Seek and Ye Shall Search

SEQu-RAmA: Search Engine Query - Results Accumulator Aggregator

http://wgilreath.github.io/sequrama.html

February 11, 2018 | by William Fletcher Gilreath (**will.gilreath@outlook.com**)

"Nothing is as powerful as an idea whose time has come."
Victor Hugo (1802–85), French poet, novelist, and playwright

Introduction

As of June 2017, 51% of the world's population has Internet access. Well over 3-billion people are on the Internet, searching for information. [1] There is an explosion of information online from sites that allow search within a website, to search engines around the world, to web directories, and web sites that allow a search of a website. More Internet users, more information, more search.

The early Internet had much information, but it was the earliest search engines that began to index and organize the information—such search engines as Altavista, findorama, webSummon, Go! (actually, Findorama, webSummon are fictitious search engines…) Web crawlers, spiders, and bots are trawling the Internet, building, indexing, and organizing webpages.

Overall the bots represent approximately 52% of the web traffic…over half the Internet activity is structuring it. [3]

But now there are many a search engine in different countries, search engines for specific resources—blogs, wikis, answer sites, and sites often have their own search—such as Amazon and Craigslist.

Hence there is an abundance of search available. In searching though, a user is looking for a specific needle in a specific haystack--in an Internet field of many haystacks. Consider one such haystack, the Internet Archive, which has stored over 310-billion webpages from web pages on the Internet. [2]

This has been and is still an ongoing process from the creation of web to now the cloud. With so much information out there, searchable, there is a plethora of search for specific information. Now with so many humongous libraries and information repositories of data, the need is to search within the indices of the the libraries--search within search.

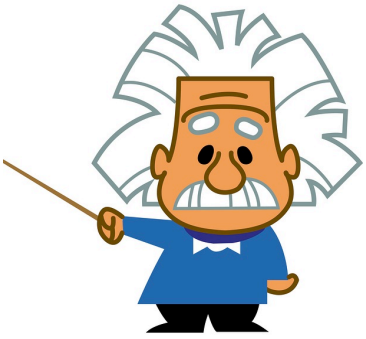
Sometimes a search is like art, you don't know what art you like, but you will know when you see it. Search is like that—don't know what you are looking for, but you will know when you find it. The problem is finding it, searching for the right search.

The major imperative for search is to *find information for the user*, no more, no less.

Searching within the search is finding information, but not from one site or a single search engine—a heterogenous search of many in tandem. Need to rise above source and the index of information—the emphasis on search and the information, not classification by site, directory, search engine.

The Problem—Oodles of Datum

The problem is: "The Many Library Sources for the Professor's Research Problem." Consider the professor…



A college professor "The Professor" with M-graduate students in different lands for an online class assigns a research problem to find N-sources at their local library. The M-graduate students go out, and e-mail the professor with the N-sources each has found.

Now, consider a continuum of results. All M-graduate students find the exact similar N-sources for the research problem—all the same. The professor gets N-sources.

Alternatively, all M-graduate students find uniquely N-sources for the research problem—all are different. The professor gets M × N-sources.

The ends of the continuum are N-results all the same or M × N-results all unique. In reality, the results of the search would be a mix of some duplicate and unique sources.

The question is: How to present and organize the results from the many searches?

Some more specific questions about presenting and organizing the results are:

1.How to handle (order and rank) duplicate results from different search engines?

2.How to handle unique results (order and rank) consistently with duplicate results?

3.What is the organizing key for the results--the contents, the title, the link, the date of the content?

4. How to handle other search factors--result time, search engine reliability, cardinality of results from a specific search engine?

5. How to integrate different ordered, ranked, and structured results from different sites--search engine, web directory, other sites?

There are other questions, but the problem is organizing and indexing the datum of the results from many search sources. This is the problem of "super search"...or searching within search...searching multiple search engines and sites.

But why super-search?

Search of Search

The reason for super-search is to search across many different sources for information, and not just have tunnel vision and the limits, idiosyncrasies of a particular search engine. When an Internet user “surfs” the Web…more often they are surfing the massive web store’s index generated from a web crawler, that leads to an actual web page.

More significantly, why SEQu-RAmA? SEQu-RAmA is an acronym for Search Engine Query—Results AccuMulator Aggregator.

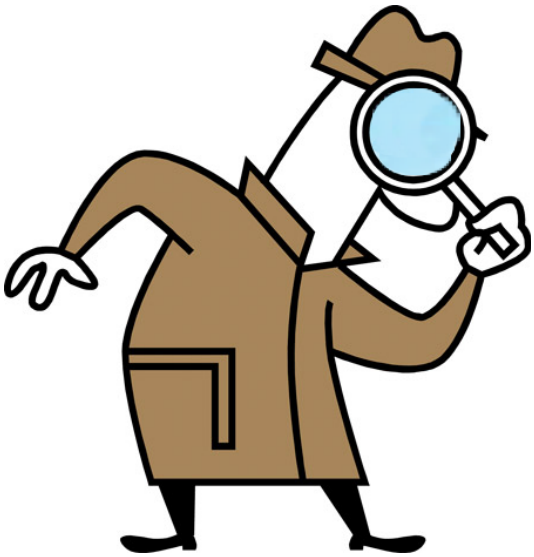
SEQu-RAmA is a super-search engine, that searches other search sources.

There are a variety of reasons for super-search or SEQu-RAmA which are:

Massive/N-dimensional search

1. SEQu-RAmA uses different search results for better search. MANY
2. SEQu-RAmA searches around the world for greater breadth. PLANET
3. SEQu-RAmA searches through different sites, search engines, directories. DIVERSE

Independence/neutral search



4. SEQu-RAmA is more efficient search not hit-or-miss with one search engine. EFFICIENT
5. SEQu-RAmA is independent of any countries specific politics of results. INDEPENDENT
6. SEQu-RAmA is not dependent upon any one specific search engine ranking. UNBIASED

Focused/heterogeneous on search

7. SEQu-RAmA uses different search engines, sites for heterogenous search. INTEGRATIVE
8. SEQu-RAmA provides a single point/nexus of access for search. CENTRAL
9. SEQu-RAmA finds results that can be hidden/deleted on other search engines. DEPTH

Anonymity of search

10. SEQu-RAmA hides the user from the various search engines. INVISIBLE
11. SEQu-RAmA hides your query/you from other search engines and sites. OPAQUE
12. SEQu-RAmA is a specialized proxy to the various search engines, sites. PROXY

How SEQu-RAmA is Different

SEQu-RAmA is different from other search engines, but simply different in comparison. Consider what SEQu-RAmA both is not and is yes as an Internet search engine.

What SEQu-RAmA is NOT in terms of Internet search…

- No massive data store a'la a web archive of web pages.
- No precomputed index on web pages.
- No storage of any web data in a massive data warehouse
- Not a web crawler, spider, or bot—no crawling the web for results

What SEQu-RAmA IS yes in terms of Internet search…

- A "web scrape" results from different sites, search engines an integrate scraps together.
- A given result could contain a link to a cache of a web page from Site A, a fragment from another Site B, and the link from Site C, a link to Site D, and etcetera.
- Each result is a nexus from the various results from the different sites and search engines.
- The search for a super-search is "stateless" as no data or state or index is saved.

Crawling versus Scraping for Search

SEQu-RAmA is a web scraping search, not a web crawling search. Huh?

So what's the big deal, or the difference?

One search engine crawls the web and builds a massive web store of indexed results, whereas SEQu-RAmA scrapes the search sources and organizes the results.

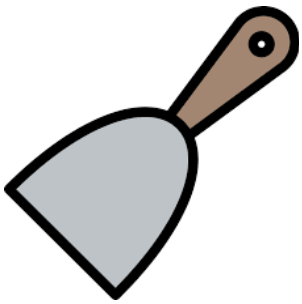
The table summarizes the contradistinction between a web crawler and web scraper for search…

	Crawler	Scraper
Information	Homogeneous	Heterogenous
Output Source	Data Warehouse	Search Results
Retrieval	Index on Data	Scraps from Search
State	Stateful—Index, Classification	Stateless—Query
Nature of Date	Static, Cached Web Page	Dynamic from Search of Site
Response Time	Instant—within a second	Delayed—many seconds

Table of Web Crawler versus Web Scraper for Search

The early web was static, but now have a dynamic web with changing content, much more heterogeneous information, and search is available from a web site (like Amazon, Craigslist), and have many search engines in different nations. There is more existing search, and search activity to fetch, cache, process, and index web information.

A web crawler "crawls" a site or search the Internet, cache the web pages, classify the web page so when queried for a search the large repository is searched. A web scraper "scrapes" together information from a search on various sites, organizes the results to a search query.

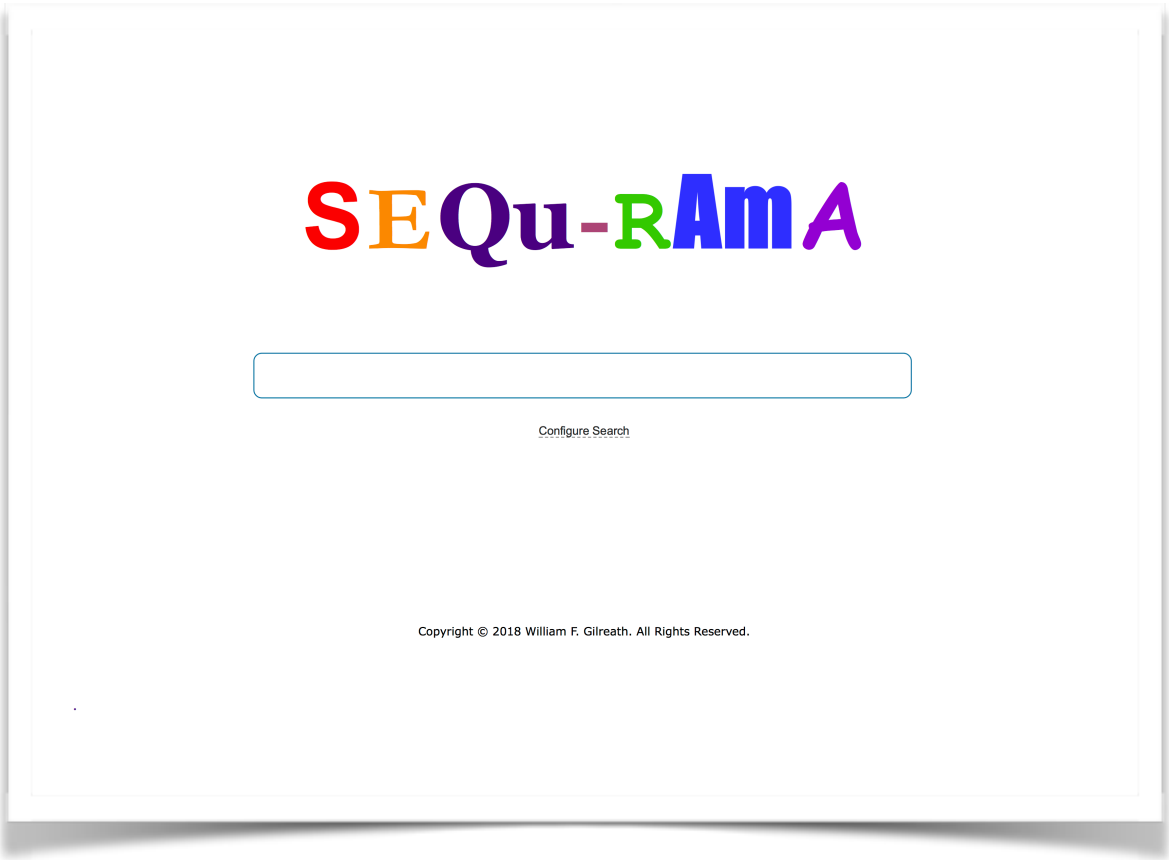


The primary difference is the web scraper uses the results of the web crawlers to search more efficiently with an emphasis on results.

Exempli Gratia of Super Search (e.g. SEQu-RAmA)

The goal of SEQu-RAmA is: create a nexus of super-search for more efficient search within a context, but **without** the need to crawl the web, and build a huge repository that is indexed. A user will not just “surf the web” but will “massively surf many a web” with super-search. Search the search, not the index of a massive web cache of information.

SEQu-RAmA is not a specific search engine (like say Kayak for best travel prices), but is a general-purpose super-search. SEQu-RAmA is “quod erat demonstrandum”—a working prototype of the super-search engine that uses the algorithm to search different search engines to find results to a user query by web scraping from various sites.



The prototype for SEQu-RAmA uses HTML5 with CSS3, using JavaScript for the browser interface on the front-end, and on the backend is a Java servlet running in Apache Tomcat on a Platform As A Service (PaaS) host for the super-search engine implementation. [4] Try it out!

References

1. http://www.internetworldstats.com/stats.htm
2. https://archive.org/
3. https://www.theatlantic.com/technology/archive/2017/01/bots-bots-bots/515043/
4. https://wgilreath.github.io/sequrama.html