I.J. Information Technology and Computer Science, **2016, 12, 47-58**
Published Online December 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2016.12.06

Journey of Web Search Engines: Milestones, Challenges & Innovations

Mamta Kathuria
YMCA University of Science & Technology, Faridabad, 121001, India
E-mail: mamtakathuria7@rediffmail.com

C. K. Nagpal and Neelam Duhan
YMCA University of Science & Technology, Faridabad 121001, India
E-mail: nagpalckumar@rediffmail.com, neelam.duhan@gmail.com

Abstract—Past few decades have witnessed an information big bang in the form of World Wide Web leading to gigantic repository of heterogeneous data. A humble journey that started with the network connection between few computers at ARPANET project has reached to a level wherein almost all the computers and other communication devices of the world have joined together to form a huge global information network that makes available most of the information related to every possible heterogeneous domain. Not only the managing and indexing of this repository is a big concern but to provide a quick answer to the user's query is also of critical importance. Amazingly, rather miraculously, the task is being done quite efficiently by the current web search engines. This miracle has been possible due to a series of mathematical and technological innovations continuously being carried out in the area of search techniques. This paper takes an overview of search engine evolution from primitive to the present.

Index Terms—World Wide Web, Search Engines, Web Search, Information Retrieval.

I. Introduction

In today's life, it has become hard to think of life without internet. It is amazing to imagine that this integral part of our current daily life was almost non-existent half a century ago and was an expensive academic luxury few decades back. An innovation which started in 1960s, with a view to connect immobile bulking computers of that time in order to avoid the postage and travel delay of storage devices, underwent tremendous scalability and started undertaking almost every communicating device into its fold. The flexible scalable network created by the various heterogeneous devices gave birth to an information repository that was commonly sharable worldwide leading to the coining of the term *World Wide Web* (*WWW*) in early 1990s.

For the purpose of information retrieval from *WWW*, an application known as web browser can be used which has to be provided with the unique identity of the resource in possession of the information known as its Uniform Resource Locator (URL). The tremendous growth of the *WWW* led to huge number of information resources with each one having its own URL(s) resulted in enormous number of websites beyond the grasp of any individual. This led to the requirement of manual directories/automated mechanisms to provide the list of desired URLs in possession of the requisite information. The crossing of total number of online websites one billion mark in September 2014 [1] combined with continuous growth has rendered it meaningless to solely manage the system through manual directories and therefore making the automated system an essentiality, though the combination of both is still going on. The "Fig. 1" shows the rise in number of websites year wise.

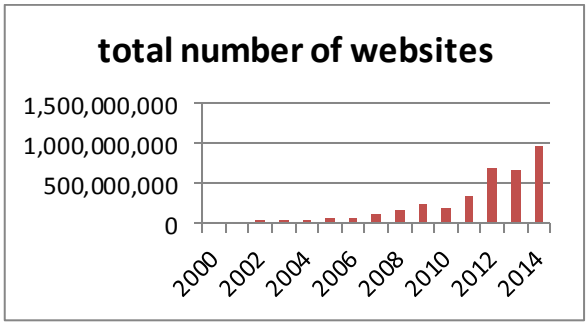

Fig.1. Proliferation in the number of web sites

The exploration for the automated mechanisms to find the desired URLs led to the creation of one of the most complex and complicated type of the software in the world known as Search Engine. Search Engines help their users in gathering and analyzing large amount of information available on various resources on the internet by presenting it in categorized, indexed and logical way. The use of the search engine is second most common activity amongst the internet users next to sending/receiving of emails [2] as depicted in "Fig. 2".

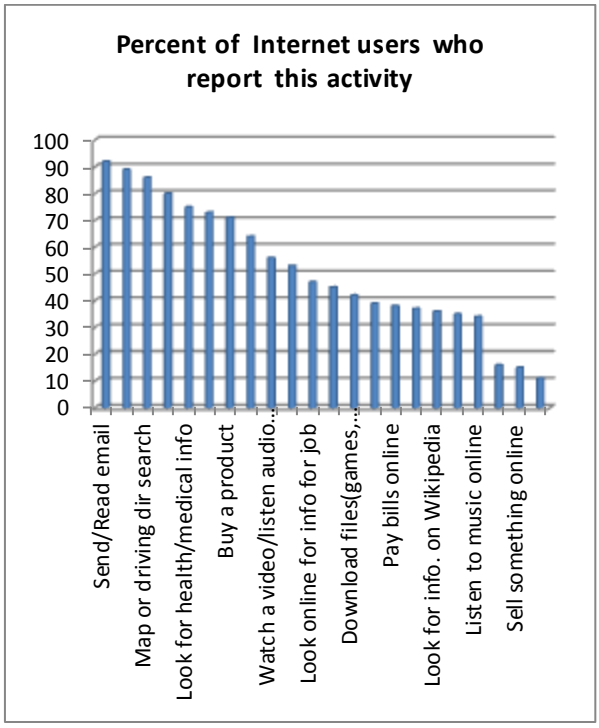

Fig.2. Internet activities of different users

With the use of mathematical, statistical and technological innovations to exploit the enormous growth of *WWW*, the search engines have been able to provide their users the requisite information in all heterogeneous domains and have proven to be indispensable information provider. Let us take a look at the rapid evolutionary process which the search engine technology has undergone with the time.

The paper contains 5 sections. Section 2 contains basic terminologies associated with the search engine technology including various types of search engines, basic architecture and search methodologies. Section 3 contains a list of few prominent search engines evolved in the journey with their salient features. Section 4 talks about the current challenges faced by search engine industry and associated innovations. Section 5 includes the persistent issues which will continue to exist in the domain of web search due to its inherent structure and operations.

II. Search Engine Basics

This section describes the various types of search engines along with their architecture and search methodologies.

A. Crawler Based Search Engine

We start our journey with the general architecture of a typical crawler based search engine as shown in "Fig. 3".

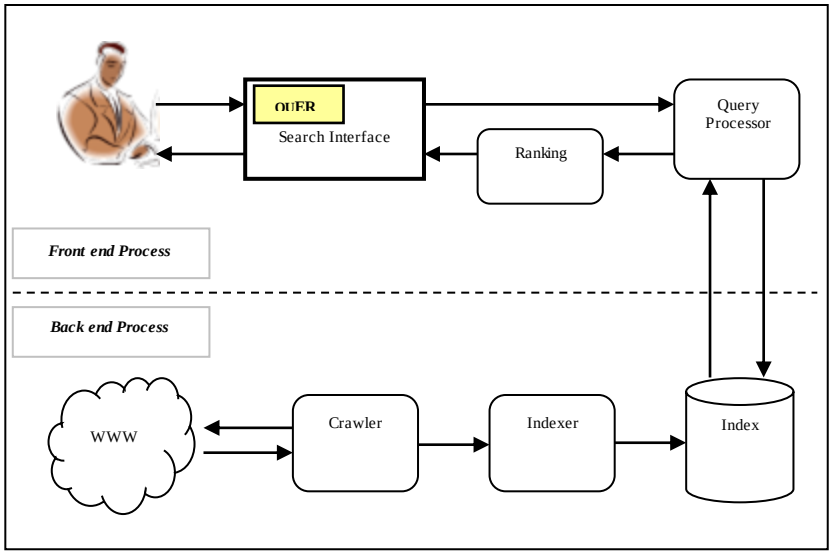

Fig.3. General Architecture of a Web Search Engine

The complete process of searching is divided into two phases:

- The back-end phase
- The front-end phase

At the *front-end*, when user submits his query in the form of keywords on the interface of the search engine, the *query processor/engine* performs its execution by matching the query keywords with the document information present in the index. A page is considered as a hit if it possesses at least one of the query keywords. The matched URLs are retrieved from the index and given to the *ranking module* so as to return a ranked list to the user.

At the *back-end*, *Crawler* is the most important component of search engine that traverses the hypertext structure of the *WWW*, downloads the web pages and parses them. The parsed pages are then routed to an *indexing module* that builds the index on the basis of

different terms present in the pages. The index is used to keep track of the Web pages fetched by the web crawler. Some of the most prevalent crawler type search engines include Google, Yahoo, Bing, Ask and AOL.

When one has a specific query in the mind then the crawler-based search engines are quite efficient in finding relevant information. However in case of generic query a crawler-based search engines may return large number of irrelevant responses.

B. Human-powered directories

Another type of search engine includes human powered directories. These search engines classify the web-pages on the basis of brief human description which can be provided by the webmasters or by the editorial group of the directory. The search engines in this category are Yahoo directory, Open Directory and LookSmart [3].

Human-powered directories are good at the searches made on the general topics where they can guide and help the searcher in narrowing down his/her search and get refined results [4]. However in case of specific search they are unable to provide an efficient way to find information.

C. Hybrid Search Engine

A hybrid search engine (HSE) uses different types of data with or without ontologies to yield algorithmically generated outcomes based on web crawling. Previous types of search engines used only text to generate their results while hybrid search engines use a combination of both crawler-based results and human-powered directories[5]. Most of the search engines these days are moving towards a hybrid-based model. The search engines in this category include Google, Yahoo and MSN Search.

D. Meta-search Engines

A meta search engine uses the services of other search engines and forwards the user's query simultaneously to several search engines working in the back. The results supplied by these search engines are then integrated and after the application of features like clustering and removal of replicates, the results are presented to the user. The search engines in this category include Dogpile[30, 31], Mamma[6] and Metacrawler[19]. Meta-search engines are good for saving time by searching only in one place and sparing the user from the need to use several separate search engines. Fig. 4 shows the architecture of a meta search engine.

E. Vertical Search Engine

Vertical search engines focus on a particular domain of search. They are also referred to as specialty or topical search engines. The common verticals of search include travel, online shopping, legal information, medical information etc. The crawler of these search engines focuses on the web pages of the particular domain and is referred to as focused crawler.

III. The Milestones in the Journey

After a brief discussion on the various types of the currently prevalent search engines, let us have a look at the journey travelled by the search engine technology over the period and talk about various milestones crossed. The journey has been presented through Table 1 which contains most of the prominent search engines evolved in the journey along with their year of development, name of developing team members/ organization, features and innovations, current activation status and Alexa rank[83,84].

IV. Challenges & Innovations

With the time, the search engines have evolved and facing novel and un-envisaged challenges. These challenges are being handled through innovations. This section takes a look at the challenges and the corresponding innovations.

A. Standardization

To markup different types of information on web-pages multiple standards and schemas are prevalent making it difficult for webmaster to choose one. A common schema supported by major search engines was required to resolve this problem. Schema.org [53] is a collaborative effort by the Bing, Yahoo, Google and Yandex to assist search engines to achieve faster and relevant search using a structured data markup schema that helps in recognizing people, events, and attributes on web resources. The on-page markups help search engines to understand the information on web pages and provide richer search results.

Schema.org is not a standard body like W3C [9, 10] or IETF [54] but is a website providing the schema and markup supported by major search engines. The common markup and schema is mutually beneficial for all the stakeholders i.e. Webmasters, Search Engines and Users.

B. Beyond Keywords

The conventional mechanism of the web search by the search engine is based upon the keywords typed by the user/ searcher [55]. With the time, the efforts are being made to extend the keyword based web search to the semantic search wherein a search engine is expected to understand the natural language using machine intelligence and identify the underlying intent of the searcher. The underlying concept of semantic search is based upon the semantic similarity being taken over documents [56], words [57, 58], terms [59], sentence [60] and entities [61]. The available search engines in this category include Powerset[62], Hakia [63] and Google hummingbirds[51].

To implement natural language search, Powerset uses natural language technology platform developed by Palo Alto Research Centre (PARC) that can encode synonyms and identify relationships between the entities. Hakia uses its own feature called QDEX that is inclined towards analyzing of the web pages rather than indexing. For

short queries it displays relevant categories and for long queries it displays relevant sentences and phrases. Google Hummingbird takes into account the entire sentence (instead of individual keywords) for understanding the underlying intent of the user.

Table 1. Milestones in the Journey

Sr.No	Year/Search Engine	Key Developer / Developed at or Owner	Features and Innovations	Current Active status/ Alexa Rank
1.	1990 Archie[7]	Alan Emtage, Peter J. Deutsch, Bill McGill University, Montreal	1. FTP Server based sharing of files 2. crawling concept 3. Script-based data gatherer 4. Regular Expression based matching retrieval of files for user query	Not Active Alexa N.A
2.	1992 Veronica & Jughead[8]	Fred Barrie, Rhett Jones University of Naveda System Computing Services group	1. Menu Driven approach 2. Ability to search plain text files 3. Keyword based search in 4. 4. Its own designed Gopher Index System	Not Active Alexa N.A
3.	1993 W3 Catalog[9,10]	Oscar Nierstrasz University of Geneva	1. Purely textual browser 2. Integration of manually maintained catalogue. 3. Dynamic querying	Not Active Alexa N.A
4.	1993 JumpStation[11]	Jonathon Fletcher University of Stirling	1. Combines crawling, searching and indexing 2. Lays the foundation for current form of search engines 3. Unable to grow because of linear search drawback	Not Active Alexa N.A
5.	1993 WWW Wanderer[12]	Matthew Gray Massachusetts Institute of Technology	1. Introduces web robots to crawl the web 2. Track the web's growth, Indexed titles and URLs 3. Did not facilitate web search, major goal to measure web size 4. Perl based web crawler	Not Active (Redirected to Yahoo)
6.	1993 Aliweb[13]	Martijn Koster United Kingdom	1. Devoid of crawling mechanism 2. Website administrator had to register with Aliweb to get their services listed & indexed 3. Capability to perform Archie Like Indexing for the web	Active(www.aliweb.com)
7.	1994 Web Crawler[14]		1. Lays the foundation for Content Based Search 2. Use of Boolean operators in user query 3. User Friendly Interface	Active, Aggregator, (https://www.webcrawler.com/) 674
8.	1994 Meta Crawler[15]	Erik Selberg, Oren Etzioni Blucora Inc.	1. Introduced the concept of meta search wherein search results of major search engines are combined to widen the search results. 2. Does't have its own search index	Active, Aggregator, (http://www.metacrawler.co.uk/
9.	MywebSearch[16]	IAC	1. Search tool compatible with Internet Explorer (4.x or above) and Netscape 4.x. 2. It is a spyware and search toolbar program 3. Displays algorithmic search results from Google, Ask.com, Yahoo and LookSmart, along with sponsored listings, primarily from Google. 4. Easy to add/remove additional software products to the Toolbar. 5. Free to use	Active but powered by google(http://home.mywebsearch.com/index.jhtml) 405
10.	1994 Lycos[17]	Mauldin Micheal L. Canegie Mellon Univ., Pittsburg	1. Prefix matching and word Proximity 2. Keyword, search on image or sound files 3. Focuses more on directory	(http://www.lycos.com/Search/) 9041
11.	1994 Inktomi[18]	Eric Brewer University of California	1. First major search engine to launch a paid inclusion service 2. Handles thousands of search queries by distributing among many servers	Not Active, Aquired by Yahoo
12.	1994 Infoseek[19]	Steve Kirsch Infoseek Corporation	1. Provided subject oriented search 2. Allowed real-time submission of the page	Not Active Alexa N.A
13.	1995	Joe Kraus,	1. Both concept & keyword based search	Active, Now

	Excite[20]	Graha spencer Garage in Silicon velley	2. Large & up-to-date index 3. Excellent summaries 4. Fast, flexible, reliable searching 5. Idea of statistical analysis of word relationship for efficient search	an internet Portal(http://www.excite.com/) 7951
14.	1995 AltaVista[21]	Louis Monier, Michael Burrows Digital Equipment Corporation's	1. Fast Multithreaded crawler & Back-end search 2. Keyword based simple or advanced search 3. Multilingual search capabilities 4. Periodic Re-indexing of sites 5. High bandwidth 6. Allow natural language query 7. Inbound link checking	Not Active, Shutdown in 2013, redirected to Yahoo
15.	1995 Yahoo[22,23]	David Filo, Jerry Yang Yahoo Corporation	1. Keyword based search 2. Web directory organized in hierarchy 3. Separate searches for images, news stories, video, maps, shopping 4. Supports full Boolean searching 5. Support Wild Card Word in Phrase	2nd largest Active SE (https://in.yahoo.com/) 4
16.	1995 AOL[24]	Bill von Meister Control Video Corporation	1. Started as Internet 2. Messenger Service 3. Subscriber based service 4. Movie & Game portal	Not Active (http://www.aol.in/)
17.	1995 MSN[25]	Microsoft Microsoft ltd.	1. Large and unique database 2. Boolean searching 3. Cached copies of Web pages including date cached 4. Automatic local search options. 5. Neural n/wadded features	Active as Bing (http://www.msn.com/en-in/)
18.	1996 DogPile[26,27]	Aaron Flin Blucora Inc.	1. Meta Search engine 2. Has its own search Index 3. Searched multiple engines, filtered for duplicates and then presented the results to the user 4. Special provisions for Stock quotes, weather forecast, yellow pages 5. etc.	Active, Aggregator (http://www.dogpile.com/) 3084
19.	1996 InfoSpace[28]	Naveen Jain Infospace Inc.	1. Meta Search Engine 2. Selects results from the leading search engines and then aggregates, filters and prioritizes the results to provide more comprehensive results 3. Instant messenger service	Active (http://infospace.com/) 2110
20.	1996 Hotbot[29,30]	Wired Magazine Inktomi Corporation	1. Extensive use of cookie technology to store personal search preference information 2. Ability to search within search results 3. Frequent updation of Database Use of parallel processing	Active(http://www.hotbot.com/) 100902
21.	1996 WOW[31]	Jeniffer Thompson Compu Serve	1. First internet service to be offered with a monthly "unlimited" rate 2. Brightly colored 3. Seemingly hand-drawn pages. 4. Find all of the breaking news articles, top videos and trending topics that matter to you. 5. Effective advertising 6. Highly communicative design 7. Budget friendly media services 8. Creative concept development	Active (http://www.wow.com/) 767
22.	1996 Ask[32,33]	David Warthen, Garrett Gruener IAC/ InterActive Corporation	1. Natural language-based Search 2. Both concept & keyword based search 3. Allows to enter query in the form of sentence for humanize the online experience 4. Question answering system	Active (http://www.ask.com/) 28
23.	1997 Daum[34]	Daumkakao Daum Corparation	1. A popular search engine in Korea 2. Besides internet search provides facilities for E-mail, Chat, Shopping etc.	Active (www.daum.net/) 140
24.	1997 Overture[35]	Bill Gross Yahoo	1. Paid search inspired from commercial telephone directory 2. Secured, pay-per-placement directory service	Not Active Alexa N.A

25.	1997 Yandex[36]	Taylor Nelson Sofres San Francisco Bay Area	1. Full-text search with Russian morphology support 2. Encrypted search 3. Multilingual	Active (https://www.yandex.com/) 20
26.	1998 Google[37,38]	Sergey Brin, Lawrence Page Stanford University, Stanford	1. Keyword based search 2. Page Rank algorithm 3. Semantic search 4. Free, Fast and easy to search 5. No programming or database skills required	Active as most popular SE (https://www.google.co.in/) 1
27.	1999 AlltheWeb[39]	Tor Egge Norwegian Univ. of Sci. & Tech.	1. Faster Database 2. Advanced search features 3. Sleek interface 4. FAST's enterprise search engine 5. search clustering 6. completely customizable look	Not Active (URL redirected to Yahoo)
28.	2000 Teoma[40]	Apostolos Gerasoulis Rutgers Univ. computer lab	1. Provide knowledge search 2. Provide subject specific popularity 3. Clustering Techniques to 4. Determine Site Popularity 5. Unique Link popularity	Not Active , Redirected to Ask.com
29.	2000 Baidu[41]	Robin Li Beijing China	1. largest internet user population 2. pay per click marketing platform 3. China's Google	Active (http://www.baidu.com/) 5
30.	2007 LiveSearch[42]	Satya Nadella Microsoft	1. Uses a drag-and-drop interface that's really simple to pick up 2. The new search engine used search tabs that include Web, news, images, music and desktop	Active as Bing, Launched as rebranded MSN search (https://www.live.com/)
31.	2008 DuckDuckGo[43]	Gebriel Weinberg DuckDuckGo Inc.	1. Offers real privacy or protecting searchers' privacy and avoiding the filter bubble of personalized search results 2. Smarter search, and stories that user likes 3. Not profiling its users and by deliberately showing all users the same search results for a given search term 4. Emphasizes on getting information from the best sources rather than the most sources	Active (https://duckduckgo.com/) 506
32.	2008 Aardvark[44]	Max Ventilla, Nathan Stoll The Mechanical Zoo, A San Francisco based startup	1. Use Social n/w facilitated a live chat or email conversation with one or more topic experts 2. Social search Engine 3. Aadvark Ranking Algorithm	Not Active Alexa N.A
33.	2009 Bing[45]	Steve Billmer Microsoft	1. Keyword based search 2. Index updated on weakly or daily basis 3. Advertised as a decision engine 4. Social integrations are stronger 5. Direct information in the area of finance & sports	Active (https://www.bing.com/) 24
34.	2009 Caffeine[46]	Matt Cutts Google	1. New web indexing system 2. Near-real-time integration of indexing and ranking 3. Allows easier annotation of the information stored with documents 4. Provide 50% fresher result 5. Find links to Relevant content much sooner 6. Update search index on a continuous basis, globally. 7. Caffeine processes hundreds of thousands of pages in parallel. 8. Nearly 100 million gigabytes of storage in one database	Active (http://googleblog.blogspot.in/2010/06/our-new-search-index-caffeine.html)
35.	2010 Google Instant[47]	Marissa Mayer & Matt Cutts Google	1. Search-before-you-type 2. Predicts the users whole query 3. Faster Searches, Smarter Prediction, Instant Result 4. User Experience 5. Provide Autocomplete Suggestion	Active

36.	2010 Blekko[48]	Rich Skrenta Blekko Inc.	1. Uses slash tags to allow people to search in more targeted categories 2. Spam Reduction 3. Provides better search results than those offered by Google Search, by offering results culled from a set of billion trusted websites and excluding material from such sites as content farms. 4. Dynamic interface graph algorithm 5. Blekko offers a web search engine and social news platform that provides users with curated links for the entered search criteria. 6. Provided downloadable search bar which was later acquired by IBM	Active, Aquired by IBM(www.blekko.com) 4518
37.	2013 Contenko[49]	Tomas Meskauskas Amerow LLC	1. Deceptive Internet Search, promoted using various browsers hijackers 2. Provides Innovative means for browsing the internet 3. Its Startup page doesn't contain any links to privacy terms or terms of use	Active (http://www.contenko.com/) 4505
38.	2013 Alhea[50]	Manuel Barrios Amazon Technologies Inc.	1. Offers a single source to search the Web, images, audio, video, news from Google, Yahoo!, and many more search engines. 2. Alhea.com compiles results from many of the Web's major search properties, delivering	Active (http://www.alhea.com/) 11225
39.	2011 GooglePanda[51]	Navneet Panda and Vladimir Ofitserov Google	1. Focuses on eliminating sites that didn't have enough quality content and were more geared at moneymaking than providing useful content. 2. Provide new Google's search results ranking algorithm 3. Quality Search results	Active (http://www.google-panda.com/)
40.	2012 GooglePenguin[50]	Matt Cutts Google	1. Web spam update 2. goal of concentrating on webspam 3. Search Algorithm update 4. Protect your site from bad links .	Active Alexa N.A
41.	2013 Google HummingBird[51]	GianlucaFiore Lli Google	1. A core algorithm update may enable more semantic search and more effective use of the Knowledge Graph in the future, Hummingbird is about synonyms but also about context Google 2. Hummingbird is designed to apply the meaning technology to billions of pages from across the web, in addition to 3. Knowledge Graph facts, which may bring back better results 4. Search Algorithm update 5. Understand the intent of the user	Active Alexa N.A
42.	2015 SciNet[52]	TuukkaRuotsalo, KumaripabaAthukorala, DorotaGlowacka, KseniaKonyushkova, Antti Oulasvirta, SamuliKaipiainen, Samuel Kaski, Giulio Jacucci Helsinki Institute for Information Technology HIIT, Finland	1. Reinforcement Learning 2. Auto-suggestion for specific topic & document 3. Interactive approach 4. A new search engine that outperforms current ones and helps people search more efficiently. 5. SciNet displays a range of keywords and topics in a topic radar	Active 2159988

This type of the search is being referred to as the conversational search by the Google [37,38] and is intended to take into the account both context and intent of the search.

One of the major difference between the keyword based search and the semantic search is that the semantic search takes into account the connecting words like in, by, for, about etc. as they are vital to the meaning of the sentence (semantic impact) while these words are simply

discarded in the keyword based search.

C. Knowledge graph and entity based search

The basic strategy of keywords based search, as used by the conventional search engines, has a major drawback that it is unable to get real sense many times as it does not explore the underlying real world connections, properties and relationships [64].

The new type of search is referred to as entity based search and in this regard a major work has been done by Microsoft's Satori [65] and Google's Knowledge Graph[66]. To accomplish the entity based search in the future, the data/ unstructured information is being extracted from the web-pages and a structured database of nouns (people, places, objects etc.) is being created that includes the relationship as well. The newly defined structure is referred to as web of concepts [67]. The transformation from unstructured web to web of concepts includes three processes namely information extraction, linking (mapping the relationship) and analysis (categorizing information about an entity)[ ] . The knowledge graph is similar to Facebook's Open Graph and derived from Freebase [68].

D. Avoiding memory recall (Option Based Search)

A novel strategy was adopted by the Scinet[52] search engine to cater to the personal needs of the user wherein the search process has been converted from memory recall process(thinking of keywords) to the recognition process(to make a selection from the given choices). Depending upon the user's past behavior, the search engine exhibits the potential topics/keywords along with the intent radar indicating the potential direction where the search will lead to.

E. Social and continuous search

A novel initiative has been taken by a new search engine called *Yotify*[69] that does not reply user's query instantly but keeps on searching the websites to find appropriate answers and send them by e-mail. For example, if somebody is looking for a house in a desired set of localities or a particular type of furniture items then the *Yotify* keeps on searching the associated websites. In contrast to Google and yahoo alerts which focus on the news and other information, Yotify is more concerned with Shopping. At present, the problem with *Yotify* is that it can scan only a small portion of web and lacks the width like Google and yahoo.

F. Deep Web search

The Web search engines are just web spiders which index webpages by following the hyperlinks one after the other. However, there are some places where a spider/crawler cannot enter e.g. the database of a library, webpages belonging to private networks of organizations etc which may normally require a password for access. Such part of web, which remains un-indexed, is referred to as *Deep Net, Deep Web, Invisible Web or hidden web*. *Despite the remarkable progress in search technology the size of the deep web is much larger (nearly 500 times) [70] than the indexed web*. The basic reasons for the non-indexing are following:

- Dynamic pages which are accessed only through filling of forms whose contents are related to domain knowledge.
- Web pages that are not linked to other pages i.e. the pages which are not having any inlinks / backlinks. Such a situation makes the webpage contents inaccessible.
- Websites requiring registration and login.
- Webpages whose content vary as per access rights and contexts.
- Websites prohibiting search engines from browsing them using by using Robot Exclusion Standards such as CAPTCHA code.
- Textual content encoded in multimedia files or other such file formats which are not conventionally readable by search engines.
- Web contents intentionally kept invisible to the standard internet. Such contents are accessible only through darknet softwares like Tor [71], I2P [72].
- Archived versions of web pages and web-site which have become time irrelevant and are not indexed by search engines.

However, with the time various search engines have come in the market which make available a certain segment of deep web resources. Some of these search engines are Infomine[73] created by a group of libraries in USA, Intute[74] created by group of universities in UK, Complete-Planet [75] containing providing access to nearly 70000 databases over heterogeneous domains, Infoplease [76] providing access to encyclopedias; atlas and other such resources, DeepPeep [77], IncyWincy [78], Scirus [79], TechXtra[80] etc..

The deep web search engines mentioned in this subsection have been created with a positive intent to provide a controlled access to databases, clubbed through authorization, to their legitimate and authorized users which need them for academic or commercial purpose.

G. Onion Search

The onion search [81] is a type of deep web search, but with a negative intent. The onion search provides a kind of opacity wherein both the persons i.e. the information provider and the one accessing the information are difficult to trace not only by others but even by each another. The onion is a pseudo Top Level Domain (TLD) host reachable via Tor network [71]. The TLD in case of .onion sites is not an actual DNS root but is an access mechanism provided through a proxy server. The addresses in the onion domain are automatically generated based upon the public key when the hidden service is created/ configured.

- The onion search is being used for undesirable purpose such as drugs and arms dealing. One such search engine is Onion.city [82].

- The onion search is also being used by investigating agencies and defense organization to penetrate into the deep web. One such search engine is Memexa by *Defense Advanced Research Projects Agency (DARPA) project* to find things on the deep web which are not indexed by major search engines [82].

H. Entity Search

Now the search has changed its way from findings "strings"(i.e., strings that is a combination of letters in a search query) to findings "things"(i.e., entities). The move from "strings" to "things" helped in data base searches where bits of data are placed on a knowledge graph to answer the who, what, when, where and how type of questions. Entity Search gives a new insight into search optimization because now google can provides direct answers to many queries within the search results. This effort increases the search results relevancy by identifying what a query term means and helps to understand the correlation between the strings of characters and real-life context. Google's entity search aims to expand the Knowledge Graph by understanding relationships through stringing together relevant data and making real-world connections between content and how users search.

I. RankBrain in Google

Table 2. Search Engine Challenges and Innovations

	Challenge	Innovation
1	Multiple standards & Schemas	Standard Schema accepted by major search engines in the form of Schema.Org
2	Search Based upon user's actual intent	Semantic Search Engines
3	To take into account the real world relationship	Entity based search
4	Relieving the user from key based thinking	Option based search
5	To keep users' query in memory and make search during a period	Social & Continuous Search
6	To explore hidden web	Authorized collaboration of data bases and their access through a deep web search engine
7	To maintain opacity between the information provider and information seeker	Onion Search

RankBrain helps in processing and refining ambiguous search query and connect them to specific topics using pattern recognition. It is a machine learning system that gives optimize results for a specific query set for executing hundreds of millions for search queries per day. It refines the query results of Google's Knowledge Graph based entity search. It uses artificial intelligence to embed massive amount of written language into mathematical entities known as vectors that is easy to understand for computer. If a word of phrase that is not familiar with RankBrain is seen, the machine can make a guess as to what words or phrases might have a similar meaning and filter the result accordingly, making it more effective to handle the queries that have not been seen earlier. After having discussed the various innovation in the web search process, let us summerise them and list the challenges intended to overcome through these innovations. Table 2 shows this summary.

V. Speed Breakers in Journey (Inherent Search Issues)

Due to its inherent huge size, diversity of users, diversity of search requirement and heterogeneity of contents, the following issues will continue to persist and search engine will have to make a compromise between various choices.

a) To simultaneously support the generic overview of topics and enabling specialist groups to drill down to their exclusively relevant items.
b) To effectively deal with invisible or deep web.
c) To offer demand anticipation, customization and personalization.
d) To go beyond the list of possible relevant web-pages and to focus on providing an exact answer.
e) To effectively deal with the web spam i.e. the web pages that exist only to mislead search engines as well as the users to certain web sites.
f) To effectively deal with noisy, low quality, unreliable and contradictory contents continuously being uploaded on the web.
g) To deal with the multiple replica of web pages.
h) To deal with the unstructured or vaguely structured contents.
i) To effectively deal with noisy, low quality, unreliable and contradictory contents continuously being uploaded on the web.
j) To deal with the multiple replica of web pages.
k) To deal with the unstructured or vaguely structured contents.

VI. Conclusion

Search engines offer their users vast and impressive amounts of information accessible with a speed and convenience few people could have imagined one/two decade ago. Yet the challenges are not over. Every advancement in search methodology/technology is leading to more and more envisaged challenges paving the way for further innovations and the cycle continues. The paper discusses the innovations that have been carried out in the past with the hope that it will encourage the researcher for further innovations.

References

[1] http://www.internetlivestats.com/total-number-of websites/
[2] http://www.infoplease.com/ipa/A0921862.html
[3] http://www.irkawebpromotions.com/webdirectories/looks

mart/
[4] http://www.yuanlei.com/studies/articles/is567-searchengine/page2.htm
[5] https://forums.digitalpoint.com/threads/hybrid-search-engines.2612207/
[6] http://websearch.about.com/od/metasearchengines/a/mamma.htm
[7] https://en.wikipedia.org/wiki/Archie_search_engine
[8] https://en.wikipedia.org/wiki/Veronica &Jughead (search_engine)
[9] http://scg.unibe.ch/archive/software/w3catalog/W3CatalogHistory.html
[10] https://en.wikipedia.org/wiki/W3Catalog
[11] https://en.wikipedia.org/wiki/JumpStation
[12] https://en.wikipedia.org/wiki/World_Wide_Web_Wanderer
[13] http://thesearchenginearchive.wikia.com/wiki/Aliweb
[14] http://www.sciencedaily.com/terms/web_crawler.htm https://en.wikipedia.org/wiki/MetaCrawler
[15] http://malwaretips.com/blogs/remove-mywebsearch/
[16] http://www.livinginternet.com/w/wu_sites_lycos.htm
[17] http://searchenginewatch.com/sew/news/2047873/inktomi-debuts-self-serve-paid-inclusion
[18] https://en.wikipedia.org/wiki/Infoseek
[19] https://en.wikipedia.org/wiki/Excite
[20] http://searchenginewatch.com/sew/study/2067828/altavista-search-by-language-feature
[21] http://www.searchengineshowdown.com/features/yahoo/review.html
[22] https://en.wikipedia.org/wiki/Yahoo!_Search
[23] https://en.wikipedia.org/wiki/AOL
[24] http://www.msn.com/en-in/
[25] https://en.wikipedia.org/wiki/Dogpile
[26] http://investor.blucora.com/releasedetail.cfm?ReleaseID=166325
[27] http://chj.tbe.taleo.net/chj04/ats/careers/requisition.jsp?org=INFOSPACE&cws=1&rid=181
[28] https://en.wikipedia.org/wiki/HotBot
[29] http://www.searchengineshowdown.com/features/hotbot/review.html
[30] https://en.wikipedia.org/wiki/Wow!_(online_service)
[31] https://en.wikipedia.org/wiki/Ask.com
[32] http://www.searchengineshowdown.com/features/ask/review.html
[33] https://en.wikipedia.org/wiki/Daum_(web_portal)
[34] http://www.search-marketing.info/search-engines/price-per-click/overture.htm
[35] https://en.wikipedia.org/wiki/Yandex_Search
[36] https://en.wikipedia.org/wiki/Google_Search#calculator
[37] http://www.telegraph.co.uk/technology/google/10346736/Google-search-15-hidden-features.html
[38] https://en.wikipedia.org/wiki/AlltheWeb
[39] http://www.seochat.com/c/a/marketing/web-directories/teoma-the-superior-search-engine/
[40] https://en.wikipedia.org/wiki/Baidu
[41] https://en.wikipedia.org/wiki/Live_search
[42] https://en.wikipedia.org/wiki/DuckDuckGo
[43] D.Horowitz, S.D. Kamvar, "The Anatomy of a Large-Scale Social Search Engine", International World Wide Web Conference Committee (IW3C2), 2010, April 26–30, 2010, Raleigh, North Carolina, USA, ACM 978-1-60558-799-8/10/04.
[44] http://www.windowscentral.com/top-bing-features
[45] http://www.telegraph.co.uk/technology/google/6009176/Google-reveals-caffeine-a-new-faster-search engine.html
[46] http://searchengineland.com/google-instant-complete-users-guide-50136
[47] https://en.wikipedia.org/wiki/Blekko
[48] https://www.aihitdata.com/company/00D2051A/CONTENKO/history
[49] https://en.wikipedia.org/wiki/Althea
[50] http://www.searchenginejournal.com/seo-guide/google-penguin-panda-hummingbird/
[51] TuukkaRuotsalo,KumaripabaAthukorala, DorotaGłowacka, KseniaKonyushkova, AnttiOulasvirta, SamuliKaipiainen, Samuel Kaski, GiulioJacucci, "Supporting Exploratory Search Tasks with Interactive User Modeling" ,Helsinki Institute for Information Technology HIIT, University of Helsinki, *ASIST 2013*, November 1-6, 2013
[52] https://schema.org/docs/faq.html
[53] https://www.ietf.org/
[54] https://www.inbenta.com/en/blog/entry/keyword- based-versus-natural-language-search
[55] R.Priyadarshini, LathaTamilselvan, "Document clustering based on keyword frequency and concept matching technique in Hadoop", International Journal of Scientific & Engineering Research, Volume 5, Issue 5, May-2014 1367 ISSN 2229-5518
[56] DanushkaBollegala, Yutaka Matsuo, and Mitsuru Ishizuka, "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words" IEEE Transactions on Knowledge and Data Engineering, vol. 23, NO. 7, July 2011
[57] D. Mclean, Y. Li, and Z.A. Bandar, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 871-882, July/Aug. 2003.
[58] Elias Iosif, Alexandros Potamianos, "Unsupervised Semantic Similarity Computation between TermsUsing Web Documents", IEEE Transactions on knowledge and data engineering, vol. 22, no. 11, november 2010
[59] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, ``Sentence similarity based on semantic nets and corpus statistics," *IEEE Trans.Knowl. Data Eng.*, vol. 18, no. 8, pp. 1138_1150, Aug. 2006.
[60] Tao Cheng, Hady W. Lauw, and SteliosPaparizos, "Entity Synonyms for Structured Web Search", IEEE Transactions on Knowledge and data engineering, vol. 24, no. 10, October 2012
[61] Tim Converse, Ronald M. Kaplan, Barney Pell, Scott Prevost, Lorenzo Thione, Chad Walters, "Powerset's Natural Language Wikipedia Search Engine",Powerset, Inc. 475 Brannan Street San Francisco, California 94107
[62] https://www.crunchbase.com/organization/hakia and Website: http://www.hakia.com
[63] http://arstechnica.com/information technology/2012/06/inside-the-architecture-of- googles-knowledge-graph-and-microsofts-satori/
[64] http://www.cnet.com/news/microsofts-bing-seeks-enlightenment-with-satori/
[65] https://en.wikipedia.org/wiki/Knowledge_Graph
[66] AdityaParameswaran, AnandRajaraman, Hector Garcia-Molina, "Towards The Web Of Concepts: Extracting Concepts fromLarge Datasets", publisher, ACM, VLDB '10, September 13-17, 2010, Singapore(http://ilpubs.stanford.edu:8090/917/1/conceptMining-Techrep.pdf)
[67] http://www.freebase.com
[68] http://www.technologyreview.com/news/410961/making-search-social/, http://www.yotify.com/
[69] https://en.wikipedia.org/wiki/Deep_web(search
[70] https://en.wikipedia.org/wiki/Tor(anonymity_network)

[71] https://en.wikipedia.org/wiki/I2P
[72] http://www.lib.vt.edu/find/databases/I/infomine- search-engine.html
[73] https://en.wikipedia.org/wiki/Intute
[74] http://websearch.about.com/od/invisibleweb/a/completepl anet.htm
[75] http://www.infoplease.com/
[76] http://content.lib.utah.edu/cdm/ref/collection/uspace/id/54 77
[77] http://www.seochat.com/c/a/search-engine-optimization-help/search-engines-for-the-invisible-web/
[78] http://searchenginewatch.com/sew/news/2065996/scirus-a-new-science-search-engine
[79] http://library.poly.edu/news/2007/10/09/techxtra-search-engine-for-engineering-mathematics-and-computing
[80] https://www.deepdotweb.com/how-to-access-onion-sites/
[81] http://thehackernews.com/2015/02/Onion-city-darknet-seach-engine.html
[82] www.alexa.com/siteinfo/
[83] http://www. ebizmba.com/articles/search-engines
[84] Deital P J and Deital H M, " Internet & World Wide Web, How to Program", Pearson International Edition, 4th edition, 2013
[85] C Jouis, I Biskri, J G Ganascia, M Roux, " Next Generation Search Engines: Advanced Models for Information Retreival", Information Science Reference,2012
[86] J.Bernard, S. Amanda," How are we searching the world wide web?: a comparison of nine search engine transaction logs" Information Processing and Management: an International Journal(Elsevier), Volume 42 Issue 1, January 2006, Pages 248-263
[87] R Aravindhan, R. Shanmugalakshmi "Comparative analysis of Web 3.0 search engines: A survey report", International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE Conference Publications, 2013, Page(s): 1 – 6
[88] Leslie S. Hiraoka ," Evolution of the Search Engine in Developed and Emerging Markets", International Journal of Information Systems and Social Change(DBLP), Vol. 5 Issue 1, January 2014, pp.30-46
[89] Capra, R.G.P. Quinones, "Using Web search engines to find and refind information" IEEE Journals & Magazines 2005, Volume: 38, Issue: 10 DOI: 10.1109/MC.2005.355, Page(s): 36 - 42
[90] YipingKe, Lin Deng, Wilfred Ng, Dik-Lun Lee, " Web dynamics and their ramifications for the development of web search engines", The International Journal of Computer and Telecommunications Networking-Web dynamics, Elsevier North-Holland, Inc. New York, NY, USA, Volume 50 Issue 10, 14 July 2006, Pages 1430 - 1447
[91] P. Metaxas, "Web Spam, Social Propaganda and the Evolution of Search Engine Rankings", SOFSEM 2007:Theory and Practice of Computer Science, Lecture Notes in Computer Science Volume 4362, 2007, pp 1-8
[92] Q. Yang, H. Wang, J. Wen, G. Zhang, Y. Lu, K. Lee, H. Zhang, "Towards a Next-Generation Search Engine", The Connected Home: The Future of Domestic Life(Science Direct) 2011, pp 79-91
[93] Monica Peshave, KamyarDezhgosha, "How Search Engines Work and a Web Crawler Application"
[94] D.Horowitz, S.D. Kamvar, "The Anatomy of a Large-Scale Social Search Engine", International World Wide Web Conference Committee (IW3C2), 2010, April 26–30, 2010, Raleigh, North Carolina, USA, ACM 978-1-60558-799-8/10/04.
[95] Stefano Ceri, Alessandro Bozzon, Marco Brambilla, Emanuele Della Valle, PieroFraternali, Silvia Quarteroni, "Search Engines", Advanced Topics in Information Retrieval, The Information Retrieval Series Volume 33, 2011, pp 27-50
[96] Ricardo BaezaYates, Alvaro Pereira Jr, NivioZiviani, "The Evolution of Web Content and Search Engines", WEBKDD'06, August 20, 2006, Philadelphia, Pennsylvania, USA. Copyright 2006 ACM 1-59593-444-8... $5.00
[97] Gray, Matthew. "Internet Growth and Statistics: Credit and Background". mRetrieved February 3, 2014.
[98] A. Ntoulas, J. Cho, C. Olston, "What's New on the Web ? The Evolution of the Web from a Search Engine Perspective", In Proceedings of the World-Wide Web Conference (WWW), May 2004.
[99] ArvindArasu, Junghoo Cho, Hector Garcia- Molina, Andreas Paepcke, SriramRaghavan, "Searching the Web", ACM Transactions on Internet Technology, 1(1): August 2001.
[100] Dirk Lewandowski, "Web searching, search engines and Information Retrieval, Information Services & Use", 25 (2005) 137-147, IOS Press, 2005.
[101] Tom Seymour, Dean Frantsvog, Satheesh Kumar, "History Of Search Engines",International Journal of Management & Information Systems – Fourth Quarter 2011 Volume 15, Number 4
[102] TuukkaRuotsalo, KumaripabaAthukorala, DorotaGłowacka, KseniaKonyushkova, AnttiOulasvirta, SamuliKaipiainen, Samuel Kaski, GiulioJacucci, "Supporting Exploratory Search Tasks with Interactive User Modeling" ,Helsinki Institute for Information Technology HIIT, University of Helsinki, *ASIST 2013*, November 1-6, 2013
[103] Marchionini, G, "Exploratory search: from finding to understanding", *Comm. ACM 49*, (2006), 41-46.
[104] Gromov, G. R.,"History of Internet and WWW: the roads and crossroads of Internet history". from http://www.netvalley.com/intvalstat.html, Retrieved December 5, 2004
[105] Holzschlag, M. E.," How specialization limited the Web", Retrieved December 4, 2004, from http://www.webtechniques.com/archives/2001/09/desi/
[106] Jansen, B. J., Spink, A. & Pedersen, J., "An analysis of multimedia searching on AltaVista", Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval, (2003) 186-192.
[107] Kherfi, M. L., Ziou, D. &Bernardi, A., "Image retrieval from the World Wide Web" issues, techniques and systems. ACM Computer Surveys, (2004),36(14), 35-67.
[108] Wall, A., "History of search engines & web history", Retrieved December 3, 2004, from http://www.search-marketing.info/search-engine-history/
[109] Jansen, B. J., Spink, A. & Pedersen, J., "An analysis of multimedia searching on AltaVista", Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval, (2003), 186-192.
[110] ArvindArasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, SriramRaghavan, "Searching the Web", (Stanford University). ACM Transactions on Internet Technology (TOIT), Volume 1, Issue 1 (August 2001).
[111] Elgesem, D, "Search Engines and the Public Use of Reason." Ethics and Information Technology, 10(4), 2008
[112] Nagenborg, M. (ed.), 2005. The Ethics of Search Engines.Special Issue of International Review of Information Ethics.Vol. 3.
[113] "Search Engines, Personal Information, and the Problem

of Protecting Privacy in Public," International Review of Information Ethics, 3: 39–45.
[114] Bruce Croft, Donald Metzler, and Trevor Strohman, "Search Engines: Information Retrieval in Practice", Addison Wesley, 2010
[115] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet Wiener, "A large-scale study of the evolution of web pages", WWW '03: Proceedings of the 12th international conference on World Wide Web, pages 669–678, 2003.

Authors' Profiles

Mamta Kathuria received her MCA degree with Honors from Kurukshetra University, Kurukshetra in 2005 and M.Tech in Computer Engineering from Maharshi Dayanand University, Rohtak in 2006 and 2008, respectively.. She is pursuing her Ph.D in Computer Engineering from YMCA University of Science and Technology, Faridabad. She is currently working as a Assistant Professor in YMCA University of Science & Technology and has eight years of experience. Her areas of interest are search engines, Web Mining and Fuzzy Logic.

Dr. Chander K. Nagpal is Ph. D (Computer Science) from Jamia Milla Islamia, New Delhi. He is currently working as a professor in YMCA University of Science & Technology and has twenty eight years of teaching experience. He has published two books. He has published many research papers in reputed international Journals such as IEEE transaction on software reliability, Wiley STVR, CSI. His academic interests include Ad hoc networks, Web Mining and Soft Computing.

Dr. Neelam Duhan received her B.Tech. in Computer Science and Engineering with Honors from Kurukshetra University, Kurukshetra and M.Tech with Honors in Computer Engineering from Maharshi Dayanand University, Rohtak in 2002 and 2005, respectively. She completed her PhD in Computer Engineering in 2011 from Maharshi Dayanand University, Rohtak. She is currently working as an Assistant Professor in Computer Engineering Department in YMCA University of Science and Technology, Faridabad and has an experience of about 12 years. She has published over 30 research papers in reputed international Journals and International Conferences. Her areas of interest are databases, search engines and web mining.

How to cite this paper: Mamta Kathuria, C. K. Nagpal, Neelam Duhan, "Journey of Web Search Engines: Milestones, Challenges & Innovations", International Journal of Information Technology and Computer Science(IJITCS), Vol.8, No.12, pp.47-58, 2016. DOI: 10.5815/ijitcs.2016.12.06

p
r