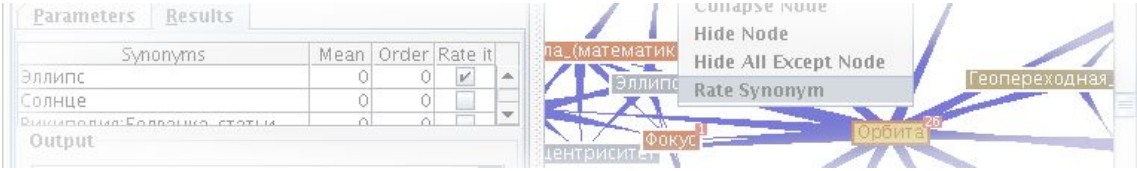


Информационный поиск

Information Retrieval

Петрозаводский государственный университет

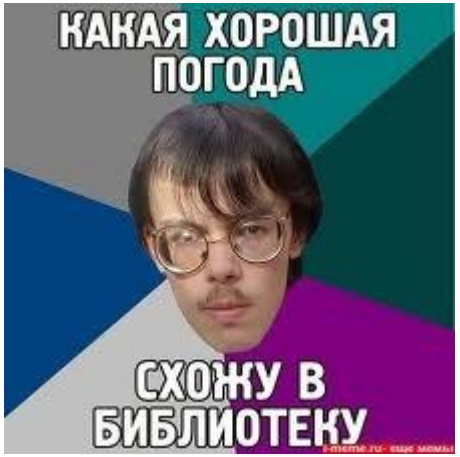
Крижановский Андрей Анатольевич

andrew.krizhanovsky@gmail.com

История ИП (1)

Каталоги библиотек,
информационные отделы

?

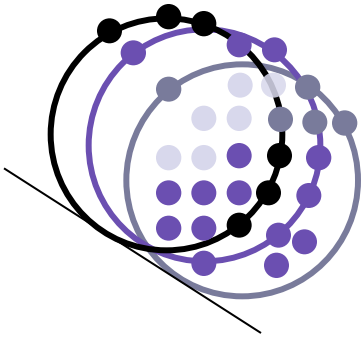


Библиотекари,
специалисты по информации

Рядовые пользователи

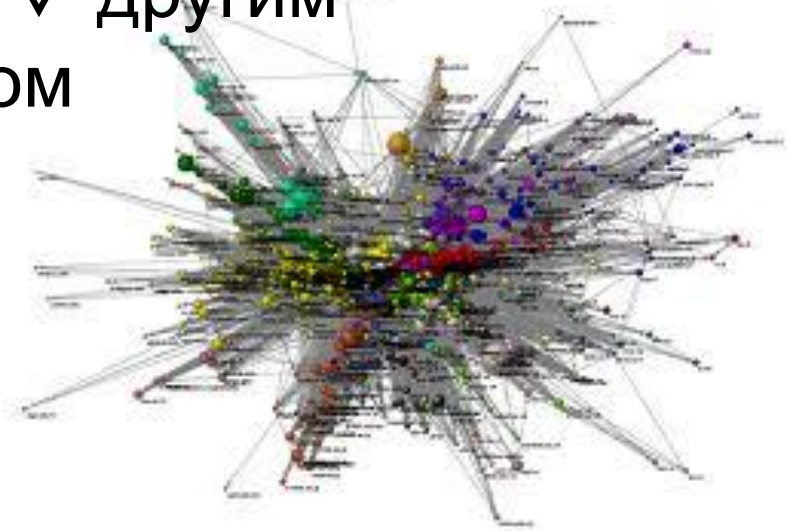
1990-е гг.

История ИП (2)



- хранилище знаний и культурных ценностей
- ∀ м. создать документ
- связать с ∀ другим документом

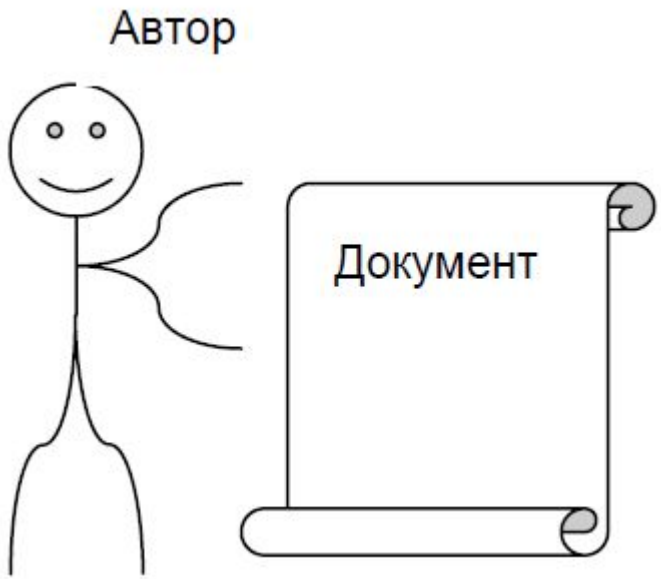
- Поиск через веб-ссылки (*hyperspace*)
- Нет чёткой модели Веба
- Интерес к ИПС

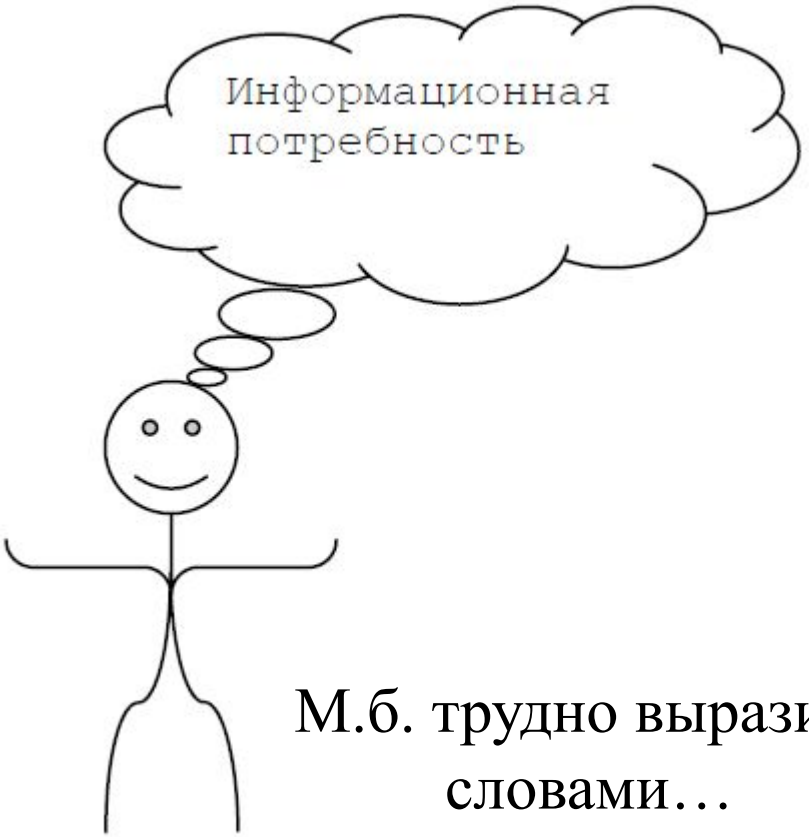


Информационный поиск

(Information retrieval) — это процесс поиска в большой **коллекции** (хранящейся, как правило, в памяти компьютеров) некоего **неструктурированного** материала ("обычно — документа), удовлетворяющего **информационные потребности**

User task – IR system (1)

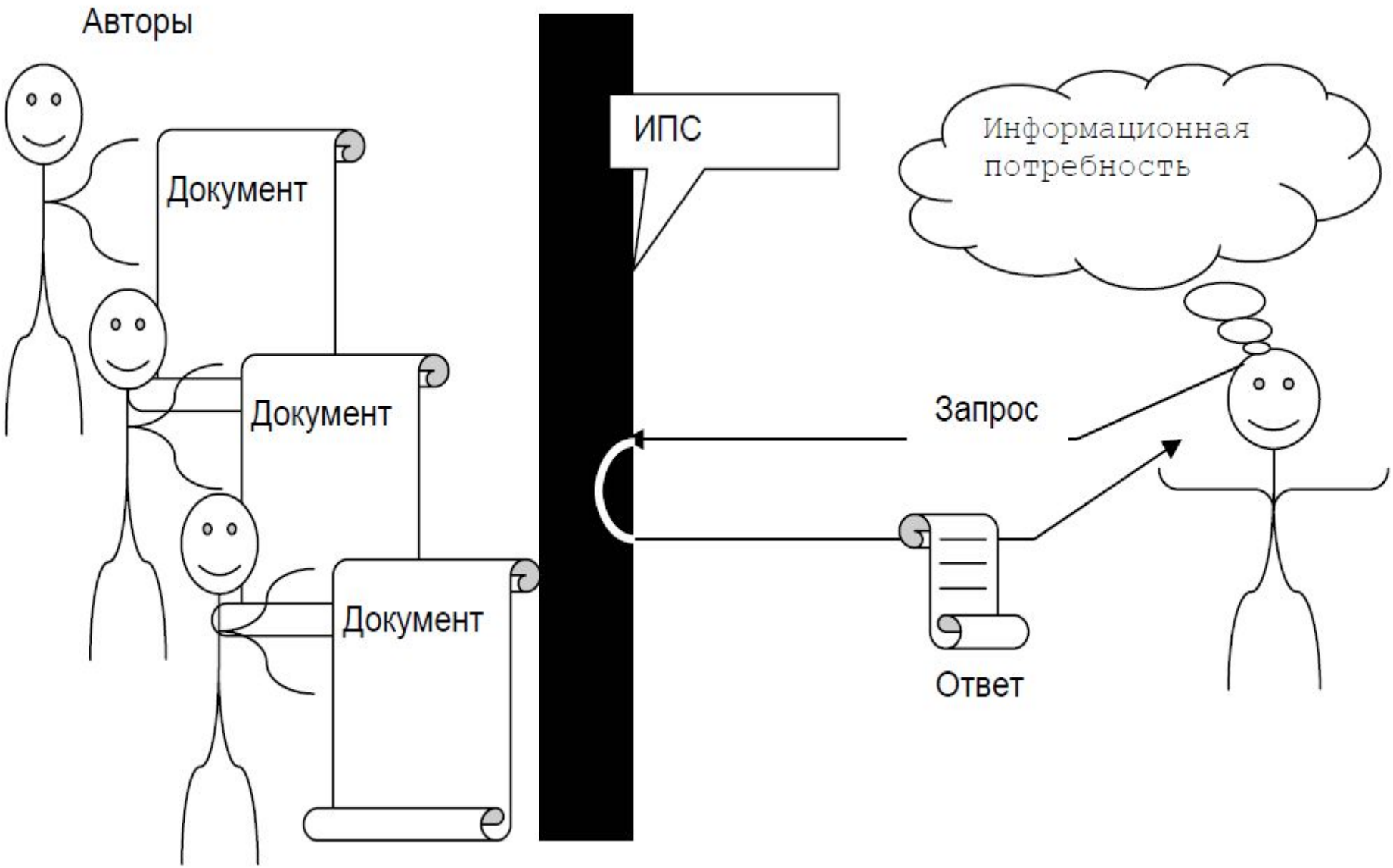




М.б. трудно выразить словами…

Капустин В. А. Основы поиска информации в Интернете. Методическое пособие. — СПб.: Институт «Открытое общество». Санкт-Петербургское отделение, 1998. — 13 с.
http://www.ict.edu.ru/ft/001919/kapustin1.pdf

User task – IR system (2)



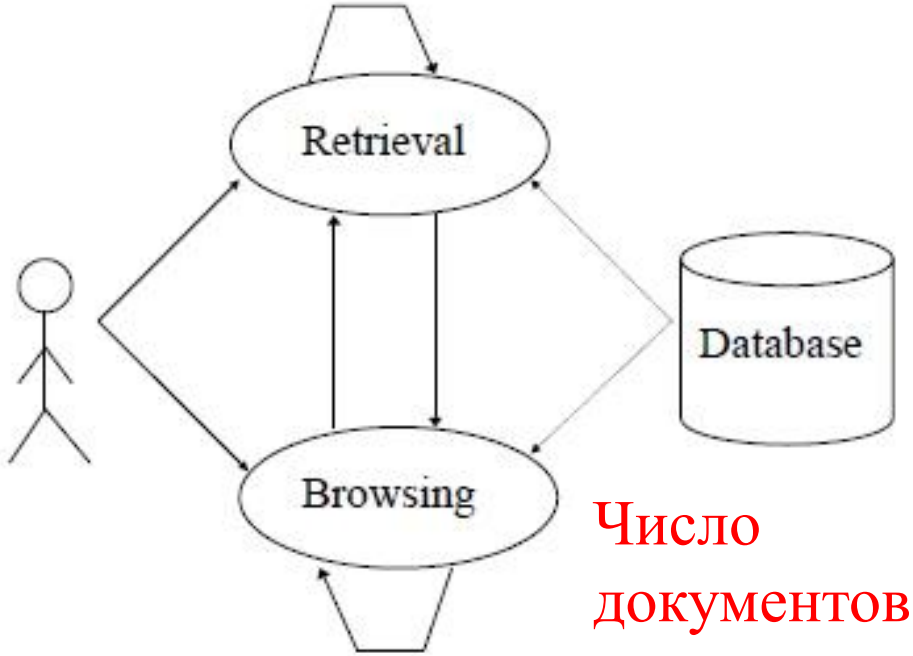
Классификация ИПС по масштабу действия

1. Веб-поиск
   1. Сбор документов
   2. Противодействие SEO
2. Персональный поиск (*personal IR*)
   1. Все форматы док-в на комп.
   2. Простота
   3. Ресурсы

- Корпоративный (Enterprise), подвед. (Institutional)
- Предметная область (domain-specific)
  - Центр-я файл-я сист.
  - Спец-е поиск-е

Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011. — ISBN 978-5-8459-1623-5

Information Retrieval (text) vs. Data Retrieval (RDBMS)



Число документов?

IR – ранжирование док-в на основе интерпретации содержимого (слово -> цифра) (релевантность)

Цель IR системы – найти все релевантные документы, и как можно меньше нерелевантных.

Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval // 1. Introduction. — Addison-Wesley, 1999. — ISBN 0-201-39829-X ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap1_introduction-modern_ir.pdf

Релевантность (1), оценка ИПС

Релевантность - степень соответствия документа запросу (инф-й потребности)

<u>Оценка ИПС</u> – эмпирическая:

1. Текстовые коллекции
2. Эксперты

Главный указатель полезности поиска?

Релевантность (2)

Удовлетворение пользователя:

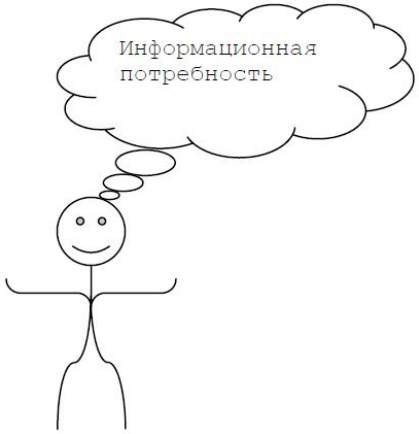
0.

1. Скорость ответа
2. Размер индекса
3. Интерфейс (удобство, наглядность, скорость отклика)

Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011. — ISBN 978-5-8459-1623-5

Релевантность (3), тестовая коллекция

1. Коллекция документов



2. Набор тестовых инф-х

потребностей (запросов), *min 50*

3. Набор оценок релевантности (обычно бинарные утверждения)

<u>Бинарная классификация</u>: эталонная оценка релевантности асессорами

Стандартные тестовые коллекции

1. Коллекция Cranfield, 1950-60
   1. 1398 аннотаций статей
   2. 225 запросов
   3. Оценки рел-ти (запрос-документ)

2. Text Retrieval Evaluation Conference (TREC). 1992…
   1. 1.89 млн док-в
   2. 450 инф-х потребностей (topics)

3. Российский семинар по оценке методов информационного поиска (www.romip.ru). 2003...

Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011. — ISBN

Набор коллекций РОМИП 2010 г.

Коллекция	Состав	Размер	Предоставлена
Narod.Ru 2003	Веб-сайты из домена narod.ru	728 000 док. 22 000 сайтов	Яндекс
Legal 2004	Законодательство РФ	60 000 док.	Кодекс
DMOZ 2003	Веб-сайты из русской части DMOZ	300 000 док. 2087 сайтов	Рамблер
News 2006	Все новости за три периода из 17 источников	31 500 док. 75 Мб	Яндекс
By.Web 2007	страницы домена .by из индекса Яндекс (май 2007)	1 524 676 док. 8 Гб	Яндекс
KM.RU 2007	~90% от объема www.km.ru на май 2007 (57 сайтов)	3 010 455 док. 13.7 Гб	КМ Онлайн
Legal 2007	Законодательство РФ, Москвы и Санкт-Петербурга (декабрь 2006)	300 000 док. 1.7 Гб	Кодекс
Flickr 2008	подмножество www.flickr.com	20 000 фот.	Flickr
ImageDupl 2008	стоп-кадры из 15 часов видеоматериала	37 800 изобр.	Оргкомитет РОМИП

Релевантность (4)

Информационная потребность

?

Релевантность

?

Запрос

Релевантность (5)

Правда ли, что красное вино более эффективно снижает риск сердечных приступов, чем белое?

Запрос

Информационная потребность

wine AND red AND white AND heart AND attack AND effective

Релевантность (6)

Правда ли, что красное вино более эффективно снижает риск сердечных приступов, чем белое?

Информационная потребность

Релевантность - степень соответствия документа …

Запрос

wine AND red AND white AND heart AND attack AND effective

+ Однословные запросы

User task – IR system

- User information need -> Query
- Keywords + regular expressions (%, *, _)
- Query example: (poorly defined or broad)

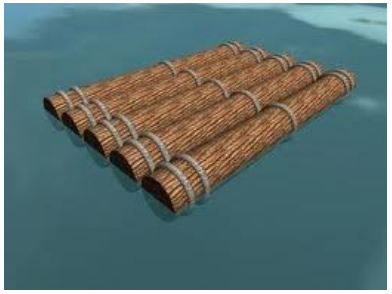
гонка

?

гонка

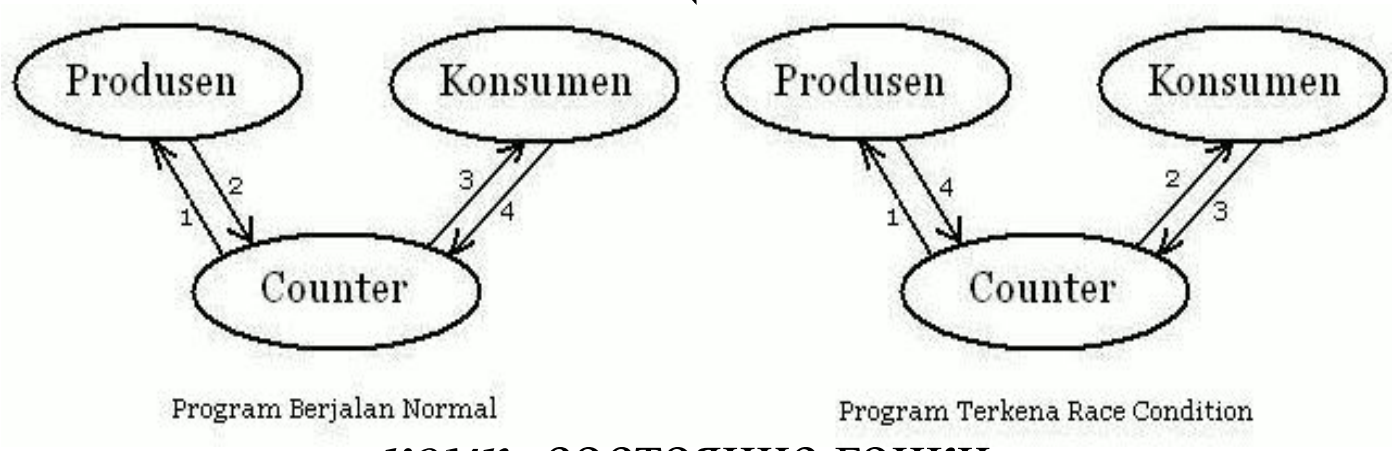
спешка,
торопливость

плоты из брёвен,
сплавляемые по
реке



автогонки,
мотогонки

компьютерная игра,
автосимулятор



комп. состояние гонки

гонка // Русский Викисловарь, 2012, http://ru.wiktionary.org/wiki/гонка

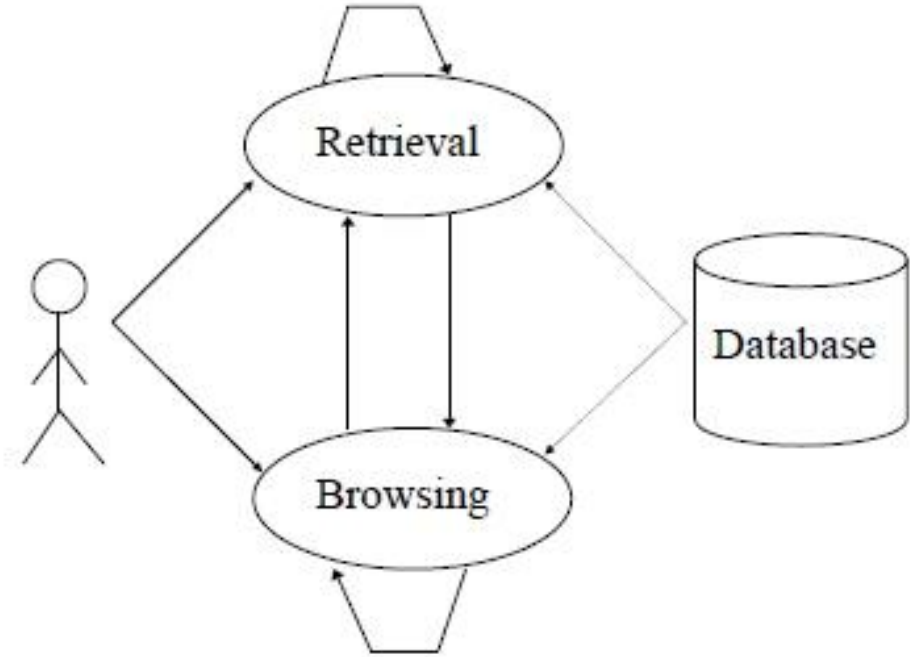
User task – IR system

- Трудность: нечёткий запрос «гонка»
  - Автоспорт
  - Париж-Дакар
  - Навигация
  - GPS

- Browsing? Сёрфинг? – тоже IR

Information (Data) Retrieval and Browsing: Pull & Push

ИПС



?

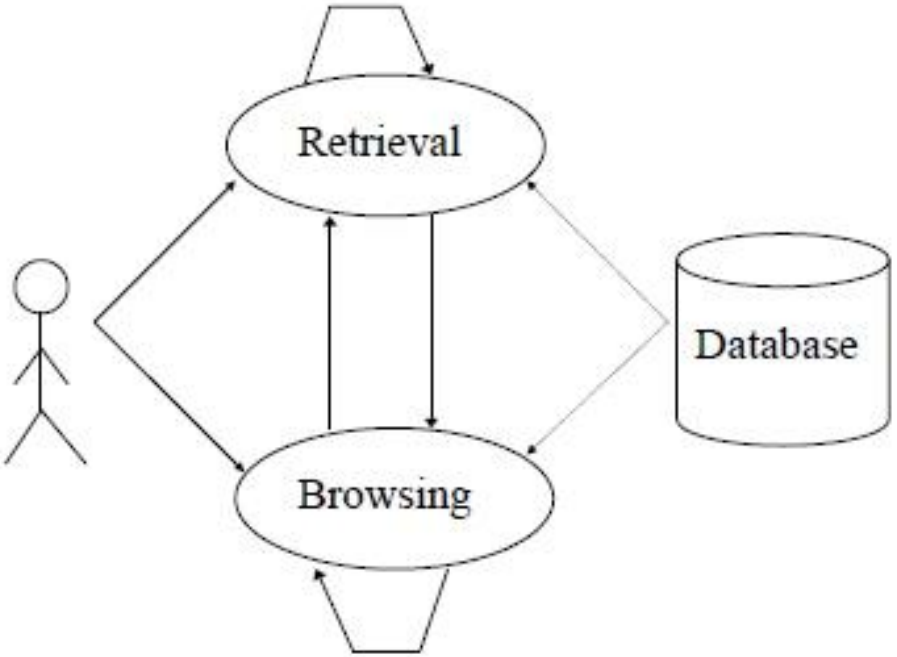
Гипертекстовые системы. Примеры?

Pull

Push

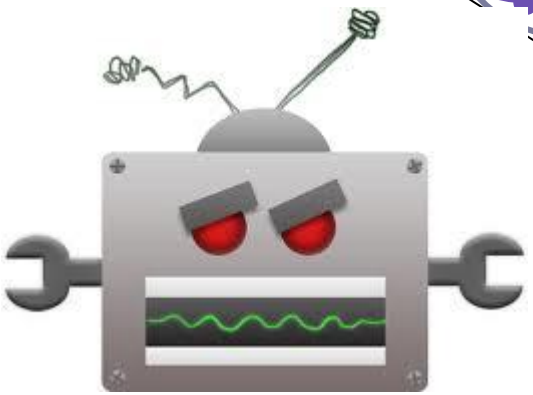
Information (Data) Retrieval and Browsing: Pull & Push

ИПС



Гипертекстовые системы. Примеры?

Pull



- новостной агент;
- Internet change detection and notification (Google alert - подписка)
- спам

Push

Представление документа

- Весь текст (самое полное предст-е)
- список **<u>всех</u>** слов
- Если упрощение, уменьш.
- Stopwords, Stemming

(Нормализация текста)

Нормализация текста (1): предобработка

- Синтез речи
- Машинный перевод
- Сохранение в базе данных
- Сравнение текстов (сортировка, индекс)

Нормализация текста (2): Этапы

1. Нормализация юникода
2. tolower() || toupper()
3. Digits → words
4. Удаление ударений, диакритики
5. Удаление знаков пунктуации
6. Раскрытие аббревиатур
7. Удаление стоп слов
8. Нормализация слов (стемминг)
9. Канонизация текста (полные синонимы)
   - "co-operation" → "cooperation",
   - «чо», «чё» → «что»
   - "should've" → "should have"

¿Словари

¿Регулярные выражения

Шумовые слова – слишком общие слова (stop words)

1. Общие

- предлоги, союзы, междометия, цифры, частицы (зависят от языка)

2. Зависимые

Словоформа - слово в узком смысле, то есть обладающая признаками слова цепочка **фонем**, формально отличающаяся от другой.

Фонема — минимальная единица звукового строя языка

Нормализация слова: «фонем» мн.ч., Род. п. -> «фонема» Им.п., ед.ч.

Нормализация слов (стемминг)

- нахождение основы слова для заданного исходного слова (не обязательно «корень»). «Портер»:

павлыч		павлыч
павлыча		павлыч
пагубная		пагубн
падает		пада
падай		пада
падал		пада
падала		пада
падаль		падал
падать		пада
падаю		пада
падают		пада
падающего		пада
падающие		пада
падеж	=>	падеж

А. Лемматизация

- Приведение словоформы к лемме — её нормальной (словарной) форме

1. Определение POS
2. Правила

кошками → кошка

бежал → бежать

дутых → дутый

В. Стохастические алг.

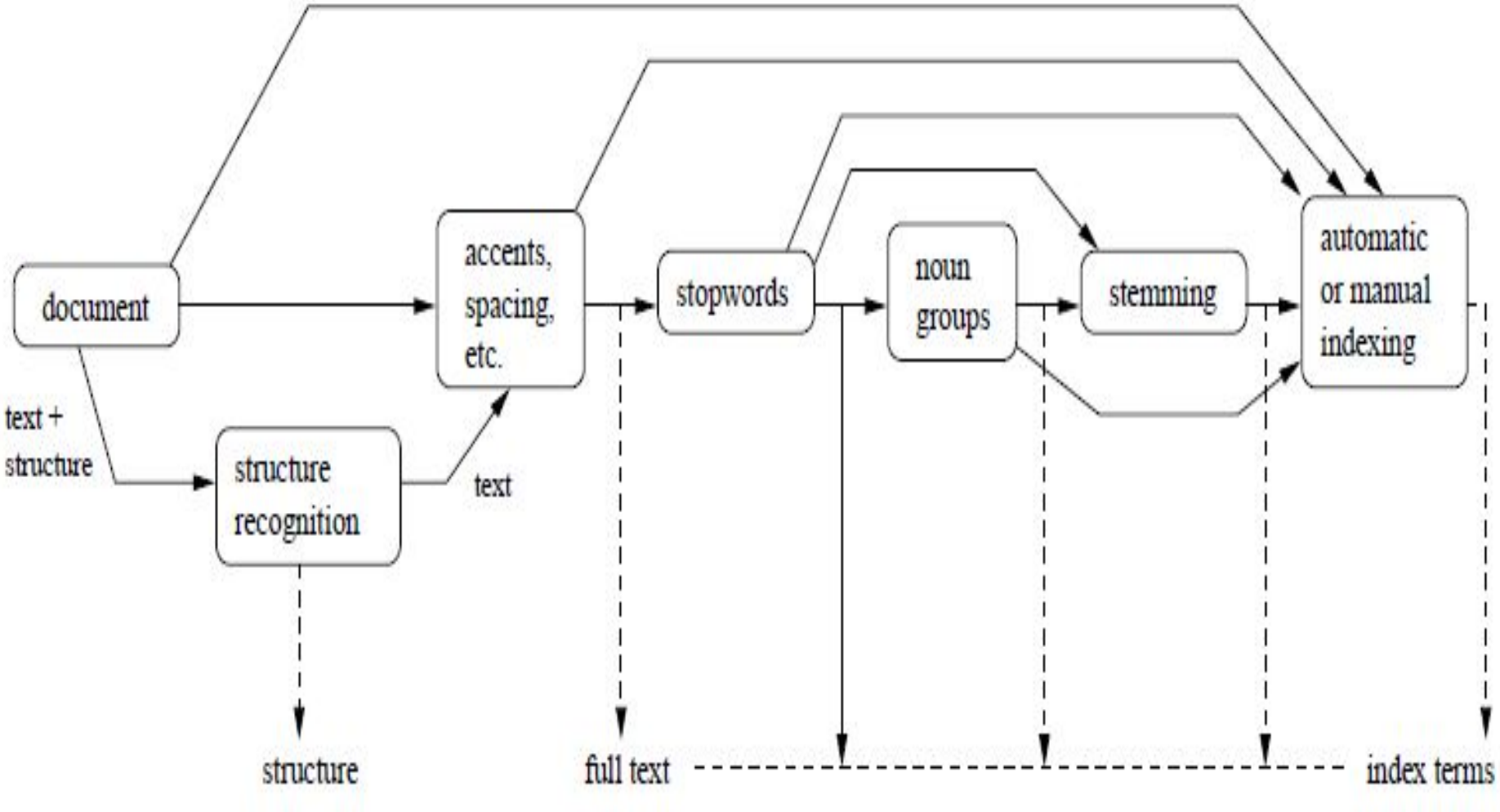
вероятность + context

Представление документа

- Весь текст (самое полное предст-е)
- Список **<u>всех</u>** слов

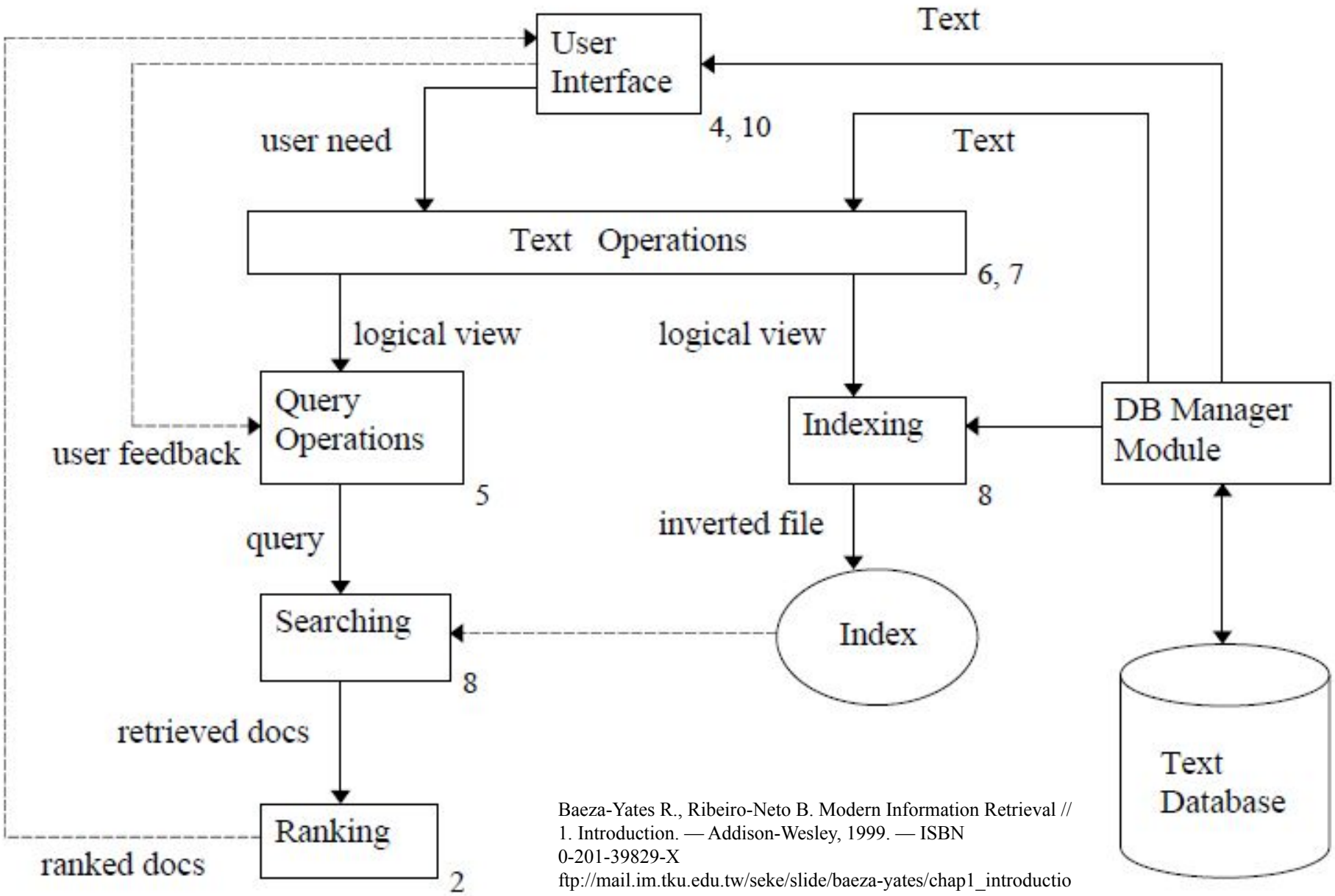
- Список нормализованных слов
- Индекс (слово -> документ)

Представление документа



Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval // 1. Introduction. — Addison-Wesley, 1999. — ISBN 0-201-39829-X ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap1_introduction-modern_ir.pdf

Архитектура ИПС (упрощ., обобщ.)



Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval // 1. Introduction. — Addison-Wesley, 1999. — ISBN 0-201-39829-X
ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap1_introduction-modern_ir.pdf

Литература

- **Маннинг** К., **Рагхаван** П., **Шютце** Х. **Введение в информационный поиск**. — Вильямс, 2011. — ISBN 978-5-8459-1623-5. http://rutracker.org/forum/viewtopic.php?t=3887364

- **Baeza-Yates** R., **Ribeiro-Neto** B. **Modern Information Retrieval** // 1. Introduction. — Addison-Wesley, 1999. — ISBN 0-201-39829-X ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap1_introduction-modern_ir.pdf

- **Капустин** В. А. **Основы поиска информации в Интернете**. Методическое пособие. — СПб.: Институт «Открытое общество». Санкт-Петербургское отделение, 1998. — 13 с. http://www.ict.edu.ru/ft/001919/kapustin1.pdf

Вопросы?

http://vk.com/club41102811

“Интернет-математика в ПетрГУ”