

Анализ гиперссылок в сети Веб: модели, подходы и алгоритмы

Лекция 2. Релевантность и гиперссылки

RUSSIR-2007. Сычев А.В.

Релевантность в WWW

- Главная особенность сети WWW состоит в том, что релевантная информация содержится не только в вершинах веб-графа (т.е. в тексте веб-документов), но и в ребрах (т.е. гиперссылках), связывающих между собой эти вершины.
- Поскольку ребра (гиперссылки) в WWW создаются пользователями для акцентирования связи между содержимым страниц в паре, их структура содержит важную информацию о содержании страниц и их релевантности.

RUSSIR-2007. Сычев А.В.

Базовые допущения при анализе гиперссылок

- Допущение о *рекомендательности*:
Если страница содержит ссылку на другую, то тем самым автор первой страницы рекомендует вторую
- Допущение о *тематической локальности*:
Если страницы связаны между собой гиперссылками, то с большей вероятностью они относятся к той же тематике нежели к разным.
- Допущение об *анкерном описании*:
Текст связанный с анкерным тэгом (*<a>*) гиперссылки описывает целевой документ, на который указывает гиперссылка.

Замечание: гиперссылки могут содержать как дополнительную полезную информацию так и шум (в т.ч. спам)

RUSSIR-2007. Сычев А.В.

Алгоритмы анализа гиперссылок

- Используются для косвенной оценки качества документов и для оптимизации работы сетевого робота.
- Принято выделять:
 - Методы глобального анализа (независящие от запроса), например **PageRank**.
 - Методы локального анализа (зависящие от запроса), например **HITS**.

RUSSIR-2007. Сычев А.В.

Алгоритм *PageRank*

- Был предложен *Сергеем Брином* и *Ларри Пейджем*, использован для ранжирования в ИПС *Google*.
- В основу заложена модель случайного блуждания по веб-графу, которая используется для вычисления веса страницы (показатель *PageRank*) как вероятности ее достижимости.
- Страница имеет высокий *PR* (показатель *PageRank*), если на нее ссылаются страницы с высоким *PR*.

RUSSIR-2007. Сычев А.В.

Алгоритм *PageRank*

- *Модель случайного блуждания*:
 - Вначале пользователь случайным образом выбирает веб-страницу
 - Далее, на каждом шаге он
 - Либо переходит на другую страницу, выбранную таким же случайным образом с вероятностью *d*
 - Либо переходит к другой случайно выбранной странице из числа тех, которые связаны с текущей гиперссылками, с вероятностью *1-d*
 - Другими словами, средняя доля шагов до страницы *a* определяется через величину *PR(a)*

RUSSIR-2007. Сычев А.В.

Расчет коэффициента *PageRank*

- Коэффициент *PR* для текущей веб-страницы *a* рассчитывается по формуле:

$$PR(a) = \frac{d}{n} + (1-d) \cdot \sum_{(b,a) \in G} \frac{PR(b)}{C(b)}$$

где

- *n* – количество страниц в веб-графе *G*
- *C(b)* – количество исходящих ссылок со страницы *b*
- *D* – коэффициент настройки, выбирается в пределах от 0.1 до 0.2.

RUSSIR-2007. Сычев А.В.

Расчет коэффициента *PageRank*

- Данная задача идентична математической задаче поиска собственного вектора *x*, являющегося решением матричного уравнения:

$$A \cdot x = \lambda \cdot x$$

где *A* – (асимметричная), построенная на основе матрицы смежности вершин веб-графа *G*.

RUSSIR-2007. Сычев А.В.

Расчет коэффициента PageRank

- Как видно, для вычисления $PR(a)$ требуется рекурсивная процедура, которая продолжается до достижения сходимости (на практике до 100 итераций).
- Следует иметь в виду, что коэффициенты PR рассчитываются только один раз и не зависят от конкретных запросов.

RUSSIR-2007. Сычев А.В.

Модификации алгоритма PageRank

- В алгоритме *Topic-Centric* [3] предложено выбирать исходящие ссылки на странице не с одинаковой вероятностью, равной $1/C(b)$, а с учетом меры близости страниц, т.е.

$$PR(a) = \frac{d}{n} + (1-d) \cdot \sum_{(b,a) \in G} \frac{sim(a,b) \cdot PR(b)}{\sum_{x \in C(b)} sim(x,b)}$$

$sim(a,b) = 1$ соответствует максимальной близости страниц a и b , а $sim(a,b)=0$ – их полному различию.

RUSSIR-2007. Сычев А.В.

Модификации алгоритма PageRank

- В алгоритме *TSPR* (Topic-Sensitive PageRank) [4] предложено предварительно группировать Web-страницы по тематике, а затем вычислять ранги страницы в каждом из тематических разделов. Сами тематические разделы отбираются из верхнего уровня ODP (Open Directory Project).
- Все страницы в тематическом разделе T_k образуют подмножество U_k . При вычислении вектора *PageRank* для тематического раздела алгоритм *TSPR* предполагает, что случайный пользователь перемещается только по страницам из множества U_k (фактически не выходит за рамки тематики).

RUSSIR-2007. Сычев А.В.

Модификации алгоритма PageRank

- Показатель *PageRank* для страницы $a \in U_k$ в этом случае рассчитывается следующим образом:

$$PR_k(a) = (1-d) \cdot \sum_{(a,b) \in G} \frac{PR_k(b)}{|C(b)|} + \begin{cases} d \cdot \frac{1}{|U_k|}, & \text{если } a \in U_k \\ 0, & \text{если } a \notin U_k \end{cases}$$

- Общая оценка для страницы a вычисляется следующим образом:

$$S_q(a) = \sum_k PR_k(a) \cdot r(q, T_k)$$

где $r(q, T_k)$ – оценка степени релевантности запроса q тематике T_k .

RUSSIR-2007. Сычев А.В.

Модификации алгоритма PageRank

- В модели *Intelligent Surfer* [5] расчет показателя *PageRank* предложено выполнять следующим образом:

$$PR_q(a) = d \cdot \frac{r(q, a)}{\sum_{x \in G} r(q, x)} + (1 - d) \cdot \sum_{(a, b) \in G} \frac{r(q, b)}{\sum_{x \in C(a)} r(q, x)} \cdot PR_q(b)$$

RUSSIR-2007. Сычев А.В.

Алгоритм *HITS*

- Hypertext Induced Topic Search*
- Поскольку короткие запросы приводят к выборке большого множества документов, то в рамках подхода, сформулированного в 1997 г. Кляйнбергом (Kleinberg), было предложено среди всех веб-страниц выделять два особых класса страниц: *авторитеты* и *концентраторы*.

RUSSIR-2007. Сычев А.В.

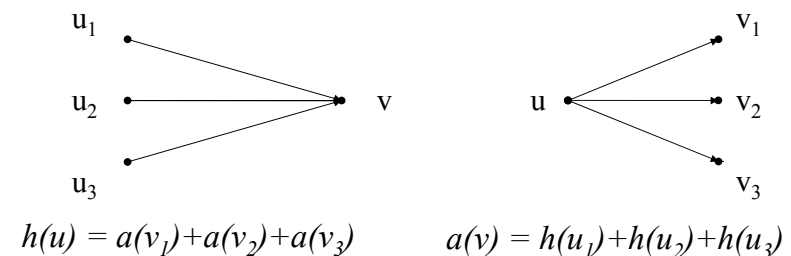
Авторитеты и концентраторы

- Хорошие *авторитеты* – страницы, которые содержат релевантную информацию (хорошие источники информации).
- Хорошие *концентраторы* – страницы, ссылающиеся на нужные страницы (хорошие источники ссылок).
- Эффект взаимного усиления:
 - Высокая авторитетность происходит из входящих ссылок от хороших концентраторов
 - Хороший концентратор имеет исходящие ссылки на хорошие авторитеты.

RUSSIR-2007. Сычев А.В.

Авторитеты и концентраторы

Показатель концентрации Показатель авторитетности



RUSSIR-2007. Сычев А.В.

Алгоритм HITS

- На основе ранжированной выборки по запросу пользователя формируется *стартовое множество* S документов (порядка двухсот первых документов из выданного списка).
- Путем использования входящих и исходящих ссылок на документы из S строится расширенное множество T документов (не более 50-ти для каждого стартового документа), находящихся на расстоянии 1 ребро от стартовых узлов в веб-графе.
- Простой учет количества входящих и исходящих ссылок на документы не является эффективным, поэтому далее следует итерационная процедура расчета показателей авторитетности и концентрации для всех узлов множества T .

RUSSIR-2007. Сычев А.В.

Процедура расчета весов авторитетности и концентрации

- Все веса инициализируются значением 1.
- Повторяется цикл до достижения сходимости:
 - Для узла v рассчитывается *вес авторитетности*

$$a(v) = \sum_{u_i \rightarrow v} h(u_i)$$

- Для узла v рассчитывается *вес концентрации*

$$h(v) = \sum_{v \rightarrow u_i} a(u_i)$$

- После каждой итерации выполняется нормализация весов

RUSSIR-2007. Сычев А.В.

Процедура расчета весов авторитетности и концентрации

- Так как алгоритм фактически вычисляет главные собственные векторы двух матриц, то векторы H и A должны сходиться, хотя точное значение числа итераций не известно. На практике вектора сходятся очень быстро.
- Математически данная задача идентична задаче поиска решения для системы уравнений:

$$A \cdot y = \lambda \cdot x, \quad A^T \cdot x = \mu \cdot y$$

- или

$$A \cdot A^T \cdot x = \lambda \cdot \mu \cdot x$$

RUSSIR-2007. Сычев А.В.

Алгоритм HITS Проблемы

- Поскольку используется относительно небольшая часть веб-графа, то добавление ребер к нескольким узлам может сильно изменить конечный результат.
- В большей степени подвержен манипулированию
- Взаимное усиление между хостами (за счет дочерних страниц)
- Динамически генерируемые ссылки
- Возможность попадания нерелевантных, но сильно связанных документов
- Как следствие - смещение темы

RUSSIR-2007. Сычев А.В.

Алгоритм HITS Расширения

- *ARC* (Automated Resource Compilation)
 - Расширение стартового подмножества за счет узлов на расстоянии 2 ребер
 - Использование текста анкерных тэгов (и их окружения) при расчете весов
- *SALSA* (Stochastic algorithm for link structure analysis).

RUSSIR-2007. Сычев А.В.

Различие между PageRank и HITS

- *PageRank* вычисляет веса для всех проиндексированных веб-страниц до запросов. *HITS* применяется только к веб-страницам, выданным по конкретному запросу пользователя.
- *HITS* находит авторитеты и концентраторы, *PageRank* – только авторитеты.
- *PageRank* – требует нетривиальных вычислений, *HITS* – простой алгоритм, но очень затратный по времени вычисления

RUSSIR-2007. Сычев А.В.

Недостатки “граф-гранулированной” (Coarse-Grained Graph) модели сети веб

- Артефакты авторства
 - ссылочный непотизм
 - клики
 - смешанные концентраторы
- Зашумление и смещение тематики

RUSSIR-2007. Сычев А.В.

Развитие моделей ранжирования

- Редукция множественных ссылок с одного сайта.
- Контентная предфилترация документов в расширенном множестве (для HITS) до начала процедуры расчета весов
- Учет текста ссылочных тэгов
- Учет структуры разметки документа

RUSSIR-2007. Сычев А.В.

Спамдексинг

- *Спамдексирование* или *поисковый спам* – это недобросовестная практика умышленного создания веб-страниц, индексируемых поисковыми системами, для повышения ранга веб-сайта (веб-страницы) в выдаче поисковой системы или воздействия на выбор категории, в которую он помещается. При этом реальное содержание документа не соответствует запросу пользователя.
- В определенной степени перекрывается с обычным стремлением веб-дизайнеров повысить доступность сайта (страницы) в Веб, т.е. *поисковой оптимизацией* (search engine optimization - SEO)

RUSSIR-2007. Сычев А.В.

Методы спамдексинга

Контентный спам:

- Внедрение в страницу неотображаемого или незаметного для пользователя текста (метатэги, комментарии, цвет шрифта и др.)
- Умышленное повышение частоты ключевых слов
- Внедрение ключевых слов/фраз в текст
- Манипулирование текстовым содержимым анкерных тэгов <a>.
- Копирование содержимого популярных страниц (по результатам поиска или напрямую) с добавлением рекламы и нужных ссылок.

RUSSIR-2007. Сычев А.В.

Методы спамдексинга

Ссылочный спам

- Ссылочные фермы – умышленное создание сообщества страниц, ссылающихся друг на друга.
- Создание невидимых для пользователя гиперссылок
- Создание нескольких веб-сайтов с разными доменными именами, ссылающихся друг на друга (спам-блог)
- Размещение бессмысленных сообщений с гиперссылками, содержащими ключевые слова, в гостевых книгах, форумах, дневниках и др.

RUSSIR-2007. Сычев А.В.

Методы спамдексинга

- Создание сетевых дневников исключительно с целью спамминга
- Инфильтрация веб-каталогов
- Скупка доменов, которые перестали использоваться
- Создание сайтов особого содержания (насилие, порнография и др.) с целью привлечения ссылок.

RUSSIR-2007. Сычев А.В.

Методы спамдексинга

Другие методы:

- Создание разных версий страницы для пользователей и сетевого робота-”паука”
- Использование автоматического перенаправления
- Создание зеркальных сайтов. Эффективно при поиске ключевых слов в URL

RUSSIR-2007. Сычев А.В.

Методы противодействия спамдексингу

- *BadRank* – вычисление понижающих коэффициентов при ссылке со страниц, оказавшихся в черном списке
- *TrustRank* – вычисление повышающих коэффициентов при ссылке со стороны доверяемых страниц (например каталогов поисковых систем)
- Обнаружение статистических аномалий (анализ частотного распределения, анализ распределения гиперссылок)

RUSSIR-2007. Сычев А.В.

Методы противодействия спамдексингу

- *SpamRank* – анализ распределения PR в подмножестве соседних узлов веб-графа.
- Использование вероятностных моделей при расчете ссылочной авторитетности (позволяет перейти от дискретной шкалы к непрерывной)

RUSSIR-2007. Сычев А.В.

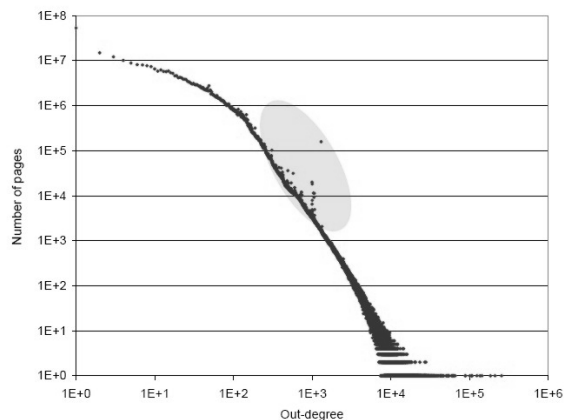
Обнаружение статистических аномалий

- В работе [10] приведена идея и результаты эксперимента по выявлению ссылочного спама на основе анализа статистических аномалий в распределении входящих и исходящих гиперссылок.
- Для эксперимента использовались данные, полученные из базы Yahoo! объемом 429 миллионов HTML страниц (в интервале между июлем и сентябрем 2002 г.).
- Среднее количество гиперссылок на странице - 62.55, уникальных – 42.74.
- Для оценки доли спама использовалась выборка размером в 1000 URL из данного тестового набора данных. Из них 465- не удалось загрузить или они не имели содержимого. Из остальных 535 страниц – 37 содержали спам (6,9%).

RUSSIR-2007. Сычев А.В.

Обнаружение статистических аномалий

Исходящие ссылки



RUSSIR-2007. Сычев А.В.

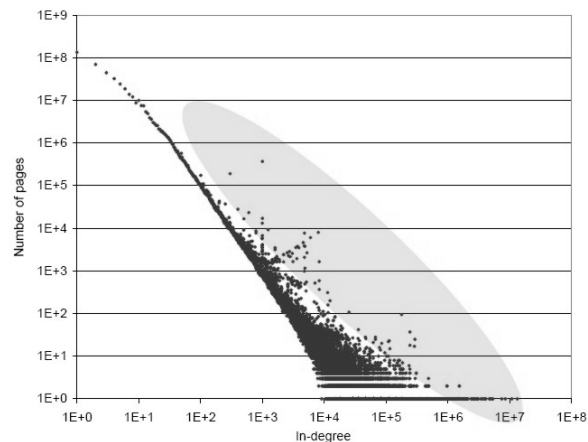
Обнаружение статистических аномалий

Исходящие ссылки

- В целом форма распределения соответствует закону Ципфа.
- Однако, на графике имеются хорошо заметные аномалии. Например, ожидаемое количество страниц, имеющих 1301 входящую ссылку равно 1700. Однако в реальности (как видно на графике) их оказалось 158290.
- Набор данных содержал 0.05% страниц, имеющих количественное превышение относительно закона Ципфа в 3 и более раз. Сравнительный анализ этих страниц показал, что все из них содержали спам.

RUSSIR-2007. Сычев А.В.

Обнаружение статистических аномалий



RUSSIR-2007. Сычев А.В.

Обнаружение статистических аномалий

Входящие ссылки

- График в еще большей степени соответствует закону Ципфа
- Однако, и аномалий наблюдается существенно больше.
- Например, страниц, имеющих 1001 входящую ссылку, вместо ожидаемых 2000 оказалось в реальности (на графике) 369457.
- Всего в тестовом наборе данных оказалось 0.19% страниц, имеющих 3-х и более кратное превышение относительно закона Ципфа. Анализ показал, что большинство из них – спам.

RUSSIR-2007. Сычев А.В.

Литература

1. S.Brin, L.Page "The anatomy of a large-scale hypertextual web search engine". Proc. 7th World Wide Web Conf. (WWW7), p. 107–117.
2. L.Page, S.Brin, R.Motwani, T.Winograd. "The PageRank citation ranking: Bringing order to the Web". Stanford, Digital Library Technologies, Working Paper 1999-0120, 1998.
3. P. Ingongngam., A. Rungsawang "Topic-centric algorithm: a novel approach to Web link analysis" //Proceedings of the 18th International Conference on Advanced Information Networking and Applications (AINA'04). -Fukuoka, Japan, 2004. - Vol.2. - P. 299 - 301.
4. Haveliwala T.H. "Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search" //IEEE Transactions on Knowledge and Data Engineering. - 2003. - Vol. 15. - №4. - P. 784 - 796.
5. Richardson M., Domingos P. "The intelligent surfer: probabilistic combination of link and content information in PageRank" //Advances in Neural Information Processing Systems. MIT Press. 2002. - Vol. 14. - P. 1441 - 1448. RUSSIR-2007. Сычев А.В.

Литература

6. M. Kleinberg "Authoritative sources in a hyperlinked environment". Journal of the ACM, 46(5):604–632, 1999.
7. Р.М.Алгулиев, Р.М.Алыгулиев "Ранжирование web-страниц с использованием взаимной информации между гиперссылками"//Проблемы управления, 2007, №.
8. T.Jones "Both Sides of the Digital Battle for a High Rank from a Search Engine". Association for Computing Machinery New Zealand Bulletin, 2005.
9. G. Weikum "Information Retrieval and Data Mining". Слайды. (http://www.mpi-sb.mpg.de/departments/d5/teaching/ws05_06/irdm/index.html)
10. D.Fetterly, M.Manasse, M.Najork "Spam, damn spam, and statistics. Using statistical analysis to locate spam web pages" // 7-th International Workshop on the Web and Databases, June17-18, 2004, Paris, France. RUSSIR-2007. Сычев А.В.