AI506: Data Mining and Search (Spring 2020)

Homework 2: Personalized PageRank

Release: April 15, 2020,
Due: April 29, 2020, 11:59pm

20203221 민향숙

[Analysis]

1. PageRank Score

1.1. PageRank with memory-based Graph

Preference_uniform

```
-----Top 10 pageranks-----
89073   0.011219708932132257
241454  0.008250330457153421
226411  0.00691142510453951
262860  0.002996886841256732
134832  0.002990483136689007
234704  0.00246201209980964 1
136821  0.002445472913930576
68889   0.0024226254395939228
69358   0.002356503411602799
105607  0.00230621338640704
```

Preference_onehot (506 = 1)

```
-----Top 10 pageranks-----
506     0.1541846581227055
213780  0.1168029262124912
129971  0.10917524546732417
118948  0.09914293826555216
194930  0.09914293826555215
24726   0.09914293826555215
152422  0.02256887605009158
52820   0.022143999574545128
223264  0.017177161060066008
252200  0.017177151894607737
```

Preference_in_degree

```
-----Top 10 pageranks-----
226411  0.011657317330549635
89073   0.009480410535379558
241454  0.006970375326880231
105607  0.005199228603519611
234704  0.0045142774632161
167295  0.0042545803440839775
38342   0.0038882737737599296
181701  0.0033606086363906196
247241  0.003292814312165141
259455  0.003292814312165141
```

1.2. PageRank with disk-based Graph

Preference_uniform

```
-----Top 10 pageranks-----
89073   0.011219708932132417
241454  0.008250330457153348
226411  0.0069114251045397325
262860  0.002996886841256752
134832  0.0029904831366890173
234704  0.002462012099809744
136821  0.0024454729139306197
68889   0.0024226254395939605
69358   0.0023565034116028076
105607  0.00230621338640716
```

Preference_onehot (506 = 1)

```
-----Top 10 pageranks-----
506     0.1541846581227055
213780  0.11680292621249137
129971  0.1091752454673243
24726   0.09914293826555226
118948  0.09914293826555226
194930  0.09914293826555226
152422  0.02256887605009158
52820   0.022143999574545173
223264  0.017177161060066008
252200  0.017177151894607737
```

Preference_in_degree

```
-----Top 10 pageranks-----
226411  0.011657317330549224
89073   0.009480410535379595
241454  0.00697037532688022
105607  0.005199228603519793
234704  0.004514277463216102
167295  0.004254580344083975
38342   0.003888273773759927
181701  0.0033606086363906357
247241  0.0032928143121651546
259455  0.0032928143121651546
```

All three different preference vector cases have different results.

Pageranks are basically influenced by in_coming nodes. If Incoming nodes have higher ranks, the nodes also can get higher ranks. Thus, pageranks can be determined by the number of incoming of nodes and the rank of incoming nodes. But in Personalized Pagerank algorithm, it assumes that random surfer jumps to a web page according to given probability distribution (we call the preference vector). According to this assumption, preference vector can also have an impact on pagerank scores. That is why all three cases have different results

with same Graphs. We saw that the list of 10 pages with the highst pagerank score change as we use different preference vector. Furthermore, if we set damping_factor smaller, preference vectors will have more impact on pagerank score.

First, we can notice that first and third cases have similar lists. Because pageranks are determined by the number of incoming nodes and incoming nodes' ranks. But third case has preference vector in proportion to incoming degree of each node. So, third case emphasizes influence of the number of incoming nodes. Consequently, first and third case have not same lists but similar lists.

(we can find that most of top-10 pages have higher in_degree.)

```
Highest in-degree
(node,In degree)
(226411, 38606)(234704, 21920)(105607, 19457)(241454, 19377)(167295, 19003)
(198090, 18975)(81435, 18970)(214128, 18967)(38342, 18958)(245659, 18935)
(34573, 18925)(89073, 15277)(69358, 13936)(67756, 13872)(134832, 10336)
(231363, 10244)(17781, 8346)(62478, 8346)(77999, 8346)(120708, 8346)
```

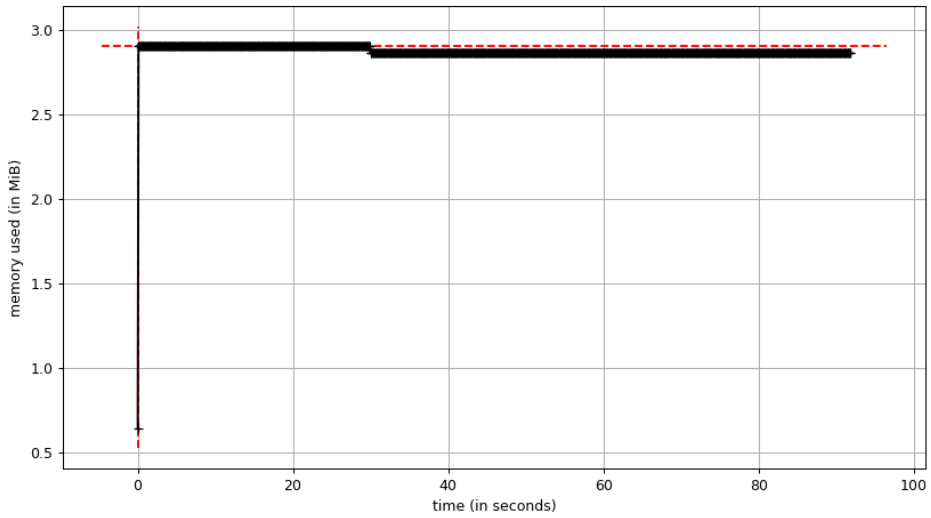
Secondly, for preference_onehot case, top 10 pageranks scores are a lot higher than other two cases. The top-1 page has 0.154 scores and the score of other 9 pages are higher than 0.015. The reason is that preference vector has probability 1 only for page 506, and other pages have probability 0. So, at every iteration, the score of page 506 always are added "(1-damping factor)" and other pages don't get added anything. That's why page 506 has a lot higher score than other pages. And pageranks are calculated recursively, it also affects other pages which are outgoing pages of page 506. We can also find that most of top-9 pages are outgoing pages of page 506.

```
Out-neighbors of page 506 : {223264, 118948, 152422, 252200, 194930, 129971, 213780, 24726}
```

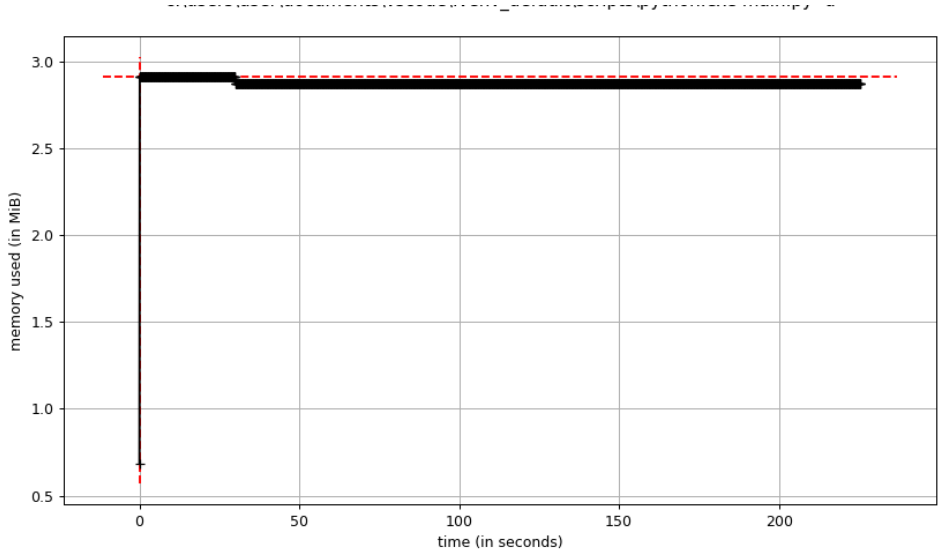
Third, comparing memory-based graph with disk-based graph, they have same results for each preference case. Because we use same graph for each case. They are just different in loading graph data.

2. Time -based Memory usage plot with both graphs.

2.1. Memory-based graph



2.2. Disk-based graph



Disk-based graph takes more time to implement pagerank algorithm. And in this case, they have similar memory usage.