Основы информационного поиска

Курышев Сергей

325 группа

Информационный поиск (IR) – это процесс поиска в большой коллекции (хранящейся, как правило, в памяти компьютеров) некоего неструктурированного материала (обычно – документа), удовлетворяющего информационные потребности.

Булев поиск

Запрос представляется логической формулой, в которой атомами могут быть термины или какие-либо дополнительные условия (поиск только в том же предложении текста, поиск точной фразы и т. п.). Поисковая машина, возвращает документы, для которых формула-запрос принимает истинные значения

Прямой поиск

Brutus AND Caesar AND NOT Calpurnia

Прямой поиск представляет собой последовательный просмотр текста всех документов. При этом отмечаются документы, содержащие слова Brutus и Caesar и исключаются документы, содержащие Calpurnia.

Прямой поиск

Прямой поиск используется на современных компьютерах для выполнения простых запросов на коллекциях данных среднего размера.

Иногда необходимо нечто большее

- Быстрая обработка больших коллекций документов
- Более сложные поисковые запросы. Например, **Romans NEAR countrymen**, где NEAR может означать «не далее 5 слов».
- Ранжированный поиск

Индексирование документов

Индексирование документов – процесс создания индексных таблиц, существенно ускоряющих обработку запроса

Пример индексирования

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Рис. 1.1. Матрица инцидентности "термин–документ". Элемент матрицы (t,d) равен 1, если пьеса в столбце d содержит слово из строки t, и 0, если не содержит

Brutus AND Caesar AND NOT Calpurnia

110100 AND 110111 AND 101111 = 100100

Ответ поисковой машины: Antony and Cleopatra, Hamlet

Инвертированный индекс

- 1 000 000 документов
- 1000 слов в каждом документе
- 500 000 различных терминов

Матрица «термин-документ» будет содержать 500 000 000 000 нулей и единиц.

Матрица будет содержать не более

1 миллиарда единиц и, как минимум, на 99.8% состоять из нулей.

Инвертированный индекс

Рис. 1.3. Две части инвертированного индекса. Словарь обычно находится в памяти вместе с указателями на каждый список словопозиций, которые хранятся на диске

Релевантность

Документ называется **релевантным**, если с точки зрения пользователя он содержит ценную информацию, удовлетворяющую его информационную потребность.

Показатели качества поиска

- Точность – доля релевантных документов выборки, по отношению ко всем документам в выборке
- Полнота – доля релевантных документов в выборке, по отношению ко всем релевантным документам коллекции

Интерпретация запросов пользователей

Запрос пользователя	Что он хочет увидеть
Подержанные авто	Подержанные автомобили Подержанные машины Автомобили с пробегом Купить подержанные автомобили Продажа автомобилей с пробегом …

(Подержанные OR с пробегом) AND (авто OR машины OR автомобили) AND (Купить OR продажа)

Расширение запросов

Суть расширения запросов заключается в дополнении пользовательского запроса новыми словами, которые могли бы улучшить релевантность выданных результатов.

При этом используется специальный словарь синонимов, который должен содержать активно применяемую в запросах лексику. Такие словари называются «тезаурусными расширениями».

Методы построения тезауруса

- Использование контролируемого словаря, поддерживаемого редакторами.
- Автоматическое создание тезауруса
- Переформулирование запроса на основе анализа лога запросов

Ранжированный поиск

Основан на вычислении релевантности через распределение частот встречаемости терминов запроса по документам коллекции.

Каждый документ коллекции представляется вектором в векторном пространстве, размерность которого равна числу токенов в инвертированном файле. Документ описывается «весами» (координатами) соответствующих токенов.

Ранжированный поиск

Для каждого фиксированного i, документ d_i представляется вектором весов:

$W_{ji}=tf_{ji} * idf_{ji}$, j=1..M

Где tf_{ji} частота встречаемости токена t_j в документе d_i, по сравнению с другими токенами документа. $idf_{ji} = \log(N/n_j)$, где N – число документов в коллекции, n_j – число документов, в которых встретился токен t_j

Ранжированный поиск

Рассматриваем запрос, как документ

$q(q_1 \dots q_M)$

$a(a_1 \dots a_N) = qW$

$w_k = W(w_{1k} \dots w_{Mk})$

$a'(a_1 \dots a_N)$, $a'_k = a_k / (|q| * |w_k|)$

В итоге вектор a’ содержит значения релевантности для каждой пары (q , d_k)

Вероятностная модель

Основные предположения:

1. Документ d либо релевантен, либо нерелевантен запросу q (т.е. для каждого события (d,q) возможно только 2 элементарных исхода (w0,w1))
2. Определение одного документа как релевантного не дает никакой информации о релевантности других документов.

Вероятностная модель

Вероятность извлечения из коллекции документа D релевантного запросу Q может быть выражена так:

$$P(R_Q = X \mid D)$$

$$\frac{P(R_Q=1 \mid D)}{P(R_Q=0 \mid D)} = \frac{P(R_Q=1)P(D \mid R_Q=1)}{P(R_Q=0)P(D \mid R_Q=0)} \approx \frac{P(D \mid R_Q=1)}{P(D \mid R_Q=0)} \approx \prod_{t \in Q} \frac{P(t \mid R_Q=1)}{P(t \mid R_Q=0)}$$

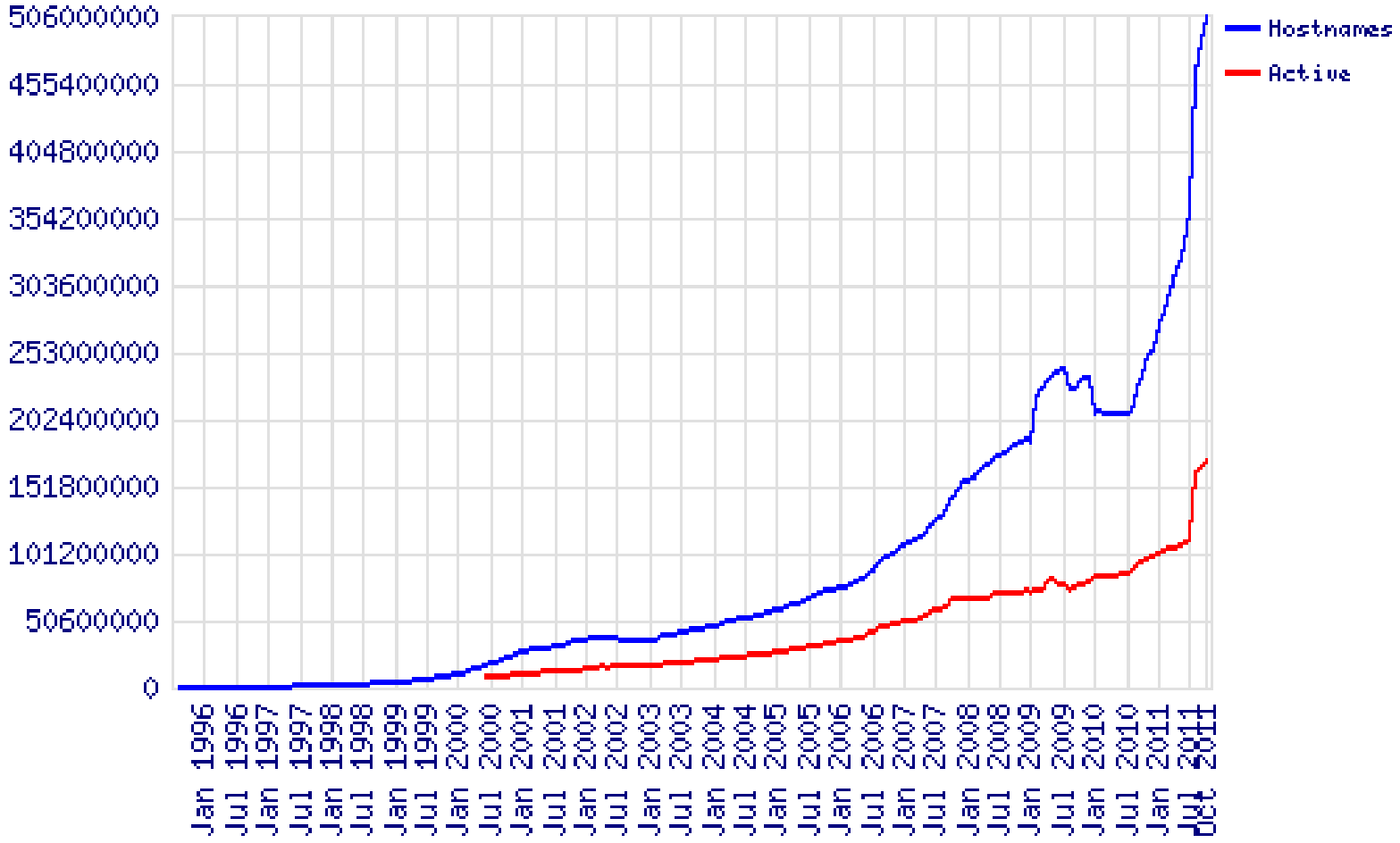
$$\prod_{t:x_t=1}^{M} \frac{P(x_t=1 \mid R=1,\vec{q})}{P(x_t=1 \mid R=0,\vec{q})} \prod_{t:x_t=0}^{M} \frac{P(x_t=0 \mid R=1,\vec{q})}{P(x_t=0 \mid R=0,\vec{q})}$$

	Документ	Релевантный (R=1)	Нерелевантный (R = 0)
Термин есть	$x_t = 1$	p_t	u_t
Термина нет	$x_t = 0$	$1 - p_t$	$1 - u_t$

Особенности поиска в web

- Масштабы web
- Поисковый спам
- Потребности пользователей
- Поисковая реклама

Масштабы web



Дорвеи

Дорвей (от англ. *doorway* — входная дверь, портал) или **входная страница** — вид поискового спама, веб-страница, специально оптимизированная под один или несколько поисковых запросов с единственной целью её попадания на высокие места в результатах поиска по этим запросам.

Клоакинг

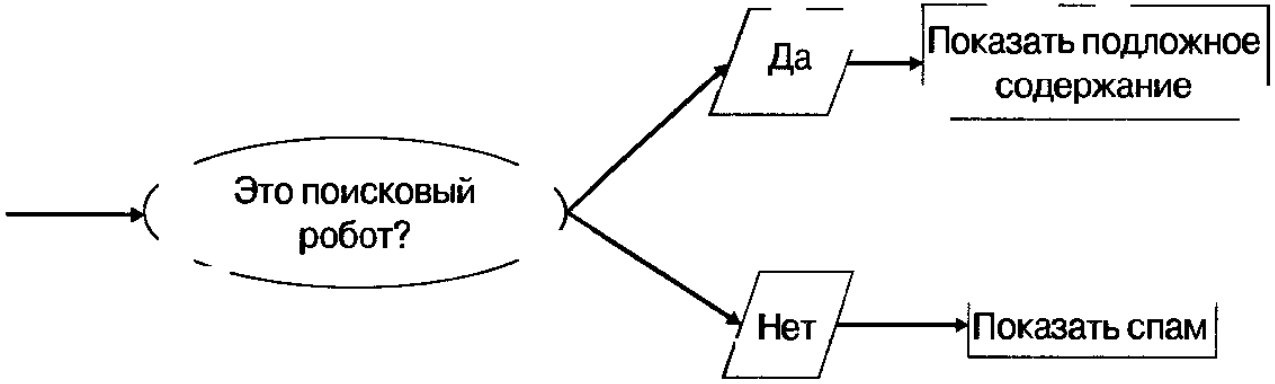


Рис. 19.5. Клоакинг

Потребности пользователей

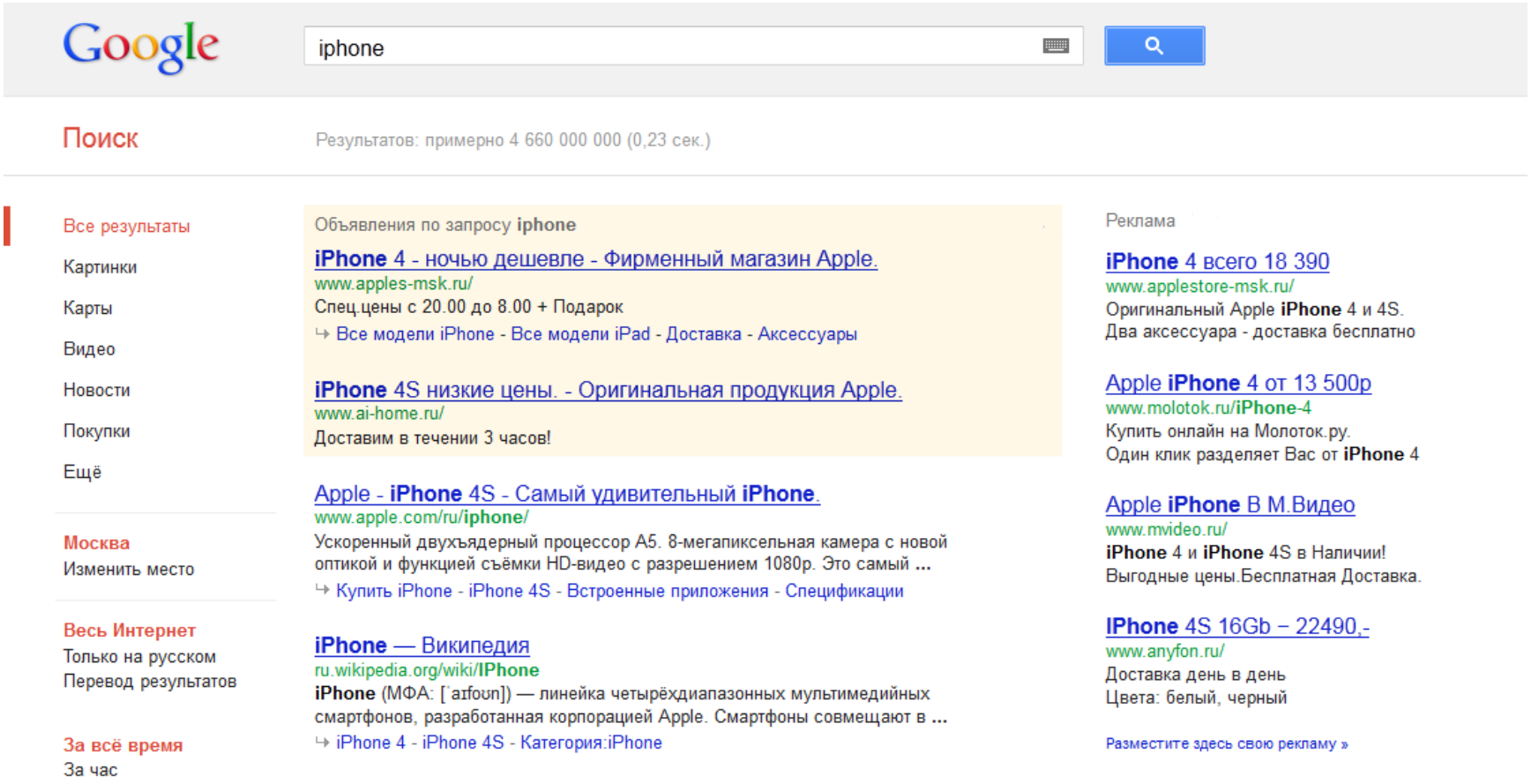
- Популярный инструмент поиска не должен предъявлять слишком высоких требований к миллиардам людей.
- 3 категории запросов

  1) Информационные

  2) Навигационные

  3) Транзакционные

Поисковая реклама



Смежные задачи ИП

- Автоматическая классификация
- Кластеризация документов
- Реферирование

Классификация в ИП

- Предварительная обработка для индексирования: идентификация кодировки, определение языка.
- Автоматическое отделение веб-спама.
- Автоматическое отделение содержания сексуального характера (Безопасный поиск)
- Выявление мнений и отзывов
- Тематический или вертикальный поиск

Кластеризация

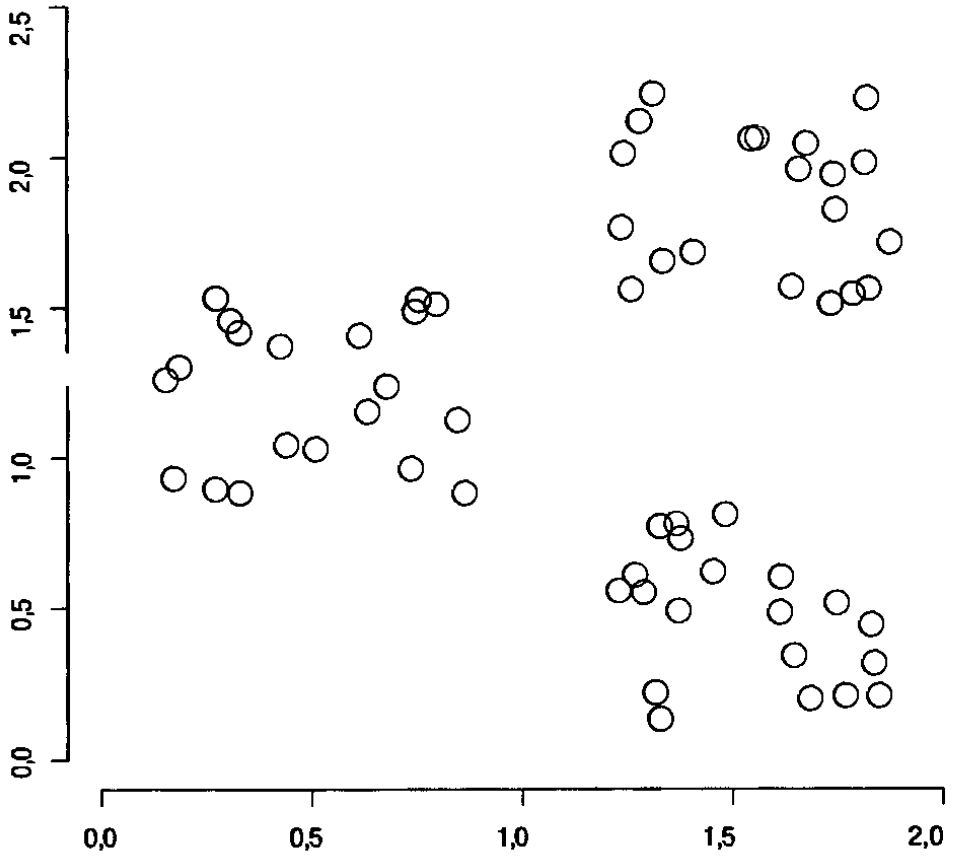


Рис. 16.1. Пример данных с четкой кластерной структурой

Кластеризация в ИП

Приложение	*Что подвергается кластеризации*	*Выгода*
Кластеризация результатов поиска	Результаты поиска	Более качественное представление информации
Разбиение и объединение	Подмножества коллекций	Альтернативный пользовательский интерфейс: "поиск без ввода слов"
Кластеризация коллекции	Коллекция	Более качественное представление информации для навигации пользователей
Языковые модели	Коллекция	Повышение точности и/или полноты
Кластерный поиск	Коллекция	Повышение производительности: скорости поиска

Виды реферирования

- Краткое изложение.

  «В этой речи Авраам Линкольн призывает вспомнить солдат, которые отдали свои жизни в битве при Геттисберге»
- Набор выдержек.

  «Восемьдесят семь лет назад наши отцы ступили на эту землю, чтобы создать новую нацию»