Name: Nada Mohamed Aboelaze ID: 20160472

**Crawling Techniques**

**Selective Crawling:**

Selective Crawling is the process to retrieve web pages based on some criteria. We are using a scoring function to determine relevant contents from a page. The fetched URLs are sorted according to a relevance score from the scoring function. Best-first search is the technique used to obtain pages with a high score first. This search leads to most relevant pages.

**Focused Crawling:**

Focused Crawling is the process to fetch pages within a certain topic. Crawler will download pages that are related to each other. It collects documents which are specific and relevant to the given topic and classify crawled pages into categories. It determines how far the given page is relevant to the particular topic and how to proceed forward.

Focused Crawling is also known as Topic Crawler because of the way it works. It is economically feasible in terms of hardware and network resources, it can also reduce the amount of network traffic and downloads. The search exposure of focused web crawler is relatively huge.

**Distributed Crawling:**

Distributed crawling is the process of partitioning of tasks or it is similar to distributed computing technique. A central server manages the communication and synchronization of the nodes, as it is geographically distributed. It basically uses Page Rank algorithm for its increased efficiency and quality search. The advantage of distributed web crawler is that it is robust against system crashes and other events. It is more scalable and memory efficient. Also have increased overall download speed and reliability.

Name: Nada Mohamed Aboelaze ID: 20160472

**Incremental Crawling:**

Incremental Crawling is the traditional crawling or the process of prioritizing and revisiting URLs, in order to refresh its collection, periodically replacing the old documents with the newly downloaded documents. The incremental crawler incrementally refreshes the existing collection of pages by visiting them frequently; based upon the estimate as to how often pages change.

It also exchanges less important pages by new and more important pages. It resolves the problem of content consistency. The benefit of incremental crawler is that only the valuable data is provided to the user, thus network bandwidth is saved and data enrichment is achieved.

**Parallel Crawling:**

Parallel Crawling is the process that runs multiple processes in parallel. That process is called C-procs which can run on network of workstations. The Parallel crawlers depend on Page freshness and Page selection. A Parallel crawler can be on local network or can be distributed at geographically distant locations. Parallelization of crawling system is very vital from the point of view of downloading documents in a reasonable amount of time.

Name: Nada Mohamed Aboelaze ID: 20160472

**Websites based on semantic web**

* **GetGlue**

## Flipboard

## Apture

#### Media Management – BBC

#### Web Search and Ecommerce