Основы обработки текстов

Лекция 9

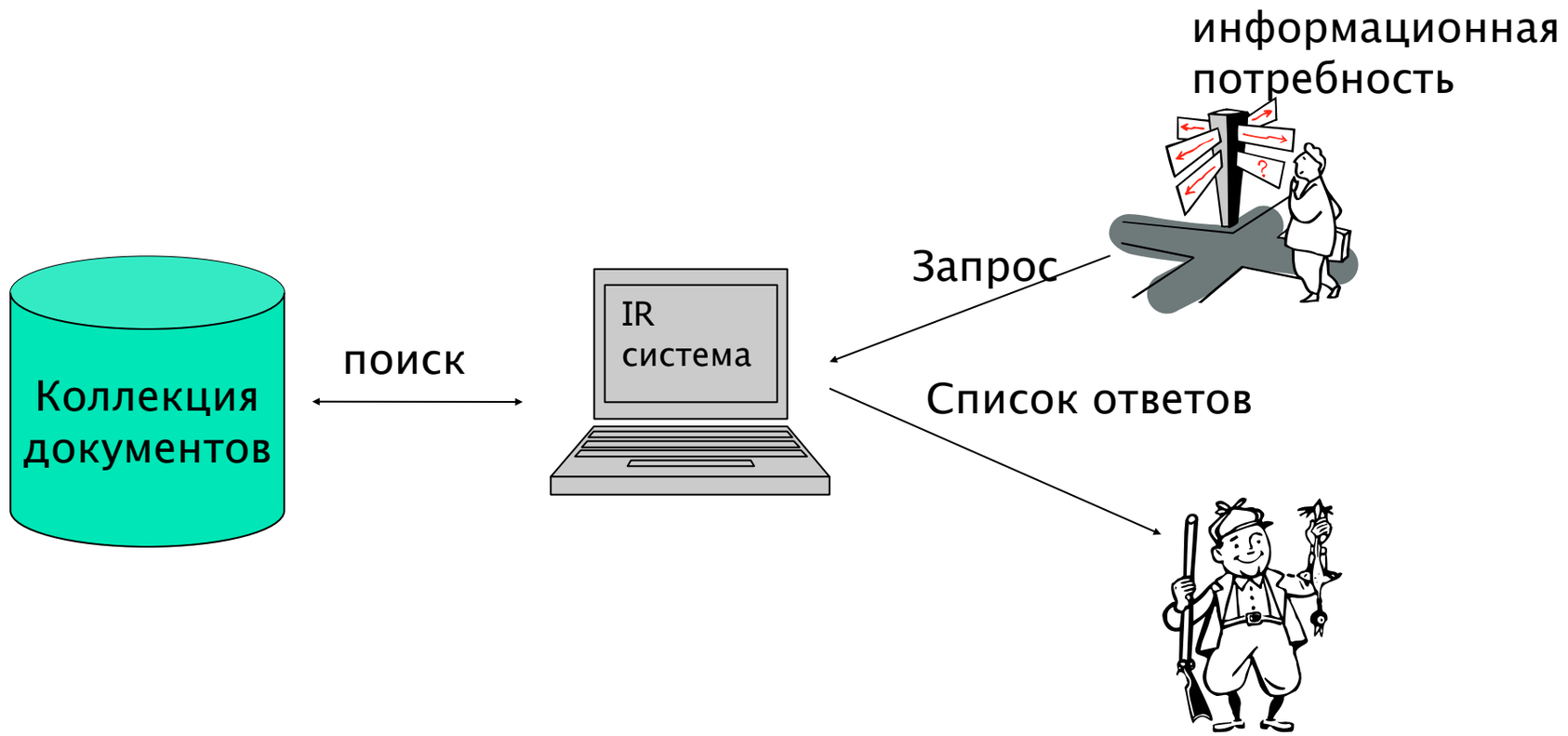
Приложения обработки текстов

План

- Информационный поиск
- Вопросно-ответные системы
- Автоматическое реферирование

Информационный поиск

- Information retrieval (IR)
- Поиск всех документов из заданного множества, отвечающих запросам пользователя



Проблема информационного поиска

- Первое приложение в библиотечном деле

ISBN: 0-201-12227-8

Author: Salton, Gerard

Title: Automatic text processing: the transformation, analysis, and retrieval of information by computer

Editor: Addison-Wesley

Date: 1989

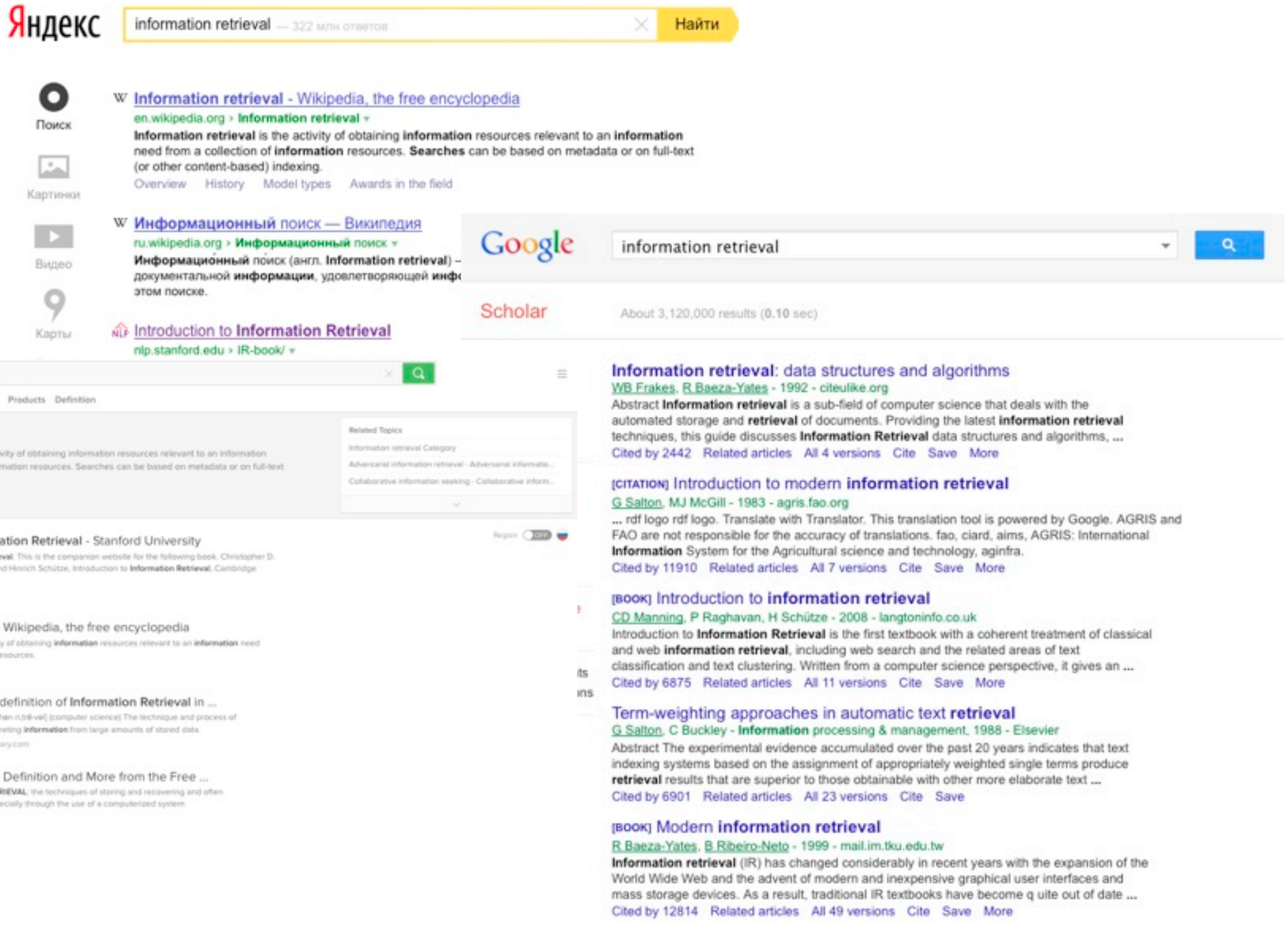
Content: <Text>

- Поиск по внешним аттрибутам - поиск в БД
- IR: поиск по контенту

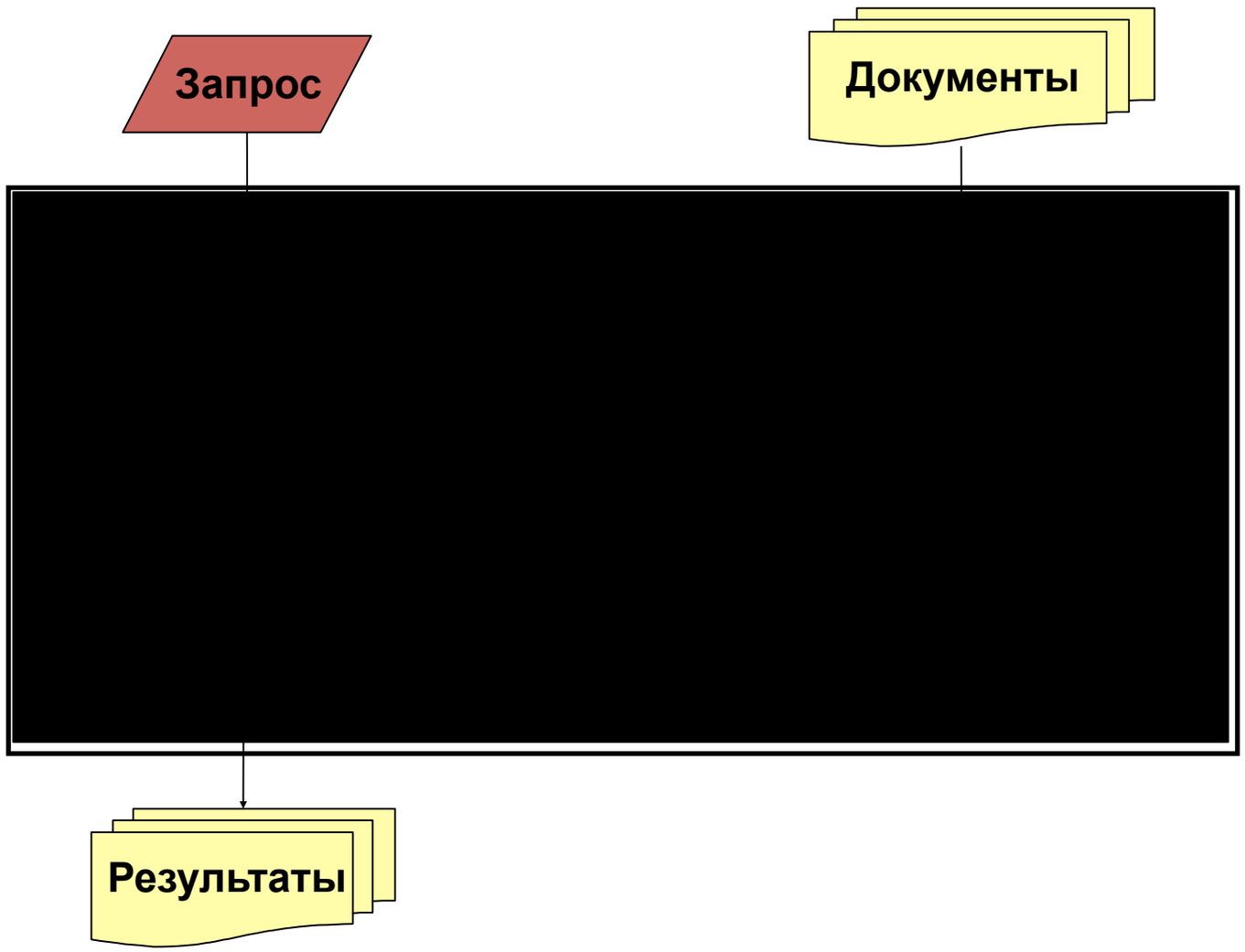
Возможные подходы

- Поиск близких строк
  - –Медленно
  - –Тяжело улучшать
- Индексирование
  - –Быстро
  - –Возможности для улучшений

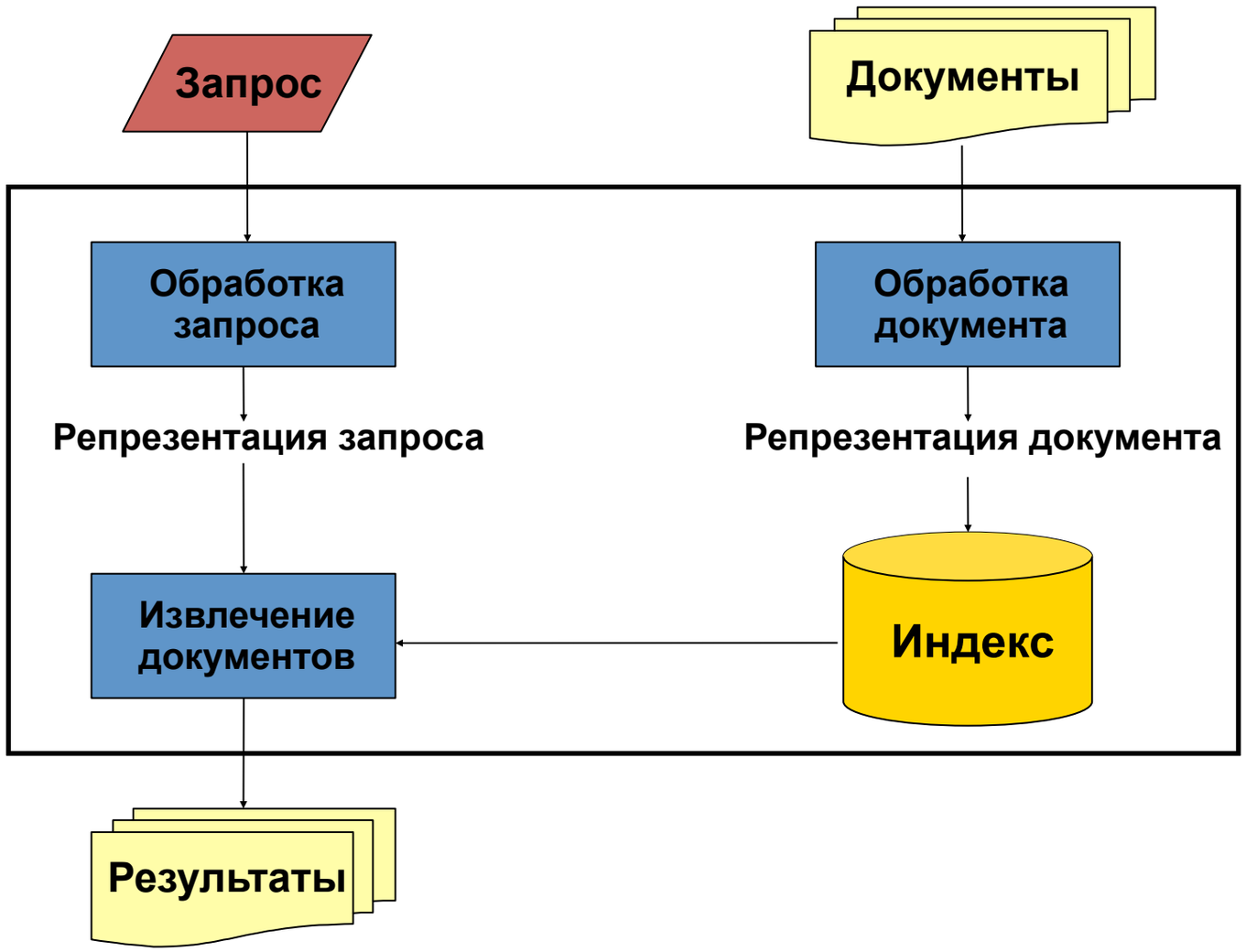
Примеры систем



Архитектура систем



Архитектура систем



Основные проблемы

- Обработка запроса и документа
  - –Какой наилучший способ представления запроса и документа
- Извлечение документов
  - –Как понять какой документ наилучшим образом удовлетворяет запросу
- Оценка систем
  - –Как понять что система работает хорошо

Представление документа

- Модель мешка слов (bag-of-words)
- Взвешивание слов (терминов)
  - tf = term frequency
    - частота встречаемости термина в документе
  - df = document frequency
    - число документов, содержащих термин
  - idf = inverse document frequency
    - специфичность термина
  - weight(t,D) = tf(t,D) * idf(t)

Варианты tf-idf

- tf(t, D)=freq(t,D)
- tf(t, D)=log[freq(t,D)]
- tf(t, D)=log[freq(t,D)]+1
- tf(t, D)=freq(t,d)/Max[f(t,d)]

idf(t) = log(N/n)

n = # документов содержит t

N = # документов в корпусе

Стоп-слова

- Функциональные слова не несут полезной инорфмации для IR систем
- Списки стоп-слов
  - –Предлоги
  - –Артикли
  - –Местоимения
  - –Некоторые частые слов (например, “документ”)
- Удаление стоп-слов часто улучшает качество IR систем
- Часто используются “стандартные” списки стоп-слов

Стемминг

- Разные формы слова имеют одно значение. Необходимо иметь для них одинаковое представление
- Стемминг: отбрасывание окончания слова до неизменяемой формы
  - –поиск
  - –поиску
  - –поисковый

  } поиск
- Porter stemmer
  - – nltk.stem.porter.PorterStemmer

Лемматизация

- Стемминг часто “схлопывает” принципиально разные слова
  - кошки -> кош
  - кошельки -> кош
- Лучше приводить слово к “нормальной” форме
  - кошки -> кошка
  - кошельки -> кошелек
- Более точные результаты, но более сложные алгоритмы

Результат индексирования

- Инвертированный индекс

Doc 1

This is a sample document with one sample sentence

Doc 2

This is another sample document

Term	# docs	Total freq
This	2	2
is	2	2
sample	2	3
another	1	1
…	…	…

Doc id	Freq
1	1
2	1
1	1
2	1
1	2
2	1
2	1
…	…
…	…

Извлечение документов

- Запрос из одного слова
  - Берем инвертированый список для слова
- Запрос из нескольких слов
  - Комбинирование нескольких списков
  - Как интерпретировать вес?
  - Модель информационного поиска

Модели информационного поиска

- Документ D = множество взвешенных ключевых слов
- Запрос Q = множество невзвешенных слов
- $R(D,Q) = \sum_i w(t_i, D)$

  t_i - слова запроса

Булева модель

- Документ - логическая конбъюнкция слов
- Запрос - Булево выражение
- $R(D, Q) = D \rightarrow Q$

$D = t_1 \wedge t_2 \wedge \ldots \wedge t_n$
$Q = (t_1 \wedge t_2) \vee (t_3 \wedge \neg t_4)$
$D \rightarrow Q$, то есть $R(D, Q) = 1$.

- Проблемы
  - R - либо 0, либо 1 (неупорядоченное множество документов
  - Сложно писать запросы

Векторная модель

- Векторное пространство всех слов

$$<t_1, t_2, t_3, \ldots, t_n>$$

- Документ

$$D = < a_1, a_2, a_3, \ldots, a_n>$$

a_i = вес t_i в D

- Запрос

$$Q = < b_1, b_2, b_3, \ldots, b_n>$$

b_i = вес t_i в Q

Матричное представление

Пространство документов	t_1	t_2	t_3	…	t_n
D_1	a_{11}	a_{12}	a_{13}	…	a_{1n}
D_2	a_{21}	a_{22}	a_{23}	…	a_{2n}
D_3	a_{31}	a_{32}	a_{33}	…	a_{3n}
…					
D_m	a_{m1}	a_{m2}	a_{m3}	…	a_{mn}
Q	b_1	b_2	b_3	…	b_n

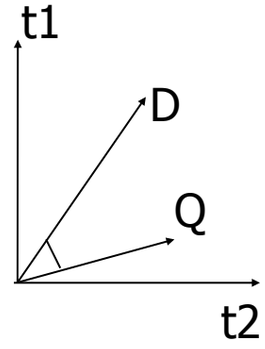
Пространство терминов

Разреженная матрица!

Подсчет близости

Скалярное произведение

$$Sim(D,Q) = \sum (a_i * b_i)$$



Косинус

$$Sim(D,Q) = \frac{\sum_i (a_i * b_i)}{\sqrt{\sum_i a_i^2 * \sum_i b_i^2}}$$

Мера Дайса

$$Sim(D,Q) = \frac{2\sum_i (a_i * b_i)}{\sum_i a_i^2 + \sum_i b_i^2}$$

Мера Жаккара

$$Sim(D,Q) = \frac{\sum_i (a_i * b_i)}{\sum_i a_i^2 + \sum_i b_i^2 - \sum_i (a_i * b_i)}$$

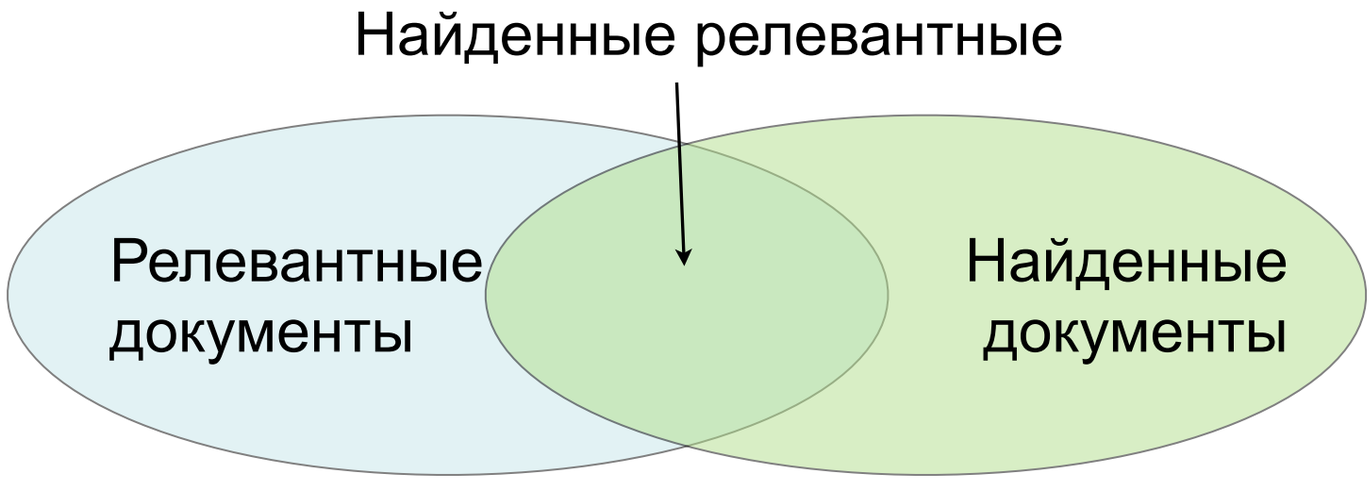
Какие еще бывают модели

Properties of the Model / Mathematical Basis	without term-interdependencies	with term-interdependencies: immanent term-dependencies	with term-interdependencies: transcendent term-interdependencies
set-theoretic	Standard Boolean; Extended Boolean		Fuzzy Set
algebraic	Vector Space; Extended Boolean	Generalised Vector Space; Latent Semantic; Spread. Activation Neuronal Network	Topic-based Vector Space; Balanced Topic-based Vector Space; Backpropagation Neuronal Network
probabilistic	Binary Interdependence; Inference Network; Belief Network	Language	Retrieval by Logical Imaging

*http://en.wikipedia.org/wiki/Information_retrieval

Оценка систем

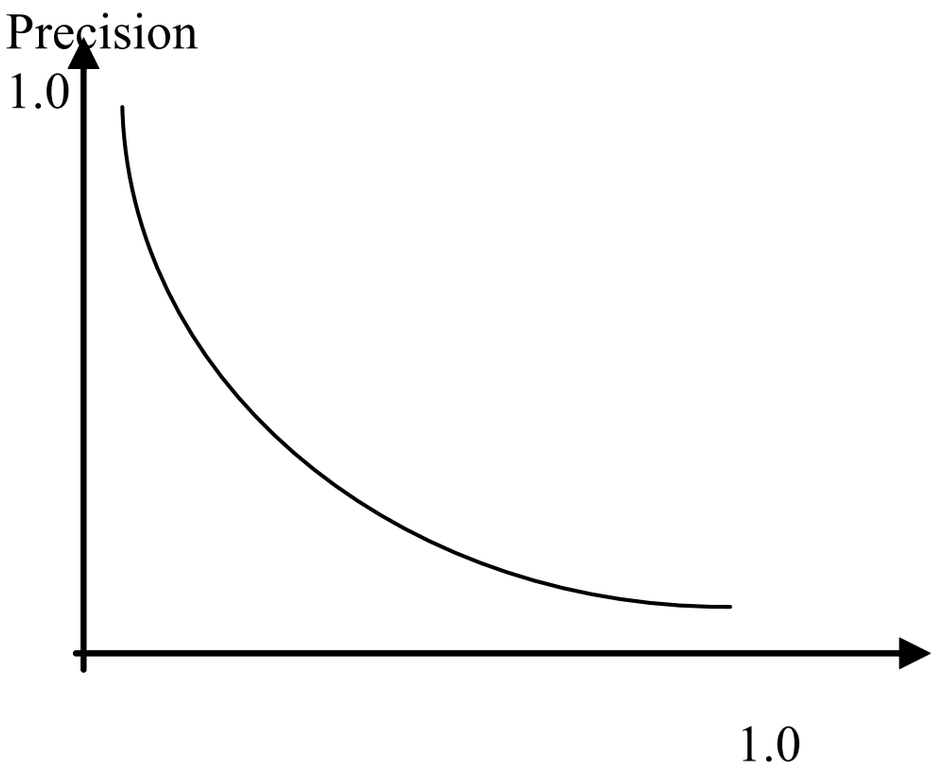
- Точность и полнота



– Точность = найденные релевантные / найденные документы
– Полнота = найденные релевантные / релевантные документы

Точность и полнота

- Общая форма зависимости
  - Точность и полнота зависимы
  - Системы нельзя сравнивать в одной точке
  - Вычисляют среднюю точность (в 11 точках с полнотой: 0.0, 0.1, ..., 1.0)



$$\text{AveP} = \int_0^1 p(r)dr$$

$$\text{AveP} = \sum_{k=1}^{n} P(k)\Delta r(k)$$

$$\text{AveP} = \frac{\sum_{k=1}^{n}(P(k) \times \text{rel}(k))}{\text{number of relevant documents}}$$

$rel(k) \in \{0,1\} = 1$, если k-й документ релевантен запросу

MAP

- Mean Average Precision

$$MAP = \frac{1}{n}\sum_{Q_i}\frac{1}{|R_i|}\sum_{D_j \in R_i}\frac{j}{r_{ij}}$$

- r_{ij} = ранг j-го релевантного документа для Q_i
- $|R_i|$ = число релевантных документов для Q_i
- n = # тестовых запросов

Ранг		
1	4	1ый рел. док.
5	8	2ой рел. док.
10		3ий рел. док.

$$MAP = \frac{1}{2}\left[\frac{1}{3}\left(\frac{1}{1}+\frac{2}{5}+\frac{3}{10}\right)+\frac{1}{2}\left(\frac{1}{4}+\frac{2}{8}\right)\right]$$

Темы для дальнейшего изучения

- Ранжирование
  - PageRank (Google), HITS, ...
- Семантический поиск
  - ключевые слова VS ключевые понятия
- IR для (полу-) структурированных данных
- Сбор данных в Вебе
- Мультимедийный поиск
- Исследовательский поиск
- Многоязычный поиск
- Сжатие и хранение данных
- Нечеткий поиск
- Учет обратной связи от пользователей
- Персонализация

Вопросно-ответные системы

Какой национальности бывший папа римский Бенедикт XVI?

Ватикан выступил во вторник, 12 мая, с опровержением информации о том, что Папа Римский Бенедикт XVI в юности состоял в гитлерюгенде. "Йозеф Рацингер (имя понтифика, **немца по национальности**) никогда не состоял в гитлерюгенде - идеологической нацистской организации.

Короткий фрагмент текста, не URL

Ответ: Немец

Примеры систем

WolframAlpha computational... knowledge engine

Enter what you want to calculate or know about:

How far is San-Francisco from Moscow?

Examples Random

Assuming Moscow (Russia) | Use Moscow (Idaho, USA) or more instead

Input interpretation:

Moscow to San Francisco, California, United States

Distance: Show non-metric units

9472 km (kilometers)

Direct travel times: More

aircraft (550 mph)	10 hours 40 minutes
sound	7 hours 40 minutes
light in fiber	44.3 ms (milliseconds)
light in vacuum	31.6 ms (milliseconds)

(assuming constant–speed great–circle path)

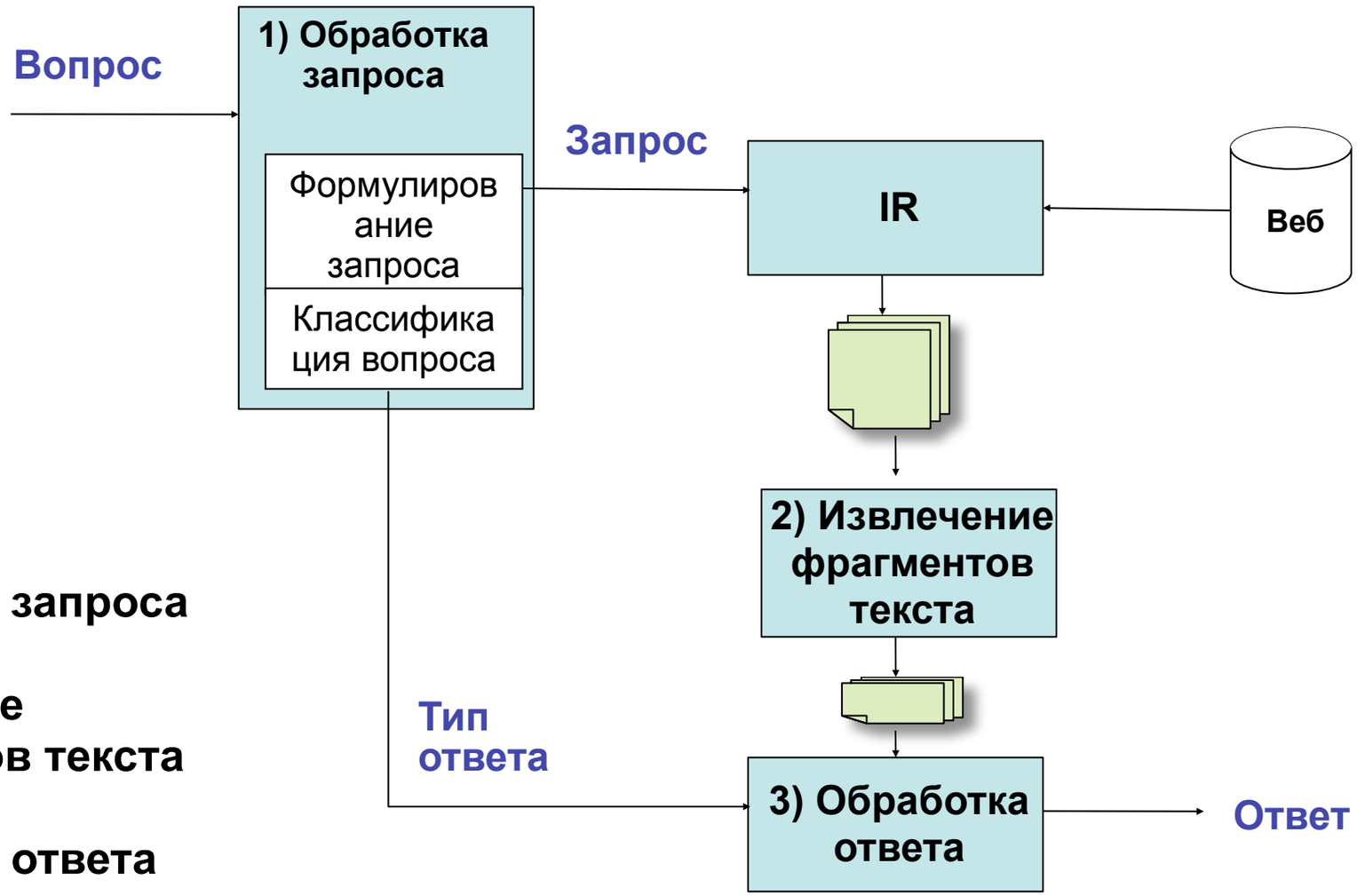
Типы вопросов

О фактах	Какая обычная высота жирафа? Где расположен главный офис Google ?
Списки	Какие страны экспортируют нефть? Какие названия имеют штаты США?
Определения	Кто такой Франсуа Томбалбай? Что такое квазар?

Вопросы о фактах

- Ответом служит простой факт
  - Примеры:
    - Где расположен Лувр?
    - Какая называется валюта Китая?
    - Какой официальный язык Алжира?
- Существует большая разница между постановкой вопроса и описанием ответа в тексте
  - Какая компания является лидером по производству открыток?
  - Компания "Арт и Дизайн" десять лет назад создала в России практически новый рынок. Теперь она является лидером среди отечественных производителей поздравительных открыток.

Типичная архитектура QA-систем



Три этапа:

1. **Обработка запроса**
2. **Извлечение фрагментов текста**
3. **Обработка ответа**

Обработка запроса

- Из вопроса на естественном языке извлекаем:
  - ключевые слова для запроса к информационно-поисковой системе
    (Формулирование запроса)
  - Тип ответа, специфицирующий класс сущности, возвращаемой в качестве ответа
    (Классификация вопроса)

Формулирование запроса

- Извлечь ключевые термины из вопроса
  - возможно расширить вопрос лексически/ семантически близкими словами
- Вопрос моделируется как множество ключевых слов

Question (from TREC QA track)	Lexical terms
Q002: What was the monetary value of the Nobel Peace Prize in 1989?	monetary, value, Nobel, Peace, Prize, 1989
Q003: What does the Peugeot company manufacture?	Peugeot, company, manufacture
Q004: How much did Mercury spend on advertising in 1993?	Mercury, spend, advertising, 1993
Q005: What is the name of the managing director of Apricot Computer?	name, managing, director, Apricot, Computer

Формулирование запроса

- Применение правил для переформулирования вопроса
  - к форме подстроки декларативного ответа
  - “когда был придуман лазер” → “лазер был придуман”
- Послать переформулированный запрос информационно-поисковой системе
- Правила (Lin 07)
  - wh-word did A verb B → A verb-ed B
  - Where is A → A is located in

Классификация вопросов

- Классификация вопросов по ожидаемому ответу

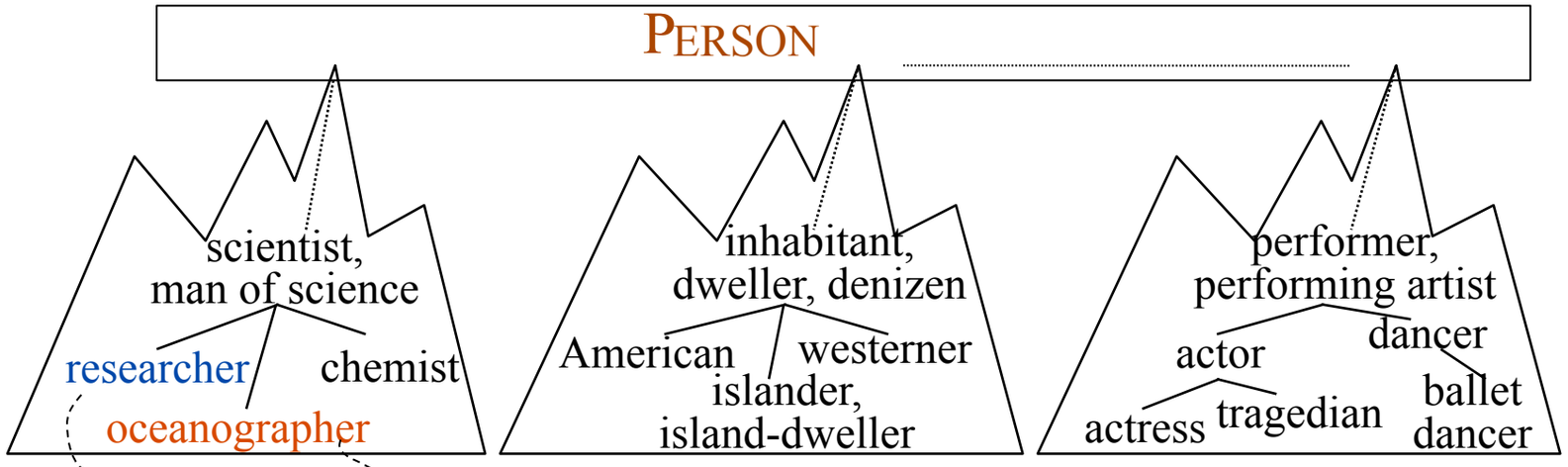
Вопрос	Основа вопроса	Тип ответа
Q555: What was the name of Titanic's captain?	What	Person
Q654: What U.S. Government agency registers trademarks?	What	Organization
Q162: What is the capital of Kosovo?	What	City
Q661: How much does one ton of cement cost?	How much	Quantity

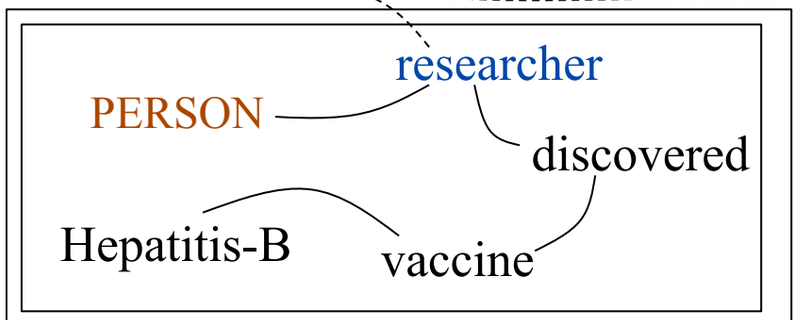
Определение типа ответа

- В некоторых случаях тип ответа можно определить по вопросу
  - Почему → Причина
  - Когда → Дата
- Для многозначных вопросов использовать дополнительные понятия в вопросе
  - What was the name of Titanic's captain?
  - What U.S. Government agency registers trademarks?
  - What is the capital of Kosovo?
- Машинное обучение (если есть размеченный корпус

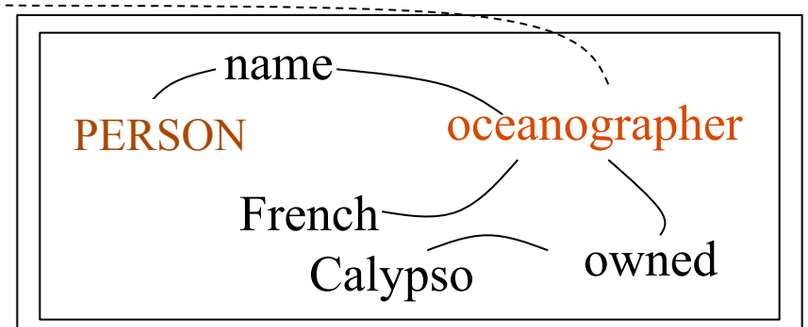
Определение типов ответов

Таксономия типов ответов (из Wordnet)





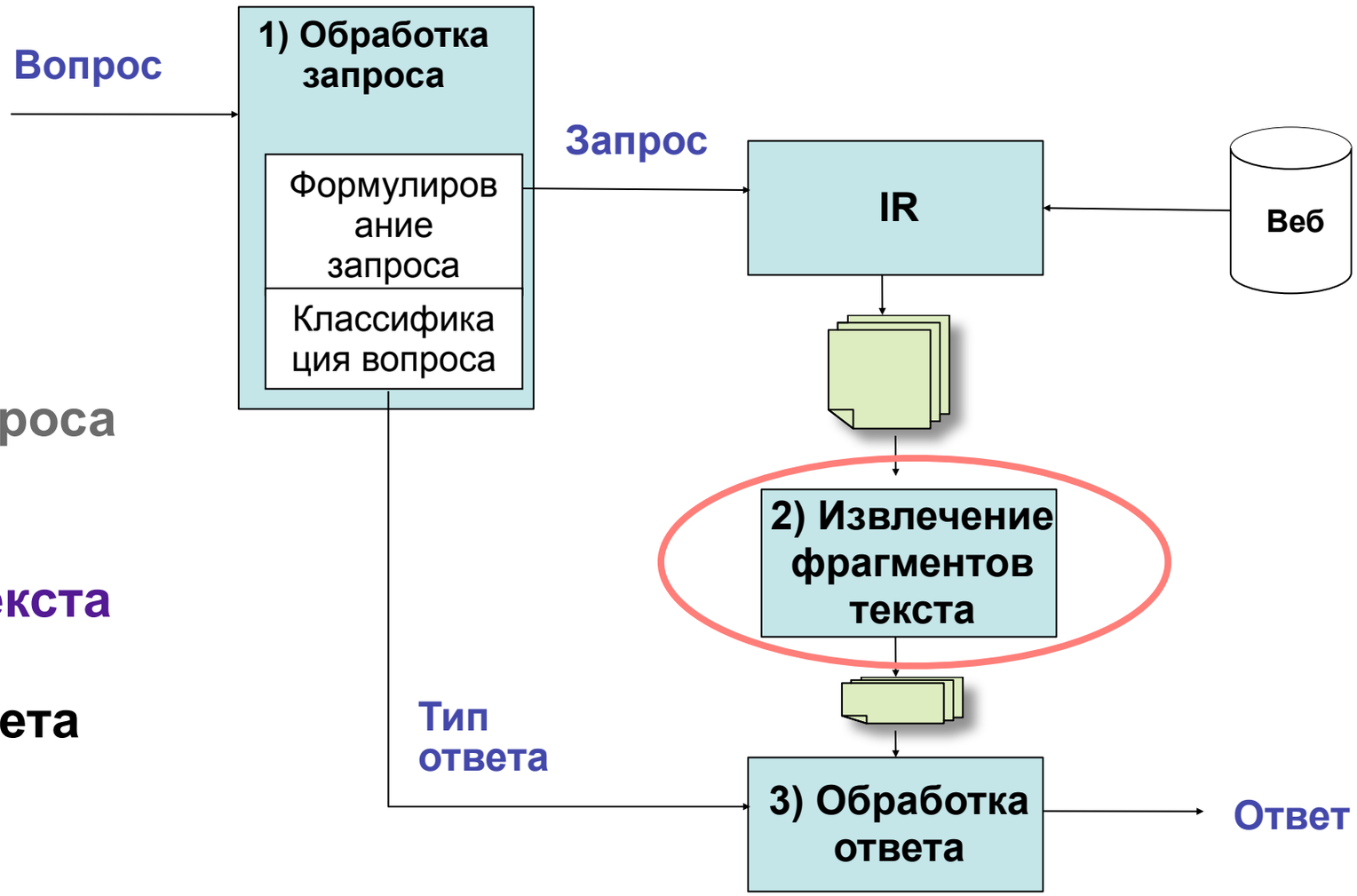
What researcher discovered the vaccine against Hepatitis-B?



What is the name of the French oceanographer who owned Calypso?

Типичная архитектура QA-систем

1. Обработка запроса
2. **Извлечение фрагментов текста**
3. **Обработка ответа**



Извлечение фрагментов текста

- IR-система возвращает список документов
- Необходимым фрагментом может быть предложение или параграф
- Необходимо выбрать фрагменты, потенциально содержащие ответ

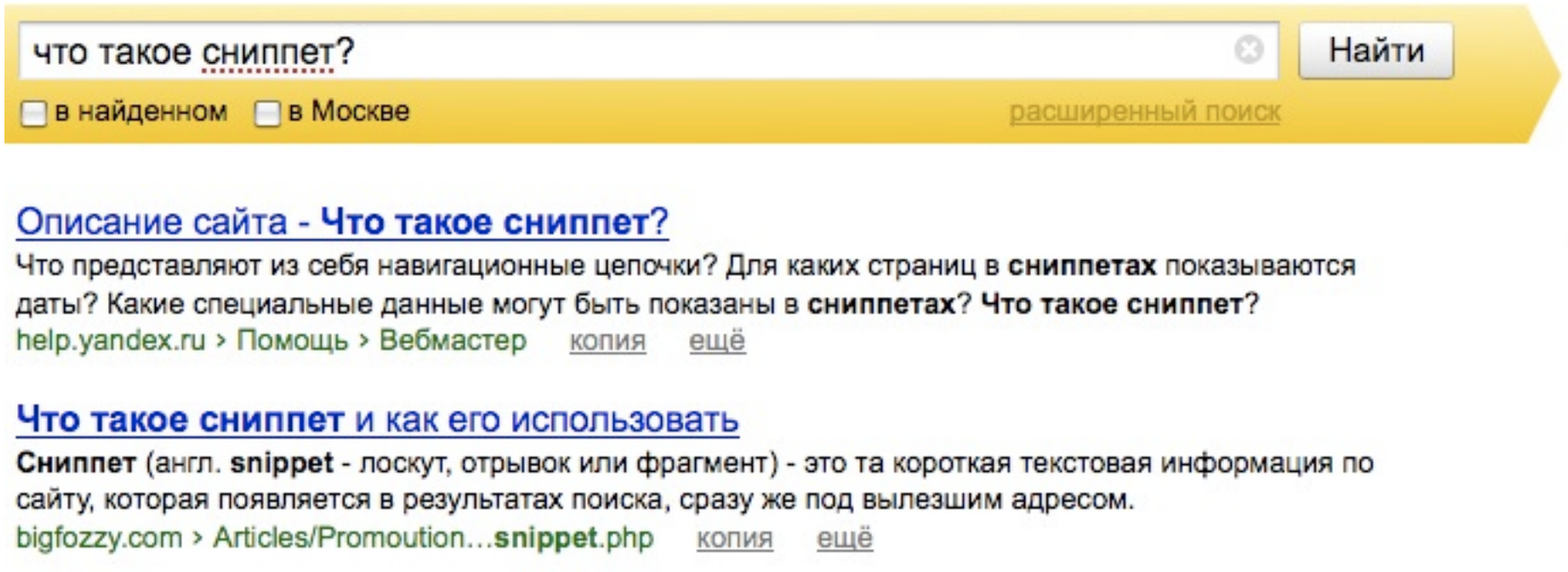
1. Отсеять фрагменты не содержащие ответ
   - распознавание именованных сущностей и классификация ответов
2. Отранжировать оставшиеся фрагменты
   - Правила, составленные вручную
   - Машинное обучение

Извлечение фрагментов текста (ранжирование)

- Признаки
  - Число именованных сущностей правильного типа в фрагменте
  - Число ключевых слов из вопроса в фрагменте
  - Наиболее длинная последовательность ключевых слов запроса в фрагменте
  - Ранг документа (IR), содержащего фрагмент
  - Плотность ключевых слов из вопроса в фрагменте
  - Пересечение N-грамм вопроса и фрагмента

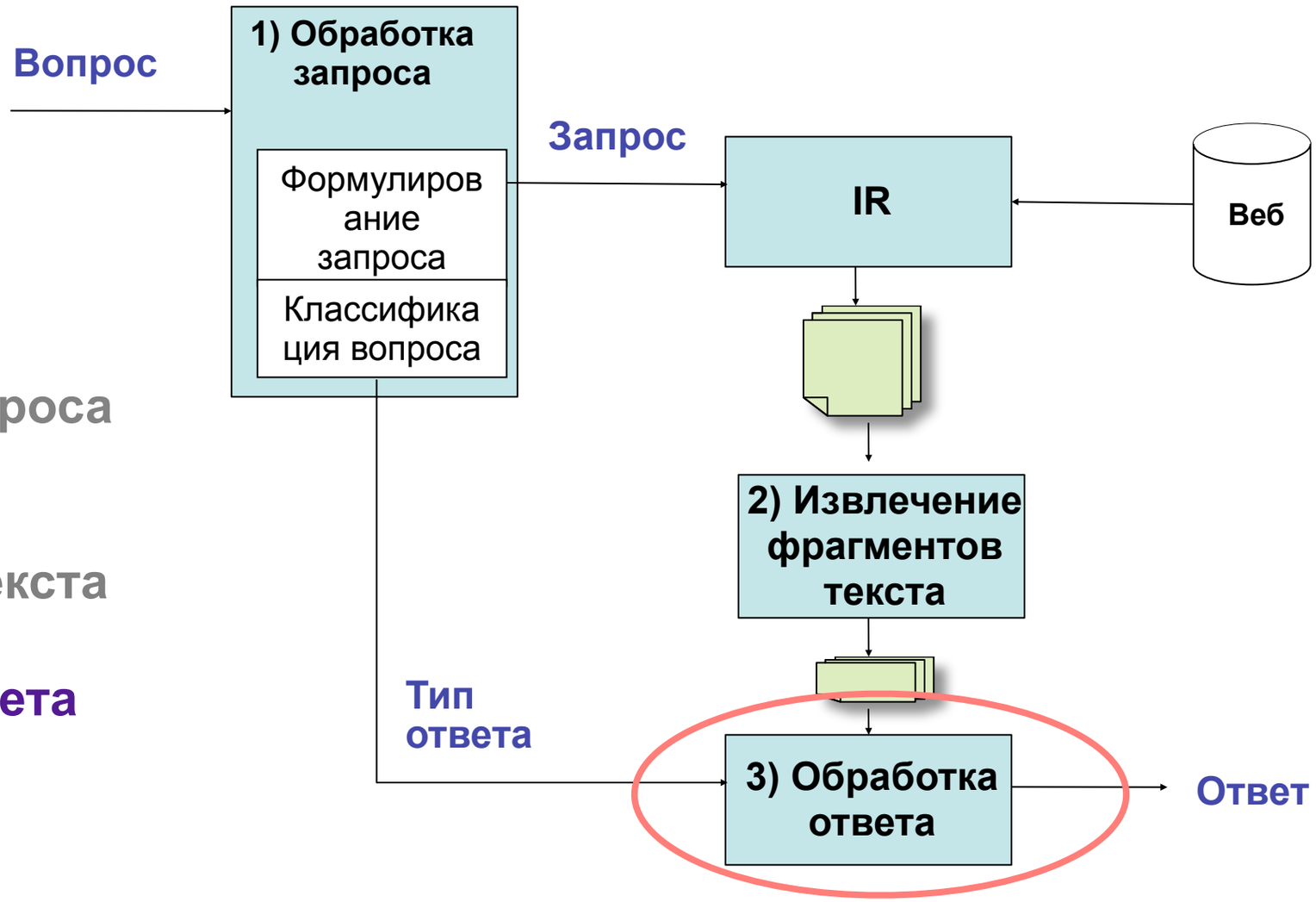
Извлечение фрагментов

- Для извлечения ответа из Веба можно пропустить шаг извлечения фрагмента и использовать сниппеты, возвращаемые информационно-поисковыми системами



Типичная архитектура QA-систем

1. Обработка запроса
2. Извлечение фрагментов текста
3. **Обработка ответа**



Обработка ответа

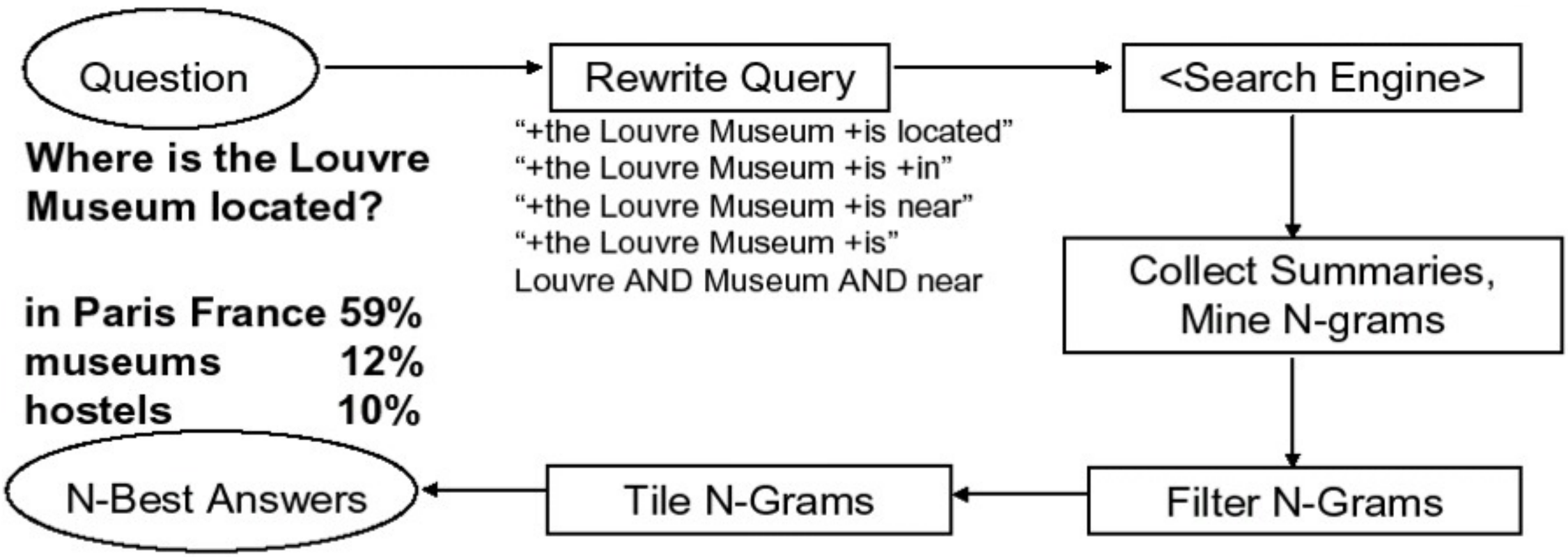
- Извлечение специфичного ответа из фрагмента
- Два основных класса алгоритмов
  - Основанные на шаблонах
  - Сбор ответа из N-грамм (N-gramm tiling)

Алгоритмы на основе шаблонов

- Использование информации о типе в регулярных выражениях
  - Если тип ответа ЧЕЛОВЕК, извлечь именованные сущности ЧЕЛОВЕК из фрагмента
- Некоторые типы ответов (например, определения) не подразумевают конкретного типа именованной сущности в ответе
  - Использовать регулярные выражения (созданные вручную или автоматически)

Pattern	Question	Answer
<AP> such as <QP>	*What is autism?*	*developmental disorders such as autism*

Сбор ответа из N-грамм
Архитектура AskMSR

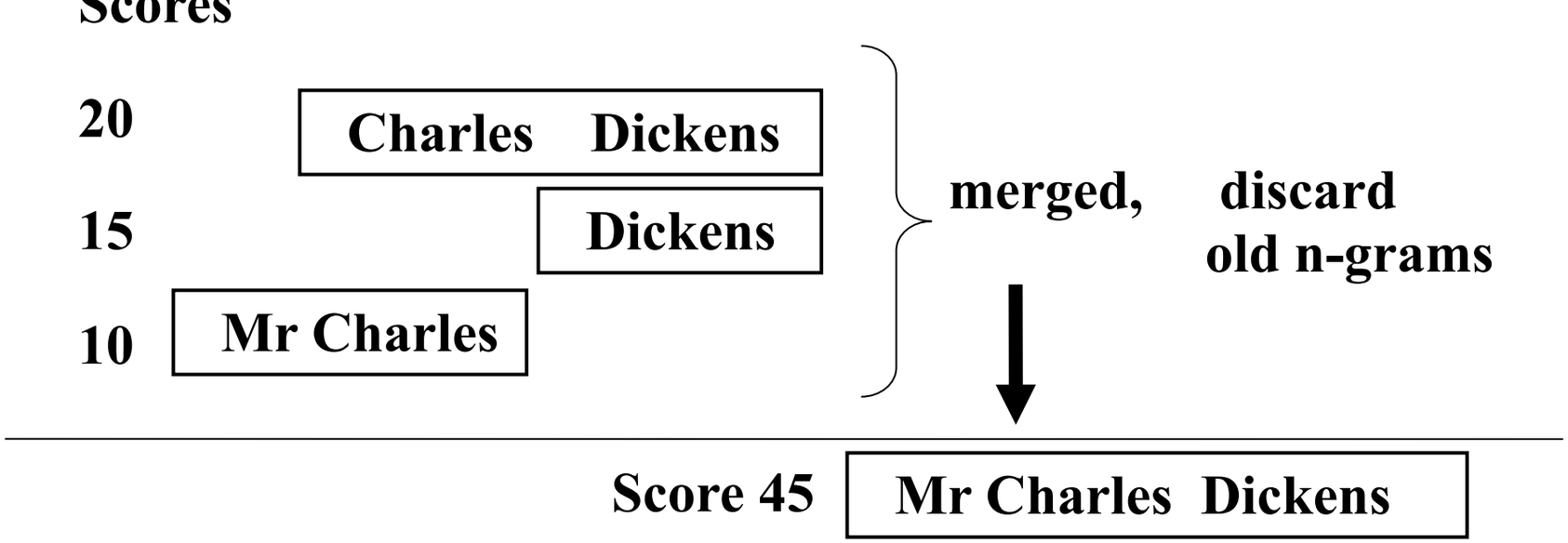


Сбор N-грамм

- Назначить вес N-грамме равный количеству сниппетов, в которых она встретилась
- Пример: “Who created the character of Scrooge?”
  - Dickens 117
  - Christmas Carol 78
  - Charles Dickens 75
  - Disney 72
  - Carl Banks 54
  - A Christmas 41
  - Christmas Carol 45
  - Uncle 31

Фильтрация и сбор ответа

- Заново взвесить N-граммы с учетом типа ответа
- Собрать ответ



Автоматическое реферирование

- Часто ответом на вопрос должен быть текст
- Пример:
  - Кто такой Франсуа Томбалбай?
- Извлечение короткого фрагмента текста является задачей автоматического реферирования

Аннотирование и реферирование

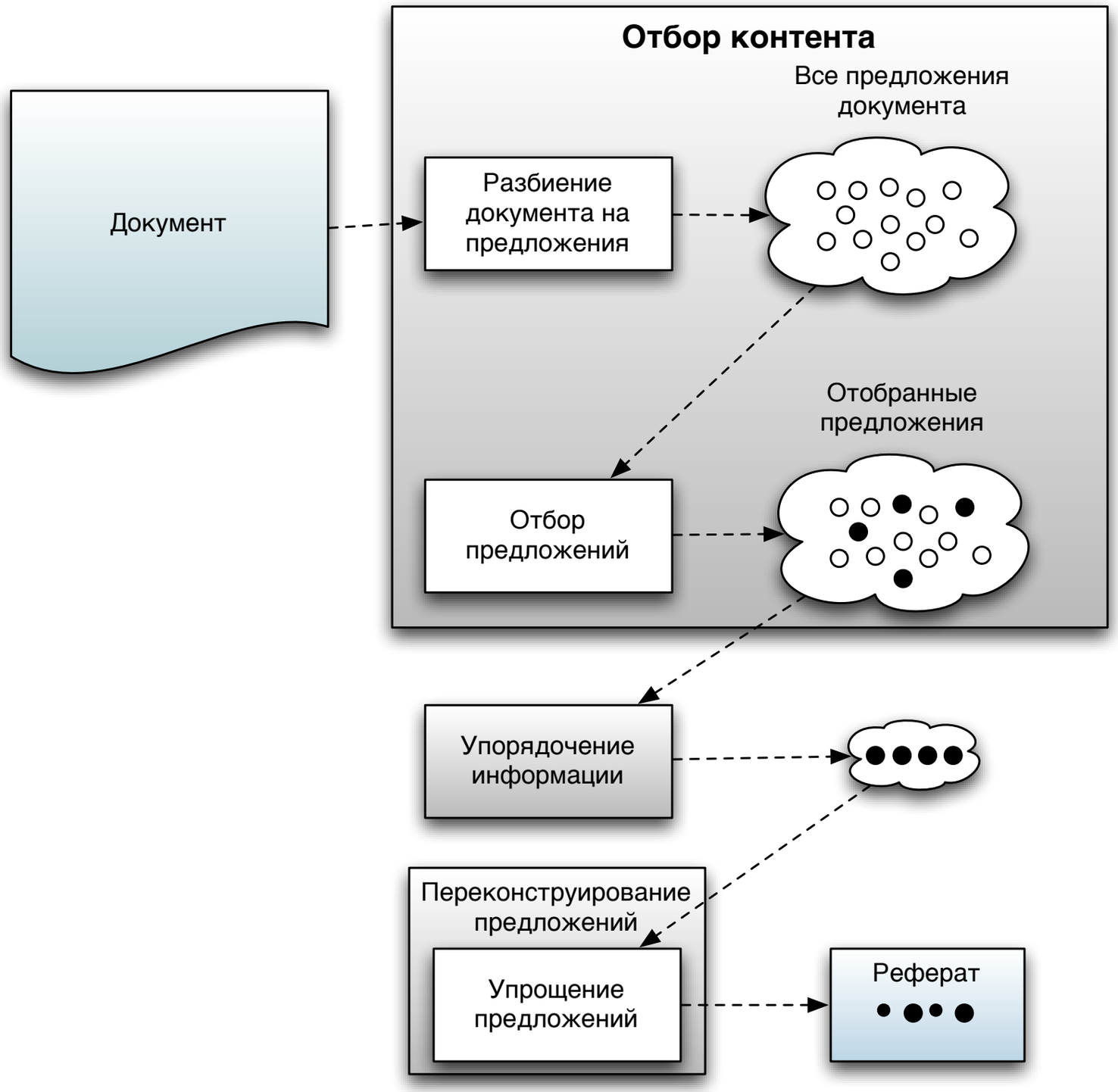
- **Реферат** состоит из частей оригинального текста
- **Аннотация** - главная мысль документа, сформулированная своими словами
  - Более компактная
  - Предполагает генерацию текста

Автоматическое реферирование

Приложения

- Аннотации и рефераты к научным и другим статьям
- Реферированное новостей (несколько документов)
- Создание сниппетов
- Текст для мобильных устройств
- Реферат встречи
- ...

Типичная архитектура



Отбор контента

- **Без учителя**
  - выбор предложений с ключевыми словами (tf-idf, логарифмическое отношение правдоподобия, ...)
  - Центральность
    - пример $centrality(x) = \frac{1}{K} \sum_{y} \text{tf-idf-cos}(x, y)$
- **С учителем**
  - бинарная классификация предложений
  - признаки: позиция, обобщающие фразы (“in summary”, “in conclusion”,...), информативность слов, длина предложения, связность

Упорядочение

- **Для одного документа**
  - –Использовать порядок внутри документа
- Для коллекции документов
  - – более сложные методы
    - кластеризация предложений

Переконструирование предложения

- Упрощение предложений
  - ~~When it arrives sometime new year in new TV sets~~, the V-chip will give parents a ~~new and potentially revolutionary~~ device to block out programs they don't want their children to see.
- Использование синтаксического разбора и удаление неинформативных частей
  - Zajic et al. 2007, Conroy et al. 2006

Заключение

- Информационный поиск
  - Обработка запроса и документа
  - Извлечение документов
  - Оценка систем
- Вопросно-ответные системы
  - Обработка запроса
  - Извлечение фрагментов текста
  - Обработка ответа
- Системы автоматического реферирования
  - Отбор контента
  - Упорядочение информации
  - Переконструирование предложений

Следующая лекция

- Машинный перевод