



ТЕХНОСФЕРА

Выявление спам сайтов на основе анализа контента страниц

Москва 2016

План лекции

- Актуальность проблемы
- Почему спам существует, в чем основная проблема
- Методы воздействия спама на поисковик и способы противодействия
- Детекция спама на основе анализа контента страниц
- Методика выявления спам-сайтов
- Антифрод (роботы, мошенничество)
- Спам в других приложениях

Актуальность проблемы

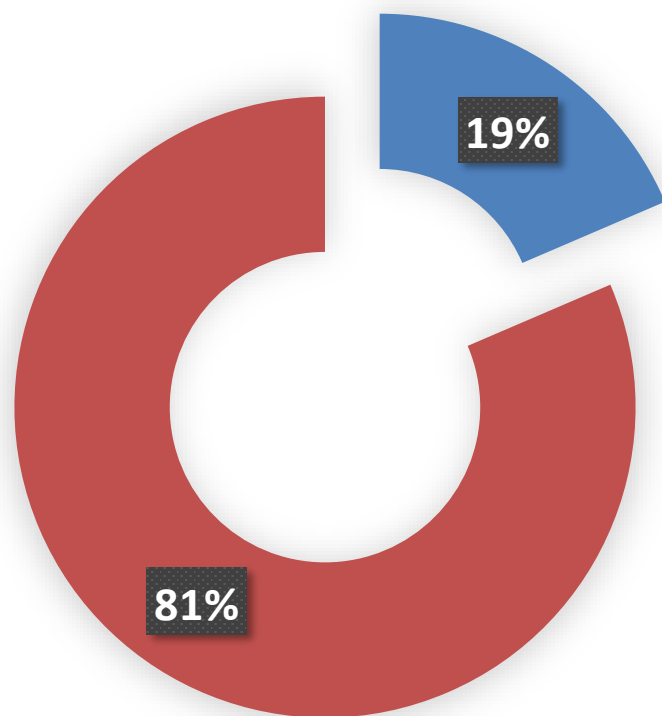
- Поисковики
- Соц.сети - VK, ОК (группы, видео)
 - Добавление в друзья
 - Сообщения
 - Спам в постах групп
 - Спам в комментариях
- Месенджеры (viber, whatsapp, instagram)
- Любые сайты где пользователи могут что-то писать (магазины, кинопоиск, форумы и тд)

Объемы рекламного рынка



Сегменты	Январь-Сентябрь 2015 года, млрд.руб.
Телевидение	90,30
Радио	9,40
Печатные СМИ	16,10
Наружная реклама	24,10
Интернет	64,70
Прочие	2,60
ИТОГО	208,50

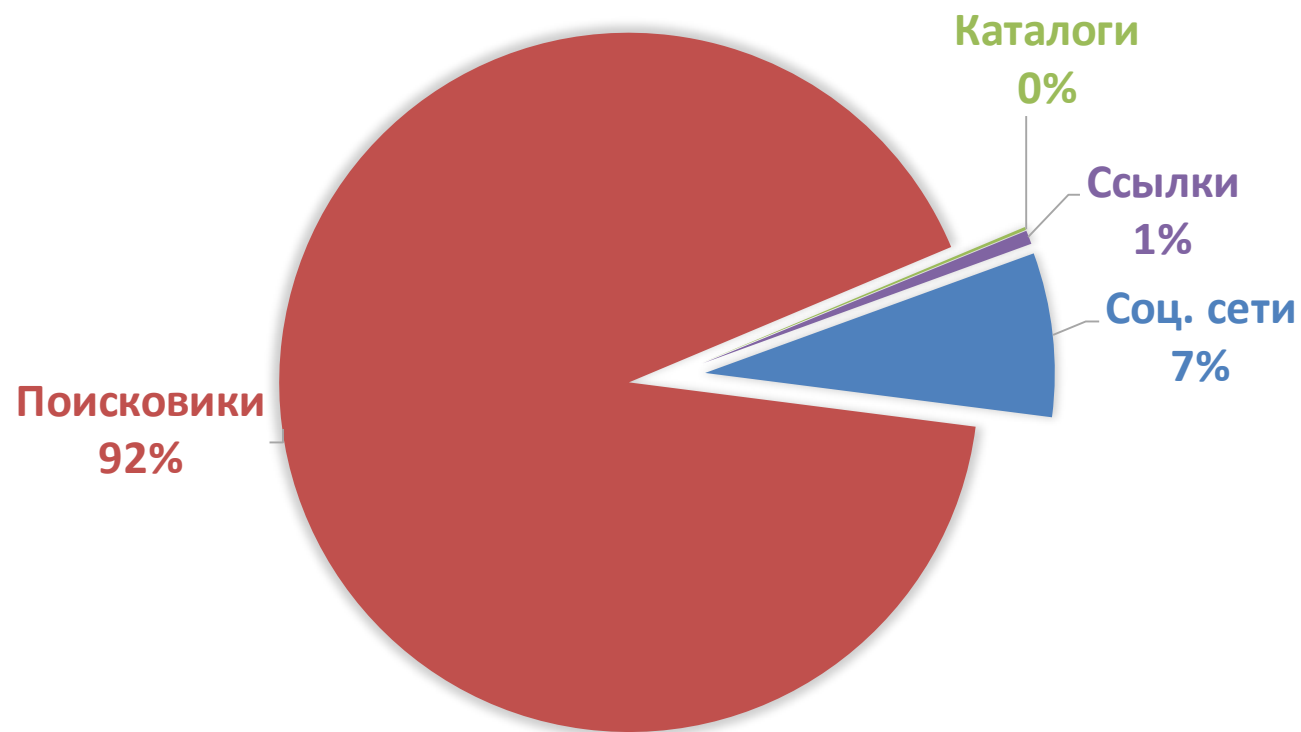
Реклама в интернете



■ Медийная реклама ■ Контекстная реклама

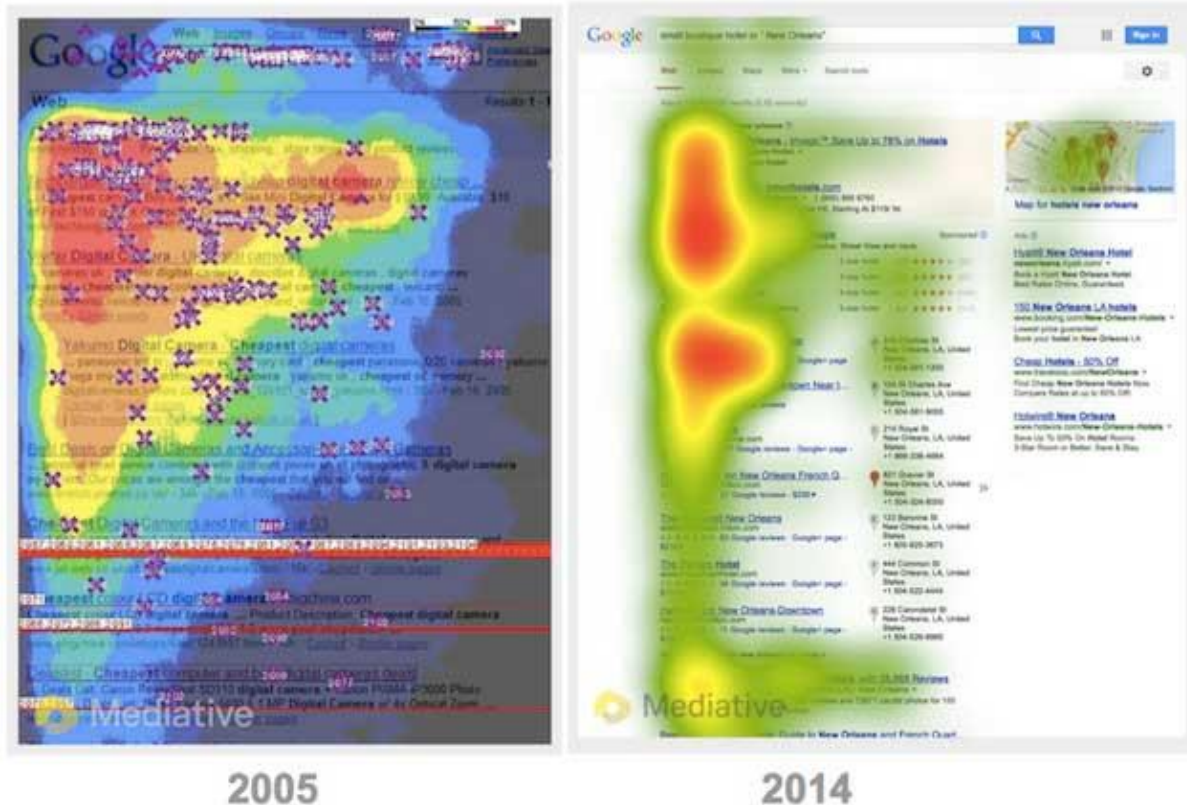
Сегменты	Январь-Сентябрь 2015 года, млрд.руб.
Медийная реклама	12,00
Контекстная реклама	52,70

Переходы с различных источников



Источник	Переходы, мл.
Соц. сети	212
Поисковики	2562
Каталоги	4
Ссылки	18

Eye-Tracking Study: Как пользователи просматривают результаты поиска



Source: The Evolution of Google Search Results Pages, Mediative, 2014

53 – участники
43 – поисковых задачи
85% запросов получают клики в 1 результат
Клики получают, в основном, первые 3 – 5 позиций

Мотивация

- Попадание в топ выдачи имеет под собой чисто экономическое обоснование
- Больше пользователей - больше выгода



А в чем проблема?



купить телевизор в ашане

[buenamebel.ru/gavprosh](#)
Тумбы под телевизор. Трельяжи, Туалетные Столики, Трюмо. ... Габаритные размеры: Спальное место: ... Наполнение спального места: пенополиуретан (ППУ)

Телевизоры ашан | Магазин АШАН
[auchan-shop.info/televizoryi-ashan](#)
Related terms: Акай, Akai Ap100с, АШАН акции, АШАН лето, АШАН 50 лет, АШАН Ассортимент, Телевизоры Akai, Кредитная... Магазины Ашан в Дисконтной программе...

Шкаф купе "Стиль-3-39" | Мебель на заказ купить стенка недорого в...
[mebelwell.ru/шкаф-купе-стиль-3-39](#) ▶ копия
Тумбы под телевизор. ... Полезная информация о недорогой мебели. ... В связи с ситуацией в стране, цены на мебель уточняйте у диспетчера!

2 группа ашан интернет магазин мебель, мебель в магазине ашан,...
[mebell-elita.ru/2-группа-0](#)
Тумбы под телевизор. ... О-Лайн овертайм - крупный план. О-Лайн опцион - крупный план. О-Лайн опера аутомн

АШАН - Группы товаров
[auchan.ru/ru/ashan_samara_...](#)
Телевизоры, домашние кинотеатры, фотоаппараты, видеокамеры, автомагнитолы, ноутбуки. Товары для отдыха и спорта

Что можно купить в Ашане?
[servstory.ru/index.php/roznitsa1...](#)
Home Розница Супермаркеты Что можно купить в Ашане? ... В электронном отделе есть часы-радио, фотоаппараты, плейеры, телевизоры.

А в чем проблема?

Генерация большого количества мусорного контента



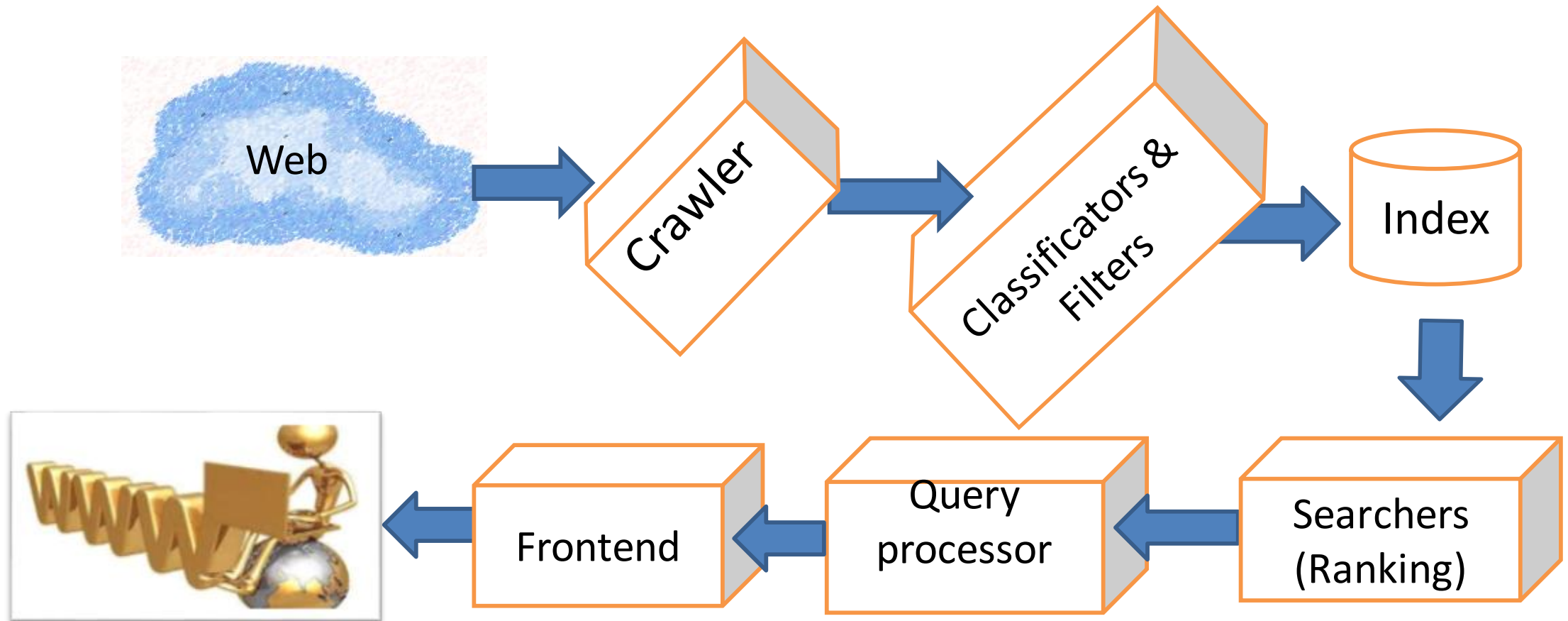
Что мы хотим получить?

- Уменьшить вероятность попадания спама в индекс
- Уменьшение количества поискового спама в выдаче поиска

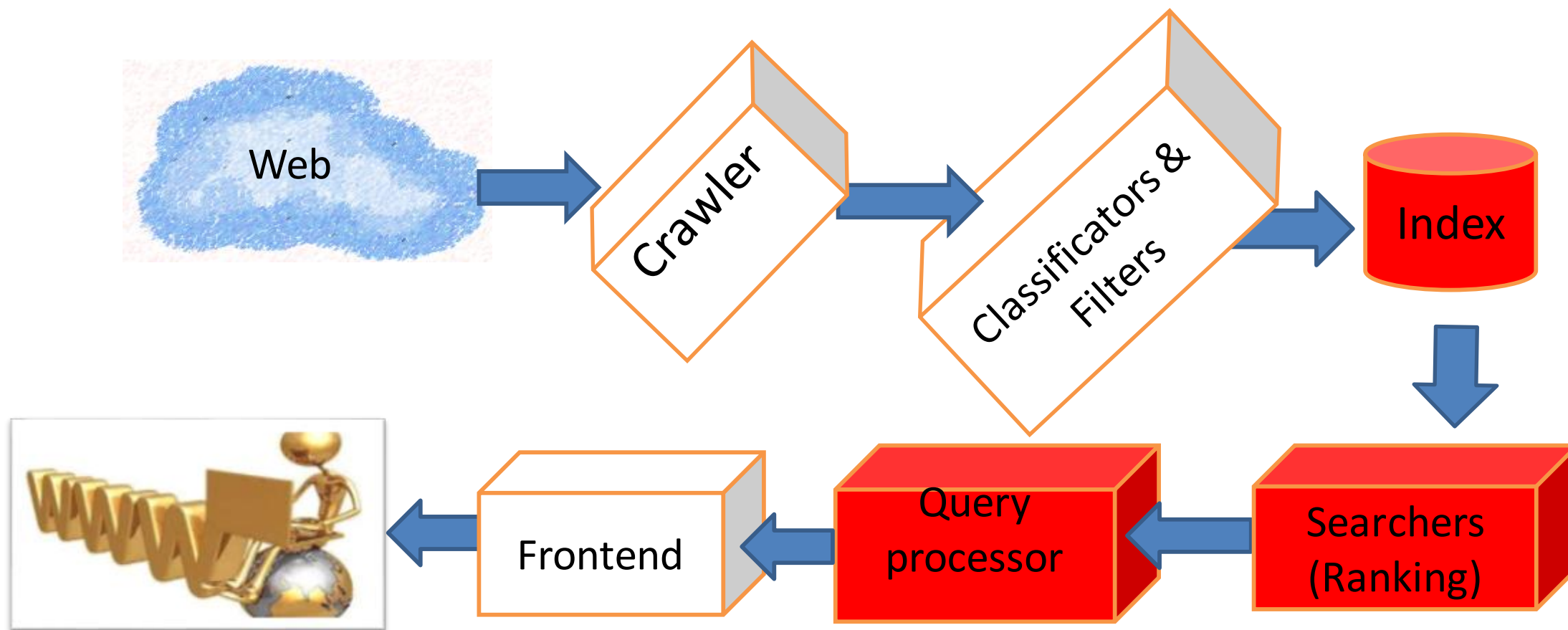
~10% - страниц в индексе это спам



Куда бьет спам?



Куда бьет спам?



Куда бьет спам?

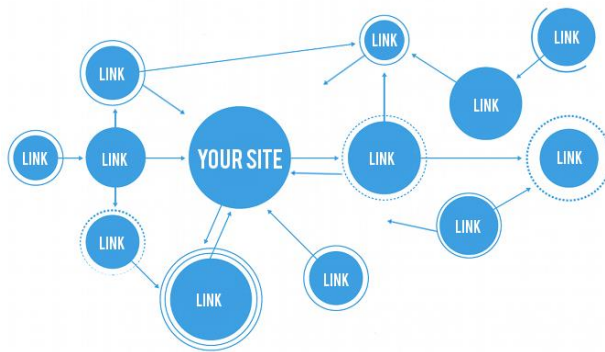
- Индекс – его объем ограничен и спам занимает место полезных документов
- Обработка запросов – накручивают нужные саджесты
- Ранжирование – пытаются пробиться в топ выдачи и мешают ранжированию
- На самом деле бьет и по всем остальным частям:
 - Crawler забивает очередь обкачки (падает актуальность)
 - Тратит ресурсы классификаторов (порно и тд)
 - Фронтенд только не задеват явно

Как воздействовать на систему ранжирования?



Основные компоненты ранжирования

1. Поведенческое ранжирование
2. Ссылочное ранжирование
3. Текстовое ранжирование



CTR



Поведенческое ранжирование CTR

Click Through Rate

$$CTR_q = \frac{C_q}{V_q}$$

C_q - количество кликов для запроса q

V_q - количество показов для запроса q

Поведенческий спам CTR

$$CTR_{qf} = \frac{C_{qu} + C_{qf}}{V_{qu} + V_{qf}} = \frac{C_{qu}}{V_{qu}} \left(\frac{1 + C_{qf}/C_{qu}}{1 + V_{qf}/V_{qu}} \right)$$

f - спам клики и показы

u – чистые клики и показы

$f/u \propto C_{qf}/C_{qu} \propto V_{qf}/V_{qu} \sim$ доля оригинальной статистики

Поведенческий спам CTR

Статистика запросов за месяц:

купить телевизор ~ 61 845,

купить холодильник ~ 50 074

...

На частотных запросах клик спам может стать экономически невыгодным, требуется сгенерировать статистику равноценную оригинальной.

На какие запросы можно повлиять

- Средне и низкочастотные, особенно если хорошего релевантного сайта нет
 - Как вставить наушники в iphone7 (и тут сайт с гаджетами, которые сверлят дырку)
- Трендовые заранее прокачанные:
 - Игра престолов 9 сезон
 - Физрук 5 сезон
 - Если заранее сделать спам под будущие тренды можно попасть временно, а если сайт не совсем гавно то и длительно в топ выдачи

Как влияют на поведение

- Владельцы сайтов накликают свой сайт и просят друзей
- Боты кликающие, боты просто сканирующие выдачу
- Люди выполняющие задания за деньги («заработок в интернете»)

Поведенческое ранжирование vs спам Learning to Rank

Поведение как признаки модели.

$X = \{x_1, \dots, c_1, \dots, c_m, \dots, x_n\}$; - полное пространство признаков

N – размер вектора, M – количество поведенческих признаков

$F(X; \{\alpha, \beta\}) = \sum_{i=0}^K \alpha_i \cdot h_i(X, \beta_i)$; - модель

$P = \{\alpha, \beta\}$ - параметры модели $\succ \arg \min F(P) \mid \arg \max F(P)$

тогда $\succ c_i = c_{iu} + c_{if}$

c_u - клики реальных пользователей

c_f - спам клики, можно считать шумом

Выбрать алгоритм устойчивый к шуму.

Ссылочное ранжирование

- PageRank (по запросу «pagerank алгоритм» куча бредовых статей SEO)
- TrustRank
- PersonalizedPageRank
- Тексты ссылок

Ссылочное ранжирование

Модель веб графа:

$$G = (V, E)$$

V – вершины графа – страницы

E – ребра графа – ссылки между страницами.

w_{ij} - вес ребра между страницами p_i и p_j ; $(i, j) \in E$

$$w_{ij} = \frac{1}{|Out(p_i)|}$$

$|Out(p_i)|$ - количество исходящих ссылок со страницы p_i

Матрица переходов

$$M = \begin{bmatrix} w_{01} & w_{01} & w_{0j} \\ w_{11} & \dots & w_{1j} \\ w_{i0} & w_{i1} & w_{ij} \end{bmatrix} \Rightarrow M = \begin{cases} w_{ij}, & \text{if } (i, j) \in E \\ 0 & \end{cases}$$

Идея PageRank

Идея заключается в том, чтобы посчитать вероятность того, что пользователь окажется на данной странице, если будет случайно блуждать по интернету (когда это придумали интернет был маленький и это было даже реалистично).

Очевидно, что эта вероятность зависит от кол-ва ссылок,

входящих на страницу A,

а вероятность что по ссылкам перейдут с B на A

обратно пропорциональна кол-ву исходящих с B ссылок

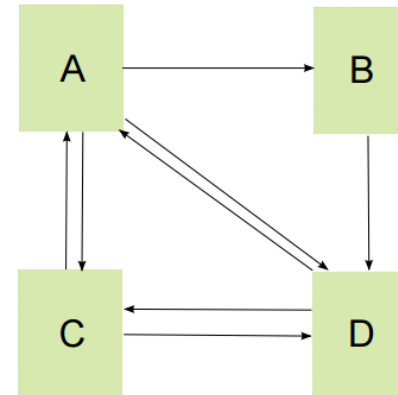
Соответственно вероятность оказаться на каждой странице

в начальный момент времени это вектор p состоящий из компонент $1/\text{кол-во страниц}$

$$p_2 = Mp_1$$

$$p_3 = Mp_2 = MMp_1$$

$$p_n = \prod_{i=1}^n M p_0$$



Let N be the total number of pages. We create an $N \times N$ matrix \mathbf{A} by defining the (i, j) -entry as

$$a_{ij} = \begin{cases} \frac{1}{L(j)} & \text{if there is a link from } j \text{ to } i, \\ 0 & \text{otherwise.} \end{cases}$$

In Example 1, the matrix \mathbf{A} is the 4×4 matrix

$$\begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1 & 1/2 & 0 \end{bmatrix}.$$

Ссылочное ранжирование PageRank

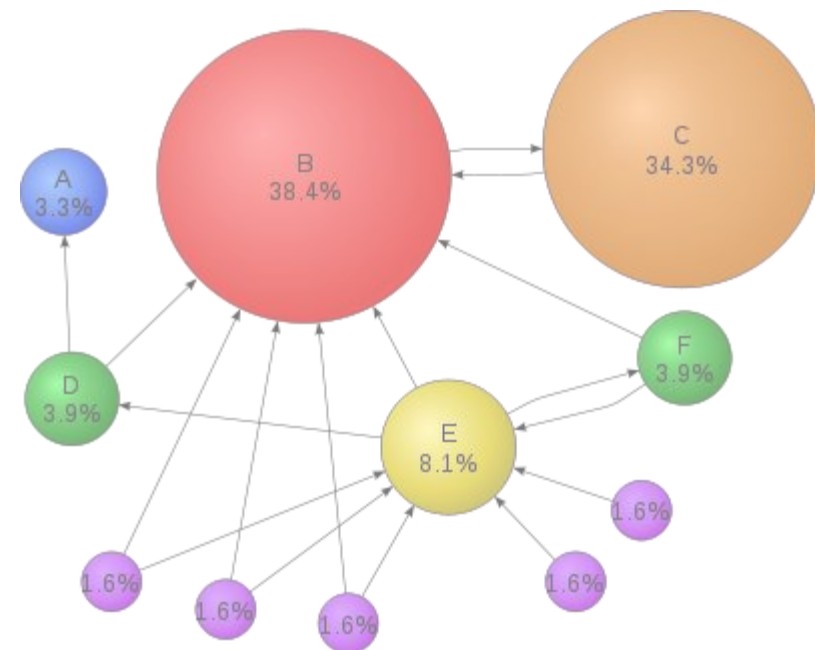
Вес страницы подбирается итерационно:

$$\vec{\pi} = (1 - c) M^T \vec{\pi} - c \vec{r}$$

c — дампинг фактор

\vec{r} - статический вектор.

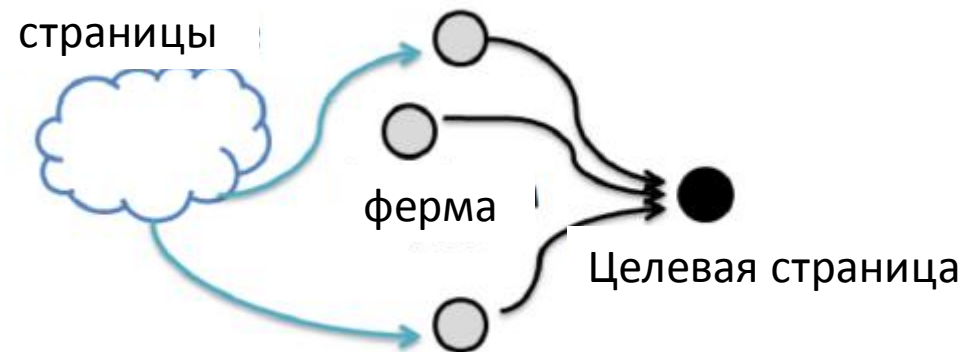
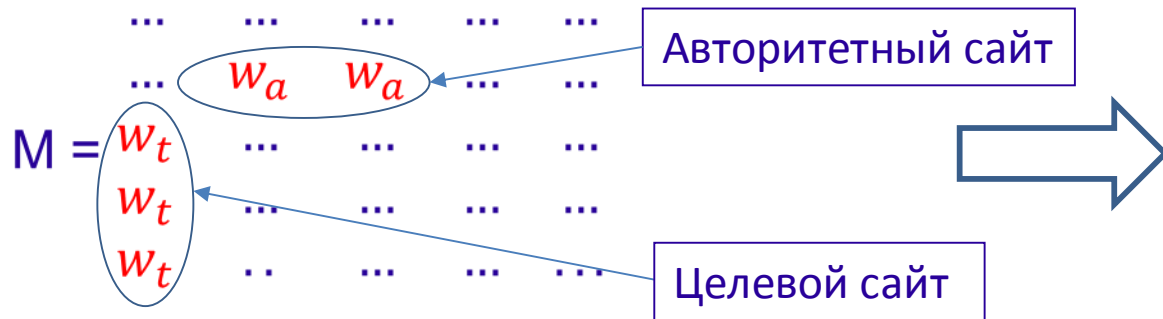
$\vec{r} = \left(\frac{1}{|V|}, \dots, \frac{1}{|V|} \right)$ - для не персонализированного PageRank. Гарантирует существование стационарного распределения $\vec{\pi}$ соответствующей цепи Маркова, описанной не стохастической матрицей M .



Ссылочный спам

M – стохастическая матрица $p_{ij} > 0$; $\sum_{j=1}^V p_{ij} = 1$; $\forall i$ для цепи Маркова
 $\vec{\pi}$ - стационарное распределение матрицы M

Для влияния на стационарное распределение расширим матрицу M



Увеличиваем ранк целевой страницы через ферму ссылок

Ссылочный антиспам

Распространение меток

Ключевая идея:

$V = \{p_1, \dots, p_v\}$ – множество страниц

$\tilde{V} \subseteq V$ - множество страниц с метками.

Цель: рассчитать значение меток для остальных сайтов через правила распространения меток.

Ссылочный антиспам

Распространение меток

- TrustRank - \tilde{V} - доверенные страницы + персонализированный PR
 - Для определения доверия используем обратный PR
 - Размечаем K лучших страниц получаем \tilde{V}
 - Делаем персонализированный вектор \vec{r}
 - Считаем PR

Ссылочный антиспам

Признаки из ссылок

Степень ссылочности (входные, выходные ссылки)
PageRank (PR, In-degree/PR, Out-degree/PR, STD(PR))
TrustRank (TrustRank, TrustRank/PR, TrustRank/In-degree)

$X = (x_1, \dots, x_n)$; - полное пространство признаков

$F(X; \{\alpha, \beta\}) = \sum_{i=0}^K \alpha_i \cdot h_i(X, \beta_i)$; - классификатор

$P = \{\alpha, \beta\}$ - параметры модели $\triangleright \arg \min F(P)$

Ссылочный антиспам

Подкрепление меток

Маркируем спам страницы используя классификатор

Кластеризуем страницы, используя граф ссылок $G = (V, E)$

Страница получает метку спам если большинство страниц в кластере спам и наоборот

Ссылочный антиспам

Признаки

- PageRunk, TrustRank, PersonalizedPageRank
- Признаки по кол-ву исходящих/входящих ссылок
- Принадлежность компоненте связности и признаки этой компоненты
- Lapel propagnation с точно спамовых страниц по ссылкам
- P.S. Надо понимать что ссылочные алгоритмы могут не работать на небольшой подвыборке

Текстовое ранжирование.

Модели для текстового ранжирования:

- Модель векторного пространства
- BM25
- Статистическая языковая модель



Текстовое ранжирование.

Модель векторного пространства

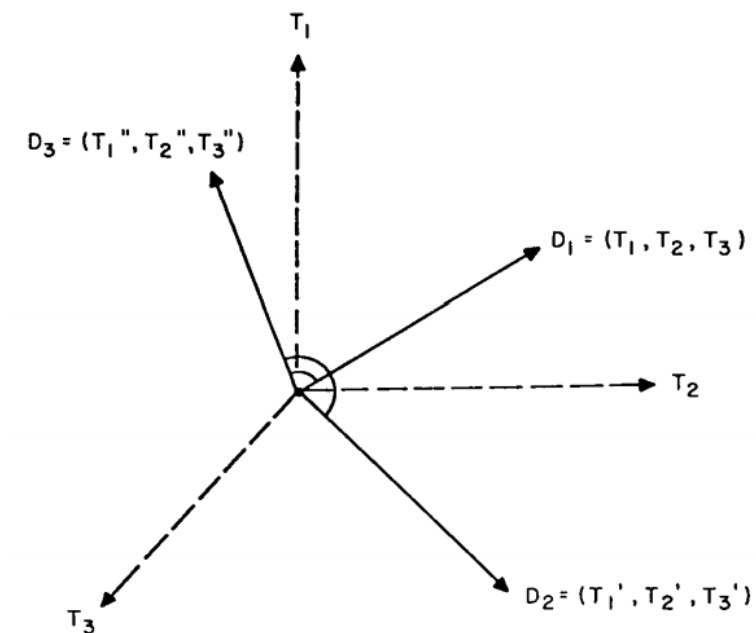
$D_i = \{w_{i1}, w_{i2}, w_{i3}, \dots, w_{it}\}$ - вектор документа
t – размерность вектора. $|V|$ - размерность словаря
 w_{ij} - вес j-го термина в документе i

$s(Q, D_i)$ - Мера сходства документа и запроса

$$w_{ij} = tf_{ij \in Q}(t_j, d_i) \cdot idf_j(t_j, D)$$

$tf_{ij \in Q}$ - частота j слова в документе i

idf_j - инвертированная частота слова j



Текстовое ранжирование BM25

Вес слова j в документе i

$$w_{ij} = \frac{(k_1 + 1) \cdot tf_{ij}}{k_1 + ((1 - b) + b \frac{l_i}{avg(l_i)})} \cdot \log \frac{|D| - df_j + 0.5}{df_j + 0.5}$$

$\propto idf_j(t_j, D)$

Вес документа d для запроса q :

$$W(q, d) = \sum_j w_j(d) \cdot q_j$$

k_1 и b – параметры
 l_i – длина документа
 tf_{ij} – частота слова в документе
 df_j – частота слова в коллекции
 $|D|$ – количество документов

Текстовое ранжирование.

Вероятностная языковая модель

Модель – Бернулли. Слово w есть или нет в документе d

$$p(q = (x_1, x_2, \dots, x_{|V|}) | d) = \prod_{i=1; x_i=1}^{|V|} p(w_i = 1 | d) \prod_{i=1; x_i=0}^{|V|} p(w_i = 0 | d)$$

Мульти-номинальная модель. Моделирование частоты слов

$$p(q = q_1 \dots q_m | d) = \prod_{i=1}^{|V|} p(w_i | d)^{c(w_i, q)}; \sum_{i=1}^{|V|} p(w_i | d) = 1$$

$q = q_1, \dots, q_m$ - слова запроса,

$c(w_i, q)$ - частота слова i в запросе q

Ранжирование на основе правдоподобия запроса

$$\log p(q | d) = \sum_{i=1}^{|V|} c(w_i, q) \cdot \log p(w_i | d) -$$

$$p(w | d) = \frac{c(w, d) + 1}{|d| + |V|} \text{ - вероятность вхождения слова в документ}$$

$C(w, d)$ - частота
слова в документе
 $\propto tf_{ij}$

Текстовое ранжирование

Что общего?

$$tfidf(q, p) = tf_{ij}(t_j, d_i) \cdot idf_j(t_j, D)$$

$$idf_j = \log \frac{|D|}{n_d} \quad - \text{инверсная частота терма}$$

$$tf_{ij}(t_j, d_i) = \frac{f_j}{\sum_{l=1}^L f_l} \quad - \text{частота терма } j \text{ в документе } i$$

Увеличивая частоту слова в документе,
увеличиваем вероятность его нахождения

Текстовое ранжирование BM25 зоны

Зоны – различные части документа, по которым можно считать ранк BM25.

Пусть документ разбит на K зон тогда суммарный ранк документа:

$$W(q, d, v) = \sum_{i=1}^K v_i W_i(q, d)$$

v – вес зоны документа

Увеличение частоты слова в различных зонах документа, по разному влияет на его вероятность нахождения.

Контекстный антиспам классификатор

$X = (x_1, \dots, x_n)$; - полное пространство признаков

$F(X; \{\alpha, \beta\}) = \sum_{i=0}^K \alpha_i \cdot h_i(X, \beta_i)$; - классификатор

$P = \{\alpha, \beta\}$ - параметры модели $\triangleright \arg \min F(P)$

Контекстный антиспам прочее

Выявление спама через нахождение дубликатов

(“Detecting phrase-level duplication on the world wide web.” Dennis Fetterly)

Выявление спама через сравнение языковых моделей

(“Blocking Blog Spam with Language Model Disagreement ” Gilad Mishne)

$$KL(\theta_1 || \theta_2) = \sum_w p(w|\theta_1) \log \frac{p(w|\theta_1)}{p(w|\theta_2)}$$

...

Методы воздействия на поисковый механизм:

- Перенасыщение заголовков ключевыми словами.
- Перенасыщение текстов ключевыми словами.
- Оптимизация текстов под одно ключевое слово.
- Оптимизация текстов под большое количество ключевых слов.
- Оптимизация анкоров ссылок под ключевые слова.
- Активный обмен ссылками.
- Фермы ссылок.
- ...



Классификация воздействий на ПОИСКОВЫЙ МЕХАНИЗМ

- Воздействие при помощи оптимизации контента страницы.
- Воздействие при помощи оптимизации ссылок.
- Воздействие на поведенческие факторы.

...

Вопрос:

Разработка в каком направлении даст
лучшие результаты?

В 2006 году в рамках материалов конференции IW3C2 была опубликована статья: «Выявление спам-страниц через анализ контента» («Detecting Spam Web Pages through Content Analysis». А. Ntoulas и коллектив авторов).



В статье показано, что **86%** спама можно вычислить на основе анализа контента страниц.



Разработка в направлении детекции контекстного спама даст лучший профит.

Нам интересны более простые методы выявления искусственности страниц.

Достаточно просто поддерживать в актуальном состоянии.

Использовать для классификации спама с высокой точностью.

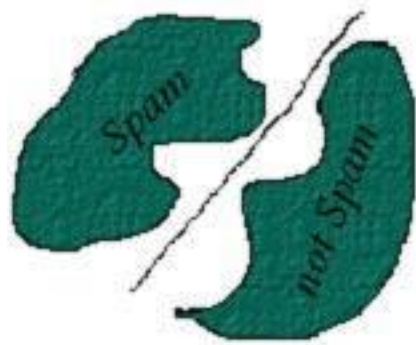
Рассмотрим проблему обнаружения спам страниц как задачу бинарной классификации.

- 1 — спам
- 0 — не спам

Требуется:

1. Определить пространство признаков.
2. Определиться с методом классификации.

Качество классификации напрямую зависит от качества признаков, описывающих пространство.



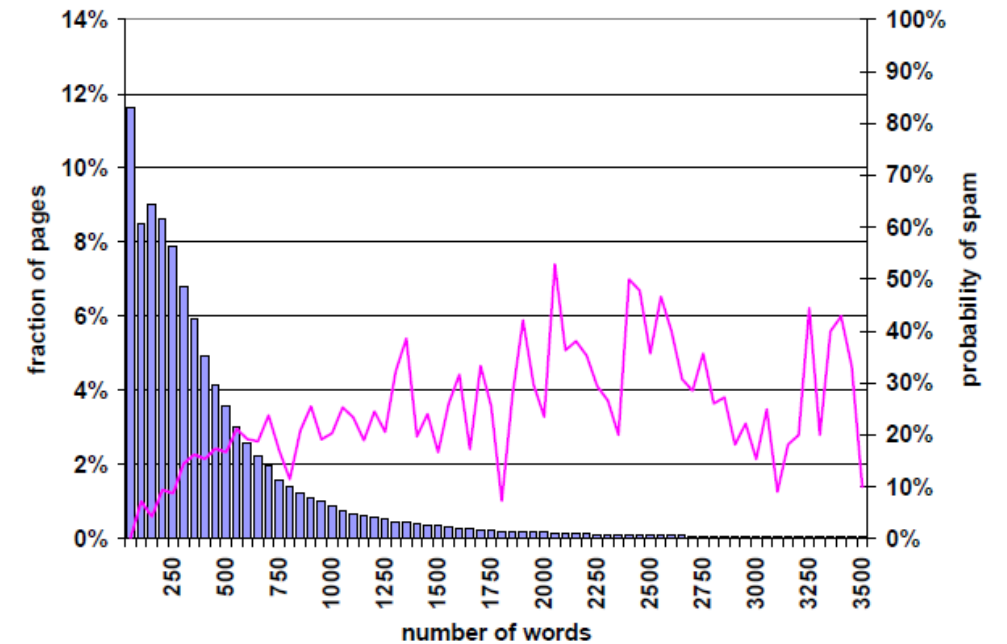
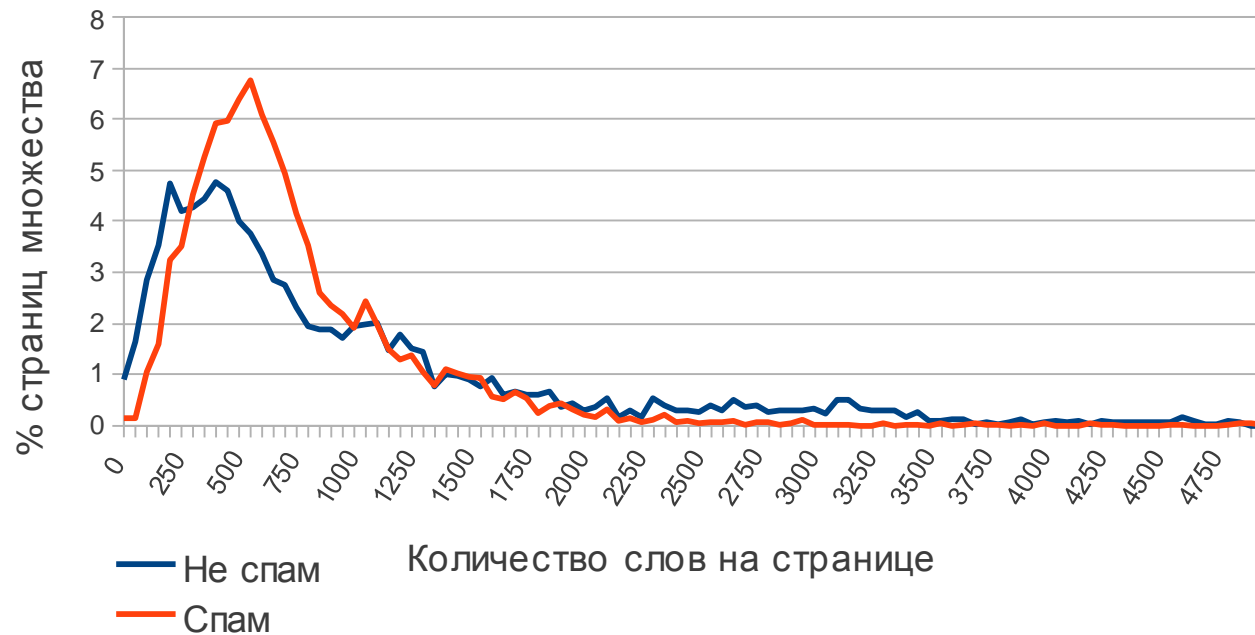
Линейно
разделимые
признаки



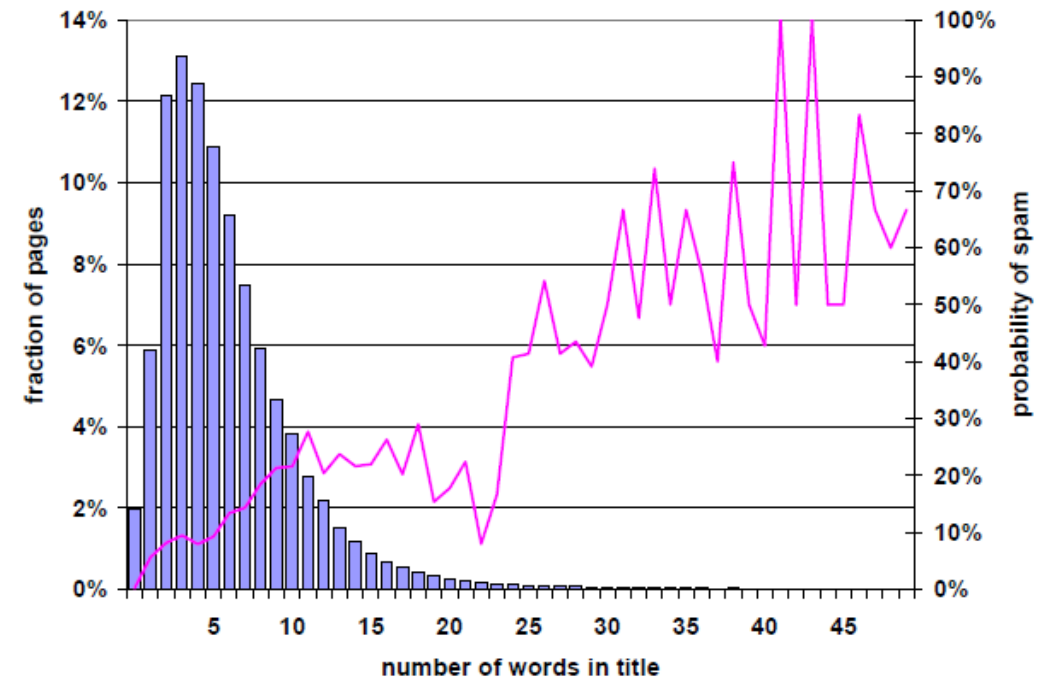
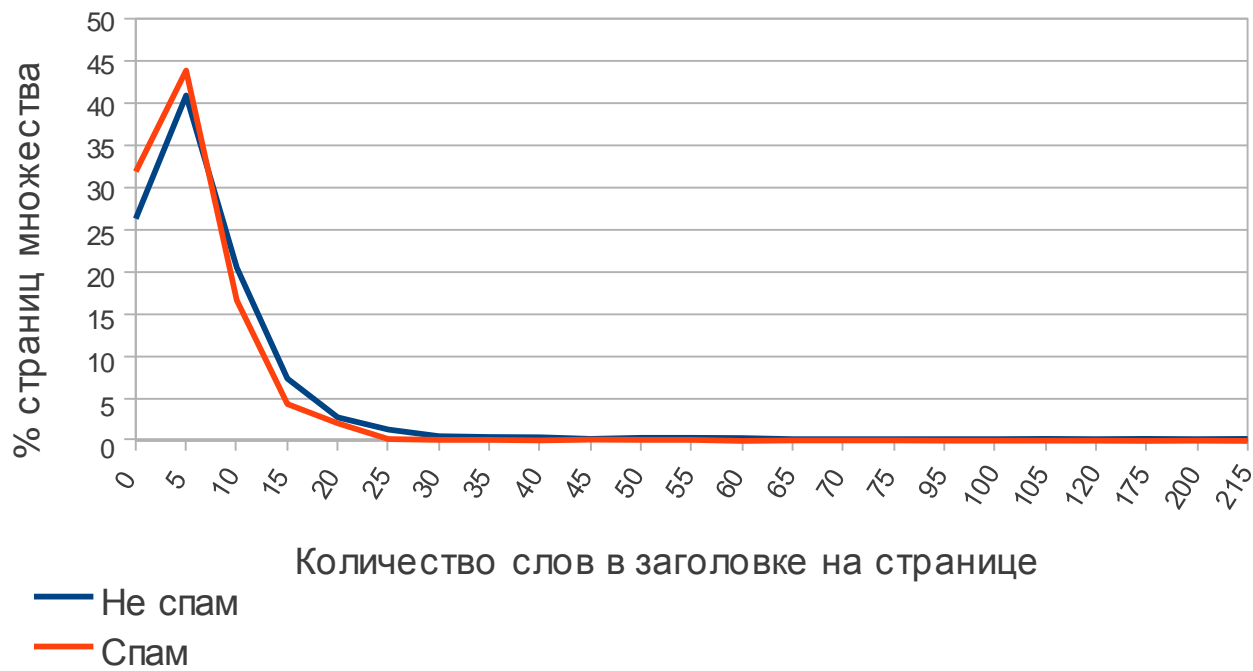
Линейно
неразделимые
признаки.

Выделение небольшого количества хорошо разделимых признаков позволит нам решить задачу классификации с большей эффективностью.

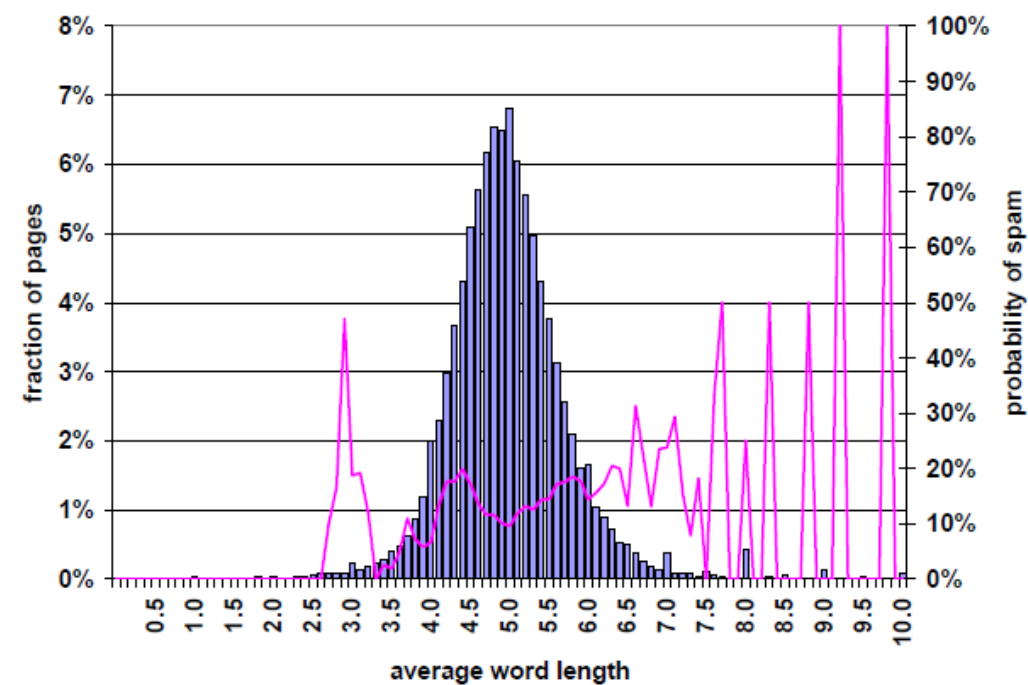
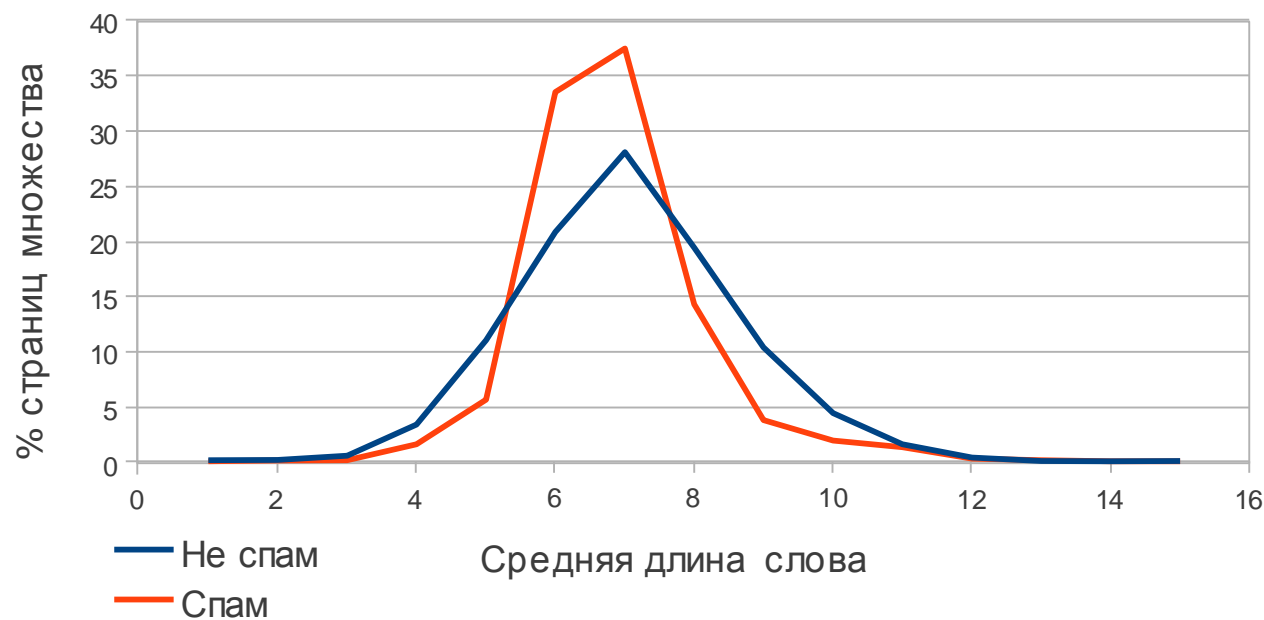
Распределение количества слов на странице в спамовых и неспамовых множествах



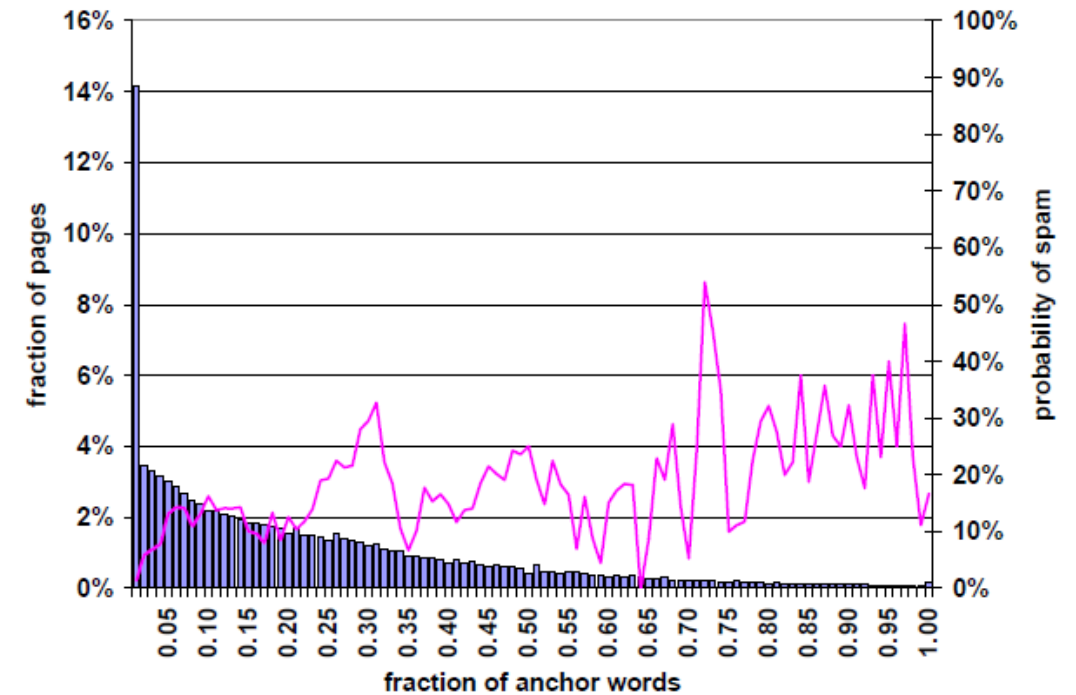
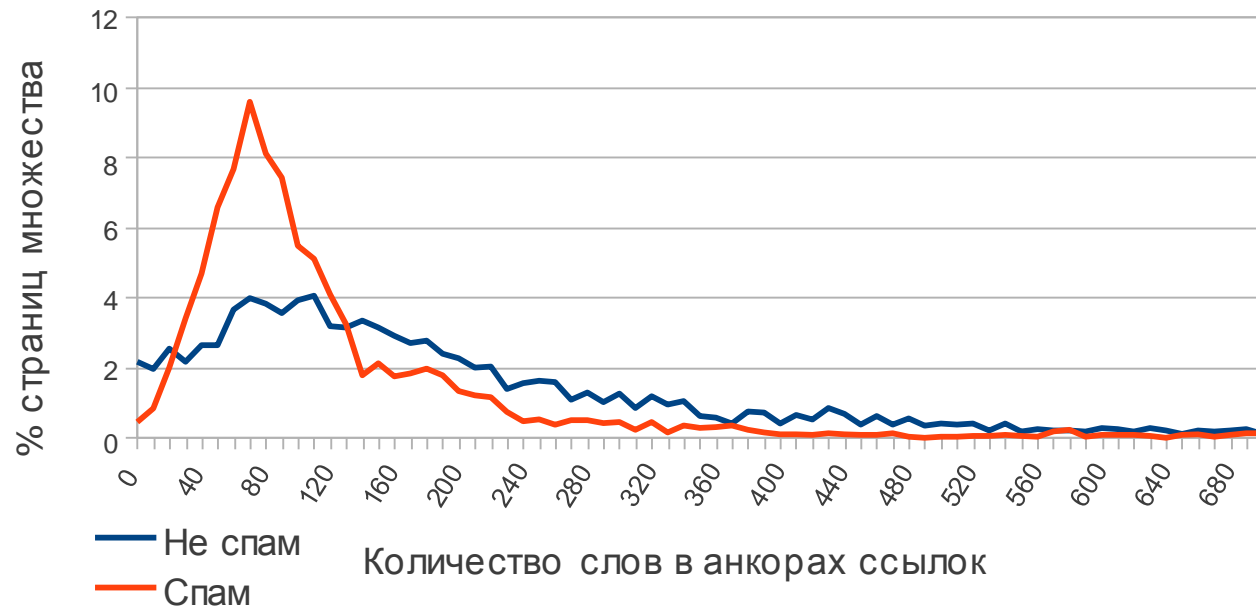
Распределение количества слов в заголовке страниц в спамовых и неспамовых множествах.



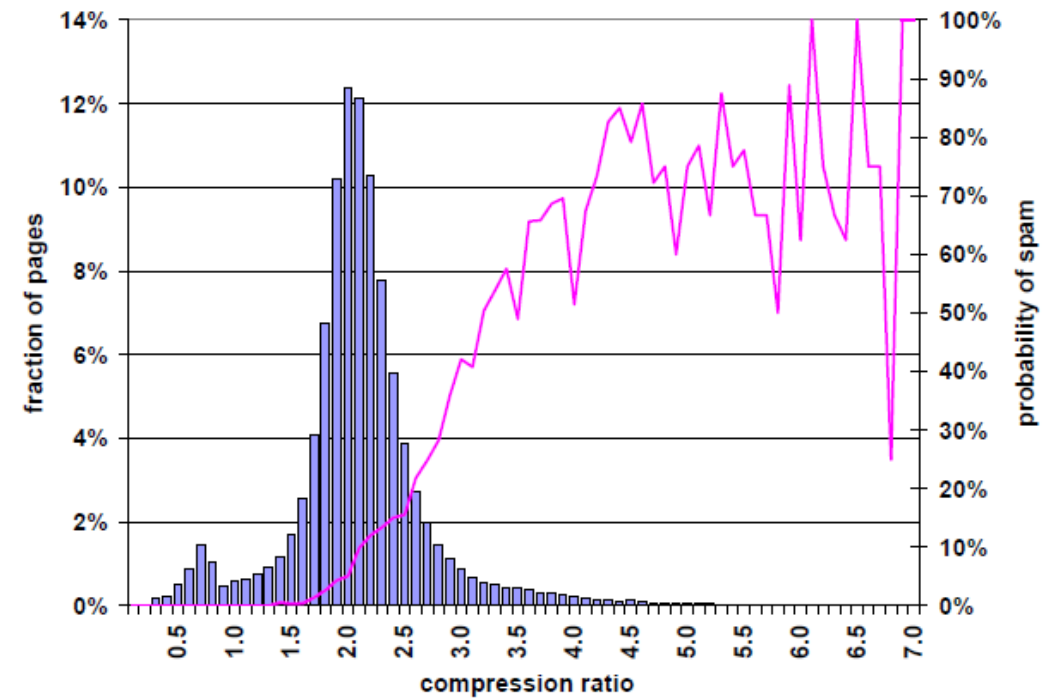
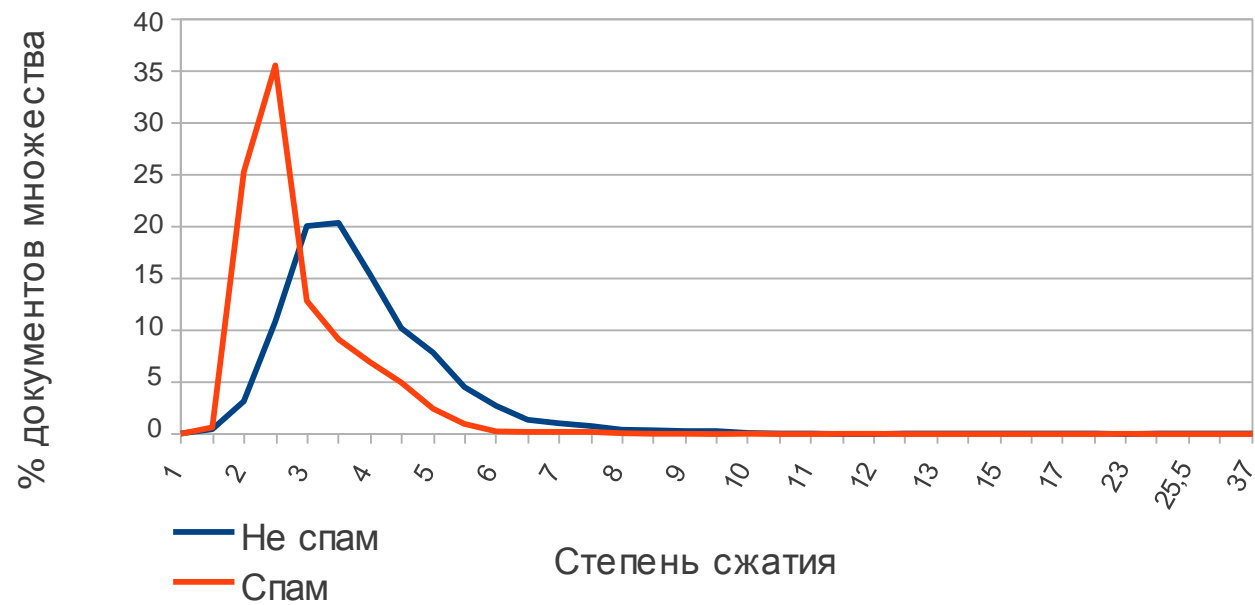
Распределение средней длины слова в спамовых и неспамовых множествах



Количество слов в анкерах ссылок для спамовых и неспамовых множеств



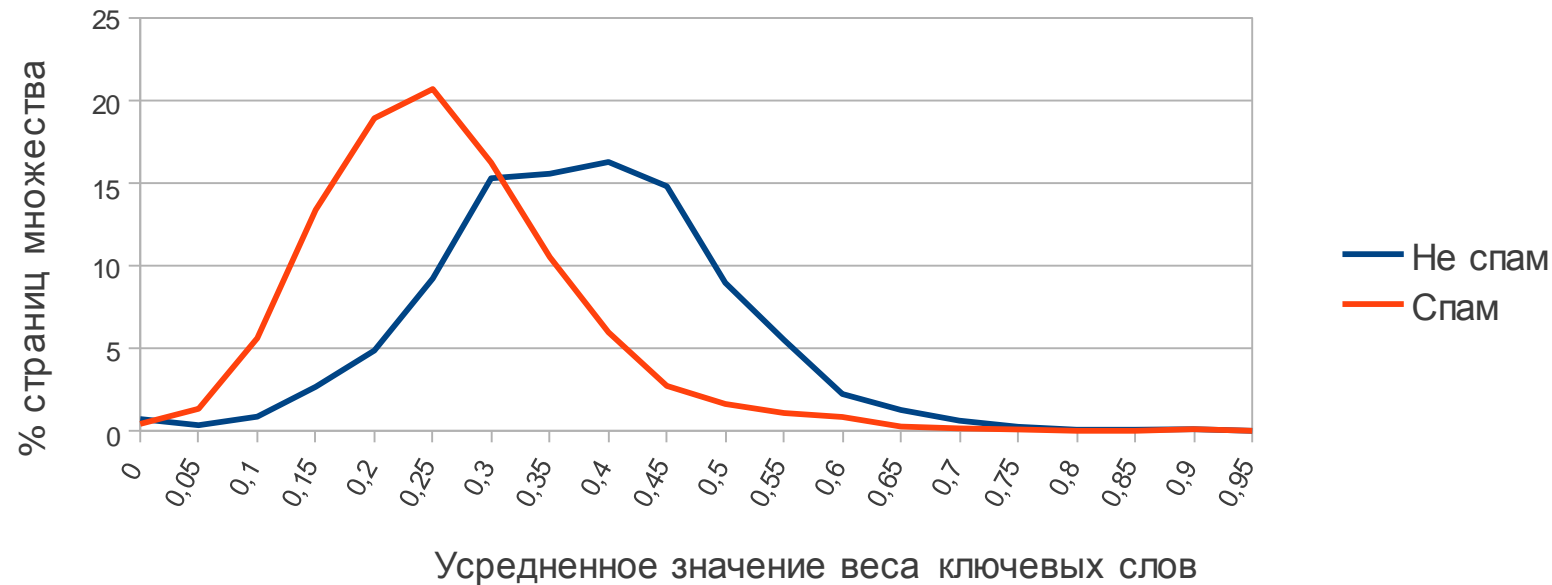
Степень сжатия документов в спамовых и неспамовых множествах



Сравнивая приведенные данные с ранними исследованиями, приходим к выводу, что спам подвергается мутациям, в сторону обычных страниц.

Хотя, в распределениях все еще присутствует явная «искусственность».

Распределение усредненного веса ключевых слов для спам- и обычных страниц



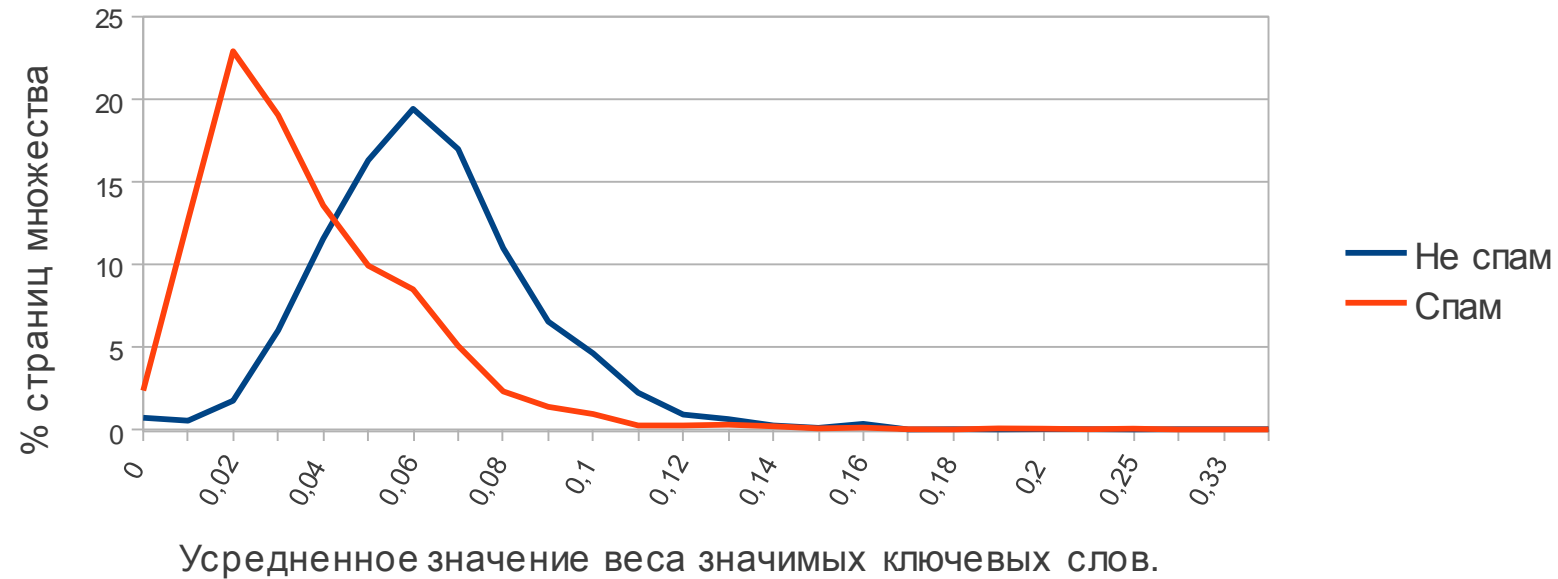
Усредненное значение веса ключевых слов документа:

$$\bar{w}^{kw} = \frac{\sum_{i=1}^N w_i}{N}$$

w_i вес ключевого слова

N количество ключевых слов

Распределение отношения веса значимых ключевых слов к общему количеству слов в спамовых и неспамовых множествах



Усредненное значение веса значимых ключевых слов документа:

$$w_d^{imp} = \frac{\sum_{i=1}^K w_i}{N}$$

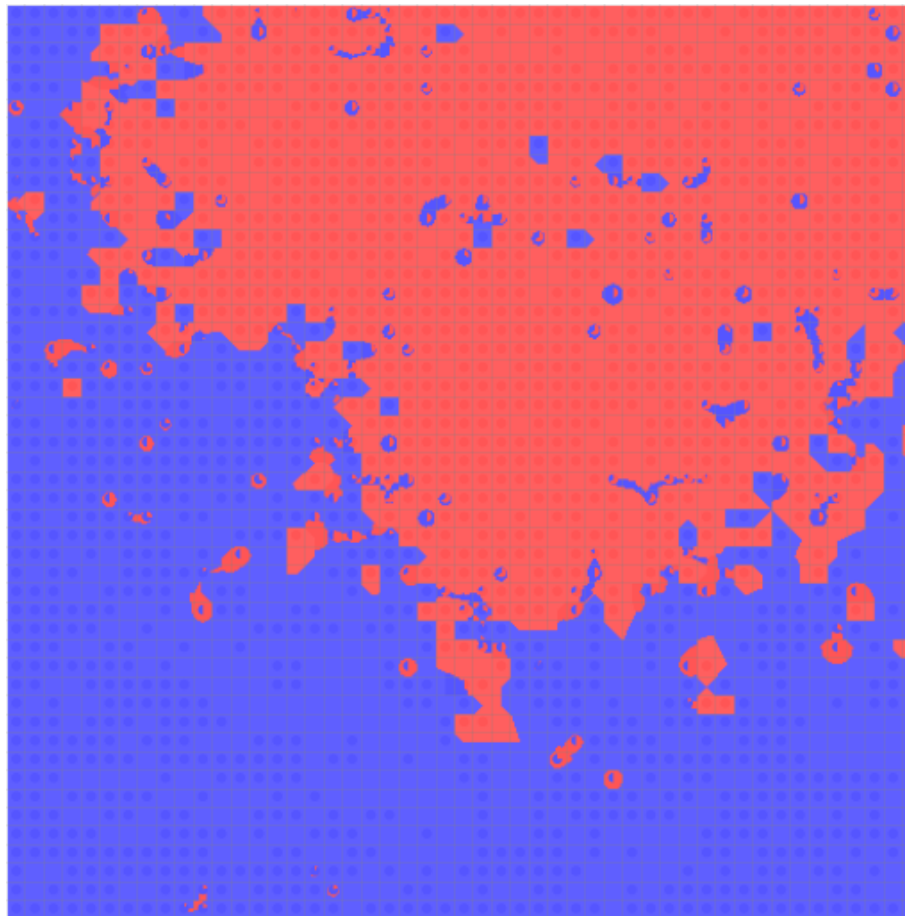
- w_i вес ключевого слова
- N количество ключевых слов
- K количество значимых слов

Мы привели несколько характеристических языковых признаков и увидели, что они дают лучшее разделение, чем признаки, полученные на основе параметров страницы.

В эксперименте мы рассчитали 10 дополнительных признаков, основанных на статистике распределения слов в текстах. Теперь, имея хороший набор факторов, перейдем к решению поставленной задачи, а именно – попробуем создать классификатор на основе описанных признаков.

Карта классов (SOM)

Class Legend
 Spam
 Not Spam



Классификатор — многослойный персептрон:

Входной слой — 80 нейронов ,

Скрытый слой — 96 нейронов

Выходной слой — 2 нейрона спам=1 и не-спам=0

Функция активации — сигмоид

Для тренировки нашего классификатора мы использовали страницы, отобранные ассессорами.



Обучающий вектор - 80 признаков.

Размер обучающего множества — 20000 страниц. Размер

тестового множества — 50000 страниц.

Точность - 0,97

Полнота - 0,94

F-мера - 0,96

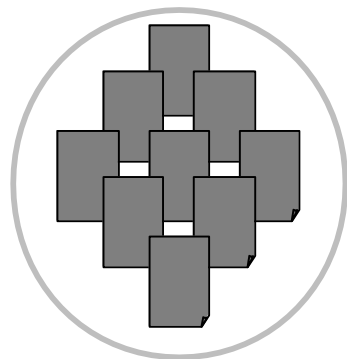
Результат показывает, что использование признаков, связанных со статистикой распределения слов и грамматических конструкций в текстах, привело к значительному улучшению качества классификации спам-страниц, даже несмотря на использование слабого алгоритма классификации.

Что делать дальше.

Можно ли использовать информацию, полученную
из контента страниц, для классификации
сайтов?

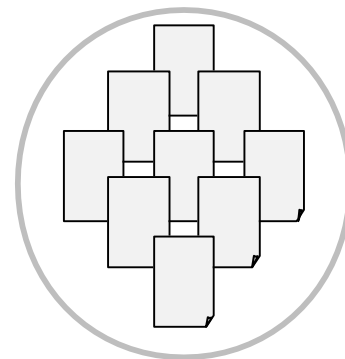
Спам или нет?

Спам сайт



100% = спам

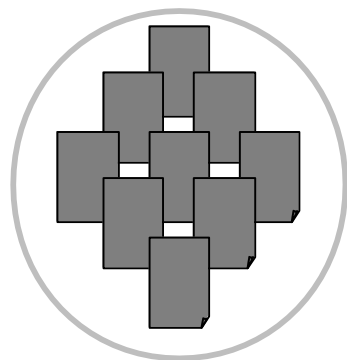
Не спам сайт



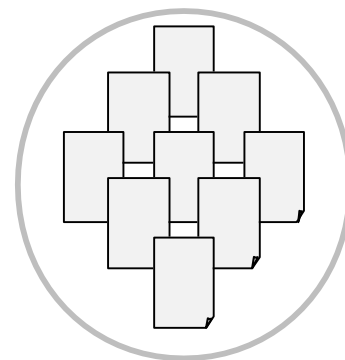
0% = не спам

Спам или нет?

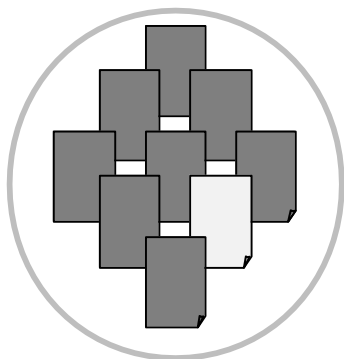
Спам сайт



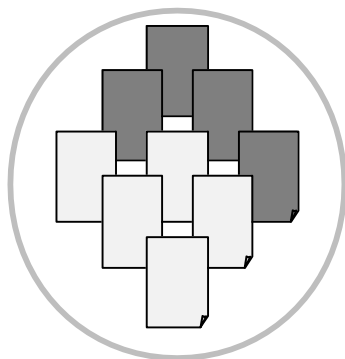
Не спам сайт



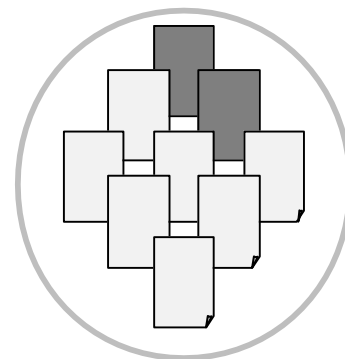
?



?



?



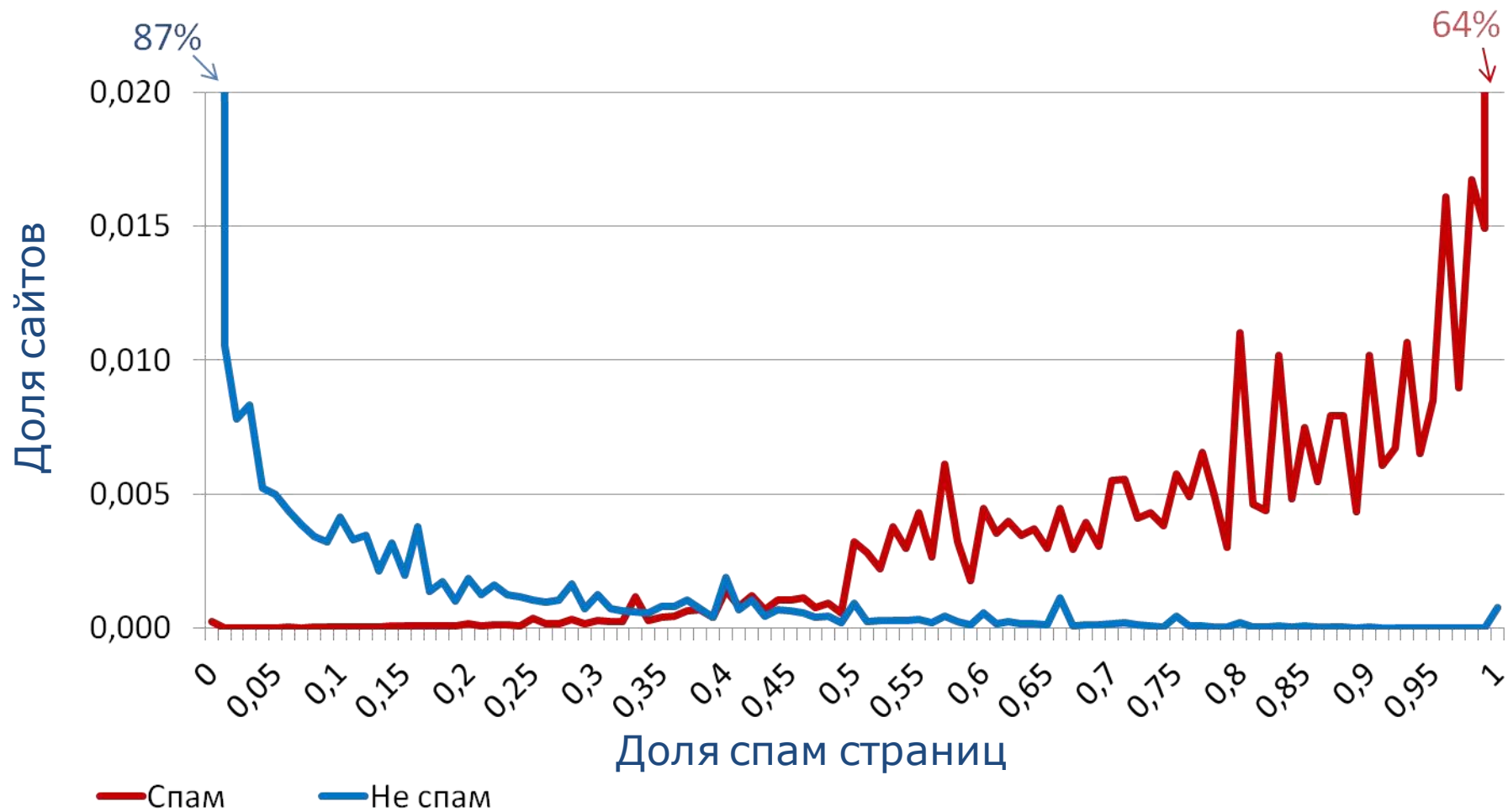
Причины:

- Хороший сайт со спам страницами:
- Ошибка классификатора.
- Взломанный сайт.
- Переоптимизированный контент.
- Спам сайт с полезными страницами:
- Ошибка классификатора.
- Разбавление спама не спам страницами.

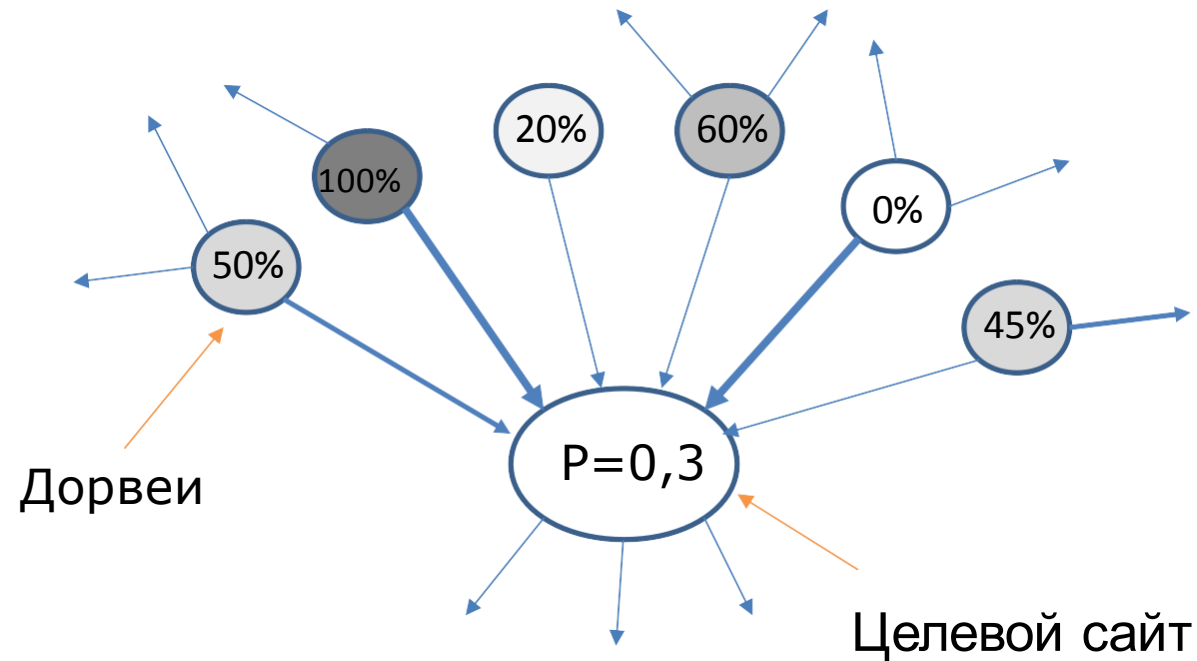
Характеристики сайта:

1. Доля спам страниц.
2. Расположение спам страниц.
3. Вероятность прихода/ухода на спам страницу с сайта.
4. На какие страницы ведут входящие/исходящие ссылки.
5. Вероятность участия в спам-ферме.

Доля спам страниц



Участие в спам ферме



1. Вычисляем вероятность того, что сайт раскручивается спам-сайтами.

2. Вычисляем вероятность участия в спам ферме

Вероятность участия в спам-ферме



На отобранных признаках строим классификатор

Всего получили 20 признаков

Используем алгоритм Expectation Maximization для выделения из множества сайтов двух центров, соответствующих классам: спам и не спам.

Используем полученные центры как исходные данные для классификации при помощи алгоритма *k-nearest neighbor*.

Результаты:

Уменьшение количества спама в выдаче
в среднем на 20%.

Точность анализатора - 90%.

Доля спам сайтов - 17%.

Другие применения антиспама и антифрода

- Покупные комментарии, как положительные так и отрицательные
- Рекламные сообщения и добавления в друзья
- Фейковые объявления (аренда, продажа и тд)
- Несоответствующие тематике объявления (интим услуги на авито и тд)
- Кредиты по поддельным паспортам (вклеивают фото - распознавание лиц)
- Спорные транзакции в банках (приложения списывающие деньги, мошенники)
- Накликивание рекламы, скликивание конкурентов
- Фейковые пользователи и посещаемость (через iframe сайты покупают фейковых пользователей – ловили интернет кинотеатры)

Как бороться с накрутками поведения в интернете

Основным инструментом в первую очередь являются различные статистические метрики:

- распределение по ip кликов/запросов/пользователей
- распределение по времени (ботов пишут люди и запускают обкачку в одно и тоже время)
- Анализ энтропии (разнообразия) действий пользователя
- Время сессий

Анализ собственной дистрибуции

Retention и тд

Тут как раз слишком хорошие показатели должны вызывать сомнения

Если высокий показатель установок при открытии ссылки или стабильное время установки после открытия и тд

Сервисы такси

Хитрые клиенты и хитрые водители

Банки

Обналичивание и тд



ТЕХНОСФЕРА

Спасибо!

Вопросы.