ISSN:0975-9646

Medical Search Engine For More Accurate and Specific Search Results in Cancer Patients Data

Venkat Reddy Korupally[#1],Subba Rao Pinnamaneni[*2]
[#]*Assistant Professor, Dept of CSE, DRK Institute of Science & Technology*
**Associate Professor, Dept of CSE, Chalapathi Institute of Engineering & Technology*

Abstract - **This paper enumerates about personalized Search Engine for doctors who are treating different cancers around the world. Till now we don't have any good search engine which can give related results of different prognosis and diagnosis techniques of treating different types of cancers. Diagnosis and Prognosis are the two major challenging aspects which are to be addressed in treating cancer. The survival of Cancer patients depend upon the diagnosis of Cancer at the early stages (either in Stage I or Stage II). If the cancer diagnosed in Stage III or later stages, the chances of survival of the patient will become more critical. Prognosis will reveal the survival pattern for different attributes i.e., for specific drug, before and after the treatment. Better the diagnosis and prognosis, better the treatment outcome For Cancer. Generally single patient records will generate a large amount of data if we manage and analyze such big data, we may solve many problems in identifying the patterns which will lead to diagnose and prognosis of the cancer. This will help the doctors to take proper decisions. In this paper I am proposing a search engine which will be implemented on Hadoop and I am proposing an algorithm which will analyze the open Cancer patient's data given by The Cancer Genome Atlas (TCGA) Data Portal, The National Cancer Institute, USA and guide the doctors in decision making in Diagnosis and Prognosis of the Cancer Patients using the proposed search engine.**

Keywords: Index **Medical Search engine, Page Ranking System.**

I. INTRODUCTION

Searching for medical information on the Web is a challenging task for ordinary Internet users. Often, users are uncertain about their exact medical situations, are unfamiliar with medical terminology, and hence have difficulty in coming up with the right search keywords. This paper explains about a surfer model which helps to improve the present web page ranking system and how to stop spam and how to give valid ranks to web pages which have the necessary and useful content for doctors. Before going into the details let us outlook the present searching system.

How a search engine is assigning ranks to the web pages on the net?

- Based on the calculations of no. of in-bound links and out-bound links and some other great factors – Google's technique
- HITS Algorithm – Developed by Jon Kleinberg
- Trust Rank – A link Analysis Technique – Described in a paper by Stanford University and Yahoo Researchers
- No. of Hits per page

What are the major problems we are facing while searching for any content on the net?

- We may get spam pages with our desired search pages.
- And the content which we will actually need may be appeared in the fourth or fifth page to which we may not go.

What are the reasons of getting spam mails or spam pages while we searching for any content on the net?

- The Spam pages are appearing on the search engine top results because of their higher page rank values.
- We are getting spam emails because the sender of them either may want to increase the page rank for his website or he may want to advertise the items which he wants to sell.

What's Wrong with Spam?

Most spam messages on the Internet today are advertisements from individuals and the occasional small business looking for a way to make a fast buck. Spam messages are usually sent out using sophisticated techniques designed to mask the messages' true senders and points of origin. And as for your email address, spammers use a variety of techniques to find it, such as "harvesting" it from web pages and downloading it from directories of email addresses operated by Internet service providers (ISPs).

But spamming today could well be undergoing a revolution. Over the past year, AT&T, Amazon.com, and OnSale.com all have experimented with bulk email. Although the companies clearly identify themselves in the mail messages, these bulk mailings can cause many of the same problems as spam messages from less scrupulous individuals and companies.

Spammers often say that spam isn't a problem. "Just hit Delete if you don't want to see it." And many spam messages carry the tagline "If you don't want to receive further mailings, reply and we'll remove you." But spam is a huge problem. In fact, junk email and junk postings are one of the most serious threats facing the Internet today.

Spam messages waste the Internet's two most precious resources: the bandwidth of long-distance communications links and the time of network administrators who keep the Internet working from day to day. Spam also wastes the time of countless computer users around the planet. Furthermore, in order to deliver their messages, the people who send spam mail are increasingly resorting to fraud and computer abuse.

The Price Users Pay

It may take a spammer just five or ten minutes to program his computer to send a million messages over the course of a weekend. Now it's true that each of these messages can be deleted with just a click of the mouse, which takes only three or four seconds: a few seconds to determine that the message is in fact spam plus a second to click Delete. But those seconds add up quickly: one million people clicking Delete corresponds to roughly a month of wasted human activity. Or put another way, if you get six spam messages a day, you're wasting two hours each year deleting spam.

The price users pay for spam increases if you include the cost to the business or organization that operates the computer that holds your mailbox. These computers, called *mail servers*, require full-time connections to the Internet that can cost anywhere from $250 to $2,000 per month or more. The cost of the connection is determined, in part, by the amount of data it can carry. If a company's Internet connection is filled with spam, that company will be forced to spend more money on a faster Internet connection in order to handle the rest of its email traffic. Likewise, the company will be forced to buy faster computers and more disk drives. These costs must eventually be passed on to end users.

This scenario is not theoretical. In July 1997, spam mail overwhelmed AT&T WorldNet's outgoing mail system, delaying legitimate email by many hours.

Now before going to the actual solution for the above all problems let us observe the present Ranking system.

II. THE PRESENT PAGE RANK SYSTEM

Google describes Page Rank as:

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".

In other words, a Page Rank results from a "ballot" among all the other pages on the World Wide Web about how important a page is. A hyperlink to a page counts as a vote of support. The Page Rank of a page is defined recursively and depends on the number and Page Rank metric of all pages that link to it ("incoming links"). A page that is linked to by many pages with high Page Rank receives a high rank itself. If there are no links to a web page there is no support for that page.

Before moving into the description we come across some frequent questions while talking about the search engine providers like GOOGLE, YAHOO and MSN etc.

They are,

- How the page results are displayed in the search engine?
- What are the factors that affect the display of results?
- Does my results provided are actually based on its Rank?
- What are the different algorithms I can use for implementing the page rank for the web pages?
- Does my page rank technique handle my current Internet Traffic?

The answers for all the above queries can be found by the end of the below description on Page Ranking System.

Google assigns a numeric weighting from 0-10 for each webpage on the Internet. This Page Rank denotes a site’s importance in the eyes of Google. The Page Rank is derived from a theoretical probability value on a logarithmic scale like the Richter scale. The Page Rank of a particular page is roughly based upon the quantity of inbound links as well as the Page Rank of the pages providing the links. Let us see them in detail.

Simplified Page Ranking Algorithm

Page Rank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. Page Rank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided between all documents in the collection at the beginning of the computational process. The Page Rank computations require several passes, called "iterations", through the collection to adjust approximate Page Rank values to more closely reflect the theoretical true value.

A probability is expressed as a numeric value between 0 and 1. A 0.5 probability is commonly expressed as a "50% chance" of something happening. Hence, a Page Rank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 Page Rank.

How Page Rank Works

Assume a small universe of four web pages: A, B, C and D. The initial approximation of Page Rank would be evenly divided between these four documents. Hence, each document would begin with an estimated Page Rank of 0.25.

In the original form of Page Rank initial values were simply 1. This meant that the sum of all pages was the total number of pages on the web. Later versions of Page Rank (see the below formulas) would assume a probability distribution between 0 and 1. Here a simple probability distribution will be used- hence the initial value of 0.25.

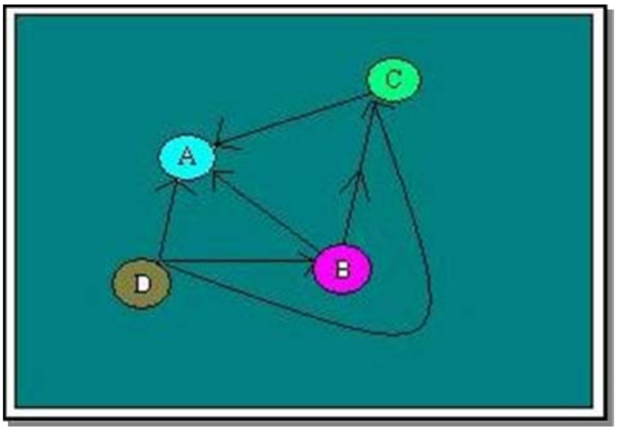


Fig. 1 Figure corresponds to the links in the WEB

If pages B, C, and D each only link to A, they would each confer 0.25 Page Rank to A. All Page Rank PR( ) in this simplistic system would thus gather to A because all links would be pointing to A.
This is 0.75.That is

$$\boxed{PR\ (A) = PR(B)+PR(C)+PR(D)}$$

Again, suppose page B also has a link to page C, and page D has links to all three pages. See the Figure 1 for correspondence. The value of the link-votes is divided among all the outbound links on a page. Thus, page B gives a vote worth 0.125 to page A and a vote worth 0.125 to page C. Only one third of D's Page Rank is counted for A's Page Rank (approximately 0.083).

$$\boxed{PR(A) =PR(B)/2+PR(C)/1+PR(C)/3}$$

In other words, the Page Rank conferred by an outbound link is equal to the document's own Page Rank score divided by the normalized number of outbound links L( ) (it is assumed that links to specific URLs only count once per document).

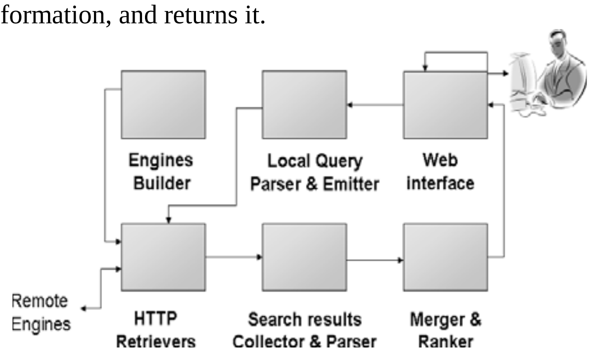
In the general case, the Page Rank value for any page u can be expressed as:

$$\boxed{PR(u) = \sum PR(u\ )/\ L(u)}$$

Now we will see the new addition to the above said ranking system to make it more accurate and reliable. I name it as “Improved Algorithm For Web Usage Mining for doctors on the cancer patients records”.

III. Improved Algorithm For Web Usage Mining for doctors on the cancers patients records

In this section we describe the architecture of our new search engine. The Web Interface allows doctors to submit their queries and to select the desired domains [related cancer] among those supported by the engine. This information is interpreted by the Local Query Parser & Emitter that re-writes queries in the appropriate format for the chosen domains. The Engines Builder maintains all the settings necessary to communicate With the remote search engines. The HTTP Retrievers modules handle the network communications. As soon as search results are available, the Search Results Collector & Parser extracts the relevant information, and returns it.



It is very difficult for the major search engines to provide a comprehensive and up-to-date search service of the Web. Even the largest search engines index only a small proportion of static Web pages and do not search the Web’s backend databases that are estimated to be 500 times larger than the static Web. The scale of such searching introduces both technical and economic problems. What is more, in many cases users are not able to retrieve the information they desire because of the simple and generic search interface provided by the major search engines. we are trying to eliminate this using the vertical search method.

It is easy to understand the working of the meta search engines. When an user sends a search query to the meta search engine, the query is automatically redirected to several search engines and the databases. The results are first aggregated and then displayed in order of their sources. The biggest benefit of the meta search engine is that it sends the results of all the search engines with only one search query and you do not have to repeat the search again and again. As the web is the largest database of information and there are lacks of new websites pouring in everyday thus a simple search engine won’t do now. Thus Meta search engines were integrated; they not only save time but also give better user oriented results for doctors.

Now we used the combined approach of both the domain specific(a particular cancer) and meta search. After gathering the domain constrained results from all the search engines in the results collector and parser they are sent to the merger and ranker module where in we propose a new Web Usage Mining Algorithm using the following factors like

- Common Results
- Sequence of the Produced Results
- Page Ranks of the Retrieved URL’S
- Frequency of search term in the gathered pages.

On the usage of the above key factors in our algorithm, we initially find out the common results which are retrieved in common from all the search engines. Then we will perform comparison of the sequence number of a particular URL in all the search engines. A short range between the sequence numbers in different search engines of a particular URL indicate us a better search result compared to that of one which are having a large range between the sequence numbers. It is then followed by evaluation of page rank given by various search engines. Finally we will be taking the count on how many time the search word in appearing in the pages.

Likewise including many other factors in our mining we implement this algorithm to alleviate more accurate and specific results.

IV Conclusion

The current major search engines are failing to provide ideal search in a number of ways for doctors on their intention to get some useful information regarding different cancers worldwide. They cover a relatively small proportion of the static Web pages, their indexes can be significantly out of date, they do not search they generally do not search the vast number of pages in the Invisible Web and can fail to provide sophisticated search when the doctor has a specialized category or topic of search in mind on a particular type of cancer. Specialized search engines alleviate these problems in a number of ways. They can search more of the Web and in a more up-to date fashion within their domain of different cancers. They can provide more search functionality, superior search in their domain versus the major search engines in terms of standard retrieval metrics and provide more structure search results. Ultimately the future of specialized medical search engines will be driven by technical and economic imperatives.

REFERENCES

1. Pavel Braslavski, Gleb Alshanski andengine , Anton Shishkin.2004 ProThes-Thesaures based Meta Search Engine for Specific Application Domain .
2. Patel, V. Adhil, M. ; Bhardwaj, T. ; Talukder, A.K. “Big data analytics of genomic and clinical data for Diagnosis and Prognosis of Cancer”, Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference
3. (CIMTA) 2013 "Home-PubMed-NCBI" , *United States National Library of Medicine (NLM)* [online] Available: www.ncbi.nlm.nih/ pubmed/
4. H. Fang and J. Gough, "The'dnet' approach promotes emerging research on cancer patient survival", *Genome Medicine 2014*, vol. 6, no. 64, 2014
5. L. Li, J. R. David, J. P. Chirag, C. W. Susan, C. Rong, P. T. Nicholas, T. D. Joel and J. B. Atul, "Disease Risk Factors Identified Through Shared Genetic Architecture and Electronic Media Records", *Sci Transl Med*, vol. 6, no. 234ra57, 2014
6. S. C. Pingle, Z. Sultana, S. Pastorino, P. Jiang, R. Mukthavaram, Y. Chao, I. S. Bharati, N. Nomura, M. Makale, T. Abbasi, S. Kapoor, A. Kumar, S. Usmani, A. Agrawal, S. Vali and S. Kesari, "In silicomodeling predicts drug sensitivity of patient-derived cancer cells", *Journal of Translational Medicine*, pp. 12-128, 2014
7. K. Goh, M. Cusick, D. Valle, B. Childs and M. Vidal, "The human disease network", *National Academy of Sciences*, vol. 104, no. 8685, 2007
8. W. Tian, L. Zhang, M. Taan, F. Gibbons and O. King, "Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function", *Genome Biology 9 Suppl*, vol. 1, no. S7, 2008
9. I. Ulitsky and R. Shamir, "Identification of functional modules using network topology and high throughput data", *BMC systems biology*, vol. 1, no. 8, 2007
10. N. Nagarajan, A. Tewari, J. O. Woods, I. S. Dhillon, E. M. Marcotte and U. M. Singh-Blom, "Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses", *PLOS one*, 2013