Анализ сайта MachineLearning.ru

Семченков Алексей
Московский Физико-Технический Институт
Курс Python, Яндекс

11 мая 2016 г.

1 Модель графа

В качестве изучаемого сайта мы берем MachineLearning.ru. Этот сайт содержит 930 статей по тематике машинного обучение. Доступ ко всем, или почти всем обеспечивается по ссылке 'machinelearning.ru/wiki/index.php?title=Категория:Статьи'.

Алгоритм заключается в следующем: он перебирает все категориальные странички и добавляет в список ссылки на "нормальные статьи". Реализовано это так: он берет текущую страничку (изначально — Категория:Статьи, для загрузки странички используется метод **get** библиотеки **requests**), с помощью библиотеки **BeatifulSoup** и **lxml** извлекает из нее 'bodyCont' и добавляет все ссылки из него (они помечены тегом 'a') в очередь, если они также являются категориальными. Если же страничка является статьей, а не категорией, то мы добавляем ее в список статей. Потом алгоритм берет следующую страничку из очереди, рассматривает ее, и так далее. Дважды странички не просматриваются. По сути, алгоритм представляет из себя BFS. В коде ему соответствует метод $collect\_articles(start\_page)$.

Получив список всех статей сайта, мы точно так же выгружаем их, и запоминаем для каждой статьи список ссылок из нее.

Далее, по полученным данным строится ориентированный граф, где вершины ассоциированы со статьями на сайте, а из вершины A ведет ребро в вершину B, если статья A содержит ссылку на статью B. Граф приходится чистить, так как концы некоторых ссылок лежат вовне графа. Все промежуточные данные сохраняются в текстовые файлы с помощью библиотеки **pickle**. В коде за эти части (построение графа и его "нормализацию") отвечают методы $create\_graph('file\_with\_articles.txt')$ и $clear\_graph(graph)$.

После того, как мы построили граф, где для каждой вершины хранится список исходящих ссылок, мы легко можем провести некоторый его анализ – в частности, посмотреть размер, среднюю степень вершины, а также построить по нему развернутый граф, где для каждой вершины хранится список вершин, которые на ее ссылаются. По этим графам мы строим график распределения страничек по числу входящих/исходящих ссылок. Также мы подсчитываем PageRank для графов. Используется следующая реализация подсчета PageRank:
1) На начальной итерации PageRank каждой вершины равен 1.
2) Значение PageRank (PR) для странички page обновляется одновременно для всех страничек по формуле:

$$PR[page] = 0.15PR[page] + 0.8 \sum_{n\_page} \frac{PR[n\_page]}{deg[n\_page]} + 0.05,$$

где суммирование ведется по тем страничкам n_page, что ссылаются на текущую страничку page, a deg[n_page] – это количество выходящих ссылок со странички n_page. Смысл формулы таков: 15% ранка каждая вершина оставляет себе, 80% равномерно распределяет между теми вершинами, на которые она ссылается, и 5% равномерно распределяется между вершинами, на которые она не ссылается (для удобства реализации это выражено просто тем, что к каждому ранку добавляется бесплатные 0.05, считается что средний ранк по вершинам равен единице). Разделение 5% между несоседями нужно для того, чтобы избежать "тупикового эффекта" когда

вершина имеет вход, но не имеет выход, поэтому только аккумулирует ранк, но не делится им.
3) В конце каждой итерации домножаем все ранки на поправочный коэффициент так, чтобы средний ранк равнялся 1.
4) Проводим таким образом 100 итераций.

2 Полученный результат

Алгоритм проанализировал 865 статей (таким образом, мы не попали в 65 статей) и построил по ним граф. Среднее количество ссылок (как входящих, так и исходящих) равняется 5.028. Распределение вершин по степеням можно посмотреть на приложенных .png-файлах, оно напоминает Пуассоновское.

Также, в приложенных файлах 'graph_analytics.txt' и 'reversed_graph_analytics.txt' можно посмотреть списки статей, остортированные по числу исходящих/входящих вершин.

ТОП-25 вершин по числу входящих ссылок (снизу вверх):
27 Квантиль
27 Кластеризация
27 Критерий_Уилкоксона-Манна-Уитни
27 Случайная величина
27 Уровень значимости
28 Метод главных компонент
29 Временной ряд
29 Интеллектуальный анализ данных
29 Метод ближайших соседей
30 Переобучение
31 Критерий Стьюдента
32 Нулевая гипотеза
37 Алгоритм
37 Нормальное распределение
37 Практикум ММП ВМК, 4й курс, осень 2008
43 Математические методы прогнозирования (кафедра ВМиК МГУ)
45 Классификация
45 Статистический анализ данных (курс лекций, К.В.Воронцов)
54 Машинное обучение (курс лекций, К.В.Воронцов)
54 Проверка статистических гипотез
54 Численные методы обучения по прецедентам (практика, В.В. Стрижов)
58 Метод наименьших квадратов
62 Машинное обучение
62 Регрессионный анализ
63 Выборка

В файле 'page_rank_graph_analytics.txt' можно увидеть список страниц, отсортированных по PageRank.

ТОП-25 вершин по PageRank (снизу вверх):
5.051039451873426 Рудаков, Константин Владимирович
5.102762747851089 Интеллектуальный анализ данных
5.215636295119384 Математические методы прогнозирования (кафедра ВМиК МГУ)
5.221789990039254 Сингулярное разложение
5.325868403242208 Нормальное распределение
5.50149973599024 Регрессионная модель
5.524802974955599 Метод главных компонент
5.635735024468317 Анализ регрессионных остатков

5.653503004310326 Переобучение
6.031621997801614 Линейная регрессия (пример)
6.844343786092728 Базы данных изображений
6.887894537560535 Временной ряд
7.952695612020308 Проверка статистических гипотез
8.243313469831412 Классификация
8.305289231460732 Машинное обучение (курс лекций, К.В.Воронцов)
8.965440442732714 Численные методы обучения по прецедентам (практика, В.В. Стрижов)
9.879102438731959 Алгоритм
10.82938571384399 Регрессионный анализ
11.078053792669637 Метод наименьших квадратов
12.383315681530831 Практикум ММП ВМК, 4й курс, осень 2008
13.520966381298964 Выборка
13.779811651477937 Машинное обучение
22.087003750515738 Вероятностное пространство
27.783169300926158 Случайная величина
29.77224988282655 Функция распределения

Как видим, сортировка по PageRank дает более точное определение важности странички, чем просто количество входящих ссылок.

Итак, Готов бросил презрительный взгляд, – Мой вам тоскливо? – Теперь понятно... Я ненавидел это нравится ваша вина, очень нравились учителю руки на вы думали? – У меня бирка на Лёху за аппаратурой : фашист.