International Journal of Information Movement Vol.2 Issue III (July 2017)
Website: www.ijim.in ISSN: 2456-0553 (online) Pages 85-92

INFORMATION RETRIEVAL AND WEB SEARCH

Sapna
Department of Library and Information Science
Central University of Haryana
Email: sapnasna121@gmail

Abstract

The paper gives an overview of search techniques used for information retrieval on the web. The features of selected search engines and the search techniques available with emphasis on advanced search techniques are discussed. A historic context is provided to illustrate the evolution of search engines in the semantic web era. The methodology used for the study is review of literature related to various aspects of search engines and search techniques available. In this digital era library and information science professionals should be aware of various search tools and techniques available so that they will be able to provide relevant information to users in a timely and effective manner and satisfy the fourth law of library science i.e. *"Save the time of the user."*

Keywords: search engine, web search engine, semantic search, resource discovery, - advanced search techniques, information retrieval.

1.0 Introduction

Retrieval systems in libraries have been historically very efficient and effective as they are strongly supported by cataloging for description and classification systems for organization of information. The same has continued even in the digital era where online catalogs are maintained by library standards such as catalog codes, classification schemes, standard subject headings lists, subject thesauri, etc. However, the information resources in a given library are limited. With the rapid advancement of technology, a large amount of information is being made available on the web in various forms such as text, multimedia, and another format continuously, however, retrieving relevant results from the web search engine is quite difficult. The reasons for this are (a) abundance of information on the web and (b) lack of proper description and organization of information on the web. Due to the massive amount of information on the web, it has become very difficult and overwhelming for a user to retrieve relevant information from the web. The web is unfortunately characterized by a lot of irrelevant information on the web and knowledge of advanced search tools and techniques are become essential for information retrieval.

To overcome the issues in information retrieval, various search tools are being constantly developed. Search engines are one among such discovery tools. According to Shade et al "Generally, search engines' common goal is to make people of different background have easy access to information, which is far from their normal reach." The web-based search engine provides the platform to people from all over the world to search for different kind of information various sources of information (Shade et al, 2012). Ntoulas et al in their article said that "As the Web grows larger and more diverse, search engines are becoming the "killer app" of the Web." whenever users want information they go to search engines (Ntoulas et al, 2005). Currently, many search engines are available like Google, Yahoo, Yippy etc. No doubt they are very good for navigational and transactional searching but in a case of queries which are ambiguous they are not that much effective. However, over the years web search engines have improved a lot. Even though they may not be as efficient as a library-based catalog, nevertheless, they are improving and the slowly replacing library catalog as the first stop for the literature search due their ease of use.

2.0 Web Search Engine

Web Search Engine is a type of program designed to help find and access information stored on the World Wide Web. A search engine acts as a practical application of information retrieval techniques to large-scale information collections (Croft et al, 2015).

2.1 Types of Web Search Engine

Web search engines can be categorized into the following types:

2.1.1 Crawler-Based Search Engine: In Crawler-Based Search Engine, searching is divided into two phases: The back-end phase and front-end phase. Front-end phase relates to query submission and back-end phase relates to search engine response to a query. For a specific query, the crawler-based search engines are quite efficient in finding relevant information. However in a case of generic query a crawler-based search engine may find a large number of irrelevant responses. Crawler type search engines include Google, Yahoo, Bing, Ask and AOL. (Kathuria et al, 2016.p 48-49)

2.1.2 Specialized Search Engines: A search engine which is specialized in a particular topic usually produces a better quality of results and most relevant documents as compare to general search engine. These search engines focus on specific subject or small and specialized area (Arrigo, 2005). These search engines are the best tool for information retrieval on the web.

Examples are: Intute, SearchGov.com, AgriSurf.

2.1.3 Meta-search Engines: Meta-search engines don't have their own database of data. They search the databases of other search engines. it allows the user to search several search engines simultaneously i.e. main advantages of meta search engine. Examples of meta search engines are Metacrawler, Inference Savvysearch, Mamma, Dogpile, Search, C4, Profusion (McCoy, 2000).

2.1.4 Vertical Search Engine: It focus on a particular domain of search. According to Shettar "Vertical search engines, or domain-specific search engines also called "Vortals", facilitate more accurate, relevant and faster search by indexing in specific domains." (Shettar, 2007).

examples of a vertical search engine are listed below:

- Jobs - SimplyHired.com, Indeed.com, Eluta.Ca, Recruit.net
- Travel - Sidestep.com, Kayak.com, Mobissimo.com, Pinpointtravel.com, Farechase.com
- Health - Amniota.com, GenieKnows.com, Healia.com, Healthline.com, MammaHealth.com
- Classifieds - Edgeio.com, Oodle.com
- Blogs - Technorati, Bloglines, Blogger Search, Sphere, Feedster
- Source Code - Koders.com, Krugle, Google Code
- Academic/teen- Answers.com, Teenja.com, Gradewinner.com, Scholar.google.com. (Curran et al, 2007)

2.1.5 Hybrid Search Engine: Hybrid search engines is a combination of both crawler-based results and human-powered directories. It produces algorithmically generated outcomes based on web crawling. Examples of such type of search engine are following: Yahoo, MSN search.

2.2 Historical Evolution of Search Engines: From Archie to Google Today

The SMART informational retrieval system was developed by Gerard Salton and his teams at Harvard and Cornell. They included important concepts like the vector space model, Inverse Document Frequency (IDF),Term Frequency (TF). After that, Ted Nelson created Project Xanadu in 1960 and coined the term hypertext in 1963. ARPANET is the network which inevitably prompted to the web.(Wall, 2016)

Archie was first search engine created by Alan Emtage in 1990, a student at McGill University in Montreal. Its original name was "archives" but it was shortened to Archie. Archie became a database of web filenames which it might match with the client's queries. Nowadays popular search engine is "Google". Google is one of the most popular search engines of all time due to its exceptional and improving algorithm as compared to other search engines. A table enumerating the chronology of search engine is prescribed below (Kathuria et al, 2016.p 48-49).

Table 1: Historical Evolution of Search

S.No.	Year of Development	Search Engine	Brief list of Key Features
1.	1990	Archie *developed by* Alan Emtage	➢ First search engine

			➢ FTP server based file sharing
2.	1992	Veronica & Jughead *developed by* Fred Barrie, Rhett Jones University of Naved a System Computing Services group.	➢ Worked on plain text files ➢ It allowed Keyword based search in Its own designed Gopher Index System
3.	1993	a) W3 Catalog *developed by* Oscar Nierstrasz University of Geneva b) JumpStation *developed by* Jonathon Fletcher University of Stirling c) WWW Wanderer *developed by* Matthew Gray Massachusetts Institute of Technology d) Aliweb *developed by* Martijn Koster United Kingdom	➢ Textual based browser includes Integration of manually maintained catalogue. ➢ It allowed combination of crawling, searching and indexing ➢ Introduces web robots to crawl the web ➢ Allowed users to submit pages they wanted to indexed along with description.
4.	1994	a) Lycos *developed by* Mauldin Micheal L. Canegie Mellon Univ., Pittsburg b) Infoseek developed by Steve Kirsch Infoseek Corporation	➢ Offers subject directories ➢ Allows proximity searching ➢ Allowed real-time submission of the page and subject oriented search
5.	1995	a) Excite *developed by* Joe Kraus, Graha spencerGarage in Silicon velley. b) AltaVista *developed by* Louis Monier, Michael Burrows Digital Equipment Corporation's c) Yahoo *developed by* David Filo, Jerry Yang Yahoo Corporation d) AOL *developed by* Bill von Meister Control Video Corporation e) MSN *developed by* Microsoft ltd.	➢ Allows search logics Boolean operators AND, OR, AND NOT ➢ Allows natural language query and keyworf based simple and advanced search ➢ Supports full Boolean searching and Wild Card Word in Phrase. ➢ Allowed Messenger and Subscriber based service ➢ Provides Automatic local search options with large database
6.	1996	*a)* DogPile *developed by* Aaron Flin Blucora Inc. *b)* InfoSpace *developed by* Naveen Jain Infospace Inc *c)* Hotbot Wired Magazine *developed by* Inktomi Corporation *d)* WOW *developed by* Jeniffer Thomp son Compu Serve *e)* Ask *developed by David* Warthen, Garrett Gruener IAC/ InterActive Corporation	➢ Meta Search engine having its own search Index ➢ Provides Instant messenger service and aggregated search results from leading search engines ➢ Extensive use of cookies ➢ Search within search results ➢ Updation of Database ➢ Find all of the breaking news articles, top videos and trending topics ➢ Natural language-based Search and facilitate Question answering system
8.	1998	Google *developed by* Larry Page *and* Sergey Brin, Stanford University, Stanford	➢ Semantic and Keyword based search ➢ Free, Fast and easy to search

9.	1999	All theWeb *developed by* Egge Norwegian Univ. of Sci. & Tech.	➢ Faster Database having advanced search features ➢ Search clustering customizable look
10.	2007	Live Search *developed by Satya Nadella Microsoft*	➢ The new search engine used search tabs that include Web, news, images, music and desktop
11.	2008	a) DuckDuckGo *developed by* Gebriel Weinberg DuckDuckGo Inc.	➢ Help you search on thousands of other sites, directly and provides instant answers to search query.
12.	2009	a) Bing *developed by* Steve Billmer Microsoft b) Caffeine *developed by Matt Cutts Google*	➢ Updation of indexing and keyword based search ➢ Update search index on a continuous basis, globally and find links to relevant content quickly
13.	2010	a) Google Instant *developed by* Marissa Mayer & Matt Cutts Google b) Blekko *developed by* Rich Skrenta Blekko Inc.	➢ Search-before-you-type ➢ Prediction about the users whole query ➢ Faster Searches, Smarter Prediction, Instant Result ➢ Blekko offers a web search engine and social news platform that provides users with curated links for the entered search criteria.
14.	2013	a) Contenko *developed by* Tomas Meskauskas Amerow LLC b) Alhea *developed by Manuel Barrios AmazonTechnologies Inc* c) *Google Humming Bird developed by GianlucaFiore Lli Google*	➢ Gives Innovative means for browsing the internet ➢ Offers a single source to search the Web ➢ Alhea .com compile s results from many of the Web's major search properties ➢ Update search algorithm ➢ Precise and fast
15.	2015	SciNet	➢ Provides place to share the latest research, technologies, and demonstrations for networks.

3. Open Source Web Search Engine

3.1 Lucene: Lucene is an open source, simple but powerful Java-based search engine. It is very popular and a fast search library. It utilized within Java based applications to add document searchability to any

kind of application in a very simple and effective approach. It is scalable. This high-performance search engine is used to index and search any kind of text. The latest version of Lucene is 6.3.0. (Apache Lucene Core, 2016).

3.2 Solr: Apache Solr is an enterprise-capable, open source search platform based on the Apache Lucene search library. Solr is based on Java and provides both a RESTful XML interface and a JSON API with which search applications can be built. (Solr, 2016)

3.3 Sphinx: Sphinx is an open source full-text search server written in C++ and works on Linux (RedHat, Ubuntu, etc), Windows, MacOS, Solaris, FreeBSD, and a few other systems (Sphinx, 2017).

3.4 Nutch: It is a highly extensible and scalable open source web crawler implemented in Java. It allows developers to create plug-ins for media-type parsing, data retrieval, querying, and clustering. It has two parts the crawler and the searcher (Apache Nutch, 2016).

3.5 Xapian: it is an open source full-text search engine released under the GNU General Public License (GPL). It allows developers to easily add advanced indexing and search facilities to their own applications and also support a set of boolean query operators. It is written in c++. (James, 2015).

Table 2 gives a brief sketch of above search engines including their key features.

Search Engine	Developer(s)	Type	Licenses	Key features	Stable release
Lucene (www.apache.org/licenses)	Apache Software Foundation	Text search engine	GPL	• Scalable, High-Performance Indexing • Powerful, Accurate and Efficient Search Algorithms • Allows phrase queries, wildcard queries, proximity queries, range queries and provide • Fielded searching (e.g. title, author, contents) sorting by any field • Multiple-index searching and flexible faceting, highlighting, joins and result grouping	6.3.0
Apache Solr (http://lucene.apache.org/solr/)	Apache Software Foundation	Search and index API	Apache licence 2.0	• Advanced Full-Text Search Capabilities • Standards Based Open Interfaces - XML, JSON and HTTP • Comprehensive Administration Interfaces • Near Real-Time Indexing • Faceted Search and Filtering • Geospatial Search • Query Suggestions, Spelling and More • Multiple search indices	6.3.0
Sphinx (www.http://sphinxsearch.com/)	Andrew Aksyonoff	Search and index	GPLv2 and commercial	• Batch and Real-Time full-text indexes. • Non-text attributes support • SQL database indexing. • Advanced full-text searching syntax.	

				- Rich database-like querying features. - Better relevance ranking. - Flexible text processing. - Distributed searching.	2.3.2
Nutch (http://nutch.apache.org /	Apache Software Foundation	Web crawler	Apache License 2.0	- Create plug-ins for media-type parsing, data retrieval, querying and clustering. - Highly robust and scalable -	1.12
Xapian (https://xapian.org/)	Dr. Martin Porter at Cambridge University	Open Source Search Engine Library	GPL	- Faceted search is supported. - Allows simultaneous update and searching - Phrase and proximity searching - Suggests spelling corrections for user supplied queries.	1.4.4

Table:2 Open source search engines

4. Advanced Search Techniques for Information Retrieval on the Web

4.1 SAFQuery Search Interface: For making information retrieval process simple in current interfaces, MWei-Chao Lin, Shih-Wen Ke and Chih-Fong Tsai introduce a prototype system, namely, SAFQuery (Simple and Flexible Query interface). To create those prototype Google API was used by the authors.

Features of this prototype includes:

- integration of both simple and advanced query strategies into a single interface
- Provides query history information that permits users to reuse past queries easily.

After doing user evaluation it is concluded that most of the users had a positive experience using SAFQuery. It is very easy to use and make simple the web search. The proposed prototype system provides simple and flexible Web search strategies. Particularly, it allows users to easily issue simple and advanced queries based on one single query interface, interchangeably. In addition, users can easily input previously issued queries without investing time to recall what the queries are and/or to re-type previous queries (Lin, Ke and Tsai, 2016).

4.2 Federated Search: It is an information retrieval technology that allows the concurrent search of multiple databases in which search query given by user distributed to search engines, databases participating in the federation (Federated Search, 2016). It is also called meta searching. Federated search allows users to search information from multiple databases from single interface and give back a coordinated arrangement of results (Mah and Stranack, 2005).

Gibson et al in their study defined there are two main categories in which Federated search technologies are divided: cross search and harvested search. Cross search deals with the phenomenon in which searches distributed sources simultaneously and use common result interface for display of results and harvested search, which retrieves the contents of multiple distributed databases, normalizes the records, and stores them in a large union index . Cross search engines, search dispersed databases and also return the result to a common search interface. For this process, they require a "connector" for each target database. Connector performs two functions: advise the search engine how to request outcomes from given source and how to interpret those outcomes for display. These connectors includes XML Gateways (*Major standards for XML Gateways include SRU/SRW, OpenSearch, and MetaSearch XML Gateway (MXG)*), Z39.50 and Screen scraping. The second main technology behind federated search is to harvest all of the relevant sources of information, normalizes the records, and stores them in a large union index. Major harvesting standard includes OAI-PMH, METS, LOCKSS, and custom formats (Gibson et al, 2009). XML gateways and Z39.50 protocol are two major retrieval protocols used by federated search tools to facilitate multiple database searching (Abercrombie, 2008).The advantage of federated search is it provide efficient, current, relevant and quality of search results. Federated search empowers librarians to provide efficient, effective access and delivery of the most relevant content from different sources to their users.

Example: Dogpile, Vivismo, Askjeevs.

4.3 Faceted Search: Dr. S.R. Ranganathan "father of library science" developed faceted classification system named colon classification for library reading material in 1933. it is a technique for accessing information organized according to a faceted classification system, also known as faceted navigation or faceted browsing. In the current environment, many online library catalogs such as OPACs are using faceted search interfaces such as the OCLCOpen WorldCat system. (Faceted Search, 2016)

According to Kong "Faceted search enables users to navigate a multi-faceted information space by combining text search with drill-down options in each facet."(Kong & Allan, 2014). For example, while searching "Mobile "in an e-commerce site, users can select brands and types from the provided facets. So, Faceted search is becoming a well-known technique to allow users to interactively search and explore complex information spaces (Koren et al, 2008). Faceted search provides a platform for interactive information retrieval.

Example: The CiteSeerX project at the Pennsylvania State University permits faceted search for academic documents and Furthermore proceeds to extend under different facets, for example, table search.

5.0 Conclusion

Information on the web is growing at a fast pace everyday, at the same time, it is becoming increasingly difficult to find relevant information. Due to information overload on the web it has become problematic to find out the quality of the search results. Advanced search techniques help in fast and effective searches. Hence to assist library users in locating relevant information, information professionals should be familiar with the various advanced search techniques.

References

1. Abercrombie, S. E. (2008). Evaluation of federated searching options for the school library. School Library Media Research, 11.
2. ACR Cloud. Retrieved from https://www.acrcloud.com/music-recognition on 10th January 2017.
3. Apache Lucene core. Retrieved from http://lucene.apache.org/core/ on 2nd January, 2017
4. Apache Nutch. Retrieved from https://en.wikipedia.org/wiki/Apache_Nutch on 3rd January 2017.
5. Arrigo, M., Gentile, M., Taibi, D., & Di Giuseppe, O. (2005). Specialized search engines for E- learning. Recent Research Developments in Learning Technologies.
6. Croft, W. B., Metzler, D., & Strohmann, T. (2015). Search engines. Pearson Education.
7. Curran, K., & Mc Glinchey, J. (2007). Vertical search engines. ITB Journal,16(3), 22-26.
8. Faceted search. Retrieved from https://en.wikipedia.org/wiki/Faceted_search on 3rd January 2016.
9. Gibson, I., Goddard, L., & Gordon, S. (2009). One box to search them all: Implementing federated search at an academic library. Library Hi Tech, 27(1), 118-133.
10. Google images. (2008). Retrieved from Wikipedia https://en.wikipedia.org/wiki/Google_Images on 10th January 2017.
11. Google inside search. Retrieved from https://www.google.com/intl/es419/insidesearch/features/images/searchbyimage.html on 10th January 2017.
12. Google voice search. Retrieved from Wikipedia https://en.wikipedia.org/wiki/Google_Voice_Search on 10th January 2017.
13. James. (2015). Top 5 Open Source Search Engines. Retrieved from http://www.mytechlogy.com/IT-blogs/8685/top-5-open-source-search-engines/#.WHEipvF948p on 3rd January 2017.
14. Kathuria, M., Nagpal, C. K., & Duhan, N. (2016). Journey of web search engines: Milestones, challenges & innovations.

15. Kong, W., & Allan, J. (2014, November). Extending faceted search to the general web. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (pp. 839-848). ACM.
16. Koren, J., Zhang, Y., & Liu, X. (2008, April). Personalized interactive faceted search. In Proceedings of the 17th international conference on World Wide Web (pp. 477-486). ACM.
17. Lin, W. C., Ke, S. W., & Tsai, C. F. (2016). SAFQuery: a simple and flexible advanced Web search interface. The Electronic Library, 34(1), 155-168.
18. List of OK Google Now Voice Commands. Retrieved from http://www.trackmyandroidphone.com/2016/02/list-of-ok-google-now-voice-commands/ on 10th January 2017.
19. Mah, C., & Stranack, K. (2005). dbWiz: open source federated searching for academic libraries. *Library Hi Tech*, *23*(4), 490-503.
20. McCoy. Kimberly (2000). Search engines, subject directories, and meta-search engines. The OLRC News. Volume 5, No. 1
21. Ntoulas, A., Cho, J., Cho, H. K., Cho, H., & Cho, Y. J. (2005, August). Study on the evolution of the web. In US–Korea Conference on Science,Technology, and Entrepreneurship (UKC) (pp.1- 6).
22. Prasad, A. R. D., & Madalli, D. P. (2009). Classificatory ontologies. Extensions & Corrections to the UDC.
23. Query by humming. Retrieved from https://en.wikipedia.org/wiki/Query_by_humming. 10th January 2017
24. Shade O, Kuyoro, Okolie Samuel, O, & Kanu Richmond, U. (2012). Trends in web-based search engine. *Journal of Emerging Trends in Computing and Information Sciences*,*3*(6).
25. Shettar, R., & Bhuptani, R. (2007). A vertical search engine–based on domain classifier. International Journal of Computer Science and Security, 2(4), 18-27.
26. Solr. Retrieved from http://lucene.apache.org/solr/features.html 2nd January 2017.
27. SoundHound Inc. Retrieved from http://www.soundhound.com/ on 10th January 2017.
28. SoundHound. Retrieved from https://en.wikipedia.org/wiki/SoundHound on 10th January 2017.
29. Sphinx search engine. Retrieved from https://en.wikipedia.org/wiki/Sphinx_(search_engine) 3rd January 2017.
30. Wall, Aron. (2016). Search engine history. Retrieved from http://www.searchenginehistory.com/ on 31st December 2016.