

Интеллектуальные информационные системы

Тема 1

Модели и методы ИП

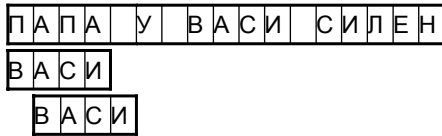
Павел Исаакович Браславский
pb@imach.uran.ru
весенний семестр 2006

План

- Индексирование
- Булевская модель
- Векторная модель
- Оценка качества поиска

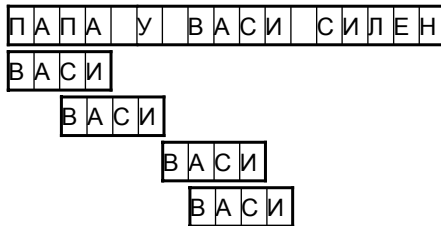
Прямой поиск

- Brute force



Средняя сложность: $O(n+m)$

- Boyer-Moore

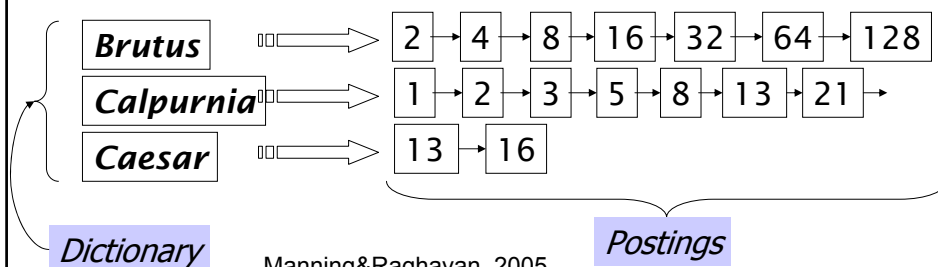


Средняя сложность: $O(n/m)$

Индексирование

Процесс создания поискового образа документа (логического представления)

Обычно – инвертированный индекс:



Manning&Raghavan, 2005

Предварительная обработка

- Извлечение текста (HTML, PDF...)
- Определение кодировки и языка
- Разбиение на слова и предложения (tokenization)
- Удаление стоп-слов
- Лемматизация (stemming) – приведение слова к словарной форме (об этом позже...)

Tokenization

Даты, числа: 13/02/2006 3,1415...

Наречия: *без устали, в упор, в конце концов*

Вводные слова: *другими словами, шутка ли (сказать), короче говоря*

Предлоги: *в преддверии, несмотря на*

Частицы: *всё ж таки, вроде как, вроде бы, к тому же, как будто*


Союзы: *в силу того что, несмотря на то что, тем более что*

<http://www.ruscorpora.ru/corpora-morph.html>

Многословные токены: Комсомольск-на-Амуре, Иван Топорыжкин, царская водка... (выделение collocations – позже)

Границы предложений:


И. И. Петров приехал в г. Екатеринбург прошлой осенью.



Стоп-слова

- Текст = неструктурированный набор значимых слов ('bag of words')
- Стоп-слова (stop-words) – служебные части речи – предлоги, союзы, частицы...

а, ага, ай, ау, ах, ба, без, близ, брр, брысь, будто, бы, быть, в, вы, ваш, вблизи, вглубь, вдобавок, вдоль, ведь, взамен, вместо, вне, внутри, во, возле, вокруг, вон...



Модель ИП

- Способ представления документов
- Способ задания инф. потребности (запросов)
- Способ вычисления близости между запросом и документом

Булевская модель

- Документ = множество слов (термов)
- Запрос = булевское выражение:
(кошка OR собака) AND корм
лебедь ANDNOT генерал
- Обработка запроса = операции со множествами, соответствующими словам (термам)

Булевская модель - пример

	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

Manning&Raghavan, 2005

Булевская модель +/-

+

- Простота
- Удобно для тех, кто знаком с лог. операторами

–

- Слишком «контрастно» (как представление документа, так и релевантность)

Векторная модель ИП

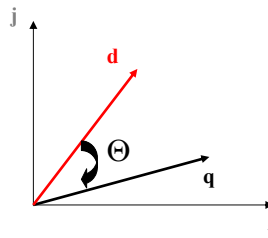
- Документ и запрос – векторы в пространстве слов (термов); компонент вектора – значимость слова для документа (запроса)
- Мера близости – косинус угла между векторами (→ ранжирование!)

Векторная модель - 2

$$\text{sim}(\bar{d}, \bar{q}) = \frac{\sum d_i \cdot q_i}{|\bar{d}| |\bar{q}|}$$

d_i – вес термина i в документе

q_i – вес термина i в запросе



Baeza-Yates&Ribeiro-Neto, 1999

Вес термина

1. Как часто встречается в документе?
2. Как часто встречается в коллекции?



Подход TF*IDF

TF – term frequency

IDF – inverse document frequency

TF*IDF – базовый вариант

$$tf_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

$$idf_i = \log \frac{N}{n_i}$$

$$w_{ij} = tf_{ij} \cdot idf_i$$

Baeza-Yates&Ribeiro-Neto, 1999

TF*IDF, Окари

$$TFIDF_D(l) = \beta + (1 - \beta) \cdot tf_D(l) \cdot idf_D(l)$$

$$tf_D(l) = \frac{freq_D(l)}{freq_D(l) + 0.5 + 1.5 \cdot \frac{dl_D}{avg_dl}}$$

$$idf(l) = \frac{\log\left(\frac{|c| + 0.5}{df(l)}\right)}{\log(|c| + 1)}$$

avg_dl – средняя длина док-та
 c – размер коллекции
 $\beta=0..1$

Цит. по Агеев и др., РОМИП-2003

Векторная модель +/-

+

- Хорошо работает на «чистых» статичных коллекциях
- Допускает частичные совпадения

–

- Легко атакуется (спам)
- Плохо работает на коротких текстах

Web

- Неконтролируемая коллекция
- Объемы
- Разные форматы
- Разнообразие (язык, темы...)
- Конкуренция (спам)
- Клики
- Ссылки! (см. *PageRank*)

Оценка качества поиска

Основа – понятие *релевантности* (соответствие информационной потребности)

- Точность (precision)

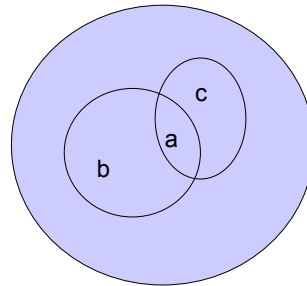
$$p = a/b$$

- Полнота (recall)

$$r = a/c$$

- *F* мера

$$F = (p+r)/2pr$$

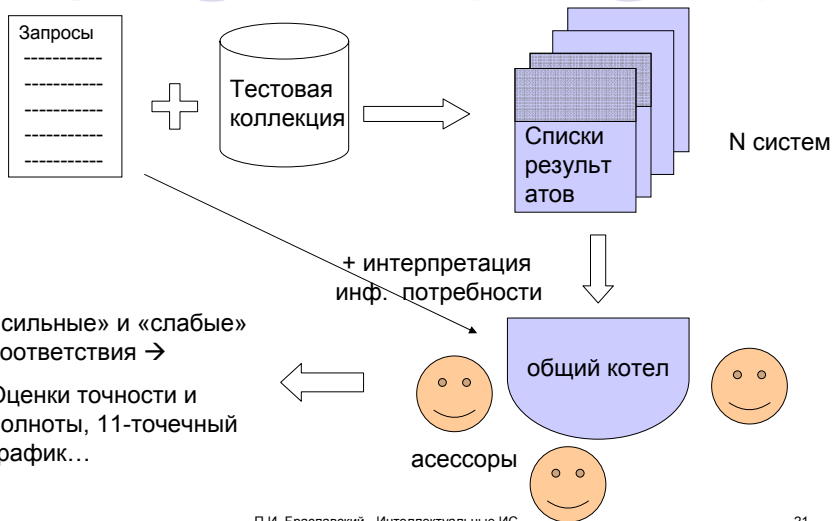


a – релевантные в отклике,
b – всего в отклике,
c – всего релевантных.

Инициативы по оценке ИП

- TREC (Text Retrieval Evaluation Conference) <http://trec.nist.gov>
- CLEF (Cross-Language Evaluation Forum) <http://www.clef-campaign.org/>
- РОМИП (Российский семинар по Оценке Методов Информационного поиска) <http://romip.narod.ru>

Метод общего котла

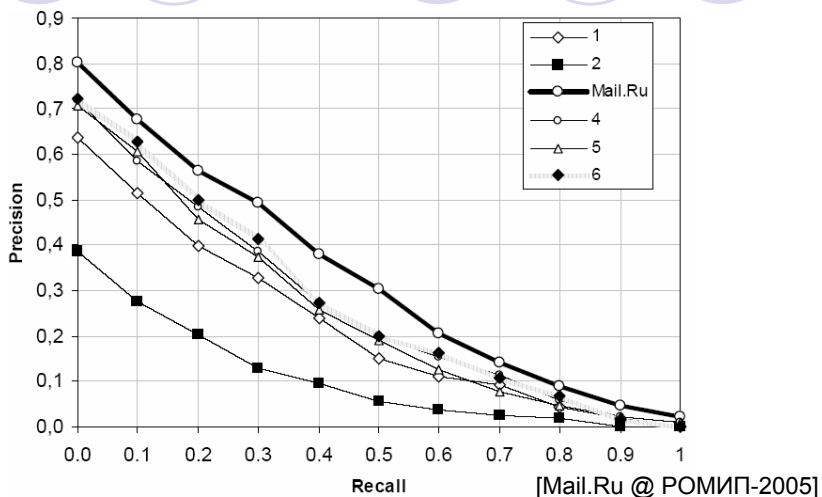


2006

П.И. Браславский - Интеллектуальные ИС

21

11-точечный график P/R



2006

П.И. Браславский - Интеллектуальные ИС

22