Отчет по домашней работе по ссылочному ранжированию

Викулин Всеволод

15 ноября 2016 г.

Задача решалась в два этапа. На первом скрипт create graph.py извлекает из данных ссылки и формирует макет графа, который записывается в graph.txt (на гитхабе выложен файл меньшего размера, чем весь граф, потому что там лимит 100 мб на файл, а финальный граф занимает 120 мб). Это **самый долгий процесс** работы, я его ставил на ночь. Проблема не в том, что я не умею кодить, (не только в этом) библиотека питоновская, которая извлекает данные из html работает очень долго, (около 9 html файлов в секунду) остальной код работает быстро, так как все реализовано через словари. На этом этапе всем урлам, которых не было в файле urls.txt присваивается свой индекс (отрицательный). Это будут висячие ссылки в нашем графе. Строка файла graph.text имет вид: <pageid + tab + output links>. Output links имеет вид списка, записанный через пробел объектов вида: [индекс исходящей вершины:сколько раз она встречалась на странице]

Общие сведения по графу. Количество вершин примерно 2.6 миллиона, висячих из низ 2.1 миллиона, то есть около 80 процентов.

Построением самого графа из файла занимает около минуты. После построения он сразу проверяется на равенство количества исходящих и входящих ребер.

1 Hits alg

. Этот алгоритм перенесет на питон без изменений с лекции. У каждого объекта класса WebPage есть параметры хабность и авторитетность, на каждой итерации пересчитывается хабность как сумма авторитетности исходящих страниц и авторитетность как сумма хабностей входящих. Помимо задания количества итераций можно задать eps, если относительные изменения хабности и авторитетности будут меньше eps, то цикл

прервется. Обычно считается около 5 итераций с моим стандартным eps 0.02. **Время работы - 1 минута. 6 итераций**

2 PageRank alg

. Этот алгоритм также перенесет практически без изменений. На каждой итерации дополнительно проверяется, что сумма PageRank‘а равна единице. Все висячие циклы стали ссылаться на себя с вероятностью единица. Чтобы PageRank не утекал в них, была использована телепортация с коэфициентом 0.3. Это значительно уменьшило утекание вероятности в висячие ссылки. Для примера: начальное распределение равномерное, следовательно у висячих ссылок 80 процентов. После всех итераций и стабилизации их вероятность растет до 90 и дальше такой остается, не повышаясь. Но если убрать телепортацию, это значение приблизится к 100. Вы также можете задать eps, если относительное изменение pagerank будет меньше eps, то цикл прервется. 5-6 итераций обычно достаточно для точность 1e-6. **Время работы (с учетом проверки равенства суммы единице) - 5 минут. 6 итераций** .

3 Топ 30

3.1 Hits alg

Авторитетность

1. ’http://lenta.ru/rubrics/ww1/’,
2. ’http://lenta.ru/rubrics/library/’,
3. ’http://lenta.ru/chronicles/ukraine/’,
4. ’http://lenta.ru/news/2008/08/26/medvedev/’,
5. ’http://lenta.ru/articles/2013/12/05/banking/’,
6. ’http://lenta.ru/russia/2003/04/17/yushenkov/’,
7. ’http://lenta.ru/chronicles/battle/’,
8. ’http://lenta.ru/sport/2002/02/09/opening/’,
9. ’http://lenta.ru/news/2006/10/02/blockade/’,

10. 'http://lenta.ru/articles/2013/03/25/cyprus/',
11. 'http://lenta.ru/articles/2013/08/27/uralkalij/',
12. 'http://lenta.ru/articles/2013/07/31/kalij/',
13. 'http://lenta.ru/russia/2004/02/14/aquapark/',
14. 'http://lenta.ru/articles/2012/10/22/rsnft/',
15. 'http://lenta.ru/vojna/2004/05/09/grozny/',
16. 'http://lenta.ru/news/2014/12/16/cbr/',
17. 'http://lenta.ru/articles/2014/08/20/cosmos/',
18. 'http://lenta.ru/articles/2013/09/23/tolokonnikova/',
19. 'http://lenta.ru/articles/2013/12/17/aid/',
20. 'http://lenta.ru/articles/2013/10/13/biryulyovo/',
21. 'http://lenta.ru/russia/2003/07/02/lebedev/',
22. 'http://lenta.ru/chronicles/again/',
23. 'http://lenta.ru/russia/2004/09/13/electoral/',
24. 'http://lenta.ru/chronicles/ua/',
25. 'http://lenta.ru/news/2008/02/17/kosovo/',
26. 'http://lenta.ru/news/2014/08/07/ban/',
27. 'http://lenta.ru/news/2012/10/22/deal/',
28. 'http://lenta.ru/video/2013/10/14/birulevo/',
29. 'http://lenta.ru/news/2006/06/13/webby/',
30. 'http://lenta.ru/world/2004/11/27/rada4/'

Хабность

1. 'http://lenta.ru/news/2006/03/24/demo/',
2. 'http://lenta.ru/most/2003/10/16/gazeta/',
3. 'http://lenta.ru/news/2007/12/07/murder/',

4. 'http://lenta.ru/news/2010/07/10/rocket/',
5. 'http://lenta.ru/vojna/2002/11/01/zakaev/',
6. 'http://lenta.ru/vojna/2004/10/29/tvabkhaz1/',
7. 'http://lenta.ru/?id=60/',
8. 'http://lenta.ru/news/2008/03/20/fsb/',
9. 'http://lenta.ru/news/2010/05/20/dtp/',
10. 'http://lenta.ru/articles/2005/04/29/kadyrov/',
11. 'http://lenta.ru/news/2005/10/25/nalchik/',
12. 'http://lenta.ru/news/2005/05/17/beslan3/',
13. 'http://lenta.ru/russia/2003/04/10/budanov/',
14. 'http://lenta.ru/articles/2010/12/13/neproidet/',
15. 'http://lenta.ru/vojna/2003/04/29/budanov2/',
16. 'http://lenta.ru/vojna/2004/04/09/visimbaev/',
17. 'http://lenta.ru/culture/2002/11/29/nost/',
18. 'http://lenta.ru/news/2006/05/29/trepashkin/',
19. 'http://lenta.ru/terror/2004/09/28/aushev/',
20. 'http://lenta.ru/russia/2002/06/17/catch/',
21. 'http://lenta.ru/news/2005/09/15/buinaksk/',
22. 'http://lenta.ru/news/2011/08/29/shift/',
23. 'http://lenta.ru/russia/2004/10/04/diploms/',
24. 'http://lenta.ru/news/2011/09/14/troitsky/',
25. 'http://lenta.ru/news/2005/11/11/zapros/',
26. 'http://lenta.ru/russia/2003/01/10/noheat/',
27. 'http://lenta.ru/most/2002/12/06/titov',
28. 'http://lenta.ru/news/2009/05/28/insolvent/',

29. 'http://lenta.ru/news/2006/01/11/duma/',
30. 'http://lenta.ru/terror/2002/10/26/foreign/'

3.2 PageRank alg

1. 'http://lenta.ru/rubrics/ww1/',
2. 'http://lenta.ru/rubrics/library/',
3. 'http://lenta.ru/',
4. 'http://lenta.ru/chronicles/ukraine/',
5. 'http://lenta.ru/news/2010/06/17/burn/',
6. 'http://lenta.ru/news/2007/01/18/tim/',
7. 'http://lenta.ru/news/2007/03/04/kirill/',
8. 'http://lenta.ru/articles/2013/12/05/banking/',
9. 'http://lenta.ru/news/2010/10/28/supercomp/',
10. 'http://lenta.ru/chronicles/battle/',
11. 'http://lenta.ru/news/2010/10/12/elvis/',
12. 'http://lenta.ru/news/2011/11/03/fine/',
13. 'http://lenta.ru/articles/2013/03/25/cyprus/',
14. 'http://lenta.ru/articles/2014/08/20/cosmos/',
15. 'http://lenta.ru/articles/2013/07/31/kalij/',
16. 'http://lenta.ru/articles/2013/08/27/uralkalij/',
17. 'http://lenta.ru/articles/2013/09/23/tolokonnikova/',
18. 'http://lenta.ru/articles/2013/12/17/aid/',
19. 'http://lenta.ru/articles/2013/10/13/biryulyovo/',
20. 'http://lenta.ru/chronicles/ua/',
21. 'http://lenta.ru/chronicles/again/',

22. 'http://lenta.ru/video/2013/10/14/birulevo/',
23. 'http://lenta.ru/news/2014/12/16/cbr/',
24. 'http://lenta.ru/articles/2012/10/22/rsnft/',
25. 'http://lenta.ru/news/2008/08/26/medvedev/',
26. 'http://lenta.ru/articles/2014/09/17/sanctionfoodprice/',
27. 'http://lenta.ru/articles/2013/08/26/escalate/',
28. 'http://lenta.ru/articles/2013/03/19/cyprus/',
29. 'http://lenta.ru/sport/2002/02/09/opening/',
30. 'http://lenta.ru/articles/2013/06/11/snowden/'

В качестве вывода можно заметить, что топ 30 по авторитетности и PageRank очень похож, что кажется логичней, ведь чем больше на тебя ссылаются хорошие хабы, тем больше вероятность, что случайный блуждатель перейдет на твой сайт. Топ 30 хабов совсем на них не похожи, но тут уже такой прямой связи нет: вероятность прихода на твой сайт никак не связана с количеством хороших ссылок с твоего сайта.

В данных топах не участвуют висячие ссылки, очевидно, что некоторых из них по Авторитености и по PageRank были бы в топе, потому что на них много ссылаются (на них кстати реально ссылаются больше чем на ссылки ленты ру). Висячие ссылки из топа были отфильтрованы по причине того, что при построении графа я забыл сдампить словарик, в котором было бы соответствие между придуманным мной айдишником висячей ссылки и ее реальным урлом. Если интересно - могу этот словарик снова сделать и сохранить.

Хорошая домашка!