**Comparison between Web Crawler using**

**Breadth-FirstSearch and Depth-First Search Algorithm**

**A perspective**

**By**

**Mohan Pothukuchi**

**For**

**CS 6200- Information Retrieval Coursework**

Contents

[Preamble: 2](#_Toc505618151)

[FIG.1 BFS\_Structure 2](#_Toc505618152)

[FIG. 2 DFS\_Structure 2](#_Toc505618153)

[URL Overlap: 2](#_Toc505618154)

[Perceived Quality , Efficiency and Coverage: 3](#_Toc505618155)

[FIG3. bfs-logger.txt 4](#_Toc505618156)

[FIG.4 – dfs-logger.txt 4](#_Toc505618157)

# Preamble:

Crawler, in this project, is developed using python programming language. All the web pages are stored in plain HTML text format and the output files can be seen below for each of the strategies:

A screenshot of a cell phone

Description generated with very high confidence

## FIG.1 BFS\_Structure

A screenshot of a cell phone

Description generated with very high confidence

## FIG. 2 DFS\_Structure

# URL Overlap:

From the above figures, we can see that the overlap between BFS Search and DFS Search. From the 999 URLs that are scraped in the depth-2,

* For DFS, only 12 links were crawled before 1000 URLs limit was reached
* For BFS, only 12 links were crawled before 1000 URLs limit was reached.

Based on the evidence, it is clear, that only 1 URL was common for both the strategies. And has ***8.3% overlap*** for the URLs crawled.

Also, based on the entire URLs in the bfs-urls.txt and dfs-urls.txt, the complete ***overlap between them 23.9%.***

# Perceived Quality, Efficiency and Coverage:

***bfs-logger.txt*** and ***dfs-logger.txt*** track the path the strategies are evolving into.

Below are the images which describe how the path has evolved before 1000 URLs are retrieved. Each crawl can be considered to understand the relevance of content to a solar eclipse.

In Figure 3, bfs-logger.txt depicts that it has crawled links very much understandable and relative to a solar eclipse. The links indicating the time frame of other solar eclipses, orbital status during the eclipse, people who have published papers on that topic are displayed.

In Figure 4, dfs-logger.txt depicts that the crawl links are not very much relative to the solar eclipse topic. Though there are few relations to the period of occurrence of the eclipse on a different planet, Passover of a solar eclipse, few other irrelevant details such Sparta and Invasion of Greece do not particularly point to direct correlation between the content.

From the understanding, the strategies have different behaviors because, in BFS, the crawl proceeds to the links derived the seed URL and until all the links are derived from the seed are traversed, the next set of URLs do not come into the picture. As the links go down the tree, it is clear, that relevance decreases. As BFS gives importance to all the links in the same level before proceeding to the next level, BFS had better relevant results.

While, DFS did not yield many relevant details when compared to BFS, because, in this strategy, the links to crawl go down to the end of the tree, before links on the same level are crawled. As said earlier, when we go down the tree of links, the relevance to seed URL decreases relatively. Hence, DFS results were not in context.

A screenshot of a cell phone

Description generated with very high confidence

## FIG3. bfs-logger.txt

A screenshot of a cell phone

Description generated with very high confidence

## FIG.4 – dfs-logger.txt