Assignment 5

Information Retrieval and Web Search

Ahmed Rekik - 790063 & Tim Henning - 789242

December 7, 2017

1. Showing Results

1.a. Think about result visualization for a comment search engine. What information could be shown in a snippet?

In general, the result is returned what the search engine considers as relevant documents in terms of a given query by the user. Certainly, the result could differ in its ranking or even in its relevance consideration in terms of the employed informational retrieval model. In all cases, a search engine aims to return an optimal result for the user in many formats (text, images, videos, link to documents), but it could also include sponsored advertisements that have a relationship to the query or most viewed/rated documents (important news, shop offer...) in sophisticated and lucrative search engine as "Google" or "Bing". In the "Comment search engine" frame those type of result could apply.

On the result snippet, it is reasonable that the result will include a ranked vertical list of matched comments at the front of the page in a textual form (a normalized snippet if the comment is very long) since they represent the main product and the potential information need to the user. Besides that, other information/suggestion could occur such as the article related to the comment, the comment author, the comment-date or something irrelevant to the user informational need such that the most rated comments of the week or month of all the newspaper-website to show interactivity in the search engine.

1.b. What elements are similar, different, or the same compared to result pages of Google, Bing, or any other web search engine?

Although all search engines have an identique aim (returning relevant information to the user), may there result visualization and their ranking differ. As a point of similarity that most of famous web search engine share, is that the query result is returned in a vertical list of snippets from the document headed with a link for the page titled by the page title and that have the same length size. The sponsored advertisments (that could have a relationship with the query or not) has been also more and more introduced by all prominent web search engine in the result visualisation since the last decade.

Otherwise, the result visualization differs from an engine to another due to their different offered feature. This difference can occur even in the matched query result. The first reason for this fact is that some search engine favorize the document shared by websites related to their company than others. For example: by "*Yahoo*" web search engine documents provided by "Yahoo *news/sports/answers*" may be ranked better than all other documents (usually on the top of the result) or "*Wikipedia*" pages are always taking the top-5 result in "*Google*". Other information could be shown also in terms of the query context due to the web searcher feature. For example: If a user types a football team name, a snippet of latest match results, current or former players and club rivals could be listed by "*Google*", otherwise by "*Bing*" they will scroll other information and may suggest tickets sellers or bets website and win probability for the next games.

2. Evaluation Measures

For a given query and a collection of 100 Web pages (which contains 40 relevant pages), a search engine produces the following ranking: R, N, R, R, N, N, R, N, R, R. R denotes a relevant document and N denotes a non-relevant document. Calculate the following evaluation measures:

2.a. Precision and recall

Precision: $\mathrm{P} = 7 \div 11 = 0,63$
Recall: $\mathrm{R} = 7 \div 40 = 0,175$

2.b. Precision at 7 and recall at 7

Precision at 7: $\mathrm{P} = 4 \div 7 = 0,571$
Recall at: $\mathrm{R} = 4 \div 40 = 0,1$

2.c. What is 'independent' in the binary independence model (BIM) and is this a reasonable assumption? Explain.

"Independence" denotes to the independent representation of each term in the document vector and each of them has no modeled relationship with the other ones (no position or terms similarity are taken into consideration). In a sense, this assumption is equivalent to the orthogonality assumption in the vector space model. In this context, the independence assumption is exaggerated and absolute which could mislead the result, as in real life (in our information retrieval framework) it is almost impossible to find a document, in which all its value have no relationship with the other. (either by their position or in the document context) A term classification (topic, synonym, similarity...) could be a more reasonable way to proceed with this formula and thus elaborating class-independence could lead to a safer result.

2.d. MAP

Table 1: MAP

i	1	2	3	4	5	6	7	8	9	10	11
Result	R	N	R	R	N	N	R	N	R	R	R
i-Precision	1	0.5	0.67	0.75	0.6	0.5	0.571	0.5	0.55	0.6	0.63

$\Rightarrow 1 + 0.67 + 0.75 + 0.57 + 0.55 + 0.6 + 0.63 = 4,77$
$MAP = 4.77 \div 7 = \mathbf{0.681}$

2.e. NDCG (assume binary gain value (relevant/non-relevant))

Table 2: Normalized DCG

i	1	2	3	4	5	6	7	8	9	10	11
Result	R	N	R	R	N	N	R	N	R	R	R
DCG	1	1	1.63	2.13	2.13	2.13	2.48	2.48	2.79	3.09	3.37
Perfect result	R	R	R	R	R	R	R	N	N	N	N
P-DCG	1	2	2.63	3.13	3.56	3.94	4.29	4.29	4.29	4.29	4.29
N-DCG	1	0.5	0.619	6.8	0.598	0.54	0.578	0.578	0.65	0.72	0.78

3. (Programming) Index compression

We are using the BM25 model to rank the results.

```
Searching using BM25:  christmas market
Found 3930 results in 27.28ms.
15.5 - who just wanted to visit Christmas market which is their
    traditions since ages.
15.5 - notice how attacks are never on Government buildings?
   just christmas markets.
15.5 - This app makes a great Christmas gift.\n\nIt will be
   marketed under the brand name Pok-e-alterboy.
13.9 - The immigrants are a tool to fuel Gernan industry, the
   German elite don't do Christmas markets, so they don't care!
13.9 - The last few years I did the Christmas markets in
   Germany and Switzerland, this year I am going to Russia...

Searching using BM25:  catalonia independence
Found 3254 results in 14.11ms.
```

```
18.3 - Catalonia will be independent, whether you want it or
   not.
17.6 - This is a pre Catalonia Independence warning from Madrid
16.9 - We should all stand behind Soros and his support for the
    independence of Catalonia!
16.9 - Independence?? That is not going to end well for
   Catalonia I think...
16.5 - the "independent" country of Catalonia full of Spanish
   flags... mmm... what a strange "independent" country :-)

Searching using BM25:  'european union'
Found 2481 results in 10.37ms.
9.0 - Well, yugoslavian union was kind of a european union, and
    we know what happened. The new union, EU, is on the same
   path.
8.9 - There already is a new soviet union under the name of
   European Union.
8.6 - European Union.
8.4 - The European Union? Each has a say.....No Divide..Good
   for All..This is NOT A UNION..\nBeen in the UNION my entire
   life...Do not disgrace it..\n Union is the collective rights
    of all individuals for better wages and better life of the
   members.
8.4 - Should the United Kingdom remain a member of the European
    Union or leave the European Union?\n\nThat's still biased
   in favour of staying in. It should read:-\n\nShould the
   United Kingdom leave the European Union or remain a member
   of the European Union?

Searching using BM25:  negotiate
Found 950 results in 4.31ms.
10.9 - How do negotiations work when you're calling for
   surrender? What is there to negotiate?
10.7 - Britain's top Brexit negotiator that can't negotiate
10.1 - Too many negotiations.
10.1 - No negotiation with terrorists !
10.1 - Opps"" Negotiation.
```