Assignment 1

Fall 2017
CS834 Introduction to Information Retrieval
Dr. Michael Nelson

Orkun Krand

September 21, 2017

1 Question 1.1

1.1 Question

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?

1.2 Methodology

Google and Yahoo! will be used for this experiment. I will be using the incognito mode in Chrome. By now, we have all heard of Psy, the Korean pop star. He is the singer of "Gangnam Style" which is the first video on YouTube that exceeded YouTube's upper view limit[1]. After his success, he collaborated with Snoop Dogg in 2014 to make a new song called "Hangover". With the search terms I use, I will actually be trying to get the link to the music video of this song. The following search queries were issued to the two search engines:

1. korean pop star hangover
2. psy drunk
3. gangnam style snoop dogg
4. drop it like it's hot singer psy song
5. psy snoop dogg

1.3 Analysis

Both search engines returned similar results. Only one of the search terms (number 4) failed to produce the result I was looking for.

1.3.1 korean pop star hangover song

Both search engines were able to figure out that I'm looking for the song but Yahoo! search's video results didn't show the song's music video, even though the first result after the video results was the link to the music video. Google's other results were for articles written about the song.

Yahoo! on the other hand showed more irrelevant results than Google including a news article on Yahoo! that was about a Korean convenience store. There is another article about K-pop stars and drugs. So if we think of a score system where the number of links that are related to what I wanted to find in any way receive a point, Google will receive 10 points while Yahoo! receives 5. Those 5 good results even include Wikipedia pages for the singer and the song. Interesting enough, the image results for this search on Yahoo! are all screenshots from the music video I'm looking for.

Google did a better job not only presenting relevant results but also ordering the results in a useful way where the music video is number 1 followed by the Wikipedia page for the song. The image results on Google seem to have more to do with Psy than the song.

1.3.2 psy drunk

Since hangover comes after drinking, I figured 'drunk' is a close enough term that should produce good results. However, good results were hard to find in Google. Funny enough, Google realized that I was looking for the song and showed me its music video/information box before the search results. The problem is that the YouTube link it showed was a link to something called "PSY IS DRUNK AND NAKED IN A TRAIN - Gangnam STYLE..." which was a video of a picture of an Asian man lying on the floor of a train, sleeping. It wasn't even Psy. It had the lyrics of the song I wanted and more information about it but the link was off by miles. Image results were all screenshots from the music video. The other videos were a live show of Psy and another video that seemed like spam. The rest of the results are about interviews with Psy. Based on the scoring system discussed earlier, Google gets 2 points out of 10.

Yahoo! on the other hand was better and worse than Google at the same time. It showed the link to the video as the first text result, had 2 links to the music video in the video results but after that, lost its track and figured I meant 'psychology' when I typed Psy and therefore showed me a bunch of websites that had to do with the psychology of drunk people. Yahoo! gets a 2 out of 10 as well not including the video results. But it also totally missed the point with the whole psychology spree.

I understand that this is a long shot when we think of search queries. Either way, I expected Google to have a higher score. Showing interviews about Psy makes sense since I didn't have anything that referred to Snoop Dogg and apparently Psy has a drinking problem.

1.3.3 gangnam style snoop dogg

Apparently, this was an easy query since both engines got 10 out of 10. Even the first 4 results were the same. Both search engines were able to associate Gangnam Style with Psy. I guess gangnam style is a very unique combination of words, especially considering I'm doing these searches from the United States.

Interestingly for this query, Yahoo! didn't choose to show me a tab of video results but instead, it showed me youtube links as text results. Google was on point with the main music video as the big video link, and the other 2 videos with thumbnails.

1.3.4 drop it like it's hot singer psy song

I figured if Gangnam Style can be linked to Psy, drop it like it's hot should be linked to Snoop Dogg, I even added the word song at the end to help the search engine out. Both engines showed me music videos for 'Drop It Like It's Hot' by Snoop Dogg and not the music video I was looking for. Google was able to show me two articles about Psy, one of which is about the song 'Hangover' and the other is an article written before 'Hangover' came out. It just says Gangnam Style's beat sounds like a fusion of Snoop Dogg's 'Drop It Like It's Hot'. The one link that is related to 'Hangover' was number 7 on the top 10 results.

There might be a couple reasons for this. For example, the search engines might have decided to ignore the word ‘Psy’ because it is not in the dictionary. Or maybe it has to do with Snoop Dogg’s song having a very long name. Assuming search engines prioritize the words in the order they were put on, it’s normal for the engines to ignore ‘Psy’ since it is only one word out of 7. Weird enough, Google didn’t bother to show the official music video for ‘Drop It Like It’s Hot’ either.

1.3.5 psy snoop dogg

This was my safe bet. I figured this would absolutely find me the music video and it did since these two singers never collaborated on anything else before or since ‘Hangover’. I was right, the first result was the link to the YouTube link of the music video. Google included the Wikipedia link for the song while Yahoo! didn’t. Since this was the safe bet, I think the performance of the search engines on this query easily shows which is the better search engine.

1.4 Results

It is safe to say that Google is the better search engine but Yahoo! wasn’t as bad as I was afraid. I would say it is a good alternative to Google if Google shuts down some day. I never realized how nice it is to not have ads on a search result page until I tried Yahoo! even though the ad is for the song. It received higher relativity scores than Yahoo! on all search results, even though Yahoo! claims to have found more results than Google (517,000 vs. 2,470,000 for the last query)

2 Question 1.2

2.1 Question

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.

2.2 Methodology

Search Engines: Information Retrieval in Practice[2] was used to answer this question.

2.3 Answer

Site search, as stated in the question, is searching a query on a single website. Forum searches can be an example of this. When I’m searching why my 1978 Honda CB400T dies after riding on the highway, searches on a variety of motorcycle forums is not ideal since I want the opinion of people who have been dealing with the same type of bike that I have, especially since carburetor technology has vastly advanced in the past 40 years and is almost not used anymore. So a site search allows me to only search within the HondaTwins forum, disregarding opinions of Kawasaki Ninja owners.

Web search is easily done using a search engine which searches the whole web (not just one website) for websites that have something to do with the search query. The search engines are rated based on how well they retrieve what the user is looking for. The more a search engine can read the user’s mind, the better it is considered. Because web search requires searching many websites, it is a much more cumbersome job than site search.

Vertical search is a search for results in the same topic. Indeed is a good example for vertical search. Indeed searches for terms in job postings from various company websites. It searches many websites so it is a search engine but all it finds are job postings, so all the results are about a single topic which makes it vertical search. Vertical search is similar to site search in that it has a smaller domain than web search. But site search is still working with a much smaller pool of data.

Enterprise search is searching the database or databases of a company intranet. For example when you go to your auto parts store, and ask them for spark plugs for a 1966 Ford Galaxie, assuming their database contains that information, they will be able to show you the spark plugs they have in stock, and ones they can order. If you decide to pick up the NGK ones that are in stock, the salesperson will then look up its location on their screen and then go into the back to find that part and bring it for check out. This type of search is similar to site search in that the pool of data is much smaller than vertical or web searches, which is supposed to provide faster processing times even though my local auto parts store's system never seems to work right. Must be something the programmers are missing.

3 Question 1.4

3.1 Question

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.

3.2 Answer

3.2.1 YamahaFZ1OA

Yamaha FZ1 Owners Association is a forum for Yamaha FZ1 owners. It's a platform where we discuss anything and everything about the FZ1. The website also features a search which I've used on many occasions. It is placed so people search for answers to their questions that may have already been discussed rather than asking the same thing all over again. For example, Ivan's ECU flash has been bought and used by many members and naturally, there are many threads already available discussing it in depth with input from Ivan himself. Creating a new thread asking for information about Ivan's flash is unnecessary if users can do a search.

This website uses a grep style search where it will bring up all the threads that have the query you searched for ranked by date (newest first) and upon clicking on one of the threads, will highlight your query for easier navigation.

3.2.2 Spotify

Spotify is a music streaming service that lets users play any songs or playlists they want. You can search for a specific song, artist, album or even a playlist that was created by another user. Spotify does a database search, shows the results in categories (song, artist, album...) and ranks them using crowdsourced votes and machine learning so when you type ‘e’, it can decide to show you Elvis or Eminem. They also use a system called search-as-you-type[3] which basically updates the result as the user types without requiring to submit the query once done typing.

3.2.3 Telegram

Telegram Messenger is a messaging app similar to Facebook Messenger and WhatsApp. It features two types of search on their website/application. I can do a search within a chat, and I can do a search amongst the people I’ve been chatting with or anybody on the application. For example, I talk to my girlfriend on Telegram often and I can find all the times we said ‘motorcycle’ simply by typing it in the search box. It would show newest first. It uses a grep style search and just shows every message where anybody in the conversation used the term ‘motorcycle’.

3.2.4 Revzilla

Revzilla is one of the biggest motorcycle accessory and gear retail websites I know of. They only have an online store so search is very important for them. Any time someone can’t find what they are looking for, they lose money. They don’t have any fancy systems like search-as-you-type or showing little thumbnails of products as you type. But if I type FZ1, I can find all the parts they have that fit my bike even if it doesn’t have FZ1 in the name (ex. Yoshimura TRC Carbon Fiber Exhaust) and most times, that’s all I want. I use Revzilla mostly for safety gear which is less specific and typing ‘Shoei’ in the search box brings me helmets ranked by popularity, followed by helmet accessories. It does a database search to see what helmets they have, I can pick models, colors, year and type of helmet once the initial results come back.

3.2.5 Partzilla

Partzilla is an online OEM and aftermarket parts seller. Similar to Revzilla, Partzilla only has an online store as well. They sell parts for lawn mowers, motorcycles, jet-skis, snowmobiles and so on. Their search function is a little primitive unless you can narrow down what you are looking for very well. I can look up part numbers, but along with the part I want, it would show me parts whose part numbers closely match my query which is a bad idea when it comes to OEM parts since manufacturers use similar part numbers for similar parts. If I’m searching for a brake caliper for a 2012 Yamaha Super Tenere and I put the part number in the search box, search results also show me similar part numbers, which are generally brake calipers that fit other bikes or sometimes ATVs which is useless to me. Especially since they don’t immediately show what vehicles those other parts fit.

The search box is not their strong suit but then again, nobody really uses it. Everyone wants to look at the OEM parts fiche for their vehicles and select parts from there. That’s why this is a great example of a badly developed search box. Trying to search the term ‘fz1 brake caliper’ brought me an aftermarket part that fits early 1980s Harley Davidson Electra Glides.

4 Question 3.8

4.1 Question

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?

4.2 Answer

This is a good initial strategy, but it is not enough to make sure the two crawlers work on different pages. For example on Wikipedia, if a user starts on any page and keeps clicking on the first hyperlink, they eventually end up in the *philosophy* page. So different starting points is a good idea, just not enough.

On the other hand if these crawlers can communicate and keep track of which pages have been crawled, then that would be a good strategy that reduces runtime. If as soon as a crawler starts working on a page, it can add that page to a list, then the other crawler would know not to access that page since it is already being crawled. The other crawler can spend it's valuable time working on other pages.

5 Question 3.9

5.1 Question

Write a simple single-threaded web crawler. Starting from a single input URL (perhaps a professor's web page), the crawler should download a page and then wait at least five seconds before downloading the next page. Your program should find other pages to crawl by parsing link tags found in previously crawled documents.

5.2 Methodology

I used Python to write the crawler, assuming by download a page, they don't mean create a local copy. My code gets a page, and then looks for all `<a>` tags and tries to get the href attribute of those tags to add to its yet to crawl set. I chose to ignore *IOErrors* since some websites don't allow crawlers to crawl their sites and I don't want that stopping my crawler from moving on.

5.3 Code

```
from bs4 import BeautifulSoup
import urllib, requests, sys
from urlparse import urlparse
from time import sleep

if len(sys.argv) != 2:
    print 'Usage: python crawler.py [url]'
    sys.exit(1)
starturl = sys.argv[1] #get start url
print "starting crawler with " + starturl
if urlparse(starturl).netloc == '':
    starturl = "http://www.cs.odu.edu/~mln/"
    print "Bad url, crawling www.cs.odu.edu/~mln/"


def crawl(crawled, tocrawl):
    while len(tocrawl) > 0:
        url = tocrawl.pop()
        if url not in crawled:
            print "Crawling " + url
            crawled.add(url)
            try:
                res = urllib.urlopen(url)
                html = res.read()
                soup = BeautifulSoup(html, "lxml")
            except IOError:
                continue
            links = soup.find_all('a')
            for each in links:
                try: #skipping unicode errors
                    newlink = str(each.get('href'))
                except:
                    continue
                if newlink not in tocrawl:
                    if urlparse(newlink).netloc != '' and urlparse(newlink).scheme != '':
                        tocrawl.add(newlink)
                        print "Adding " + newlink

            sleep(5)

tocrawl = set() #set that holds links not crawled yet
tocrawl.add(starturl)
crawled = set() #set that holds the crawled links
crawl(crawled, tocrawl)

print "Crawled " + str(len(crawled)) + " sites"
```

References

[1] “Gangnam Style Music Video ‘Broke’ YouTube View Limit.” *BBC News*, BBC, 4 Dec. 2014, www.bbc.com/news/world-asia-30288542.

[2] Croft, William Bruce, et al. *Search Engines: Information Retrieval in Practice.* Pearson, 2010.

[3] Isaakson, Marcus. “Personalizing Spotify Search By Learning To Rank.” *YouTube*, 20 Sept. 2017, www.youtube.com/watch?v=IFRKLBuuXCA.