Введение в анализ данных: Анализ ссылок 2

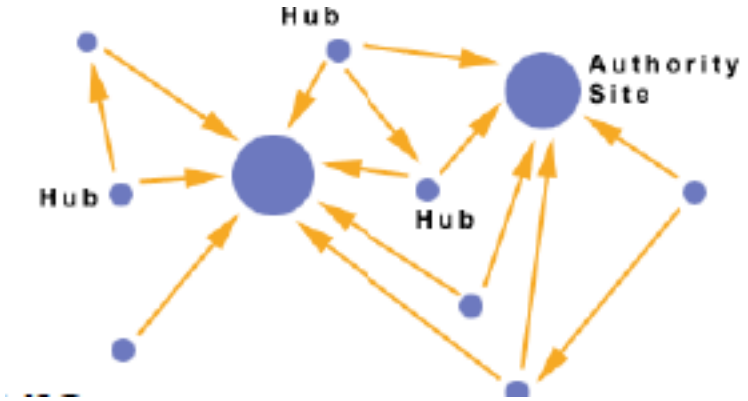
Юля Киселёва
juliakiseleva@yandex-team.ru
Школа анализа данных

План на сегодня

- Hub and authorities
  - HITS (hyperlink-induced topic search)
- Link Spam
  - История спама
  - «Фермы» спама
  - Trust Rank

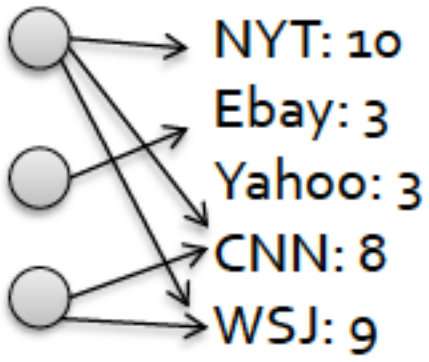
Hub and authorities

- Существует два типа интересных страниц:
  - Authorities - это страницы, которые содержат интересную информацию
    - Стартовая страница газеты
    - Стартовая страница производителя авто



Hub and authorities(2)

- Hub – это страницы, которые ссылаются на authorities
  - Список газет
  - Список производителей автомобилей



Рекурсивное определение

- Хороший hub ссылается на много хороших authority страниц
- На хороший authority ссылаются много hub страниц

- Модель использует 2 оценки:
- Для веб-страниц *v* в рассматриваем срезе Интернета*: h(v) –* ***hub score****, a(v) –* ***authority score***

Рекурсивное определение (2)

- ***h(v)=a(v)=1*** для всех *v*
- ***v->y*** – условие при котором существует гиперссылка между **v** и **y**

$$h(v) \leftarrow \sum_{v \to y} a(y)$$

$$a(v) \leftarrow \sum_{y \to v} h(y)$$

Матрица переходов A

- HITS (hyperlink-induced topic search) использует матрицу смежности
- A[i, j] = 1 если i ссылается на j
- 0 иначе

$$\vec{h} \leftarrow A\vec{a} \qquad \vec{a} \leftarrow A^T\vec{h}$$

- A^T похожа на матрицу ***M*** из PageRank, но в матрице ***M*** были дроби а в A^T единицы

Hub and authorities (3)

- HITS (hyperlink-induced topic search) использует матрицу смежности
- A[i, j] = 1 если i ссылается на j
- 0 иначе

- похожа на матрицу ***M*** из PageRank, но в матрице ***M*** были дроби а в единицы

Hub and authorities (4)

- Нотация:
  - Вектора: $a = (a_1, a_2, \ldots, a_n)$ $h = (h_1, h_2, \ldots, h_n)$
  - Матрица смежности (n x n):

$$\mathbf{A}_{ij}=1 \text{ if } i \rightarrow j \text{ else } A_{ij}=0$$

$$h_i = \sum_j A_{ij} a_j$$

$$h = A \cdot a \qquad a = A^T \cdot h$$

Уравнение Hubs and authority

- Hub score страницы ***i*** пропорционально сумме authority score, страниц на которые она ссылается: $h = \lambda A a$
  - ***Где*** λ ***– это масштабирующий коэффициент***

    $\lambda = 1/\sum h_i$
- Authority score страницы ***i*** пропорционально сумме hub score страниц, которые на нее ссылаются: $a = \mu A^T h$

Итеративный алгоритм

- HITS алгоритм
  - Инициализируем h и a (все 1)
  - Повторяем
    - $\mathbf{h} = \mathbf{A}\,\mathbf{a}$
  - Сумма элементов h = 1
    - $\mathbf{a} = \mathbf{A}^{\mathsf{T}}\,\mathbf{h}$
  - Сумма элементов a = 1
  - До тех пор пока h and a не сойдутся

Hubs and authorities Алгоритм

- Algorithm:
  - Set: $a = h = 1^n$
  - Repeat:
    - $h=\mathbf{M}\ a$, $a=\mathbf{M}^T h$

Внимание вопрос?

Какие можем сделать модификации?

Hubs and authorities Алгоритм(2)

$$a = \mathbf{M}^T (\mathbf{M}\, a)$$

new h

new a

$$a = (\mathbf{M}^T \mathbf{M})\, a$$

$$h = (\mathbf{M}\, \mathbf{M}^T)\, h$$

Существование и единственность

- HITS итеративный алгоритм сходится к векторам *h** и *a**
- *h* - это собственный вектор матрицы* $\mathbf{M}\,\mathbf{M}^{\mathrm{T}}$
- *a* - это собственный вектор матрицы* $\mathbf{M}^{\mathrm{T}}\,\mathbf{M}$

PageRank and HITS

- PageRank **и HITS –** это 2 решения одной проблемы:
  - Какова значимость ссылки со страницы ***v*** на страницу ***u***?
  - В модели PageRank, значимость ссылки зависит от ссылок **на** страницу ***u***
  - В модели HITS значимость ссылки зависит от ссылок **со** страницы **u**

План на сегодня

- Hub and authorities
  - HITS (hyperlink-induced topic search)
- Link Spam
  - История спама
  - «Фермы» спама
  - Trust Rank

Что такое спам?

- Процесс спама (spamming) = любое преднамеренное действие
  с целью повышения позиции веб-страницы в результатах поиска,
  путем увеличения реальной стоимости страницы
- Спам (spam) = страницы созданные для spamming
- SEO = search engine optimization
- Примерно 10-15% всех страниц – это спам

Ранние поисковые машины

1. Crawl Интернет (следуя ссылкам со страницы на страницу)
2. Индексация страниц на основе слов, которые они содержат
3. Ответ на запрос пользователя (список слов) страницами, которые содержат слова из запроса

Первые алгоритмы ранжирования

- Пытались ранжировать страницы на основе важности страницы для конкретного запроса
- Поисковые машины использовали:
  1. Количество раз когда слова встретилось на странице
  2. Местоположение слова (title, header, …)

Первые спамеры

- Люди начинали использовать поисковые машины для поиска по интернету. Коммерческие компании пытались оптимизировать контент своей страницы, с целью повысить позицию страницы на популярные поисковые запросы.
- Пример: страницы, посвященные продаже футболок, показывались на запрос: «про фильмы»

Первые спамеры(2)

- *Как Вы можете сделать так, чтобы Ваша страница показывалась в ответ на запроса «о фильмах»?*
- *Ответ: вставить слова «фильм» 1000 раз на свою страницу. Так, что только поисковая машина может их видеть.*

Первые спамеры(3)

- Еще варианты?

- Ответ:
- задайте запроса «фильм» в вашей целевой поисковой машине.
- Посмотрите, какие страницы показались первыми.
- Скопируйте их на вашу страницы (скройте от пользователей)

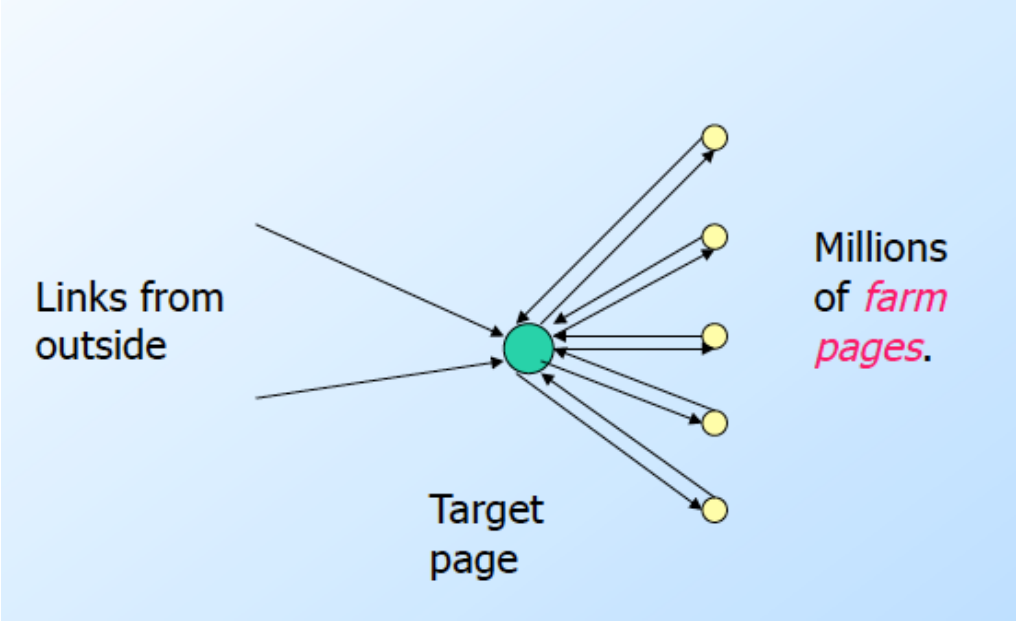
- Подобные техники называются term spam

Решение Google для борьбы с term spam

- *Верить тому, что люди говорят о тебе, а не то, что ты говоришь о себе сам*
  - *Текст в anchor text и близлежащий текст*
- *PageRank – это способ измерить «важность» страницы*

Раунд 2: Ссылочный спам

- Как только google стал один из самых популярных поисковых машин, спамеры стали находить пути, чтобы его обмануть.
- Были созданы «фермы спама» (spam farms)



Внешние ссылки

- Откуда приходят внешние ссылки?

- Страницы блога позволяют спаммерам добавлять комментарии типа «I agree. See www.mySpamFarm.com .»

Борьба с ссылочным спамом

- Определение структуры, которая выглядит как спам

- TrustRank = topic-specific PageRank с телепортацией только на «проверенные» страницы
- Пример: .edu домены

Term Spamming

- Повторение:
  - Одного или нескольких специфических термов – free, cheap, viagra
- Dumping:
  - Большое количество несвязанных термов
- Weaving:
  - Вставка спама в случайные места нормального текста
- Phrase Stitching:
  - «Склейка» предложений и фраз из разных источников

Определение спама

- Анализ текста с использование статистичеких методов (например Naïve Bayes Classifier)
- Также полезно определение страниц-дубликатов

Ссылочный спам

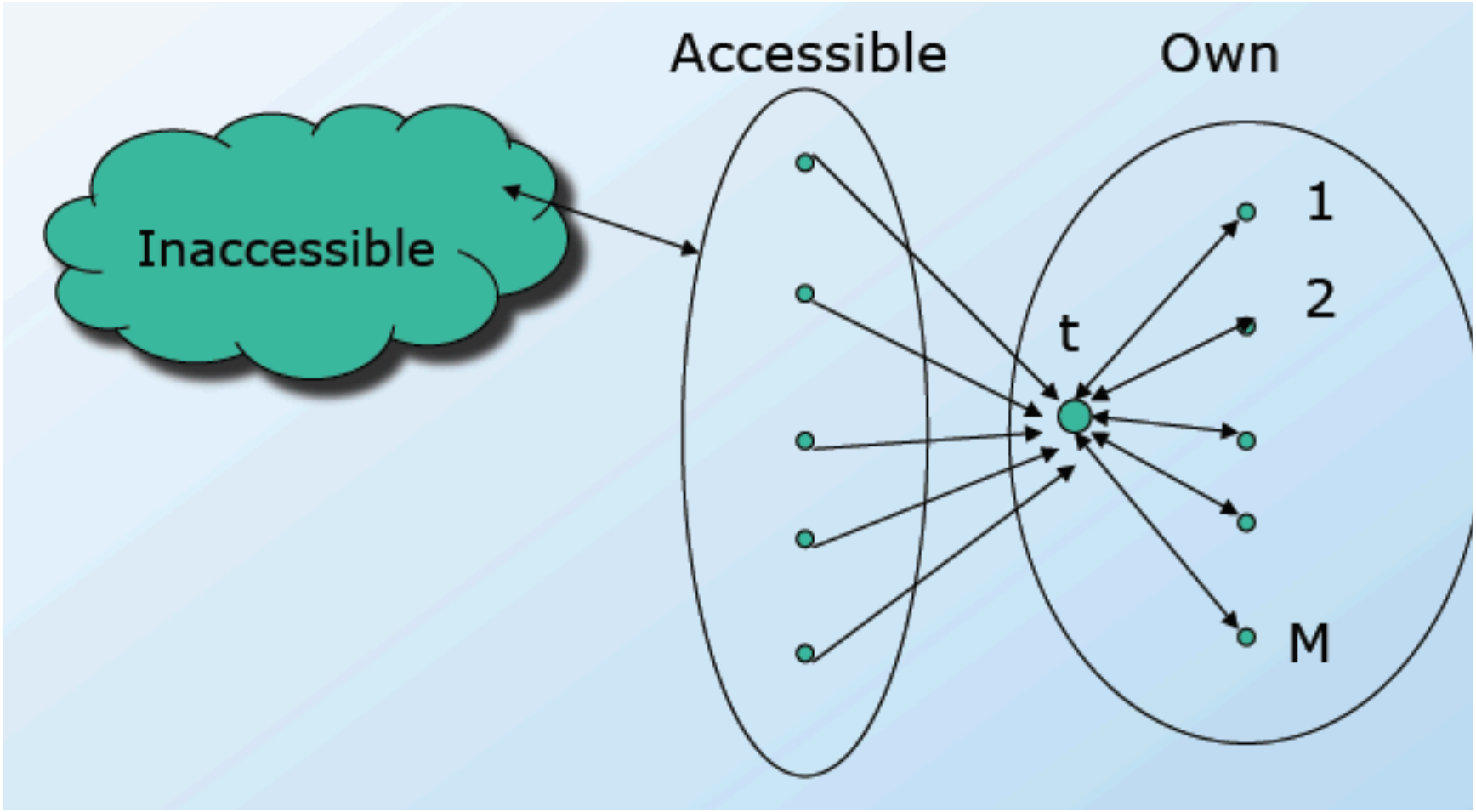
С точки зрение спамера существует 3 типа страниц:

- **Недоступные страницы**
- **Доступные страниц:**
  - **Комментарии в блогах**
- **Их собственные страницы**

«Фермы» спама

- Цель спаммеров:
  - Увеличить pageRank
- Способы:
  - Сделать как можно больше ссылок с доступных страниц на целевую страницу t
  - создать «ферму» спама, чтобы увеличит pageRank

«Фермы» спама (2)



Идея TrustRank

- Основной принцип – приблизительная изоляция:
  - Это редкость, если «хорошая» страница ссылается на «плохую»
- Набор seed страниц в интернете
- Человеческая экспертиза для определения «плохих» страниц в наборе seed страниц
  - Дорогое удовольствие, вы должны сделать набор seed как можно меньшим

Выбор seed набора страниц

- Выбрать топ страниц, согласно PageRank
  - Предположение: не может «плохим» страницам быть присвоен высокий pageRank

- Можно использовать домены, которые контролируются, как .edu, .gov и др.

Резюме

- Узнали про Hub and authorities
- Узнали про HITS (hyperlink-induced topic search). Поняли чем отличается от PageRank
- И затронули интересующий многих вопроса про Link Spam
- Вспомнили историю спама
- Побывали на «Ферме» спама
- Поверили в Trust Rank ☺