groups (i.e., simple, compound and complex) are formulated using information from the relevant literature [9, 13, 15]. The three groups are formulated to address the simple and advanced search features of the selected WSEs and all possible combinations of the selected search queries. Therefore, the selected search queries are classified into the three groups for examining strengths of the selected WSEs' for handling and controlling single word and phrased terms.

- Simple search queries are composed of single word (e.g., ontology, sensor, etc.) and issued in natural manner (i.e., enclosing in double quotations or not).

- Compound search queries are composed of more than one words (e.g., information retrieval, Internet of Things, web mining, etc.) and are issued following the synthetic rules suggested by the respective WSEs (i.e., mainly by enclosing words in a phrase in double quotations such as "web semantics" and "wireless sensor networks", etc.).
- Complex search queries are formed by using Boolean operators (i.e., AND, OR, and NOT) offered by the selected WSEs between a search query term, which could be either simple or compound or any combination of them and issued for performing more special and specific searches. For example, "Algorithm" AND "Machine Learning", "Web Semantics" AND Ontology, Smartphone AND Lifelogging, "Augmented Reality" OR "Context Awareness" could be the potential complex search queries.

The selected sample search queries are fairly distributed across the three groups (i.e., simple, compound and complex) that 10 sample search queries are assigned in each group and are shown in the Table. I.

Candidate Web Search Engines

The market of WSEs has been revolutionized and new WSEs are appearing frequently in the market. Resultantly, we have a very big list of WSEs and it is increasing in number day by day. To accomplish the objective of the study, the WSEs of diverse range are used to evaluate various search algorithms for obtaining benchmark results of retrieving computer science research articles. However, care has been taken to select WSEs with acceptable level of usability (user friendly), popularity, versatility (general and multi languages support), size (number of web pages indexed), and global availability.

Table.1: List of Selected Search Queries in the Groups

Simple	Compound	Complex
Sensor	"Software Engineering"	Algorithm AND "Machine Learning"
Lifelogging	"Human Computer Interaction"	"Sentiment Analysis" AND "Social Media"
Malware	"Data Mining"	OPENCV AND "Image Processing"
Ontology	"Cloud Computing"	"Semantic Web" AND Ontology
Algorithm	"Privacy and Security"	Sensor AND "Internet of Things"
Smartphone	"Augmented Reality"	Security AND "Computer Networks"
Crawler	"Information Retrieval"	"Public Key" AND "Block Chaining"
Web	"Development Process Model"	"Ubiquitous Computing" OR "Context Awareness"
Database	"Information System"	"Wireless node" AND "Wireless Sensor Networks"
Cryptography	"Virtual Reality"	"Personalization" OR "Customization"

Using the considerations, the selected WSEs are Google, Bing, Yahoo, Baidu, and Yandex. The selected WSEs are general, active, and have independent crawlers. According to statistia, which is statistics portal (i.e., aggregate statistics and studies from the market, researchers, organizations, specialist publications, and from government agencies), the market share of the selected WSEs by June 2022 is shown in Fig. 3. A top level comparison of the selected WSEs is shown in the Table. II. However, the former three (i.e., Google, Bing, and Yahoo) are most popular having users from different regions of the world and the latter two (i.e., Badiu, and Yandex) are not most popular and mainly having users from China and Russia respectively. In addition, this study is limited to five WSEs due to lacking of resources and time. However, results of the study are expected to increase knowledge of the users about capabilities of the different WSEs and can potentially increase usage of the better performing WSEs.

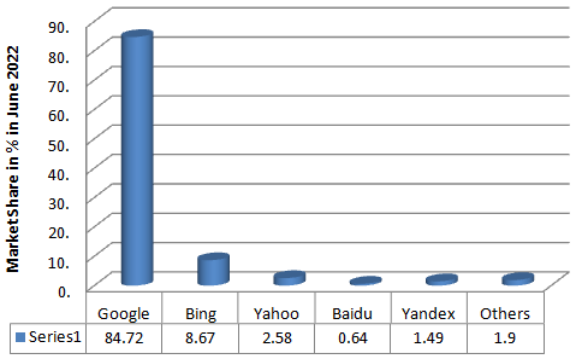

Fig. 3: Worldwide market share of the top five selected WSEs for the study.

Table. II: Top Level Comparison of the Selected WSEs for the Study

Test Computer	
Test Computer	DELL Inspiron 20 3059 All-in-One Desktop Core i3
CPU	Intel(R) Core (TM) i3-6100U CPU @ 2.3 GHz (4CPUs), ~203GHz
RAM	4 GB DDR3L RAM
Hard Drive	1 TB
Operating System	Microsoft Windows 10 Enterprise 64-bit
Internet Connection	
Digital Subscriber Line (DSL)	4MB (Bandwidth)
Web Browser	
Google Chrome	Version 84.0.4147.135 (Official Build) (64-bit)

Test Environment Setup

To perform objective comparison and obtain widely acceptable results, a fair playfield by means of a consistent testing environment is needed to be provided during the entire comparison study. All of the comparison tests on the selected WSEs are to be performed on the same computer and internet connection to eliminate any ambiguity about the results. The specifications of the test environment used in this study are depicted in Table. III. The Google Chrome browser is used for the study because of its wide usage and compatibility with the selected WSEs. The selected WSEs provide searching in two modes: simple and advanced. The selected WSEs are examined by configuring them in advanced mode throughout the study to effectively utilize their best features for producing effectively refined and precise results. For example, Google and Yahoo have almost same advance search options and "all of these words" is used for simple search queries, "the exact phrase" is used for complex and compound search queries, file format is select to PDF (because research articles are published in PDF format on the Web), etc. The Bing, Baidu, and Yandex have different advanced settings methods and are used accordingly. For example, the "filetype:pdf" is added after a search query in Bing and Baidu to limit searching and retrieving of web documents available in the PDF format. The selected WSEs are configured to return results in English language (Baidu supports translation into English using Google Translator).

Table. III: Specifications of the Test Environment for the Study

Search Engine	**Launched**	**Pages Indexed***	**Daily Direct Queries***	**Result Count**	**Language**	**Alexa Ranking**
Google	1998	Hundreds of billions	9.022 billion	Yes	Multilingual	1
Bing	1998/ 2009	13.5 billion	Unknown	Yes	Multilingual	47
Yahoo	1995	Unknown	Unknown	Yes	Multilingual	12
Baidu	2000	Unknown	Unknown	Yes	Chinese**	6
Yandex	1997	>2 billion	Unknown	Yes	Multilingual	55

* By December, 2022 (https://en.wikipedia.org/wiki/Comparison_of_web_search_engines).
** Supports English translation using Google Translate.

Performance Evaluation

The WSEs performance evaluation using traditional IR metrics can be helpful in determining their accuracy and credibility for retrieving computer science research articles. Primarily, the performance of an IR system is to be determined in efficiency and effectiveness [17, 33]. The efficiency can be computed automatically in response time (also called latency) representing the time taken by a WSE between submission of a search query and returning of results and normally depends on the amount of data collected and size of index file of a WSE. The effectiveness is difficult to be computed automatically and mainly depends on human judgment. The manual calculation is of high or at least same level of accuracy as automatic [10]. The effectiveness is measured in relevance measurement meaning a document is relevant if it contains information (partially or completely) represented in a search query [18]. However, the two main and frequently used metrics for computing effectiveness of an IR system are recall and precision [33]. The recall is measured in number of the relevant documents retrieved, whereas, precision is measured in number of retrieved documents are in fact relevant [18]. Technically, precision decreases with the increased number of retrieved documents and recall increases with the increased number of retrieved documents [34]. In this study, performance of the selected WSEs is measured using the standard IR metrics: efficiency in terms of response time (i.e., lower will be better), and effectiveness in terms of recall (i.e., higher will be better), and precision (i.e., higher will be better).

- *Response Time Measurement:* As discussed earlier the response time in this study is calculated as the time period between submission of a search query and retrieving of search results by a WSE. There is no pre-defined formula or automatic tool available for measuring response time of a WSE. Therefore, the response time is either measured by the information provided by a WSE or measured manually. In this study, the response time information provided by Google are considered acceptable, whereas, for the other candidate WSEs the response time is measured by a stopwatch (featuring 0.01 second precision). The response time is accessed for each of the search queries in the search queries tests list.
- *Relevancy Measurement:* The relevancy measurement quantifies the effectiveness of an IR system by determining whether a system respond to a query by retrieving documents containing the required information or not [18]. The typical standard methods used for relevancy measurement in IR systems are recall and precision [33]. The standard methods for measuring recall and precision encompasses the concepts of relevant and non-relevant documents and could be measured using Eq(1) and Eq(2) [33]. The Eq (1) and Eq (2) can be made clear using the contingency information shown in confusion matrix in Table IV. Therefore, using the contingency information, recall and precision will be as shown in Eq (3) and Eq (4). However, the Eq (3) and Eq (4) could be applied to set-based methods where the number of relevant and non-relevant documents is priory known, which is not possible in case of the Web. Therefore, the recall and precision in this study is measured using the methods defined by Clarke [13] for the WSEs. The modified recall and precision criterions give the same effects and results as of the standard measures.

$$\text{Recall} = \frac{\#(\text{relevant documents retreived})}{\#(\text{relevant documents})} \quad \text{eq}(1)$$

$$\text{Precision} = \frac{\#(\text{relevant documents retreived})}{\#(\text{retreived documents})} \quad \text{eq}(2)$$

Table. IV: Confusion Matrix for Contingency Information

Result	**Relevant**	**Non-Relevant**
Retrieved	True Positive (TP)	False Positive (FP)
Non-Retrieved	False Negative (FN)	True Negative (TN)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{eq}(3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{eq}(4)$$

Recall of a retrieval system signifies its strength to retrieve almost all of the relevant documents in the collection. To effectively calculate absolute recall of a retrieval system requires complete knowledge of relevant documents which are retrieved and relevant documents which are not retrieved [13]. Therefore, no scientifically agreed method exists to calculate absolute recall of a WSE because of lacking of accurate information of total relevant documents in the haystack of documents on the Web. However, a traditional recall measurement method is proposed and adopted by Clarke [13] for using in the Web scenario by giving it a relative flavor. Therefore, to calculate the relative recall for each of the search queries, the method defined by Clarke [13] and shown in Eq (5) will be used. This method suggests formation of denominator for a calculation by pooling relevant results of individual searches. The denominator could be either without overlapping or with overlapping. For five WSEs (e.g., V, W,X, Y, Z) with retrieval results of V1, W1, X1, Y1, Z1. In case, if there is no overlapping between the WSEs results (i.e., $V \cap W$, $V \cap X$, $V \cap Y$ and $V \cap Z$ is zero) then the relative recall of the WSE V would be calculated as V1/(V1+W1+X1+Y1+Z1). In case, if the overlapping between the WSEs exists (i.e., $W2 = V \cap W$, $X2 = V \cap X$, $Y2 = V \cap Y$ and $Z2 = V \cap Z$) then the relative recall of the WSE V would be calculated as V1/(V1+W2+X2+Y2+Z2). The relative recall would be more in case of overlapping between the WSEs.

$$\text{Relative Recall} = \frac{\text{Total number of articles retrieved by a WSE}}{\text{Sum of articles retireved by all five WSES}} \quad \text{eq}(5)$$

A portion of a search result that is relevant to a peculiar search query defines precision. However, accurate estimation of precision requires precise knowledge of relevant and non-relevant documents in the collection of documents under consideration [13]. The Web encompasses gigantic collection of computer science research articles document; thus, making precise prior knowledge of relevant documents practically impossible. To determine the relevancy of a WES, the absolute precision of a WSE for a search query needs to be calculated by examining the complete set of relevant document returned, which is not possible in case of millions of results returned by a WSE. However, the precise precision can be calculated by restricting to examine the first considerable number of results returned by a WSE. Therefore, the first thirty results will be examined to calculate precise precision

of a WSE. The precise precision of a WSE for a search query in this study will be calculated using the method defined by Shafi [21] and shown in Eq (6). This method suggests formation of numerator for calculation of precise precision for individual search queries. Therefore, a four-point scale criterion defined by Shafi [21] will be used in this study to find the numerator. The criterion used is as follows:

$$\text{Precise Precision} = \frac{\text{Sum of scores of articles retreived by a WSE}}{\text{Total number of results evaluated}} \quad \text{eq(6)}$$

1. A search hit containing full text of a research article (i.e.. either journal or seminar/conference proceedings) or of a patent will be scored with three.
2. A search hit representing abstract of a research article (i.e.. either journal or seminar/conference proceedings) or of a patent will be scored with two.
3. A search hit corresponding to a book or a database will be scored with one.
4. A search hit which do not corresponds to any of the three above (i.e.. company web pages, dictionaries, encyclopedia, organization, etc.) will be scored with zero.
5. A search hit which is duplicate but with different URL will be scored with zero.
6. A non response from a WSE server for subsequent three searches will bes scored with zero.

V. Results and Discussion

According to the methodology, the data for the empirical study is collected from the participants. The 10 interested, enthusiastic and expert participants are selected voluntarily from the MS and PhD students of the Department of Computer Science, University of Peshawar. However, the criterion for the selection of participants is to have substantial theoretical and practical knowledge of computer science, data mining and text mining, information systems and WSEs. In addition, a comprehensive 2- days workshop is conducted to educate the participants about the scope of the study, selected WSEs, execution of the search queries and calculation of the response time, relative recall and precise precision values form the search results. The number of participants is kept small due to limitations of resources and time. However, it is large enough to conduct tests and derive conclusions. The tests are performed for five weeks in the months of October and November 2022. The participants are split into five groups (i.e., 2 participants in each group) where each group is asked to use one WSE in each week (shown in Table V) and to execute each of the search queries on the WSEs using the test environment (discussed in Section IV – C) and to collect results about response time, relative recall, and precise precision. In addition, the participants are advised to be impartial and unaffected by preconceived notions about which WSE is superior.

Table. V: Groups Assigned to Different Wses Each Week

Group	Week 1	Week 2	Week 3	Week 4	Week 5
Group 1	Google	Bing	Yahoo	Baidu	Yandex
Group 2	Bing	Yahoo	Baidu	Yandex	Google
Group 3	Yahoo	Baidu	Yandex	Google	Bing
Group 4	Baidu	Yandex	Google	Bing	Yahoo
Group 5	Yandex	Google	Bing	Yahoo	Baidu

A search test is performed by submitting a search query to a selected WSE which could return numerous results. However, to accomplish the study, for each of the search queries the first thirty results (hits) are considered and evaluated. The thirty sample size is optimum because of the fact that users usually go through the first ten hits for receiving the relevant and expected search results otherwise get frustrated and try another WSE [9]. To the best efforts, the same search query is executed on all of the selected WSEs at the same time (using separate computers of the same configuration) to avoid any conflicting advantage for a WSE by indexing new web resources meanwhile. All of the tests are performed at the Web Engineering Laboratory of the Department of Computer Science, University of Peshawar.

Response Time

As discussed earlier, the response time (in seconds) for each of the search queries in the search queries groups is measured using either from the time information provided by a WSE (i.e., Google) or manually using stopwatch with high precision for the WSEs not providing time information (.i.e., Bing, Yahoo, Baidu, and Yandex). The mean response time comparison for the groups and among the groups for the selected WSEs is shown in Table. VI and Fig. 4. The mean response time for the WSEs is found within the range 0.42s and 0.75s. Comparing the group-wise mean response time, Google is having the lowest mean response time for simple (0.42s), compound (0.56s), and complex (0.64s) search queries groups. The Baidu has the highest mean response time for simple (0.57s) and complex (0.75s) search queries groups, and the Yandex has the highest mean response time for compound (0.64s) search queries group. Comparing the overall mean response time, Google is at the top position by having the lowest mean response time (0.54s) followed by Yahoo (0.59s).

The Baidu is at the lowest position with the highest mean response time (0.65s). However, overall mean response time difference is found not very much significant and all of the WSEs have shown acceptable level of response times.

Table. VI: Mean Response Time in Seconds of the WSEs for the Search Queries Groups

Criteria	Search Queries Groups	Google	Bing	Yahoo	Baidu	Yandex
Response Time	Simple	0.42	0.52	0.49	0.57	0.53
	Compound	0.56	0.58	0.59	0.62	0.64
	Complex	0.64	0.73	0.69	0.75	0.69
	Mean	0.54	0.61	0.59	0.65	0.62

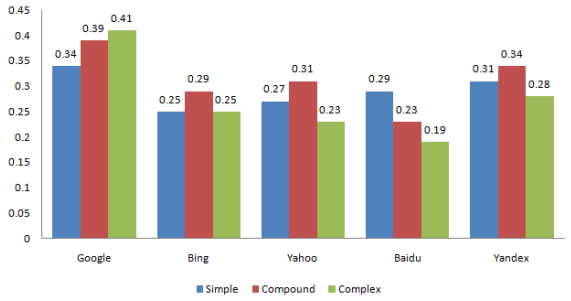

Fig. 4: Mean response time in seconds of the search queries in each of the search quereis groups for the selected WSEs.

Recall and Precisions

As discussed earlier, the relative recall of the WSEs for all of the search queries in the search queries groups is calculated using the formula shown in Eq(5) above. The mean relative recall comparison for the groups and among the groups for the WSEs is shown in Table. VII and Fig. 5. The mean relative recall for the WSEs is found within the range 0.25 and 0.41. Comparing the group-wise mean relative recall, Google has the highest mean relative recall for simple (0.34), compound (0.39), and complex (0.41) search queries groups. The Bing has the lowest mean relative recall for simple (0.25) search queries group and Baidu has the lowest relative recall for compound (0.23) and complex (0.23) search queries groups. Comparing the overall mean relative recall, Google is at the top position by having the highest mean relative recall (0.38) followed by Yandex (0.31). The Baidu has lowest position with the lowest mean relative recall (0.24).

Table. VII: Mean Relative Recall of the WSEs for the Search Queries Groups

Criteria	Search Queries Groups	Google	Bing	Yahoo	Baidu	Yandex
Relative Recall	Simple	0.34	0.25	0.27	0.29	0.31
	Compound	0.39	0.29	0.31	0.23	0.34
	Complex	0.41	0.25	0.23	0.19	0.28
	Mean	0.38	0.26	0.27	0.24	0.31

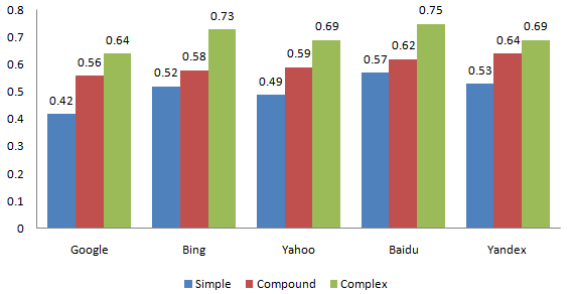

Fig. 5: Mean relative recall of the search queries in each of the search queries groups for the selected WSEs.

As discussed earlier, the precise precision of the WSEs for all of the search queries in the search queries groups is calculated using the formula shown in Eq(6). The mean precise precision comparison for the groups and among the groups for the WSEs is shown in Table. VIII and Fig. 6. The mean precise precision for the WSEs is found within the range 0.28 and 0.68. Comparing the group-wise mean precision, Google has the highest precise precision for simple (0.32), compound (0.76), and complex (0.68) search queries groups. The Yahoo has the lowest mean precise precision for simple (0.28) search queries group and Bing has the lowest mean precise precision for compound (0.54) and complex (0.47) search queries groups. Comparing the overall mean precise precision, Google is at the top position by having the highest mean precise precision (0.59) followed by Yandex (0.52). The Bing has the lowest position with the lowest mean precise precision (0.43).

Overall Perforamce

To measure the overall performance of the WSEs, they are ranked using the response time, relative recall and precise precision mean scores. The Table. IX depicts the WSEs ranked in terms of their response times (from shortest to longest), relative recall scores (from highest to lowest), and precise precision scores (from highest to lowest), with a rank of 1 denoting the best performer and so on. The mean of the three rankings shows the overall performance of the WSEs with the lowest mean the highest performance and vice versa.

Comparatively, Google has the lowest mean rank showing the Google has highest overall performance and Baidu has the highest mean showing the lowest overall performance to retrieve computer science research articles.

Table. VIII: Mean Precise Precision of the WSEs for the Search Queries Groups

Criteria	Search Queries Groups	Google	Bing	Yahoo	Baidu	Yandex
Precise Precision	Simple	0.32	0.29	0.28	0.34	0.45
	Compound	0.76	0.54	0.55	0.63	0.53
	Complex	0.68	0.47	0.51	0.48	0.59
	Mean	0.59	0.43	0.45	0.48	0.52

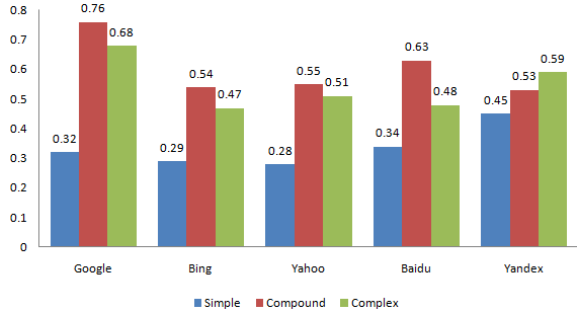

Fig. 6: Mean precise precision of the search queries in each of the search quereis groups for the selected WSEs.

Table. IX: Ranking of the WSEs According to their Response Time, Relative Recall, and Precision Scores and Overall Performance

Criteria	Search Engines				
	Google	Bing	Yahoo	Baidu	Yandex
Response Time	1	3	2	5	4
Relative Recall	1	4	3	5	2
Precise Precision	1	5	4	3	2
Mean Rank	1.0	4.0	3.0	4.33	2.67
Ranking*	**1st**	**4th**	**3rd**	**5th**	**2nd**
* Overall performance ranking based on mean ranks					

VI. Conclusion and Future Work

This paper has presented an empirical performance comparative analysis of the top WSEs (i.e., Google, Bing, Yahoo, Baidu, and Yandex) to retrieve computer science research articles using the IR approach (i.e., response time, relative recall, and precise precision). The results of the study have depicted the overall better performance of Google in searching and retrieving computer science research articles and could be the best choice for the scholars having access to various online journals or databases like IEEE Explore, and ACM etc. The Yandex is ranked second by showing acceptable response time, relative recall, and precise precision performances. The Yahoo is powered by Bing and both have shown almost the same performances. Therefore, they are ranked third and fourth respectively. The Baidu has shown satisfactory precise precision performance but worst response time and relative recall performances. Therefore, ranked fifth in the study. The findings have also established inverse proportionality of relative recall and precise precision (i.e., increase in relative recall produces decrease in precise precision and vice versa). The study has also shown that using a best WSE could maximally retrieve half of the scholarly articles. However, the study has certain limitations due to time and resources constraints. Firstly, the relative recall and precise precision are calculated by examining the first thirty results returned by a WSE. Secondly, the search queries list is restricted to 30 search queries. Lastly, the narrow segment of the participants' population (i.e., 10 students). Despite the limitations of the study, we believe the results could be valuable for computer science scholars to select relatively useful WSEs for searching and retrieving computer science research articles to find new research avenues.

In the future work, we are expected to compare the leading WSEs and bibliographic databases for measuring their performance for retrieving research articles in other domains of research with large search queries list, search results count and participants population, and come up with valuable suggestions to improve quality of the WSEs and bibliographic databases for searching and retrieving research articles.

References

[1] Lewandowski D. Understanding Search Engines. Springer Nature; 2023 Mar 7.

[2] Schlichting C, Nilsen E. Signal detection analysis of WWW search engines. 1996. Online available: http://www.microsoft.com/usability/webconf/ schlichting/schlichting.htm [Accessed 02/05/2022]

[3] Lam S. The overview of web search engines. University of Waterloo. 2001 Feb 9.

[4] Oppenheim C, Morris A, McKnight C, Lowley S. The evaluation of WWW search engines. Journal of documentation. 2000 Apr 1; 56(2):190-211.

[5] Ali S, Khusro S, Khan M. Desktop Search Engines: A review from user perspectives. Protecting User Privacy in Web Search Utilization. 2023:63-96.

[6] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems. 1998 Apr 1; 30(1-7):107-17.

[7] Duka M, Sikora M, Strzelecki A. From Web Catalogs to Google: A Retrospective Study of Web Search Engines Sustainable Development. Sustainability. 2023 Apr 17;15(8):6768.

[8] Goel S, Yadav S. An overview of search engine evaluation strategies. International Journal of Applied Information Systems. 2012 Feb;1(4):7-10.

[9] Edosomwan J, Edosomwan TO. Comparative analysis of some search engines. South African Journal of Science. 2010 Nov 1;106(11):1-4.

[10] Li L, Shang Y. A new method for automatic performance comparison of search engines. World Wide Web. 2000 Dec; 3:241-247.

[11] Mohsin M. 10 Google Search statistics you need to know in 2020. 2020. Online available: https://www.oberlo.com/blog/google-search-statistics [Accessed 15/04/2022]

[12] Gezici G, Lipani A, Saygin Y, Yilmaz E. Evaluation metrics for measuring bias in search engine results. Information Retrieval Journal. 2021 Apr;24:85-113.

[13] Clarke SJ, Willett P. Estimating the recall performance of web search engines. In Aslib proceedings. 1997 Jul 1; (Vol. 49, No. 7, pp.184-189). MCB UP Ltd.

[14] Moody G. Searching the web for gigabucks. New scientist. 1996 Apr;150(2024):36-40.

[15] Chu H, Rosenthal M. Search engines for the World Wide Web: A comparative study and evaluation methodology. In Proceedings of the Annual Meeting-American Society for Information Science 1996 Oct 19 (Vol. 33, pp. 127-135).

[16] Dong X, Su LT. Search engines on the World Wide Web and information retrieval from the Internet: A review and evaluation. Online and CD-ROM review. 1997 Feb 1; 21(2):67-82.

[17] Buttcher S, Clarke CL, Cormack GV. Information retrieval: Implementing and evaluating search engines. MIT Press; 2016 Feb 12.

[18] Khusro S, Ali S, Alam I, Ullah I. Performance evaluation of desktop search engines using information retrieval systems approaches. Journal of Internet Technology. 2017 Sep 1;18(5):1043-55.

[19] Batrimenko AV, Denisova S, Lisovskii D, Orlov S, Soshnikov S. The Internet search engines as an additional tool in public health research in the context of disease outbreaks. International Journal of Health Governance. 2022 Jan 27;27(2):194-207.

[20] Khalil A, Alrub F. A comparison of search engine's features and mechanisms. Advanced Database Systems. 2013;3:343-54.

[21] Shafi S M, Rather R. Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. Webology. 2005;2:42-47.

[22] Leighton H. Performance of four WWW index services, Lycos, Infoseek, Webcrawler and WWW Worm. 1996. Online available: http://course1.winona.edu/vleighton/webind.htm [Accessed 20/04/2022]

[23] Ding W, Marchionini G. A comparative study of web search service performance. In Proceedings of the ASIST Annual Meeting 1996 (Vol. 33, pp. 136-42).

[24] Heighton H, Srivastava J. Precision among WWW search services (search engines): AltaVista, Excite, HotBot, Infoseek and Lycos. 1997. Online available: http://course1.winona.edu/vleighton/webind2/webind2.htm [Accessed 10/04/2022]

[25] Bar-Ilan J. On the overlap, the precision and estimated recall of search engines. A case study of the query "Erdos". Scientometrics. 1998 Jun 1;42(2):207-28.

[26] Wishard L. Precision among Internet search engines: An earth sciences case study. Issues in science and technology librarianship. 1998;18.

[27] Lebedev A. Best search engines for finding scientific information in the Web. Rapport interne. 1996: 17.

[28] Anuyah O, Milton A, Green M, Pera MS. An empirical analysis of search engines' response to web search queries associated with the classroom setting. Aslib Journal of Information Management. 2020 Jan 8;72(1):88-111.

[29] Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Research synthesis methods. 2020 Mar;11(2):181-217.

[30] Martín-Martín A, Thelwall M, Orduna-Malea E, Delgado López-Cózar E. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via

citations. Scientometrics. 2021 Jan;126(1):871-906.

[31] CheshmehSohrabi M, Sadati EA. Performance evaluation of web search engines in image retrieval: An experimental study. Information Development. 2022 Nov;38(4):522-34.

[32] Cleverdon CW. The significance of the Cranfield tests on index languages. In Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval 1991 Sep 1 (pp. 3-12).

[33] Clough P, Sanderson M. Evaluating the performance of information retrieval systems using test collections. Information research. 2013 Jun 1;18(2):18-2.

[34] Manning CD. An introduction to information retrieval. Cambridge university press; 2009.