Ranked Keyword Search in Cloud Computing

Ramya Majeti[#1], Mahalakshmi Tejaswi Palvadi[#2], P. Venkata Naresh[#3], Dr. S. Satyanarayana[*4]

Computer Science and Engineering, KL University
Green Fields, Vaddeswaram, PO Dt-522 502, Andhra Pradesh, India

[*]*Associate Professor of Computer Science and Engineering, KL University*
Green Fields, Vaddeswaram, Andhra Pradesh, India

***Abstract*— This document provides a way towards modularized and a light weight approach towards the search engine process using the merits of cloud computing. The cloud based search architecture enables customization of search process as per requirements of the stake holders. This new approach provides effective and personalized search models using cloud platform for low cost. It overcomes the pitfalls of traditional search engine optimization and hence has a tremendous scope for future development. Ranked keyword search is an active practice of optimizing a web site by improving internal and external aspects. This paper describes all areas of ranked keyword search-from discovery of terms and phrases that will generate traffic.**

***Keywords*— Indexing, Ranking, Multi Cloud Interactions, Web Crawling, Search Query Optimization.**

I. Introduction

The current approach towards design of search engine is monolithic and infrastructure heavy. The majority of web traffic is driven by some of the major commercial search engines. Search engines are the primary method of navigation for almost all internet users. Hence they play a key role in displaying one's own site when it is searched in the search engine. Experiences have shown that search engine traffic can make or break organization's success. Investing in Ranked Keyword Search, whether through time or finances, can have an exceptional rate of return. A search engine mainly works using crawling, indexing, storage and ranking. Search engines are always working towards improving their technology to crawl more and the web more deeply and return increasingly relevant results to users. However there is always a limit how search engines can operate. The right moves can net you thousands of visitors and attention and the wrong moves can hide or bury your site deep in the search results where visibility is minimal. In addition to making content available to search engines Ranked Keyword Search can also help boost rankings, so that the content has been found placed where searchers will more readily see it.

II. How Search Engines Operate

Search engines have a short list of critical operations that allow them to provide the relevant web results when searchers use system to find information.

1.) Crawling the web: Search engines run automated programs called bots or spiders that use a hyperlink structure on the web to crawl the pages and documents that make up the World Wide Web. It had been estimated that there had been approximately 20 billion existing pages, search engines have crawled between 8 to 10 billion.

2) *Indexing Documents:* Once a page has been crawled, its contents can be indexed where indexes are stored in a giant database of documents that makes up a search engines index. The index needs to be tightly managed, so that the requests which must search and sort billions of documents can be completed in fractions of a second.

3) *Processing Queries:* When a request for information comes into the search engine, the engine retrieves from its index the entire document that matches the query. A match is determined if the terms or phrase is found on the page in the manner specified by the user. For example a search for car and driver magazine *at* Google would result 8.25 million results, but a search for the same phrase in quotes "car and driver magazine" would result only in 166 thousand results. The first system was commonly called "Find all" mode the Google returned all the documents containing car, driver and magazine but in the second system only the results that contain the exact phrase car and driver magazine are retrieved.

4) *Ranking Results:* Once the search engine has determined match the given query, the engine algorithm runs calculations on each of the results to determine which is the most relevant to the given query. Then sorting takes place on the result pages in order from most relevant to least so that users can make a choice about what to select.

III. Speed Bumps And Walls Of Traditional Search Engines

Certain types of navigation may hinder or entirely prevent search engines from reaching website's content. As search engine spiders crawl the web, they rely on the architecture of hyperlinks to find a few documents and revisit those that may have changed. Bumps are referred as complex links and deep site structures with little unique content and data that cannot be accessed by spiderable links is referred as walls.

A. Possible Speed Bumps for SE spiders:

- Spiders are reluctant to crawl complex URL because they often result in errors with non human visitors
- Unless there are many other external links pointing to the site, spiders will ignore deep pages.
- Spiders may not be able to retain session id or cookie to enable navigation.

- Pages are split into frames that hinder crawling and cause confusion about pages to rank the results.

B. Possible walls for SE spiders

- Pages are accessible only via a select form and submit form
- Pages representing a drop down menu to access them.
- Documents are accessible only via a search box.
- Pages requiring login
- Pages that redirect before showing content

IV.MEASURING RELEVANCE AND POPULARITY

Modern commercial search engines rely on the science and information retrieval. The IR scientists realized two critical components made up the majority of search functionality:

Relevance: It is the degree to which the content of the documents returned in the search matched the users query intention and terms. The relevance of the document increases if the terms or phrase queried by the user occurs multiple times and shows up in the title of the work or in important headlines or sub headers.

Popularity: It is the measurement of citation of a given document that matches the users query. The popularity of the given document increases with every other document that references it.

V. DOCUMENT AND LINK ANALYSIS

In document analysis, search engines look at whether the search terms are found in important areas of document – the title, the Meta data, the heading tags and the body of the text document.

In link analysis, search engines measures not only who is linking to the page or site, but what they are saying about the page or site. By this analysis we can know about the contextual data about the site the page is hosted on. We can also have a good grasp of who is affiliated with whom and who is worthy of being trusted.

Link and document analysis combine and overlap hundreds of factors that can be individually measured and filtered through the search engine algorithms. The algorithm then determines scoring for the documents and lists results in decreasing order of importance.

As search engines index the web's link structure and page contents, they find two distinct kinds of information about a given site or page - attributes of the page/site itself and descriptive about that site/page from other pages. Since the web is such a commercial place, with so many parties interested in ranking well for particular searches, the engines have learned that they cannot always rely on websites to be honest about their importance. Thus, the days when artificially stuffed Meta tags and keyword rich pages dominated search results (pre-1998) have vanished and given way to search engines that measure trust via links and content.

The theory goes that if hundreds or thousands of other websites link to you, your site must be popular, and thus, have value. If those links come from very popular and important (and thus, trustworthy) websites, their power is multiplied to even greater degrees. Links from sites like NYTimes.com, Yale.edu, Whitehouse.gov and others carry with them inherent trust that search engines then use to boost your ranking position. If, on the other hand, the links that point to you are from low-quality, interlinked sites or automated garbage domains (aka link farms), search engines have systems in place to discount the value of those links.

Engines have systems in place to discount the value of those links.

The most well-known system for ranking sites based on link data is the simplistic formula developed by Google's founders - Page Rank. Page Rank, which relies on log-based calculations, is described by Google in their technology section:

> *Page Rank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."*

Page Rank is derived (roughly speaking), by amalgamating all the links that point to a particular page, adding the value of the Page Rank that they pass (based on their own Page Rank) and applying calculations in the formula

Google's toolbar includes an icon that shows a Page Rank value from 0-10

Page Rank, in essence, measures the brute link force of a site based on every other link that points to it without significant regard for quality, relevance or trust. Hence, in the modern era of RANKED KEYWORD SEARCH, the Page Rank measurement in Google's toolbar, directory or through sites that query the service is of limited value. Pages with PR8 can be found ranked 20-30 positions below pages with a PR3 or PR4. In addition, the toolbar numbers are updated only every 3-6 months by Google, making the values even less useful. Rather than focusing on Page Rank, it's important to think holistically about a link's worth.

Here's a small list of the most important factors search engines look at when attempting to value a link:

The Anchor Text of Link - Anchor text describes the visible characters and words that hyperlink to another document or location on the web. For example in the phrase, "CNN is a good source of news, but I actually prefer the BBC's take on events," two unique pieces of anchor text exist - "CNN" is the anchor text pointing to *http://www.cnn.com*, while "the BBC's take on events" points to *http://news.bbc.co.uk*. Search engines use this text to help them determine the subject matter of the linked-to document. In the example above, the links would tell the search engine that when users search for "CNN", RANKED KEYWORD SEARCHmoz.org thinks that *http://www.cnn.com* is a relevant site for the term "CNN" and that *http://news.bbc.co.uk* is relevant to "the BBC's take on events". If hundreds or thousands of sites think that a particular page is relevant for a given set of terms, that page can manage to rank well even if the terms NEVER appear in the text itself (for example, see the BBC's explanation of why Google ranks certain pages for the term "Miserable Failure").

Global Popularity of the Site - More popular sites, as denoted by the number and power of the links pointing to them, provide more powerful links. Thus, while a link from RANKED KEYWORD SEARCH may be a valuable vote for a site, a link from bbc.co.uk or cnn.com carries far more weight. This is one area where Page Rank (assuming it was accurate), could be a good measure, as it's designed to calculate global popularity.

Popularity of Site in Relevant Communities - In the example above, the weight or power of a site's vote is based on its raw popularity across the web. As search engines became more sophisticated and granular in their approach to link data, they acknowledged the existence of "topical communities"; sites on the same subject that often interlink with one another, referencing documents and providing unique data on a particular topic. Sites in these communities provide more value when they link to a site/page on a relevant subject rather than a site that is largely irrelevant to their topic.

Text Directly Surrounding the Link - Search engines have been noted to weight the text directly surrounding a link with greater important and relevant than the other text on the page. Thus, a link from inside an on-topic paragraph may carry greater weight than a link in the sidebar or footer.

Subject Matter of the Linking Page - The topical relationship between the subject of a given page and the sites/pages linked to on it may also factor into the value a search engine assigns to that link. Thus, it will be more valuable to have links from pages that are related to the site/pages subject matter than those that have little to do with the topic.

These are only a few of the many factors search engines measure and weight when evaluating links. For a more complete list, see RANKED KEYWORD SEARCH engine ranking factors article.

Link metrics are in place so that search engines can find information to trust. In the academic world greater citation meant greater importance, but in a commercial environment, manipulation and conflicting interests interfere with the purity of citation-based measurements. Thus, on the modern WWW, the source, style and context of those citations is vital to ensuring high quality results.

VI. CONCLUSION

In this paper, as an initial attempt, we motivate and solve the problem of supporting efficient ranked keyword search for achieving effective utilization of remotely stored encrypted data in Cloud Computing.

REFERENCES

[1] M. Li, S. Yu, K. Ren, and W. Lou, “Securing personal health records in cloud computing: Patient-centric and fine-grained data access control in multi-owner settings,” in *SecureComm’10*, Sept. 2010, pp. 89–106.

[2] H. L¨ohr, A.-R. Sadeghi, and M. Winandy, “Securing the e-health cloud,” in *Proceedings of the 1st ACM International Health Informatics Symposium*, ser. IHI ’10, 2010, pp. 220–229.

[3] M. Li, S. Yu, N. Cao, and W. Lou, “Authorized private keyword search over encrypted personal health records in cloud computing,” in *ICDCS ’11*, Jun. 2011.

[4] “The health insurance portability and accountability act.” [Online]. Available: http://www.cms.hhs.gov/HIPAAGenInfo/01 Overview.asp

[5] “Google, microsoft say hipaa stimulus rule doesn’t apply to them,” http://www.ihealthbeat.org/Articles/2009/4/8/.

[6] “At risk of exposure – in the push for electronic medical records, concern is growing about how well privacy can be safeguarded,” 2006. [Online]. Available: http://articles.latimes.com/2006/jun/26/health/he-privacy26

[7] K. D. Mandl, P. Szolovits, and I. S. Kohane, “Public standards and patients’ control: how to keep electronic medical records accessible but private,” *BMJ*, vol. 322, no. 7281, p. 283, Feb. 2001.

[8] J. Benaloh, M. Chase, E. Horvitz, and K. Lauter, “Patient controlled encryption: ensuring privacy of electronic medical records,” in *CCSW ’09*, 2009, pp. 103–114.

[9] S. Yu, C. Wang, K. Ren, and W. Lou, “Achieving secure, scalable, and fine-grained data access control in cloud computing,” in *IEEE INFOCOM’10*, 2010.

[10] C. Dong, G. Russello, and N. Dulay, “Shared and searchable encrypted data for untrusted servers,” in *Journal of Computer Security*, 2010.

[11] V. Goyal, O. Pandey, A. Sahai, and B. Waters, “Attribute-based encryption for fine-grained access control of encrypted data,” in *CCS ’06*, 2006, pp. 89–98.

[12] M. Li, W. Lou, and K. Ren, “Data security and privacy in wireless body area networks,” *IEEEWireless Communications Magazine*, Feb. 2010.