INTRO. TO INFO RETRIEVAL: CS 734: A1

Due on Thursday, September 21, 2017

Dr. Nelson

Udochukwu Nweke

Contents

Problem 1

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 at least 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?

Solution 1:

In order to compute precision of queries from different search engines, I have choosen the following three queries:
Query 1: When is the next UEFA champions league competition? My expectant result is the date of the next UEFA champions league competion.

Item	Google Search Engine Result	Bing Search Engine Result
1.	http://www.uefa.com/uefachampionsleague/...	https://en.wikipedia.org/...
2.	http://www.uefa.com/uefachampionsleague/...	http://www.uefa.com/...
3.	https://en.wikipedia.org/.../	http://www.uefa.com/...
4.	https://en.wikipedia.org/...	https://en.wikipedia.org/wiki/...
5.	https://en.wikipedia.org/wiki/...	https://www.ft.com/...
6.	http://www.foxsports.com/...	https://www.nytimes.com/...
7.	http://www.foxsports.com/...	https://www.playstation.com/...
8.	https://play.google.com/store/apps/...	https://ide.uefa.com...
9	https://www.premierleague.com/european...	https://www.premierl-...
10.	http://bleacherreport.com/uefa-...	http://larrybrownsports.com/s...

Query 2: Hurricane maria 2017 updates. My expectant result is the affected cities.

Item	Google Search Engine	Bing Search Engine
1.	https://www.nytimes.com/...	http://www.cnn.com/...
2.	http://www.telegraph.co.uk/...	https://www.cbsnews.com/...
3.	https://www.theguardian.com/...	http://www.cnn.com/...
4.	https://www.cbsnews.com/...	http://coed.com/...
5.	http://www.cnn.com/...	5.http://www.npr.org/...
6.	http://www.cnn.com/...	http://www.businessinsider.com/...
7.	http://www.express.co.uk/...	http://coed.com/...
8.	https://www.wunderground.com/...	https://www.cbsnews.com/...
9	https://www.vox.com/...	https://www.nytimes.com/...
10.	http://www.businessinsider.com/...	https://reliefweb.int/report/...

Query 3: DACA.
I have included the complete links in querylinks.txt file.
Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances. In the case, precision will be the fraction of relevant links among retrieved links.

Problem 1 continued on next page...

Item	Google Search Engine	Bing Search Engine
1.	https://www.uscis.gov/archive...	https://www.uscis.gov/archive/...
2.	https://en.wikipedia.org/wiki/...	https://en.wikipedia.org/wiki/...
3.	https://undocu.berkeley.edu/...	http://www.cnn.com/2017/09/04/...
4.	http://www.bbc.com/news/av/...	http://www.immigrationequality.org/...
5.	http://www.latimes.com/local/...	https://www.nilc.org/issues/daca/
6.	http://www.nbcbayarea.com/...	https://undocu.berkeley.edu...
7.	http://thehill.com/latino/351365-...	http://www.npr.org/2017/..
8.	http://www.politico.com/story/...	https://www.uscis.gov/...
9	http://time.com/daca-dream-...	https://www.theguardian.com/...
10.	https://www.nytimes.com/...	http://www.politico.com/...

Precision = $^{\text{relevance instances}}/_{\text{retrieved instances}}$

Overlap = $^{\text{intersection}}/_{\text{minimum between both sets}}$

For query one which is : When is the next UEFA champions league competition? The relevant result should have the dates the next UEFA champions league will be played.

Precision for Google search engine result for query 1 is: 8/10 = 0.8
where 8 is the number of relevant links from Google search engine, 10 is the first 10 queries retrieved

Precision for bing search engine result for query 1 : 7/10 = 0.7

Overlap for query 1 is: 4/10 = 0.4

The two links that are not relevant to me are : https://en.wikipedia.org/wiki/2017%E2%80%9318_UEFA_Champions_League and https://en.wikipedia.org/wiki/UEFA_Champions_League.
This is because these queries did not contain my expected result.
Links that appeared in both Google search and Bing search are:
1.http://www.uefa.com/uefachampionsleague/index.html#/
2.https://en.wikipedia.org/wiki/UEFA_Champions_League
3. https://www.premierleague.com/european-qualification-explained
4. https://en.wikipedia.org/wiki/2017%E2%80%9318_UEFA_Champions_League_qualifying_phase_and_play-off_round

Precision for query 2:
Google Search Precision: 10/10 = 1.0

Bing Search Precision: 10/10 = 1.0

Overlap = 4/10 = 0.4

The two search engines gave relevant results for query 2:
Links that appeared both on Google and Bing search engines are:
1. https://www.nytimes.com/2017/09/20/us/hurricane-maria-puerto-rico.html?mcubz=1
2. https://www.cbsnews.com/news/hurricane-maria-category-2-path-latest-track-models-2017-09

3. http://www.businessinsider.com/hurricane-maria-path-track-update-2017-9
4. http://www.cnn.com/2017/09/19/us/hurricane-maria-latest/index.html

Precision for query 3:

Google search Precision 10/10 =1.0

Bing Search Precision 10/10 = 1.0

Overlap = 3/10 =0.3
Both search engines retrieved Trump's move to end DACA.
Links that appeared both on Google and Bing search engines are: 1. https://www.uscis.gov/archive/consideration-deferred-action-childhood-arrivals-daca 2. https://en.wikipedia.org/wiki/Deferred_Action_for_Childhood_Arrivals 3. http://www.politico.com/story/2017/09/03/trump-dreamers-immigration-daca-immigrants-242301

Based on the result of the 3 queries, I cannot clearly say that any search enine is better than the other. Although google gave a higher precision for my first query, the difference is not much.
Short queries retrieve more results that long queries.

Problem 2

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.

Solution 2:

Site Search
Site search is an application of information retrieval that is designed to retrieve information within a website. In site search, the users intention is to filter a particular website and get a specific result. An example will be an online store domain where a user uses the website to retrieve a particular product. User Intention: In site search, a user query is restricted to the website with the intention of finding specific information. The purpose of site search is to direct the search to a specific topic and get a filtered result within a website

Web Search
Web search is the most common application of information retrieval. In web search, a user issues a query to the search engine and the result comes back in form of documents in a ranked order. The result of a web search is not restricted to the query, but other topics. An example will be searching for "protein and DNA interaction" on Google or any other search engine. The result of this query will range from protein and DNA interaction to molecular biology, TF-DNA binding, etc. leaving the user with the responsibility of deciding what is relevant.
Site search and web search aim at providing relevant result to a user and results are both returned in a ranking order. The difference between site search and Web search is that site search is restricted to a given website while web search searches across the search engine and returns results as a ranked document. In site search, the users search is specific and the result is also specific.

Enterprise Search
An enterprise search is the application of information retrieval used in identifying and allowing definite content across the enterprise to be indexed, searched, and displayed to authorized users. In enterprise search,

the user tries to retrieve an information from several computer files scattered across corporate intranet. Any content that cannot be accessed is useless. With this in mind, an organization tries to provide an easy access to an information in any format, especially for the employees of a particular organization.
Site search and enterprise search provide relevant result to users, they are both information retrieval applications. The major difference between enterprise search and site search is security. In enterprise search, information is only available to those that need it and have been granted the privilege. The purpose of an enterprise search is to provide organizations information to the employee that will need the information
Enterprise search searches across a local network. Site search, searches within a website.

Vertical Search
Vertical search is an application of information retrieval where the domain of the search is directed to a particular topic. It is similar to web search except that the serach engine is restricted to a specific category. The information on the web is increasing enormously and the content seems unlimited. Vertical search will save users the huddle of searching the entire web for a particular query that can be gotten from the specialized domain. For example, if a user that wants to book a flight, instead of going through Google search engine and getting thousands of response, leaving the user with the difficulty of finding the particular airline and time they want, the user can search for airline domains and get specific results.

Vertical search is similar to site search in terms of specialization. In both search applications, the user's goal is defined.
The difference between site search and a vertical search is that site search is done within the website while the focus of vertical search is on a topic.

Problem 3

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

Solution 3:

The purpose of setting up multiple crawling machine is to reduce load on a single crawling machine. Setting up multiple crawling machines and starting them with different seed URLs is equivalent to setting up multiple single machines because, these two machines are not interacting with each other, they are not helping each other. They are doing totally different jobs. There is no division of labor. It still leaves each machine with the responsibility of ensuring that there are no duplicate links in the queue, remembering the links to be crawled, checking the links crawled against the links in the queue. This will paralyze the crawlers speed. Distributed means the machines have to communicate with each other and they also share the crawling load.

Problem 4

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 is using some type of ranking.
Solution 4:

Five websites that I use that appear to use site search are:

1. `https://www.odu.edu/directory`
2. `https://www.amazon.com/`
3. `https://www.ebay.com/`
4. `https://www.kayak.com/tracker`
5. `http://www.ikea.com/us/en/`

The role of search for these sites is to help users get specific result. It also helps the site owners to plan around what users are interested in when they visit the website (for e-commerce web sites).

1.The ODU student directory uses grep style of matching. searching for a student with last name "John" not only returned student with "John" as last name, but "Johnson", "Johns" etc. The image in Figure 1 shows the search result of "John" from `https://www.odu.edu/directory`. The rankings are done alphabetically, names are displayed in alphabetical order.

	Name*	Department	Contact Info
	Kimberley Johnson	Academic Affairs	kajohnso@odu.edu
	Heide Johnson		hjohnson@odu.edu
	Amy Johnson	Nursing	2152 HEALTH SCIENCES BLDG 757-683-4297 aejohnso@odu.edu
	Amanda Johnson	Housing & Residence Life	1208 VIRGINIA HOUSE 757-683-4283 ajohnson@odu.edu
	Vista Johnson	STEM Education & Professional Studies	228 EDUCATION BUILDING 757-683-4305 vgjohnso@odu.edu
	Freda Johnson		1004 ROLLINS HALL 757-683-3603 fashley@odu.edu

Figure 1: student search result

2. The second website is `https://www.amazon.com/`. This website also uses grep style of matching. It

Problem 4 continued on next page...

displays products by matching substrings. Amazon website ranks its products based on relevance . Figure 2 shows a search performed with "shoe" as a query and the result is a different shoe information based on grep style matching.

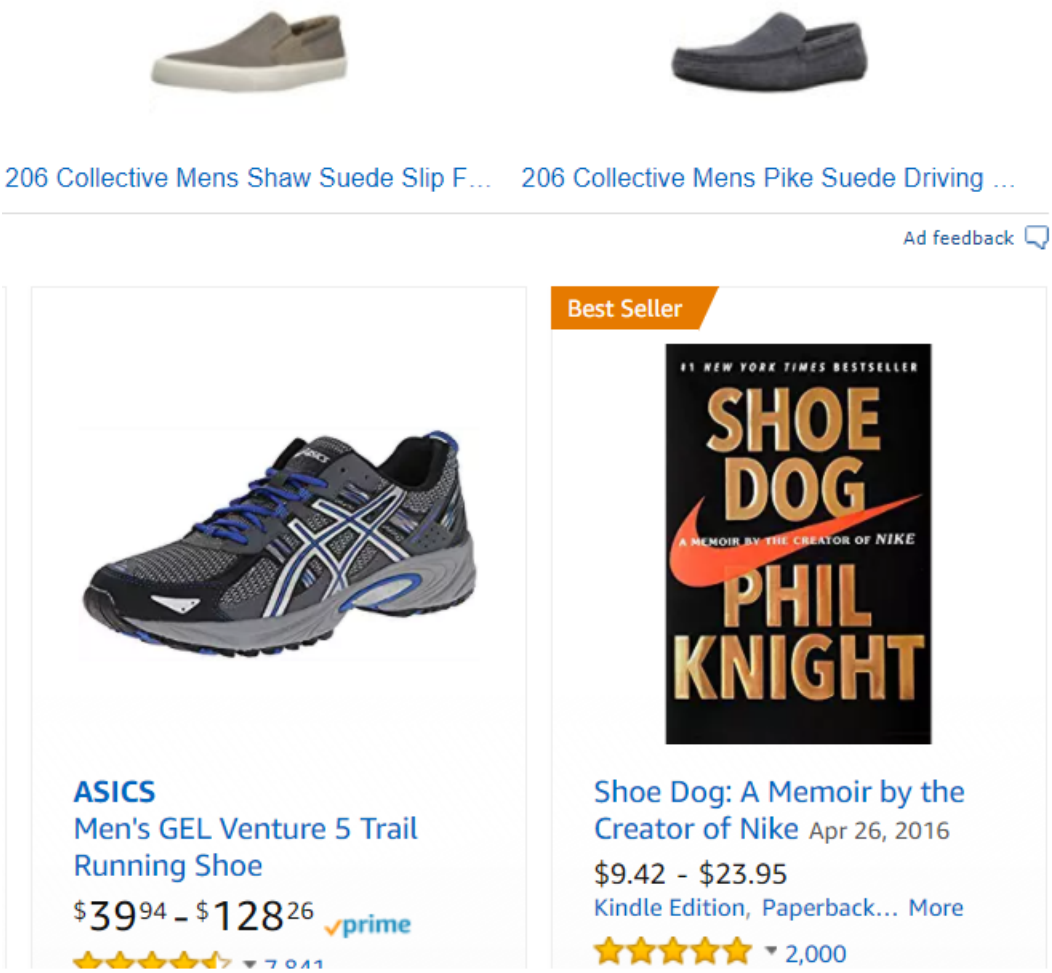

Figure 2: product search result

3. The next website is `https://www.ebay.com/`. This site also uses grep style of matching and ranking is done based on category.

4. This kayak flight tracking website uses database style of matching. Here, the user searches with a flight number and this number is matched with the database table to retrieve relevant result. Because search is done with a number, there are no substrings to match. Figure 3 shows how flight is tracked by filling out the relevant field. Figure 3 gives the itenary of an American Airline flight with flight number 2192.

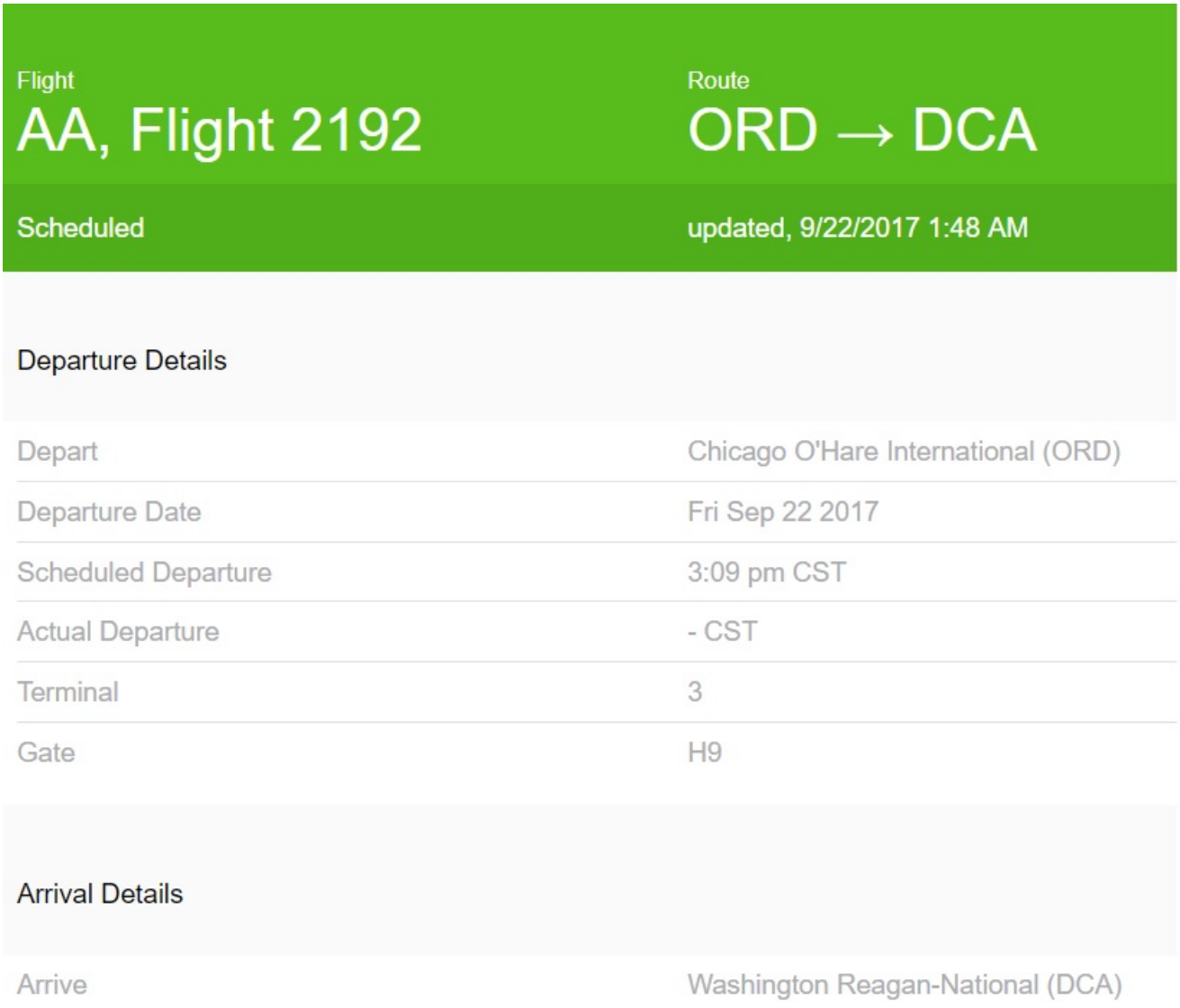

Figure 3: flight tracker

5. The last website is `http://www.ikea.com/us/en/`. This website uses grep style of matching and products are ranked based on relevance.

Problem 5

Give a high-level outline of an algorithm that would use the DOM structure to identify content information in a web page. In particular, describe heuristics you would use to identify content and non-content elements of the structure.
Solution 5:

Algorithm title: Extract text content
Input : HTML Page, P (Collection of HTML tags)
Output : Text Content, P

Description: Using a predefined list L, of HTML tags, that are known to contain text for HTML pages, extract all tags of kind same as L from the HTML page. Then if the tags do not contain elements inside, extract text enclosed by tags: $< tag >$ Text to extract $< /tag >$. If the tag contain other tags inside, recursively visit every element in the tag until you reach the final tag, then extract text.

```
Given L,
   L = [div, span, p, h1, h2, h3, h4, title, a, strong, em, table, etc.]
   Procedure extractContent(HTML page P)
       G = extract all tags from P that is in L

       for tag t in G do
           if t has element then
               extractContent(t)
           else
               T.add(t.text content)
           endif

       endfor
   endProcedure
```

References

[1] Aiim. http://www.aiim.org/What-is-Enterprise-Search#. Accessed: 2017-20-09.

[2] Trevor Strohman Bruce Croft, Donald Metzler.