ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 3, pp: (24-29), Month: July - September 2019, Available at: www.researchpublish.com

Information Retrieval of Distributed Databases A Case Study: Search Engines Systems

[1]Sarah Hamed Alahmadi, [2]Dr.Fahad AlQurashi

[1]Faculty of Computing and Information Technology, Aziz University Jeddah, KSA-King Abdul

[2]Faculty of Computing and Information Technology, Aziz University Jeddah, KSA-King Abdul

Abstract: **This research investigates information retrieval from distributed databases. The most famous system for this work is the regular web search engines. Thus, this paper discusses in depth the search engines systems and rch engines based on popular queries searched. Before going into have made an evaluation of different sea evaluation phase, we explore the structure of the search engines and explain its functions behind the scenes, how ines is the ranking algorithms used to they can get the results for a specific query. The key idea behind search eng bring the appropriate results. The paper chooses Google, Yahoo, and Bing to do evaluation experiments, since they .nowadays are top search engines used in**

Keywords: **ritms , EvaluationInformation Retrieval ; Search Engines; Ranking Algo.**

I. INTRODUCTION

can access the data located in distributed we This paper will be about Distributed Information Retrieval Systems and how s [1], the research will system. Since the search engines are popular examples of distributed information retrieval system a distributed system. study this concept using these systems as a case study and describe how to retrieve information from and ranking Information retrieval systems IR consists of many stages starting from the user's query to searching process the results and finalizing by show the revised results. One important step is the ranking process and this what the paper o will focus on particularly. Ranking algorithms in general used for prioritizing the retrieved webpages for some queries t are various algorithms nowadays for page ranking as make them more relevant to what users search about. There PageRank, Hups and Authorities, Hilltop, and others [2]. The paper will make comparison on several algorithms, study valuate the performance based on experimental results by evaluation of several search engines. their ways of ranking and e The study will focus on the evaluation of the ranking algorithms based on selected evaluation criteria. The paper out background of information retrieval and its related aspects. These aspects architecture will be as follow: section 3 is ab include definitions of distributed databases, search engines, and overview of existing ranking algorithms. Then section 4 architecture of this paper and the results of the evaluation is about the experimental results which shows the work . investigate the results at discussion part we , experiments. After that

II. LITERATURE REVIEW

A. *Distiributed Databases*

searched about located on different computers or we A distributed database is a database in which the documents or data different storage devices among a large geographical areas or simply distributed across internet [1]. Such architecture of d these data. For this purpose, the distributed data need system to manipulate these data and search systems to retrieve search engines systems appeared. These systems that exist on the World Wide Web (WWW) are examples of distributed .[1] information retrieval systems

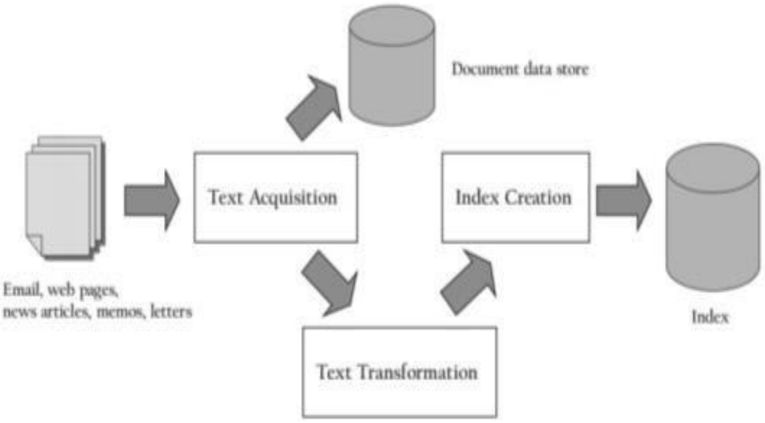
B. *Retrieval Information*

there is ,complexity and size in growing expanding is it and web pages Since the World Wide Web consists of millions of need to retrieve the best web pages that are more relevant in terms of information for the query entered by the user. Thus, web pages. Ranking algorithms explained in detail in next search engines require ranking algorithms to prioritize the .[3] sections. There is many issued related to information retrieval such as relevance and evaluation

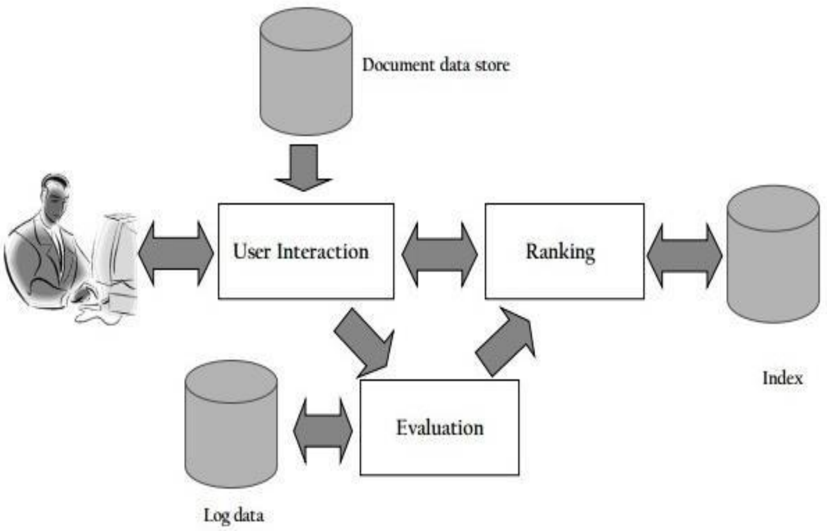
the contains document ncereleva a ,Freely .retrieval data idea in major a is Relevance .relevance is aspects these of One
data that a user was searching for when he/she presented a question to the web crawler [3]. Just contrasting the content of
ramework, a question and the content of a document and searching for a correct match, as may be done in a database f
delivers exceptionally poor outcomes regarding relevance. To address the relevance issue retrieval models proposed [3].
foundation the is It .[3] document a and inquiry an A retrieval model is a formal portrayal of the way toward coordinating
the ranking algorithms that is utilized as a part of a web searcher to create the ranked list of documents. Furthermore, of
these retrieval models used in search engines (specifically on ranking algorithms) have to focus on user relevance instead
elevance [3]. Since the user relevance satisfied user expectations and it used statistical properties of a given of topical r
of text rather than linguistic structure. The evaluation viewpoint is another center issue for data retrieval. Since the quality
king relies on upon how well it coordinates a user's desires, it was fundamental to create assessment a document ran
.[measures and experimental strategies for getting this information and utilizing it to think about ranking algorithms [3
work engines search the verify to methods many is There like precision and recall measure. This illustrated in detail in
users that queries Text .needs information their and on users experimental results section. As well there is an issue focus
oor descriptions of what the user actual wants. Thus, methods such as inquiry typed usually on search engines are often p
recommendation, query extension, and others utilized to refine the underlying query with a specific end goal to deliver
.[3] better ranked results

C. ENGINES SEARCH

A search engine is the actual application of data retrieval procedures to huge text gatherings [3]. A web search engine is
the evident example. Web crawlers can be found in a wide range of applications, for example, desktop search or inside a
inside a web search m and others. It progressively came to be used in predilection to "information ,document seek
retrieval system" as the name for the product framework that contrast inquiries to documents and delivers ranked result
example, Google and Yahoo! must have the capacity to catch, or for ,h enginesarrangements of webpages. Web searc
second reaction times to a huge number of inquiries sub crawl, numerous terabytes of information, and after that give in
ome issues regarding to search engines systems. These majors presented each day from around the world [3]. There are s
speed, scalability, incorporating new data, and adaptability are performance, response time, query throughput, indexing
g the outcome list, throughput measures the Response time is the lateness between presenting a question and gettin .[3]
quantity of inquiries that can be handled in a given time, and indexing speed is the rate at which text webpages can be
earch. Seek applications enhances the speed of s that structure data an is index An .looking for changed into indexes
commonly handle dynamic, regularly altering data. Coverage measures the amount of the current data in, say, a corporate
the recency or freshness measures the "age" of and ,crawler the web in away put and ordered been data condition has
n saved data. Scalability is plainly an essential issue for internet searcher design. Outlines that work for a given applicatio
ought to keep on working as the measure of information and the quantity of clients grow. Adaptable implies that a wide
the have must ,strategy indexing the or ,example, the ranking algorithm, the interface for ,of parts of the web crawler range
be tuned and adjusted to the necessities of the application. The search engines architecture consists of two to capacity
The indexing process shown in figure 1 determines and make .*query process* and *indexing process* main components
produce the index terms in form of inverted index. The query process shown in available of searching documents and
.results figure 2 take the user's query and produce a refined ranked list of



:Figure 1 Indexing Process



:Figure 2 Query Process

D. ALGORITHMS OVERVIEW OF MAJOR RANKING

is a fundamental part of any information retrieval system. In the instance of Web inquiry, due to the extent of the Ranking
ranking become ticklish. It is regular for Web search inquiries to of role the ,clients Web the of way special the and Web
a large number of results. Moreover, Web clients they cannot try these huge amounts of results and have thousands or
.they usually look at to the top few list. [1] shows that most Web clients don't look on the far side of first page of results
rative for the ranking process to yield the coveted outcomes inside the main couple of pages, Along these lines, it is impe
.unusable else the internet searcher is rendered

merous The top retrieved relevant pages often are results of ranking algorithms based on strong retrieval model [2]. Nu
retrieval models have been proposed throughout the years. Two of the most established are the Boolean and vector space
models[2]. The Boolean retrieval model was utilized by the oldest web indexes and is still in utilize today. It is likewise
calledmatch retrieval since reports are retrieved if they precisely correspond the query detail.Vector space models -exact d
naturally engaging system for executing term weighting, ranking, and relevance and has the benefit of being a basic
.feedback

y classifications of web ranking algorithms which found through the search. According to [4] the ranking There are man
algorithms divided into Link analysis algorithm, Personalized web search ranking algorithms, and Page Segmentation
based algorithms -y sort them into two types of ranking algorithms which are: Contentalgorithms. And according to [5] the
and Link analysis algorithms. Each of them has its own algorithms [6][7]. In next subsections, we will describe some
.[7][6] Weighted PageRank, and Page Content Rank ,important ranking algorithms which are PageRank, HITS

algorithm is used by Google to rank the pages. The PageRank is a numerical calculation based on web [6] **PageRank**
specific page. A hyperlink to graph i.e. the pages as nodes and connections as edges. Rank value shows a significance of a
a page considers a vote of support. The PageRank of a page relies on upon the quantity of connections it has. A page that
ns to a web page is connected to by many pages with high PageRank. gets a highrank itself. if that there are no connectio
.then there is no support for that page

also known as hubs and authorities) interduced by Jon Kleinberg) [7] (**HITS**) Induced Topic Search-Introduced Hyperlink
ates Web pages [6]. The centers are served is a popular algorithm on ranking process.It is a link analysis algorithms that r
as enormous directories that were not really trustworthy in the data that it held, however were utilized as aggregations of
er words, a legitimate page is a page a expansive list of data that drove users specifically to other legitimate pages. In oth
that indicated to numerous different pages, and also a page that was connected by a wide range of pages. This plan
its :alocates two scores for each page e page, and its node authority, which evaluates the estimation of the substance of th
.value, which evaluates the estimation of its connections to different pages

algorithm does not partition the rank estimation of a page equitably among its leaving connected **Whighted PageRank** values to more imperative pages[6][7]. Each outgoing connection gets a value pages, rather it allocates bigger rank corresponding to its prominence or significance. The prevalence of a page is measured by its number of incoming and .outcoming connections

heuristics that appear to be essential for resolving the substance of website strategy joins various **Page Content Rank** pages[6][7]. Here, page significance is decided on the foundation of the significance of terms contained in the page; while to a given inquiry q. PCR utilizes a neural system as its internal the significance of a term is determined with regard .characterization structure

III. COMPARISON DISCUSSION EXPERIMENTAL RESULTS AND

A. *dataset Selection of search engines and query*

n several categories and they have different goals of using Users of the internet based on study made by[8] are diveded i the inernet. The precentage of number of users for internet usage are 72% for educational purposes, 44% for repose time, like for News, blogging, Products, Lifestyle, for sports purpose, and only 5% user access net for some other purposes %15 .Current Affairs and General awareness

The search engines chosen for evaluation are Google,Yahoo and Bing. Google is chosen since it is the biggest freely it has 80.65% of usage statistics [8][9].Bing and Yahoo are on ,accessible web index and has the most astounding use .[9][8] second and third location in terms of usage statistics, 8.76%, 7.77% respectively

ased on depth search Selecting of the search queries used in evaluation process on this paper is not random process. It is b in this field conclude that, there is three types of queries used on search engines navigational, informational, and transactional queries [10]. Navigational where user try to navigate to a specific website like YouTube. Informational where the user goes for acquiring a few documents on a subject he/she is interested in. Transactional where user try to find a website for some actions e.g. downloading, playing, or so on. Navigational and transactional queries usually found t page because it seeks actually for a specific website so the other results often nonrelevance. But in one relevan informational queries are ambiguous sometimes for a search engine. Thus, the focus of this paper will be on informational .results s with scientific phrases and study the rate of relevancequires, it mixes daily expression

According to google trend the most searched words list on 2016 shows in [11] , we pick randomaly a three options from Ther is also a list on [12][13] have . **(hone 7Pokemon Go , Donald Trump,Ip)** this list to do our experiment. These are **cheap airline)** world of 2016, and similarly we pick a randomly the following terms the the most searched terms on n as a dataset as we explore The top 19 retrieved results by each search engine are saved the . **(tickets, search engine list** . early

B. *description Evaluation*

During this research evaluating of various search engines by test some of most searched terms on the WWW of 2016 is pages saved on a database. This dataset is handled. Each query has searched on all selected search engines and the top 19 .parsed and examined to classified as relevance or nonrelevance page. The results of this process is shown in next sections

ave tested each term on every have three search engines and five searched terms to do the experments. We h we ,Actually have 285 samples to check the we results were taken and saved on a data set. So in practice 19 search engines and the top nonrelevance. Precision and relevance and make the evaluation process. Each sample either have 1 for relevance or 0 for .engines recall evaluation tool is used to calculate the relevance percentage for the selected search

C. *Discussion Evaluation*

esults to do our have 285 r we Table 1 below describes the results collection process. As we said on the previous section experiments on. Each search engine has 95 pages to test, the table shows clearly the following: Google provides 49 relevant, Yahoo 54, and Bing 48 relevant pages. The relevance judgment made manually on each webpage to decide is it should use some we ce page or nonrelevance page. To evaluate these search engines based on these resultsrelevan .used evaluation tools. Most popular tool in search engines systems is precision and recall, and it is what we have

Table Results Acquirement :1

arch Se engines used	Language used in search	Type of query ,Informational transactional, ,navigational	Query terms	No. of samples to test	Relevant pages
Google	English	Informational	Pokémon Go, Donald Trump, IPhone cheap ,7 airline tickets, search engine list	95	49
Yahoo	English	Informational		95	54
Bing	English	Informational		95 **Total = 285**	48

measures **precision** measures how well the search engine is doing at finding all the relevant pages for a query, and **Recall** :[follow[3 relevant pages [3]. Recall and precision are calculated as-how well it is doing at dismissing non

$$Recall = \frac{|A \cap B|}{|A|} \quad (1)$$

$$Precision = \frac{|A \cap B|}{|B|} \quad (2)$$

A is for relevant set of a page not retrieved, B is set of retrieved pages (19 in our study), A intersection B means pages know number of relevant documents (we have millions of we and relevant. One of challenges is how can that retrieved results on a search engine query) that is not retrieved or not include on the rank position we used (19 pages considered as is to assume this number, let's consider 30 as indicated documents that are relevant and not [3] retrieved set), one solution .study values for each query on all search engines used on the **R** and recall **P** retrieved. Table 2 shows the precision

Table 2: Precision and Recall Values

Query terms	Google	Yahoo	Bing
Pokémon Go	P=0.32 R=0.17	P=0.47 R=0.23	P=0.47 R=0.23
Donald Trump	P=0.26 R=0.14	P=0.21 R=0.11	P=0.16 R=0.09
Iphone7	P=0.63 R=0.29	P=0.53 R=0.25	P=0.74 R=0.32
cheap airline tickets	P=0.68 R=0.30	P=0.79 R=0.33	P=0.58 R=0.27
search engine list	P=0.68 R=0.30	P=0.84 R=0.35	P=0.58 R=0.27
Average	AP=0.51 AR=0.24	AP=0.57 AR=0.25	AP=0.51 AR=0.24

see from above tables the results is a bit converged from each other and Yahoo got high points and thus it gives we As results than Google and Bing. In details using precision and recall terms, table 2 shows that Google high accuracy achieved high precision which means it gives more relevant retrieved pages for a specific query. The precision measure is most t measure the ratio of relevant pages from the retrieved set . For example, in "Search important for user because i engines List" query it gives 0.84 precision that's indicate 16 out of 19 pages are relevant. The precision measure is most .o of relevant pages from the retrieved setimportant for user because it measure the rati

IV. CONCLUSION

We have taken an overview of the structure of search engines as information retrieval system from distributed data bases. engines, and ranking algorithms are all discussed since they First of all, distributed databases, information retrieval, search represent the key aspects and basic structure of such systems. Then evaluation part take place we have made a query test ot chosen randomly but after exhaustive on the most popular search engines Google, Yahoo, and Bing. Queries have n search explained in detail in evaluation section. The experiment using precision and recall measures summarize that user's expectation Yahoo achieves good searches among popular searched terms, and provides most relevant results to the .than Google and Bing

REFERENCES

[1] A. Borodin, G. Roberts, J. Rosenthal and P. Tsaparas, "Link analysis ranking: algorithms, theory, and experiments", .2005 ,297-ACM Transactions on Internet Technology, vol. 5, no. 1, pp. 231

[2] R. Jindal, "An overview of ranking algorithms for search engines", Proceedings of the 2nd national A. Gupta and .2008 ,2008-conference; INDIACom

[3] W. Croft, D. Metzler and T. Strohman, Information retrieval in practice, 1st ed. Upper Saddle River, N.J.: Pearson .2009 ,nEducatio

[4] A. Kehinde and E. Daniel, "Ranking of Relevant Context Information Based on Content and User Preferences Via .2013 ,DROPT Technique", GESJ: Computer Science and Telecommunications

[5] Web Page Ranking Algorithms", International M. PaulSelvan, A. Chandra Sekar and A. Priya Dharshini, "Survey on .2012 ,7-Journal of Computer Applications, vol. 41, no. 19, pp. 1

[6] N. Duhan, A. K. Sharma and K. K. Bhatia, "Page Ranking Algorithms: A Survey," 2009 IEEE International .1537-1530 .pp ,Advance Computing Conference, Patiala, 2009

[7] doi: 10.1109/IADCC.2009.480924

[8] Goel and S. Yadav, "Search engine evaluation based on page level keywords," 2013 3rd IEEE International Advance IAdCC.2013.6514341/10.1109 :doi .876-Computing Conference (IACC), Ghaziabad, 2013, pp. 870

[9] ,Usage News", Statowl.com Browser Statistics, OS Market Share, and Technology" :Online].Available].2017 .[2017 -May-23 :Accessed] .http://www.statowl.com/search_engine_market_share.php

[10] search engines using a representative query sample", web D. Lewandowski, "Evaluating the retrieval effectiveness of .2015 ,1775-Information Science and Technology, vol. 66, no. 9, pp. 1763 for Journal of the Association

[11] /https://trends.google.com/trends :ds.google.com, 2017. [Online]. AvailableTop Charts", Tren -Google Trends " cc2c21d9e08&geo=SA&date=2016&cat=. 1-ab15-a004-f211-topcharts#vm=trendingchart&cid=770e4d53 .[2017 -May -Accessed: 23]

[12] /https://trends.google.com/trends/yis/2016 :lableGoogle's Year in Search", Google Trends, 2017. [Online]. Avai" .[2017 -May -GLOBAL. [Accessed: 23

[13] All thing Search, Content & Social", -Most popular keywords on search engines | PageTraffic Blog " -search-on-keywords-popular-http://www.pagetraffic.com/blog/most :Pagetraffic.com, 2017. [Online]. Available .[2017 -May -Accessed: 23] .engines