Assignment 7

Information Retrieval and Web Search

Ahmed Rekik - 790063 & Tim Henning - 789242

January 23, 2018

1. Remerge or Rebuild?

A search engine's index of two year's news comments contains 1 million documents. Each day, about 5000 new comments are posted and about 100 comments are deleted (because of hate speech, insults, or defamation). The index is therefore updated with a Remerge strategy on a daily basis. However, once only, a large amount of outdated comments shall be deleted and the question is:

1.a. How many outdated comments would need to be deleted as part of the next daily update so that the Rebuild strategy becomes more efficient than Remerge?

In order to improve the running time of index rebuild such that the rebuild time becomes more efficient than the remerge time, we need to adapt the amount of out-dated indexed comments that need to be deleted. We know that:

$time_{rebuild} = c.d_{new} = c.(d_{old} - d_{delete} + d_{insert})$ for a constant c.

$time_{remerge} = c'.d_{new} = c'.(d_{insert} + \frac{1}{4}(d_{old} + d_{insert}))$ for a constant c'.

Let's suppose that constants here are negligible for our two running time functions, thus we take into consideration only our variables. We assume that $Time_{rebuild} = Time_{remerge}$ if:

$(d_{old} - d_{delete} + d_{insert}) = (d_{insert} + \frac{1}{4}(d_{old} + d_{insert}))$

In our case we want to get the minimum amount of outdated comments that need to be deleted in order to get a rebuild time time at least fast as the remerge running time, thus the following function holds:

$(d_{old} - d_{delete} + d_{insert}) \leq (d_{insert} + \frac{1}{4}(d_{old} + d_{insert}))$

by subtracting d_{insert} and $\frac{1}{4}d_{old}$ in the both functions we get:

$\frac{1}{4}d_{old} - d_{delete} \leq d_{insert}$, which leads us to the following inequation

$d_{delete} \geq \frac{3}{4}d_{old} - \frac{1}{4}d_{insert}$

We know that daily, we delete at least 100 undesired comments, let's suppose that $d_{outdated}$ is the number of outdated comments that need to be deleted to improve the running time of rebuild s.t $d_{delete} = d_{outdated} + 100$, therefore:

$d_{outdated} + 100 \geq \frac{3}{4}d_{old} - \frac{1}{4}d_{insert}$

We know that : $d_{old} = 1000000$ and the daily inserted comments $d_{insert} = 5000$, thereby:

$d_{outdated} \geq 750000 - 1250 - 100$

$d_{outdated} \geq 748650$

Since we want a **more** efficient rebuild running time, we have to delete **more** than 748 650 comments. $d_{outdated} > 748650$. ⇒ 748651 outdated comments to delete

2. PageRank

2.a. Compute the pageRank score for the following network. Start by writing down the adjacency matrix. Assume a random jump probability of 0.15.

Let's define the following adjacency matrix of the given graph as follow:

$$\begin{array}{c|ccc} & A & B & C \\ \hline A & 0 & 1 & 0 \\ B & 1 & 0 & 0 \\ C & 0 & 1 & 0 \end{array}$$

We define the *pageRank* of a page u $PR(u)$ as follow:

$PR(u) = (1-p) + p.(\frac{PR(T_0)}{C(T_0)} + \frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)})$, where p is the probability ratio, $C(T_i)$ is the count of the outgoing edges from a page T_i and $T = \{T_0, T_1, ..., T_n\}$ is the set of pages that link to the page u.

In our case, all pages have only one outgoing edge, so $\forall u$ in our network $C(u) = 1$. To calculate the rankPage of our network, following equations hold:

1. $PR(A) = (1-p) + p.(PR(B))$
2. $PR(B) = (1-p) + p.(PR(A) + PR(C))$
3. $PR(C) = (1-p)$, since no vertex point to C, thereby $\mathbf{PR(C) = 0.85}$

For PR(B) we get after (1) and (3):

$PR(B) = (1-p)+p.((1-p)+p.(PR(B))+0.85) = (1-p)+p.(1-p)+p^2.PR(B)+p.0.85$

$(1-p^2).PR(B) = (1-p) + p.(1-p) + p.0.85 = 1.105 \Rightarrow PR(B) = \frac{1.105}{(1-p^2)}$

$\mathbf{PR(B) = 1.130}$

To finish, we get $\mathbf{PR(A) = 1.019}$.

2.b. PageRank is initialized with identical scores for each web page. What would be a better method to initialize pageRank so that less iterations are necessary?

PageRank is an algorithm that calculates the score of a page in terms of the scores (in pageRank) of pages that link to that page. Thus, several iterations have to be done in order to compute the pageRank for each document and that task may be expensive, knowing that in practice the amount of pages is extremly high. Therefore some optimizations could be done in order to decrease the computation time and to improve the ranking quality.

A possibly large optimization could be to initialize the algorithm with estimates for the pageRank for each page instead of identical values. The estimation could be based on simple factors as the document length or historical results, or even just a random normal distribution. This will most likely lead to much less iterations needed to get a steady, final result.

Another optimization could be the removal of navigational links between pages. Most pages nowaday include navigational links such as a menu that help the web surfer to navigate into the website. Those links increase the runtime of the pageRank calculation by introducing additional links between pages, besides the fact that those links could manipulate the rank as they are no recommendations and decrease the ranking quality.

An other reduction could be the removal of nepostic links. Several people tend to share links to their own pages in other publications (forums, comments...). Those nepostic links could be regarded as superfluous not only for the pageRanking computation time but also harm the ranking. Their removal would be a big boost for the pageRank computation, even though it is hard (and expensive) to detect those links.

We could also reduce the kern of pages (documents) by considering that $pageRank(u) = pageRank(v)$, if u and v have the same in-comming edges and therefore they are similar. This reduction could save time in the pageRank computation by eliminating "duplicate" pageRank functions.

3. (Programming) ReplyTo

3.a. Choose comment ids that lead to at least two search hits for the following queries. Print not only the search hits but also the comment that is specified in the query. The returned comments should be replies to the comment with the specified id.

3.a.1. "ReplyTo:300748"

```
Searching for replies to the following comment:
The EU was a CIA initiative, the purpose of which is to make it
    easy for Washington to exercise political control over
   Europe. It is much easier for Washington to control the EU
   than 28 separate countries. Moreover, if the EU unravels,
   so likely would NATO, which is the necessary cover for
   Washingtons aggression.
Found 3 results in 0.23ms.
-> bs. The US want to divide europe in order to keep the status
    of the dollar. The only menace to $ in all these years has
   been the euro, and it's surprising so many people don't get
   it. That UK don't belong in europe is a fact, but the EU
   must not divide now.
```

```
-> its not always the CIA as much its not always Putins fault.
   There is nothing wrong with the idea of a United States Of
   Europe . The politicians running the show in Brussels need
   to be replaced. Nobody would want a government like these
   blood suckers.
-> No matter what the new's, you'd still blame the nosey jews
   or the CIA for any outcome.\n\nOf course the rest of the
   world is so irrelevant and naive that only 2 groups of
   people in this world control everything.
Materialized 3 results in 0.10ms.
```

3.a.2. "ReplyTo:26252"

```
Searching for replies to the following comment:
"and civilians using suicide bombings and rocket strikes". I
   went to Palestine before the recent conflict and all the out
    going flights were cancelled. I stayed there during the war
   . There was no hamas suicide bombers. Nor was there hamas
   human shields. It's what we feared my fellow RTers, RT has
   turned zionist just like every other media. I'm disappointed
   .
Found 2 results in 0.43ms.
-> Tyrana, I, too have been there before and during periods of
   war.  Let me tell you, there were bombs and casualties
   everywhere.  And it was NOT just zionists.
-> And as for your disappointment in RT, try viewing news from
   more than one source in order to make informed and less
   biased opinions.
Materialized 2 results in 0.08ms.
```

3.a.3. "ReplyTo:157515"

```
Searching for replies to the following comment:
Probably the largest number of protesters protesting Modi in
   London were Nepalis. You can see Nepali flags here an there
   in RT footage too. They were protesting against India
   installed Blockade against Nepal, but not officially
   accepted by Modi. There is ongoing humanitarian crisis in
   Nepal, as it is trying to cope with recent earthquake.\nNot
   a single word about this on RT...question more. Becoming
   more like CNN?
Found 2 results in 0.37ms.
-> Nepal is currently being punished by Modi for making Nepal a
    secular country while he is trying his best to make India a
    Hindu only country.
```

```
-> Only muslims have right to spread their terror cult?
   otherwise jehadi cowards  will shout fake complains, bigger
   fools will just raise with guns. there can be no peace
   without decimating them. Nepalis are wise enough to decide
   their future.
Materialized 2 results in 0.06ms.
```