Assignment 5 - TDT4117

Hermann Owren Elton, Olaf Rosendahl

November 04, 2022

Task 1: Precision and recall

a) The information retrieval system should provide a user interface with lots of opportunities for an advanced search, but also make it easy to make a search fast. This way, one are able to quickly try and find relevant results, but also make an more advanced search with larger possibility of finding good results. An eCommerce company looking for previous campaigns might be looking for images (posters), while a researcher might be more interested in PDF's and articles. The system should for example allow the user to select what types of content they are looking for.

b) Operations such as stemming, removal of commonly used words i.e stop-words, lexical analysis where you remove stuff like punctuation, and letter casing, index term selection where you group nouns or give more weight to more relevant nouns in the document, and Thesauri where you give more weight to words that are relevant to the documents domain, and also find synonyms.

Task 2: Page rank and HITS

a) Both approaches uses the idea of using the number of inbound and outbound links to a page as a measure of quality for a page. In HITS pages with a lot of outbound links are called hubs, while pages with many inbound links are called authorities.

PageRank is an evolution of HITS which also takes into consideration the rank of the page at the other side of the link. For example, if a page is linked to from a low ranked page, the link does not contribute as much to the PageRank as it would have done as if it was linked from a high ranked page.

Another difference is that for a given query HITS invokes traditional search engine to find a set of relevant pages and then attempts to find hubs and authorities. PageRank on the other hand is computed at crawl-time and then combined with traditional information retrieval score at query time. This makes HITS much slower and not feasible for today's search engines which handles billions of queries every day compared to PageRank's much greater efficiency.

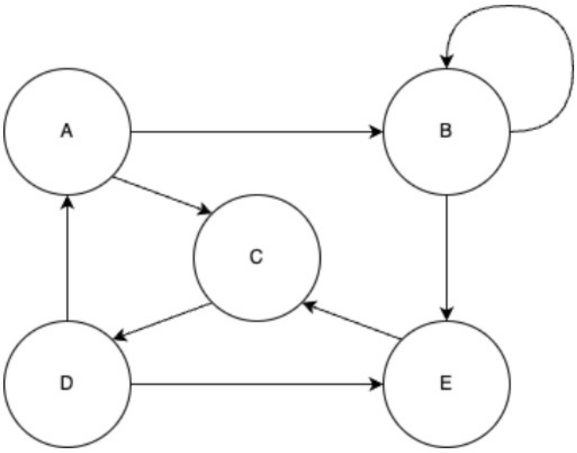

Figure 1: Graph of websites connected by links.

b) *1. iteration:*

Page	Initial		1. iteration	
	Auth	Hub	Auth	Hub
A	1	1	$D_{Hub} = 1$	$C_{Auth} + B_{Auth} = 2 + 2 = 4$
B	1	1	$A_{Hub} + B_{Hub} = 1 + 1 = 2$	$B_{Auth} + E_{Auth} = 2 + 2 = 4$
C	1	1	$A_{Hub} + E_{Hub} = 1 + 1 = 2$	$D_{Auth} = 1$
D	1	1	$C_{Hub} = 1$	$A_{Auth} + E_{Auth} = 1 + 2 = 3$
E	1	1	$B_{Hub} + D_{Hub} = 1 + 1 = 2$	$C_{Auth} = 2$

3 iterations:

Page	Initial		1. iteration		2. iteration		3. iteration	
	Auth	Hub	Auth	Hub	Auth	Hub	Auth	Hub
A	1	1	1	4	3	14	10	49
B	1	1	2	4	8	15	29	54
C	1	1	2	1	6	1	20	1
D	1	1	1	3	1	10	1	35
E	1	1	2	2	7	6	25	20

c)

Page	Initial		1. iteration		2. iteration		3. iteration	
	Auth	Hub	Auth	Hub	Auth	Hub	Auth	Hub
A	1	1	1	3	3	8	8	21
B	1	1	1	2	3	5	8	13
C	1	1	2	1	5	1	13	1
D	1	1	1	3	1	8	1	21
E	1	1	2	2	5	5	13	13

Task 3: Structured Indexing and Retrieval in Elasticsearch

a) (a) We decided that the fields: ”Message-ID”, ”Date”, ”From”, ”To” and the content of the email.

b) (a) State some basic statistics:

- *Who sent most of the emails?* - ”steven.kean@enron.com” sent 4674 emails
- *What is the most common subject?* A blank subject i most common with 3852 occurences. Only ”Re:” has 2241 occurences and ”Schedule Crawler: HourAhead Failure” has 239 occurences at third place.
- *How many emails are not from members of the Enron company?* 16149

(b) Further questions regarding “debra.perlingiere@enron.com”: (all of these are taken from the kibana web interface)

- *What are the top 5 contacts Debra is mostly communicating with?*
  Sent to:
  i. russell.diamond@enron.com : 442
  ii. genia.fitzgerald@enron.com: 393
  iii. veronica.espinoza@enron.com: 372
  iv. stacy.dickson@enron.com : 328
  v. tana.jones@enron.com: 324
  Received from:
  i. cheryl.johnson@enron.com : 143
  ii. veronica.espinoza@enron.com: 98
  iii. janette.elbertson@enron.com : 30
  iv. joanne.rozycki@enron.com: 18
  v. stacey.richardson@enron.com: 15
- *What are the top 5 contacts Debra is mostly communicating with, with no subject?*
  Sent to:
  i. jworman@academyofhealth.com: 35
  ii. janette.elbertson@enron.com: 22
  iii. nony.flores@enron.com: 22
  iv. russell.diamond@enron.com: 22
  v. allison.mchenry@enron.com: 20
  Received from:
  i. felipe.ibarra@enron.com: 1
  ii. genia.fitzgerald@enron.com: 1
  iii. sean.riordan@enron.com: 1
  iv. sheri.luong@enron.com: 1
- *How many emails did Debra send with no subject?* Debra sent 447 email with no subject

(c) Further questions regarding Howard University:

- *What subject is mostly used in any email sent by Howard University and contains "University" in the body? (Hint date information should be ignored)*
  In our 100.000 emails, there were non that where sent from an email ending with "@howard.edu".
- *What do the emails contain to Howard University that contain "University" in the body?*
  We found no emails

(d) Explore the dataset and report on additional findings, such as "What institutions are part of the emails?"

MiT, University of Texas, Stanford, University of Michigan, Washington State University, Rice University, Trinity University, New York University and others.