Assignment 1

CS 734: Introduction to Information Retrieval

Fall 2017

Hrishikesh Gadkari

Finished on October 3, 2017

1

Question

1. Suppose that, in an eort to crawl web pages faster, you set up two crawling machines with dierent starting seed URLs. Is this an eective strategy for distributed crawling? Why or why not? .

Answer

I think this is not a effective strategy for distributed crawling. Here the machines are unable to share information with each other. Although the two machines are seeded with different URLs, this will result in alot of duplicated effort, as the two machines would eventually crawl many of the same URLs. This strategy could be made more effective by sharing the URL request queue between the two machines, thereby eliminating the duplicated effort.

2

Question

2. How would you design a system to automatically enter data in to web forms in order to crawl deep Web pages? What measures would you use to make sure your crawlers actions were not destructive(for instance, so that it doesnt add random blog comments). .

Answer

The system would use two strategies. The first, called FTF (Filling Text Fields), is how to fill the fields efficiently, specially the text fields, which do not have a set of predetermined values. The second strategy, called ITP (Instance Template Pruning), is how to select queries to submit to a particular form in order to retrieve more data with fewer submissions. The strategy to minimize the set of queries, i.e., the number of form submissions, involves pruning the set of all possible queries. As each query is submitted, data extracted from the resulting page is used to identify wasteful queries and prune them.

3

Question

3. List five web services or sites that you use that appear to use search, not including web search engines. Describe the role of search for that service. Also describe whether the search is based on a database or grep style of matching, or if the search isusing some type of ranking.

Answer

1. Alexa : I use this website to montor the website traffic, statistics and analytics. It uses the grep style of matching.

2. Youtube: It is based on index-database style matching. YouTube is essentially a search engine for videos. Not surprisingly, it uses a sophisticated ranking algorithm to surface content to viewers. If you want to gain a following and rank your videos higher in YouTube search, uploading fresh content is extremely important.

3. Amazon: It is based on grep+ ranking style search. Amazon Searches on Amazon are not based solely on keywords. Amazon's search is more of a marketing tool; displaying more popular and successful products over those that are less, even if the search words used better match new product pages. If you plan on just creating a new product page for your items and expect sales, that is not going to happen unless you invest time in marketing your items. As your product generates more page views, sales and product reviews, your product will rise within the search results. Also, you should try and see the difference within the results when searching 'All' as compared to searching within the category. You will likely see your item ranked higher when doing 'in-category' searches.

4. https://www.worldcat.org/ : I use this website to find items in libraries near my loaction. It uses grep style matching. Anything that is a product or catalogue search is a grep-like search. The thing about grep style search is that its not very efficient, its normally sequential and you have no progress idicator (generally speaking).

5. www.odu.edu. For on-site searches it uses grep style of matching .

4

Question

4. Think up and write down a small number of queries for a web search engine. Make sure that the queries vary in length (i.e., they are not all one word). Try to specify exactly what information you are looking for in some of the queries. Run these queries on two commercial web search engines and compare the top 10 results for each query by doing relevance judgments. Write a report that answers atleast the following questions: What is the precision of the results? What is the overlap between the results for the two search engines? Is one search engine clearly better than the other? If so, by how much? How do short queries perform compared to long queries?

Answer

The folllowing queries were run on Google and Bing web search engines. 1. Query name: Washington - In this particular query I wanted to know about the Washington state. The overlap between the results for the two search engines is 3.

Search Engine	**Precision**
Google	0.59
Bing	0.33

Table 1: Precision results for the query

2. Query name: Top 10 queries to test precision - In this query I wanted to know about the queries which can be used to test precision on different search engines. The overlap between the results for the two search engines is 1.

Search Engine	**Precision**
Google	0.49
Bing	0.28

Table 2: Precision results for the query

3. Query name: How to become a successful businessman? - In this query I was looking for the qualities to be a businessman. The overlap between the results for the two search engines is 6.

Search Engine	Precision
Google	0.7
Bing	0.64

Table 3: Precision results for the query

4. Query name: Chinese dishes - In this query I was looking for some good chinese delicacies. The overlap between the results for the two search engines is 4.

Search Engine	Precision
Google	0.69
Bing	0.57

Table 4: Precision results for the query

5. Query name: Chinese dishes a person should definitely have in his life - In this query I was looking for some good chinese delicacies. The overlap between the results for the two search engines is 0.

Search Engine	Precision
Google	0.4
Bing	0.33

Table 5: Precision results for the query

From the precision results of the above queries, I think that Google search engine is better than Bing search engine. Also the short queries perform well than the long queries.

5

Question

5.Site search is another common application of search engines. In this case, search is restricted to the web pages at a given website. Compare site search to web search, vertical search, and enterprise search.

Answer

1. Sometimes we need something so insanely specific that a general keyword phrase doesn't really do it for you, especially for marketers on the hunt for a specific piece of content. Enter the Google site:search. This allows you to search just one domain – not the entire internet – for a particular search term. So, for instance, if you wanted to see what kind of content HubSpot, and only HubSpot, had on marketing automation, doing a site:search would limit the results to only HubSpot's marketing automation content. 2. Site search can be contrasted with web search, which applies search technology to documents on the open web, enterprise search, which involves finding the required information in the huge variety of computer files scattered across a corporate intranet and vertical search, which focuses on a specific segment of online content.