

# Информационный поиск: модели и методы

Игорь Некрестьянов  
Санкт-Петербургский Университет

[igor@meta.math.spbu.ru](mailto:igor@meta.math.spbu.ru)

# Информационный поиск (ИП)

Цель: удовлетворить информационные потребности пользователя

Обсуждаемые темы:

- Модели ИП
- Критерии оценки качества поиска
- Поиск в Интернет
- Поиск по значимости

# Модель ИП:

- Логическое представление документов
- Логическое представление запросов
- Framework моделирования представлений документов и запросов, их взаимосвязей
- Ранжирующая функция

# Классификация моделей

- Булева модель  
(теория множеств и булева алгебра)
- Векторная модель  
(векторные пространства и линейная алгебра)
- Вероятностная модель  
(множества, теория вероятностей)

Классические модели подразумевают  
независимость слов (термов)

# Булевы модели

- Модель на нечетких множествах  
(с термом запроса ассоциировано нечеткое множество документов)
- Расширенная булева модель
  - Расширяет булеву модель для использования весов термов
  - Обобщает модель на нечетких множествах и векторную модель (выбирая метрику)

# Векторные модели

- Обобщенная векторная модель  
(учет корреляции между терминами)
- Латентно-семантический анализ  
(отображение документов и запросов в пространство *концепций*)
- Нейронные сети  
(нейроны — термины документов и документы, многошаговое распространение сигнала)

# Вычисление весов термов

- Частота терма  $t_i$  в документе  $d_j$

$$tf_{i,j} = f_{i,j} / f_j$$

- Обратная частота термов в коллекции

$$idf_i = \log(N / n_i)$$

- Вычисление весов

$$w_{i,j} = tf_{i,j} \cdot idf_i \cdot norm_j$$

$$w_{i,q} = \frac{1}{2} (1 + tf_{i,q}) \cdot idf_i$$

# Нормализация весов

Преимущества длинных документов:

- Больше различных термов
- Выше частоты термов

Методы нормализации:

- по максимальной частоте
- по длине вектора весов всех термов в данном документе
- по длине в байтах



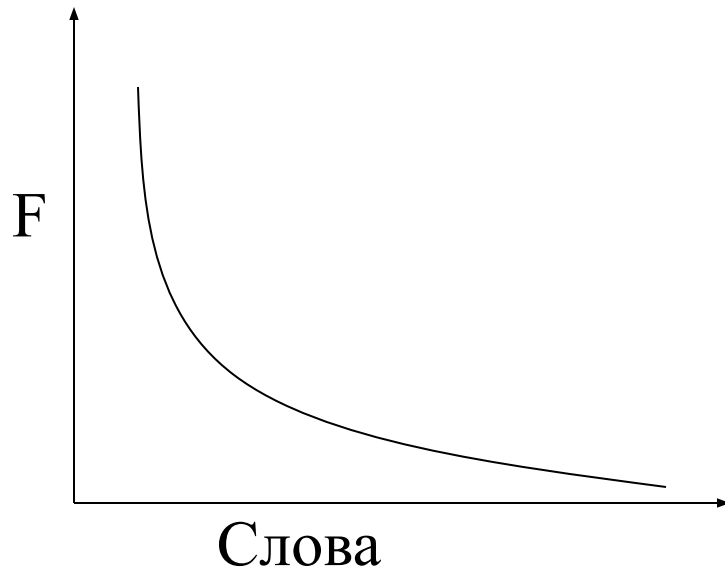
# Вероятностные модели

- Вероятностный принцип  
Оценить вероятность того, что документ будет интересен пользователю
- Модель сетей вывода (inference networks)
  - На основе сети Байеса
  - Могут имитировать булеву модель, некоторые векторные модели, обратную связь
  - Реализована в Inquiry

# Моделирование языка

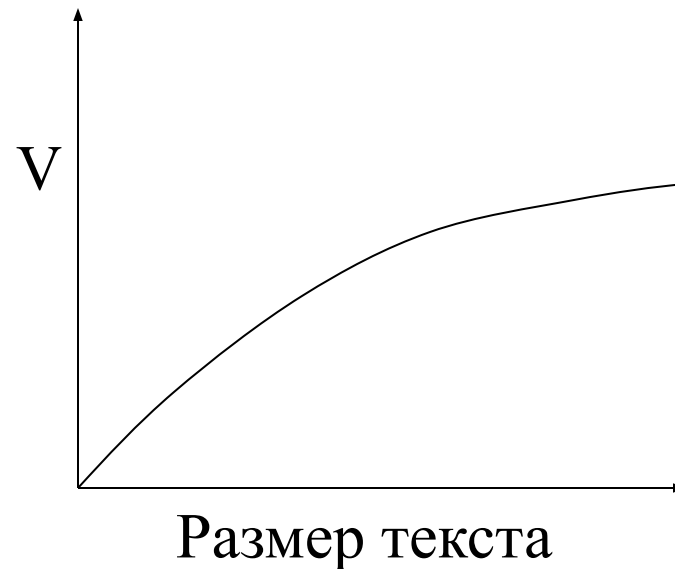
Zipf's Law

$$F(i) = i^{-\alpha}, \quad \alpha \approx 1.7$$



Heaps' Law

$$V = Kn^{\beta}, \quad \beta \approx 0.5$$



# Предварительная обработка текста

- Лексический анализ
- Исключение стоп-слов
- Выделение основ слов (stemming)
- Выбор термов для индексирования  
(например только существительных)
- Тезаурусы  
(выделение категорий термов)

# Языки запросов

- Запросы по ключевым словам
  - однословные
  - контекстные
  - логические
  - на естественном языке
- Запросы по шаблонам
- Протоколы запросов (Z39.50, WAIS)

# Уточнение запросов:

- Изменение весов термов запроса
- Добавление новых термов в запрос

## Основные подходы:

- Обратная связь (Relevance feedback)
- Автоматический локальный анализ
- Автоматический глобальный анализ

# Критерии оценки

- Точность

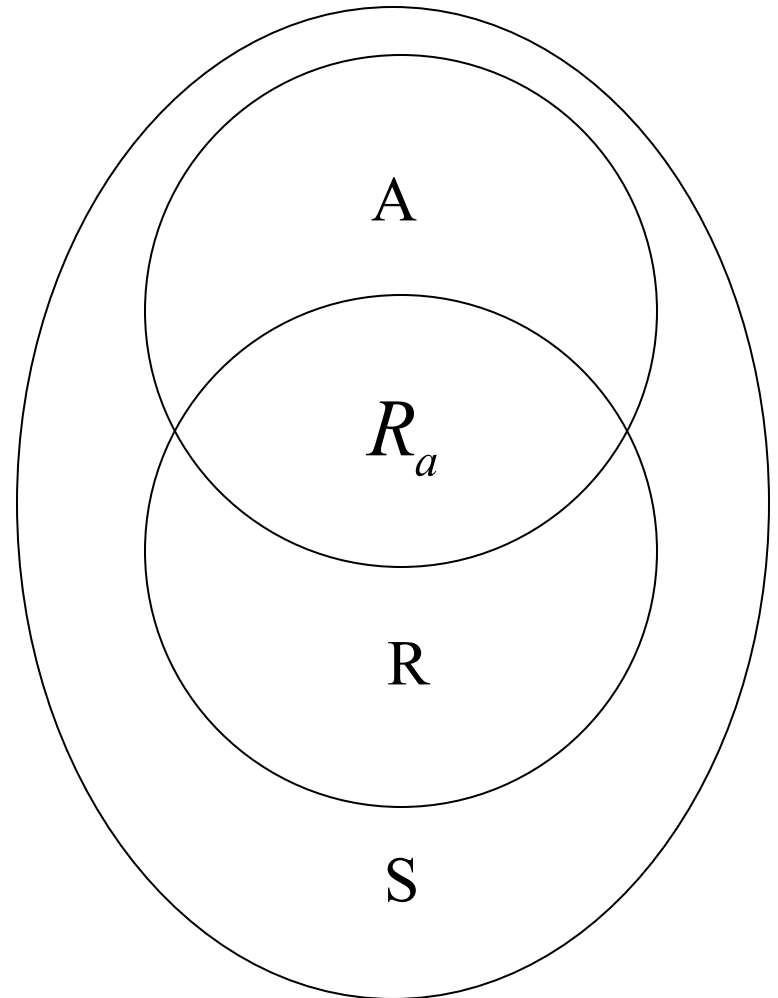
$$\text{Precision} = \frac{|R_a|}{|A|}$$

- Полнота

$$\text{Recall} = \frac{|R_a|}{|R|}$$

- Процент мусора

$$\text{Junk} = \frac{|A/R_a|}{|S/R|}$$



# Критерии оценки (2)

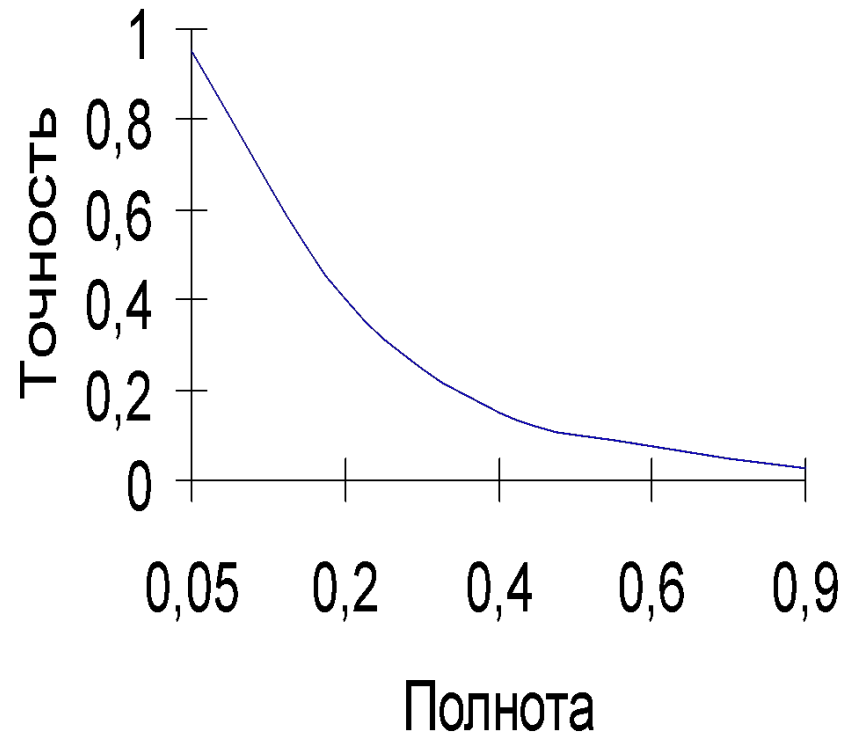
- Средняя точность

$$\bar{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$$

- Интерполяция

$$P(r_j) = \max_{r_j \leq r \leq r_{j+1}} P(r)$$

- По числу документов  
 $P@20$ ,  $P@50$ ,  $P@100$



# Критерии оценки (3)

- Точность на уровне обнаружения заданного числа релевантных документов
- Точность среди первых  $R$  возвращенных, где  $R$  – число реально релевантных (R-precision)
- Гистограммы точности (сравнение эффективности двух алгоритмов на каждом запросе)



# Критерии оценки (4)

- Среднее гармоническое ( $b = 1$ )  
(компромисс между точностью и полнотой)
- E-мера  
(учет предпочтений пользователя)

$$E(j) = \frac{1 + b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}}$$

# Критерии оценки (5)

Пользовательские критерии:

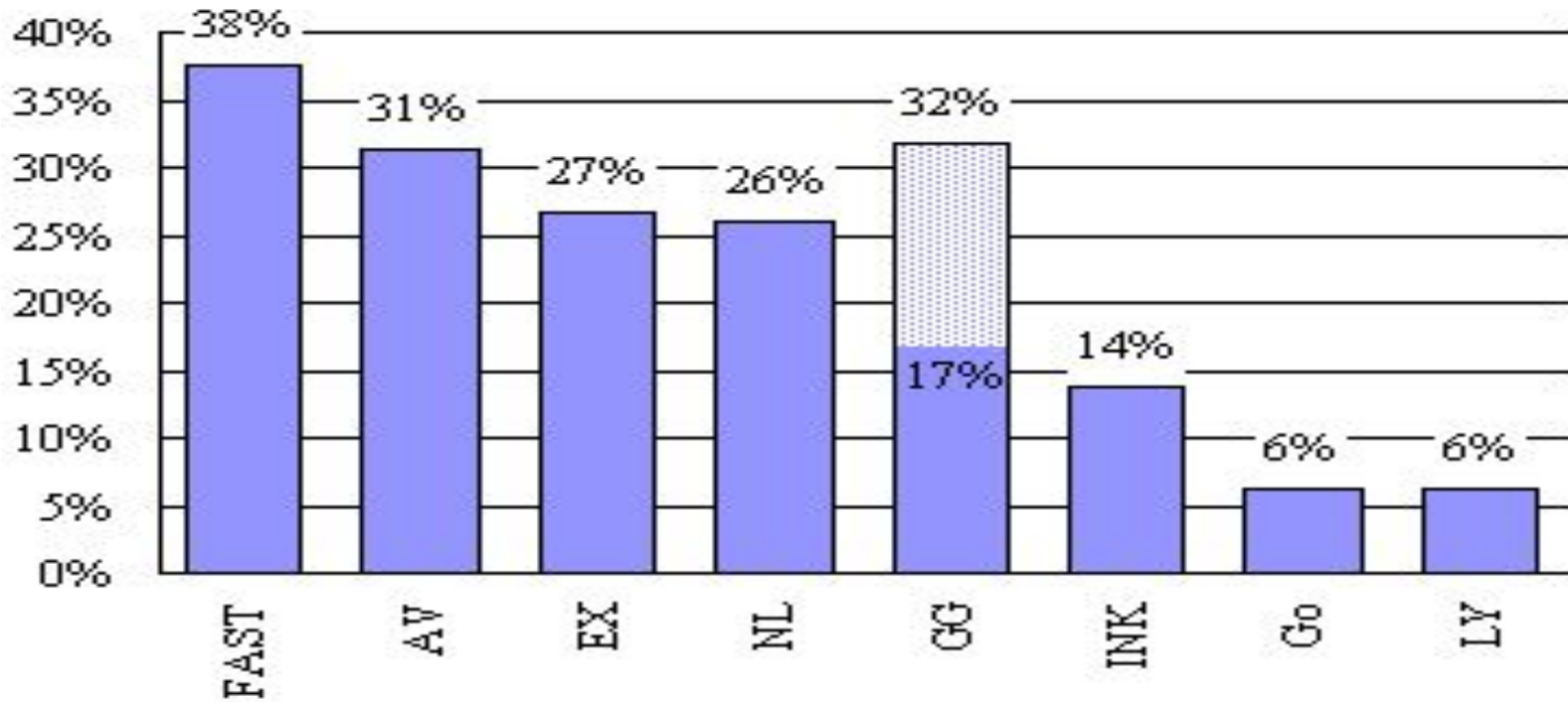
- Коэффициент покрытия  
(процент уже известного среди найденного)
- Коэффициент новизны  
(отношение нового к уже известному)
- Относительная полнота  
(отношение нового к ожидаемому)

# Особенности поиска в Интернет

- Огромный размер
  - > 1 миллиарда документов (февраль 2000)
  - > 75 миллионов узлов
- Короткие запросы ( < 2 слов)
- Почти не используются продвинутые возможности языка запросов
- Много изменений в данных (40% в месяц)

# Размер поисковых систем

**% Of Web Indexed**  
(Est. 1 billion total pages)



# Оценка качества индекса

- Не все страницы одинаково важны
- Метрики
  - Суммарная значимость всех страниц
  - Средняя значимость страниц
- Метод случайной прогулки
  - Как проверить наличие страницы в индексе?
  - Как оценить значимость страницы?

# TREC и Web

- Коллекция VLC2:  
100Гб, 18.5 миллионов документов
- Короткие запросы из TREC (2.5 слова)
- 5 поисковых систем

	Web		TREC	
	диапазон	среднее	диапазон	среднее
P@20	.23-.37	.29	.34-.44	.40

# Задачи ИП в Интернет

- Обнаружение дубликатов
- Интеллектуальные сетевые роботы
- Борьба с опечатками
  - $\text{Levenshtein}(\text{survey}, \text{surgery}) = 2$
  - $\text{LCS}(\text{survey}, \text{surgery}) = \text{surey}$
- Метапоиск
  - Как делать слияние результатов (data fusion)?

# Классификация типов копий

Повторение содержания (1) и структуры (2)

1, 2 – совпадает	
1 – эквивалентно, 2 – совпадает	
1 – похоже, 2 – совпадает	
2 – похоже, 2 – частично совпадает	1 – близко, 2 – совпадает



# Оценка повторения структуры

- Построение описаний URL  
yahoo.com/ref/art/news.htm  
(yahoo,com) (ref, art, 0) (art, news, 1) (news, htm, 2)
- Сокращение числа кандидатов  
(стоп-слова, нехарактерные пары)
- Вычисление весов (IDF)
- Вычисление оценки схожести узлов
- Уровни схожести: 100%, 50%, 0%

# Оценка повторения содержания

- Выбор страниц для проверки
- Проверка на полную идентичность
- Вычисление оценки близости:

$S(A)$  — множество  $k$ -грамм документа  $A$

$$r(A, B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|}$$

# Сетевые роботы

## Области применения:

- Построение индексов
- Сбор статистики
- Поиск ресурсов
- Исследование структуры Интернет
- Проверка целостности ссылок

## Стратегии обхода:

- Простые
- Учет структуры URL (уровень вложенности)
- Учет структуры графа (PageRank)
- Учет содержимого страницы (OASIS Crawler)

# Поиск по значимости

- Дополнение к оценкам близости
- Значимость не зависит от запроса
- Значимость бывает не только у текстовых ресурсов
- Более значимые ресурсы имеют приоритет при прочих равных

# Значимость из содержимого

- Анализ документа  
(PHOAKS, определение жанра текста)
- Анализ коллекции  
(Google, SCAM, PHOAKS)
- Информационный контекст  
(Referral Web)
- Внутренние метки в документе  
(PICS, RDF)

# Значимость из действий

Явные указания:

- Обратная связь  
(групповая)
- Триггеры на данные  
(почтовые фильтры)
- Синтезированные фильтры  
(пользователь задает высокоуровневое описание)

# Значимость из действий (2)

Неявные указания:

- Коллективное поведение пользователей (Web Watcher, Hotbot)
- Индивидуальное поведение пользователя
  - Какие документы/коллекции он посещает?
  - Что он делает с документом?  
(время просмотра, новая закладка, запись на диск, ответ на письмо)

# Основные конференции

- SIGIR (<http://www.acm.org/sigir/>)
- Digital Libraries
- Text Retrieval Conference (TREC)  
(<http://trec.nist.gov>)
- WWW Conference (<http://www9.org>)
- Электронные Библиотеки  
(<http://www.protvino.ru/dl2000>)