ABSTRACT SHEET FOR PAPER

Type of Paper (Fill the appropriate box with black colour)

- **Technical Paper** ■
- **Review Paper** □

CATEGORY: Engineering – Computer Science

TITLE OF PAPER: Comparing Search Methodologies by Varying Query Algorithm with a Search Engine Prototype

AUTHOR 1

NAME: **Pranav Siddharth S**

INSTITUTE'S NAME: Birla Institute of Technology and Science, Pilani

CONTACT NUMBER: +91-9865166625

EMAIL ADDRESS: pranavsid98@gmail.com

Workplace(s) involved and the work done there:

N.A

Supervisor/Scientist under whom the work was done:

N.A

Experiment work/simulation done: Yes

If yes, where and with whom:

Simulation done on my personal PC in the form of a Python program

Details about the experiment/simulation:

A Python script was developed, which would crawl the web and index the webpages by assigning PageRank along with other parameters. Another Python script was developed which functions like a mini search engine by using the indexed data of the previous script.

ABSTRACT:

Even slight nuances in any algorithm can substantially alter the results obtained. This paper deals with comparing the accuracy of search results generated by varying the query algorithm. This comparison primarily deals with the deviation of search results from an absolute reference value. The data for this study is obtained by executing a prototype programmed using Python. The data is used to identify and optimise the query algorithm by minimising deviation. In order to have a measure of variation, a formula for deviation was computed by using the parameters in hand. Our prototype is a python script which functions as a scaled down version of a search engine. This search engine script incorporates multiple versions of query algorithms to identify varying levels of

deviation. A web crawler was developed, which traverses URLs from a starting point and stores the required data in JSON format, which is used in the search engine script. All the crawled links are assigned a PageRank as well. Using 'Google' results as the absolute reference, we compare the deviations of a few other large scale search engines along with our prototype to have an idea of the magnitude of change. The results of this comparative study demonstrates a method that can be followed while developing the perfect query algorithm, the deviation hypothetically tending to zero.

KEYWORDS: Query, Algorithm, Search engine, Deviation, PageRank

REFERENCES: Books/articles/web page/journals/sources of study:

1) Sergey Brin, Lawrence Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine". Journal - Computer Networks and ISDN Systems. Elsevier. ISSN: 0169-7552(p). Pages: 107-117

2) The Beginners Guide to SEO – http://d2eeipcrcdle6.cloudfront.net/guides/Moz-The-Beginners-Guide-To-SEO.pdf

3) How Search Works - https://www.google.com/insidesearch/howsearchworks/

4) Scraping the web - http://web.stanford.edu/~zlotnick/TextAsData/Web_Scraping_with_Beautiful_Soup.html

5) How PageRank works - http://www.optimizationtheory.com/how-pagerank-works/

BACKGROUND AND SCOPE OF THE PAPER:

When Larry Page developed the PageRank theory for Google, it was the primary method used for ordering query results. Even though the query algorithm now depends on more than two hundred parameters which include Word count, Permutated multiple word count, Trust, Proximity etc., the PageRank remains a very important parameter. "PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites."(http://www.optimizationtheory.com/how-pagerank-works/) My paper explores the possibility of separating the parameters in the query algorithm and comparing the deviation from the reference point for each one. This would enable the identification of the impact of various parameters on the query results. The prototype developed uses the PageRank algorithm on top of varying parameters. This comparison, even though currently limited to a fraction of its potential due to hardware availability, would have immense scope for future development. This could also become a statistical reference point which will help to optimise the present query algorithm of large-scale search engines and help to, one day, develop a more efficient algorithm. Future plans include incorporating HITS algorithm and TrustRank algorithm to the comparison charts to obtain a wider spectrum of data.