Assignment 3

Information Retrieval and Web Search

Ahmed Rekik - 790063 & Tim Henning - 789242

November 26, 2017

1. Boolean Retrieval

1.a. One of the drawbacks of the Boolean retrieval model lies in the size of the returned result set. Why is the size typically difficult to control?

The fact that the IR with a Boolean retrieval model returns an out of hand result set is due to its absolute Boolean property. While the result depends on query terms, these could occur in relevant as non relevant and superfluous documents, thereby the size of the result will be hard to control and unpredictable. As an example by querying "House AND President" one can expect documents about the president's house but he can get instead a document in which the word "President" occurs in the first page and "House" in the last page.

1.b. According to query log statistics, most queries do not contain search operators. Why are Boolean operators nevertheless necessary in professional search? Are consumer and professional search precision or recall oriented?

Although they might not be frequently employed, search operators represent an efficient tool to elaborate queries in a more expressive way in order to increase chance to hit relevant documents. Since professional search might generally intend very specific documents, those could be not easy to reach with the regular queries and without specifying multiple predicates for their search.

A very specific document aimed for a professional use may contain frequent word that occur in other documents often, therefore by querying without employing several predicates irrelevant documents will submerge the relevant one. Thus, by using operator ("AND", "OR") for providing very specific key-word and eliminating the undesirable document subjects (using "NOT" operator), professionals could reach the precise intended documents faster.

As expressed above, the professional use in information retrieval tend more to the effective and pertinent result. Since their searchs look for (very) specific information (source, format...) and they want to find *all* documents containing it, their search should

rather be recall-oriented. Otherwise, the consumer in general is more tolerant about missing documents but wants to find at least one that is relevant, thus the consumer search would be more precision-oriented.

2. Boolean Retrieval in Practice

2.a. Evaluate the query: $q_1 = (t_1\ OR\ t_5)\ NOT\ t_2$

t_1 OR t_5 =D_1, D_2, D_4, D_5, D_6
$t_2 = D_2, D_6$
$q_1 = D_1, D_4, D_5$

2.b. $q_2 = (t_1\ AND\ t_5)\ OR\ (t_3\ AND\ t_2)$

t_1 AND $t_5 = D_1, D_2$;
t_2 AND $t_3 = \emptyset$
$q_2 = D_1, D_2$

3. Vector Space Model

Compute the vector space similarity between the query "digital cameras" and the document "digital cameras and video cameras" by filling out the empty columns in the Table 1. Assume N=10,000,000, logarithmic term weighting (wf columns) for query and document, idf weighting for the query only and cosine normalization for the document only. Treat and as a stop word. Enter term counts in the tf columns. What is the final similarity score?

	Query				Document			
word	tf	wf	df	q_i = wf-idf	tf	wf	d_i = Normalized wf	q_i . d_i
digital	1	1	10000	3	1	1	0.520	1.561
video	0	0	100000	0	1	1	0,520	0
cameras	1	1	50000	2.301	2	1.301	0.677	1.558

The final similarity score is 3.119.

4. (Programming) Boolean Queries

4.a. Print the number of comments that match the following queries and one example comment per query (if there are any matches):

(The search results were chosen randomly. They do not reflect our opinion in any way.)

```
Searching:  party AND chancellor
Found 1 results in 1.52ms.
```

```
The AfD is not "far-right". Conservative, liberal, but not "far
    -right". It is best comparable to the CDU under (Chancellor)
     Helmut Kohl,- before Angela Merkel has taken over the party
     and turned it into a hotpot of neo-liberalism, leftist
    ideology, multi-cultural fancies and mass-immigration policy
    .

Searching:  party NOT politics
Found 1475 results in 49.48ms.
The STATE sponsored TERROR and HORROR inflicted upon (White,
    Orthodox Christian) people in Russia since 1917 was because
    of the direct actions of the JEWISH BOLSHEVIKS of Lenin (and
     his Party competitor) Trotsky - who took TOTAL power,
    financed by their Satanist bosses, namely the Rothschild
    Banking Cartel.  Stalin had three Jewish wives.

Searching:  war OR conflict
Found 8215 results in 289.46ms.
The axis powers were  doing the same thing. It doesn't excuse
    the allies for doing it, but that's why they call it Total
    War.

Searching:  euro* NOT europe
Found 2184 results in 83.64ms.
The creation of the Euro was an act of political incest. And
    like the progeny of all incestuous couplings the result is
    doomed to a short and painful life. The sooner it is put out
     of its misery the better.

Searching:  publi* NOT moderation
Found 2249 results in 73.62ms.
These people are patriots and should be supported by the public

Searching:  'the european union'
Found 106 results in 8.85ms.
Hungary working hard like many other European Countries to Stop
     Migration and,\nthe European Union working hard to let more
     Illegal Migrants in !\n\nTime for the European Union to
    change its views, else Referendums are a MUST !\n\nVote for
    Leaders Who Promise to Tackle Illegal Migration and
    Invasions ! Future is in your VOTE!

Searching:  'christmas market'
Found 4 results in 34.41ms.
```

Thus it indicates what the motive and agenda is. Governments attack their own people, unarmed and unaware to further their own agenda, they never attack the military or police or politicians. Good rule of judgement, police bus bombed >> real, Christmas market or nightclub shot up >> shadow government.