

Основы обработки текстов

Лекция 9

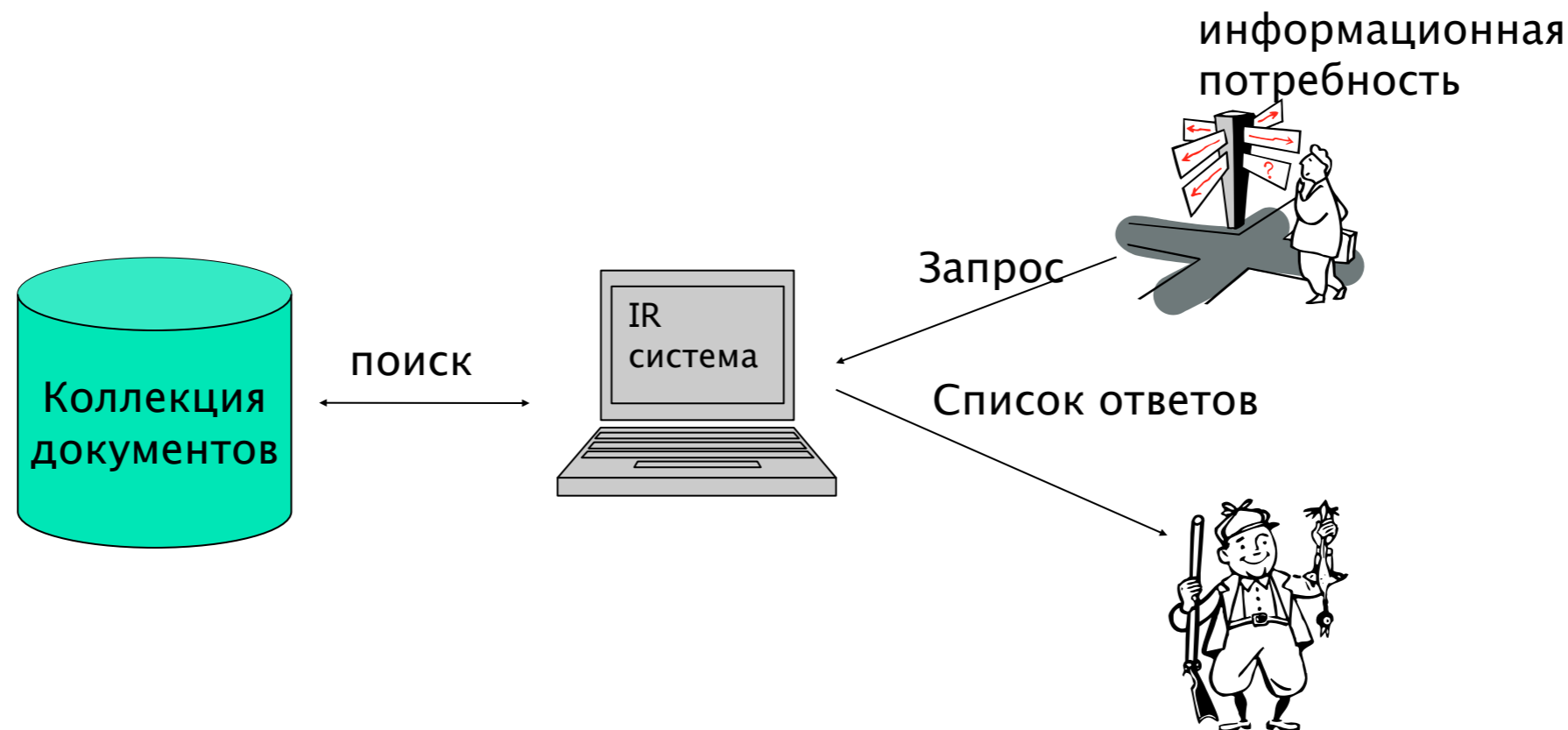
Приложения обработки текстов

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 атрибутам - поиск в БД
- IR: поиск по контенту

Возможные подходы

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

Примеры систем

Яндекс — 322 млн ответов



Поиск



Картинки



Видео



Карты

W [Information retrieval - Wikipedia, the free encyclopedia](#)

en.wikipedia.org > [Information retrieval](#)

Information retrieval is the activity of obtaining **information** resources relevant to an **information** need from a collection of **information** resources. **Searches** can be based on metadata or on full-text (or other content-based) indexing.

[Overview](#) [History](#) [Model types](#) [Awards in the field](#)

W [Информационный поиск — Википедия](#)

ru.wikipedia.org > [Информационный поиск](#)

Информационный поиск (англ. **Information retrieval**) – документальной информации, удовлетворяющей инфэ этом поиске.

MLP [Introduction to Information Retrieval](#)

nlp.stanford.edu > [IR-book/](#)

Google



Scholar

About 3,120,000 results (0.10 sec)



[About](#) [Images](#) [Videos](#) [Products](#) [Definition](#)

Information retrieval

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing.

W [More at Wikipedia](#)

Related Topics

[Information retrieval Category](#)

[Adversarial information retrieval - Adversarial informato...](#)

[Collaborative information seeking - Collaborative inform...](#)

Region

Introduction to Information Retrieval - Stanford University

Introduction to **Information Retrieval**. This is the companion website for the following book, Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, **Introduction to Information Retrieval**, Cambridge University Press, 2008.

nlp.stanford.edu

Information retrieval - Wikipedia, the free encyclopedia

Information retrieval is the activity of obtaining **information** resources relevant to an **information** need from a collection of **information** resources.

W en.wikipedia.org

Information Retrieval definition of Information Retrieval in ...

information retrieval [in-fər-'nā-shən rī-'brē-val] (computer science) The technique and process of searching, recovering, and interpreting **information** from large amounts of stored data.

encyclopedia2.thefreedictionary.com

Information retrieval - Definition and More from the Free ...

Definition of **INFORMATION RETRIEVAL**: the techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system.

merriam-webster.com

Information retrieval: data structures and algorithms

[WB Frakes, R Baeza-Yates - 1992 - citeulike.org](#)

Abstract **Information retrieval** is a sub-field of computer science that deals with the automated storage and **retrieval** of documents. Providing the latest **information retrieval** techniques, this guide discusses **Information Retrieval** data structures and algorithms, ...

Cited by 2442 [Related articles](#) [All 4 versions](#) [Cite](#) [Save](#) [More](#)

[CITATION] **Introduction to modern information retrieval**

[G Salton, MJ McGill - 1983 - agris.fao.org](#)

... rdf logo rdf logo. Translate with Translator. This translation tool is powered by Google. AGRIS and FAO are not responsible for the accuracy of translations. fao, ciard, aims, AGRIS: International **Information** System for the Agricultural science and technology, aginfra.

Cited by 11910 [Related articles](#) [All 7 versions](#) [Cite](#) [Save](#) [More](#)

[BOOK] **Introduction to information retrieval**

[CD Manning, P Raghavan, H Schütze - 2008 - langtoninfo.co.uk](#)

Introduction to **Information Retrieval** is the first textbook with a coherent treatment of classical and web **information retrieval**, including web search and the related areas of text classification and text clustering. Written from a computer science perspective, it gives an ...

Cited by 6875 [Related articles](#) [All 11 versions](#) [Cite](#) [Save](#) [More](#)

Term-weighting approaches in automatic text **retrieval**

[G Salton, C Buckley - Information processing & management, 1988 - Elsevier](#)

Abstract The experimental evidence accumulated over the past 20 years indicates that text indexing systems based on the assignment of appropriately weighted single terms produce **retrieval** results that are superior to those obtainable with other more elaborate text ...

Cited by 6901 [Related articles](#) [All 23 versions](#) [Cite](#) [Save](#)

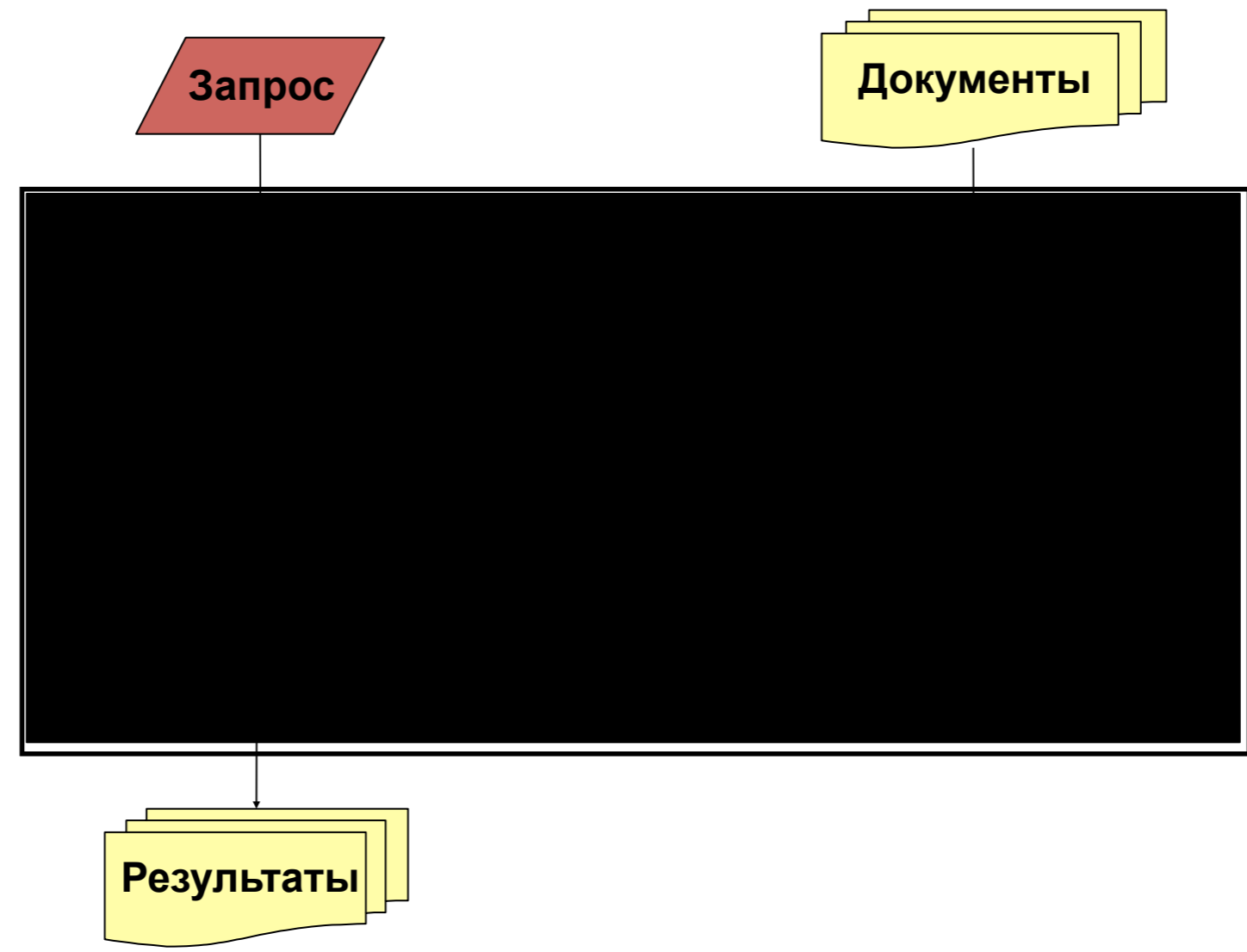
[BOOK] **Modern information retrieval**

[R Baeza-Yates, B Ribeiro-Neto - 1999 - mail.im.tku.edu.tw](#)

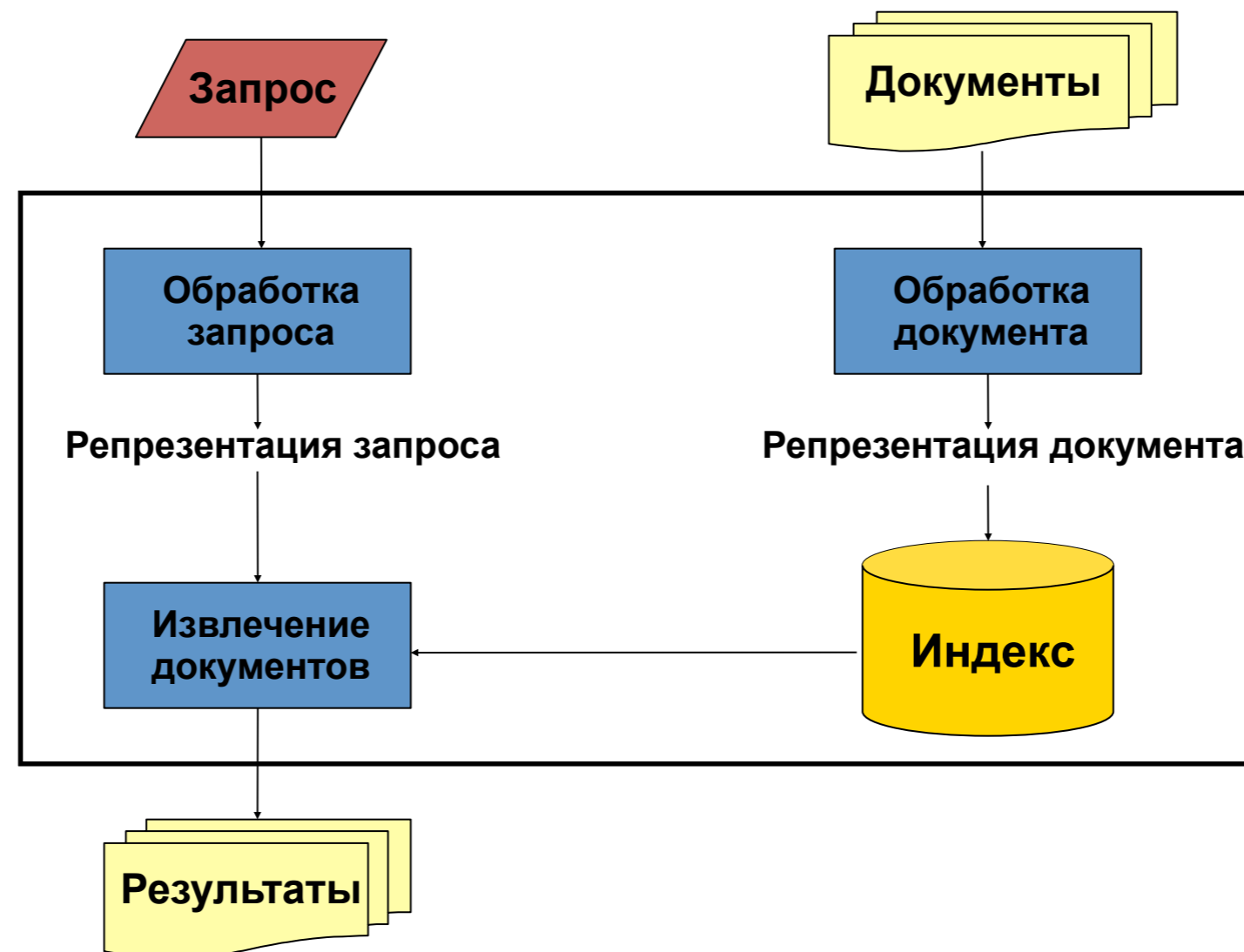
Information retrieval (IR) has changed considerably in recent years with the expansion of the World Wide Web and the advent of modern and inexpensive graphical user interfaces and mass storage devices. As a result, traditional IR textbooks have become quite out of date ...

Cited by 12814 [Related articles](#) [All 49 versions](#) [Cite](#) [Save](#) [More](#)

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

- $tf(t, D) = \text{freq}(t, D)$
 - $tf(t, D) = \log[\text{freq}(t, D)]$
 - $tf(t, D) = \log[\text{freq}(t, D)] + 1$
 - $tf(t, D) = \text{freq}(t, d) / \text{Max}[f(t, d)]$
- $idf(t) = \log(N/n)$
 $n = \#$ документов содержит t
 $N = \#$ документов в корпусе

Стоп-слова

- Функциональные слова не несут полезной информации 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

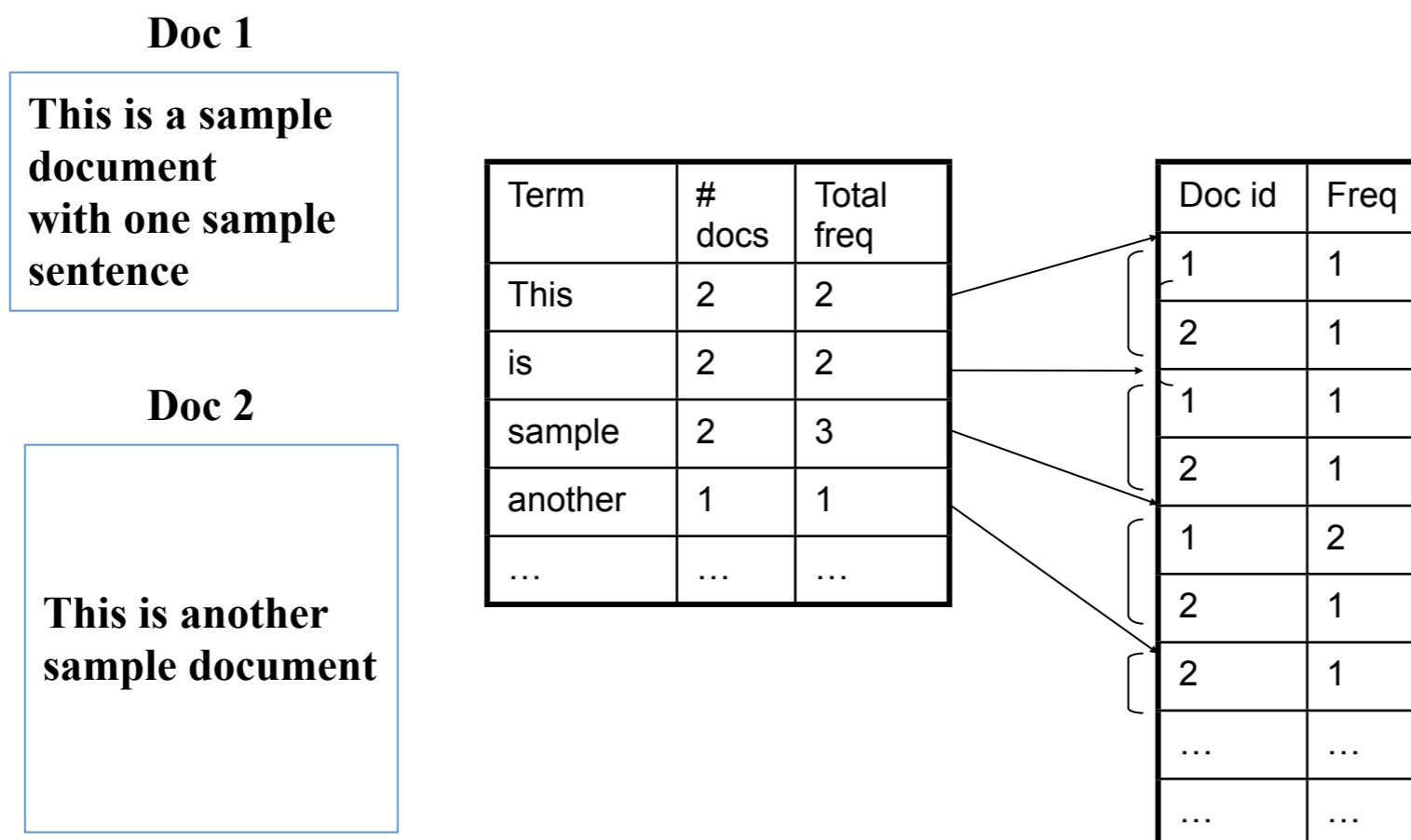
} поиск
- Porter stemmer
 - `nltk.stem.porter.PorterStemmer`

Лемматизация

- Стеemming 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



Извлечение документов

- Запрос из одного слова
 - Берем инвертированный 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 конъюнкция слов
- Запрос - Булево выражение
- $R(D, Q) = D \rightarrow Q$

$$D = t_1 \wedge t_2 \wedge \dots \wedge t_n$$

$$Q = (t_1 \wedge t_2) \vee (t_3 \wedge \neg t_4)$$

$$D \rightarrow Q, \text{ то есть } R(D, Q) = 1.$$

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

Векторная модель

- Векторное пространство всех слов

$$\langle t_1, t_2, t_3, \dots, t_n \rangle$$

- Документ

$$D = \langle a_1, a_2, a_3, \dots, a_n \rangle$$

$$a_i = \text{вес } t_i \text{ в } D$$

- Запрос

$$Q = \langle b_1, b_2, b_3, \dots, b_n \rangle$$

$$b_i = \text{вес } t_i \text{ в } Q$$

Матричное представление

Пространство
документов

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

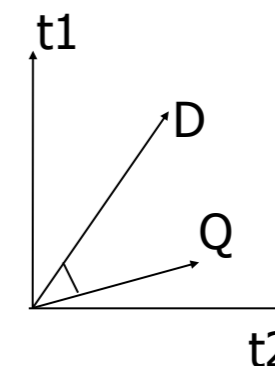
Пространство
терминов

Разреженная матрица!

Подсчет близости

Скалярное
произведение

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



Косинус

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

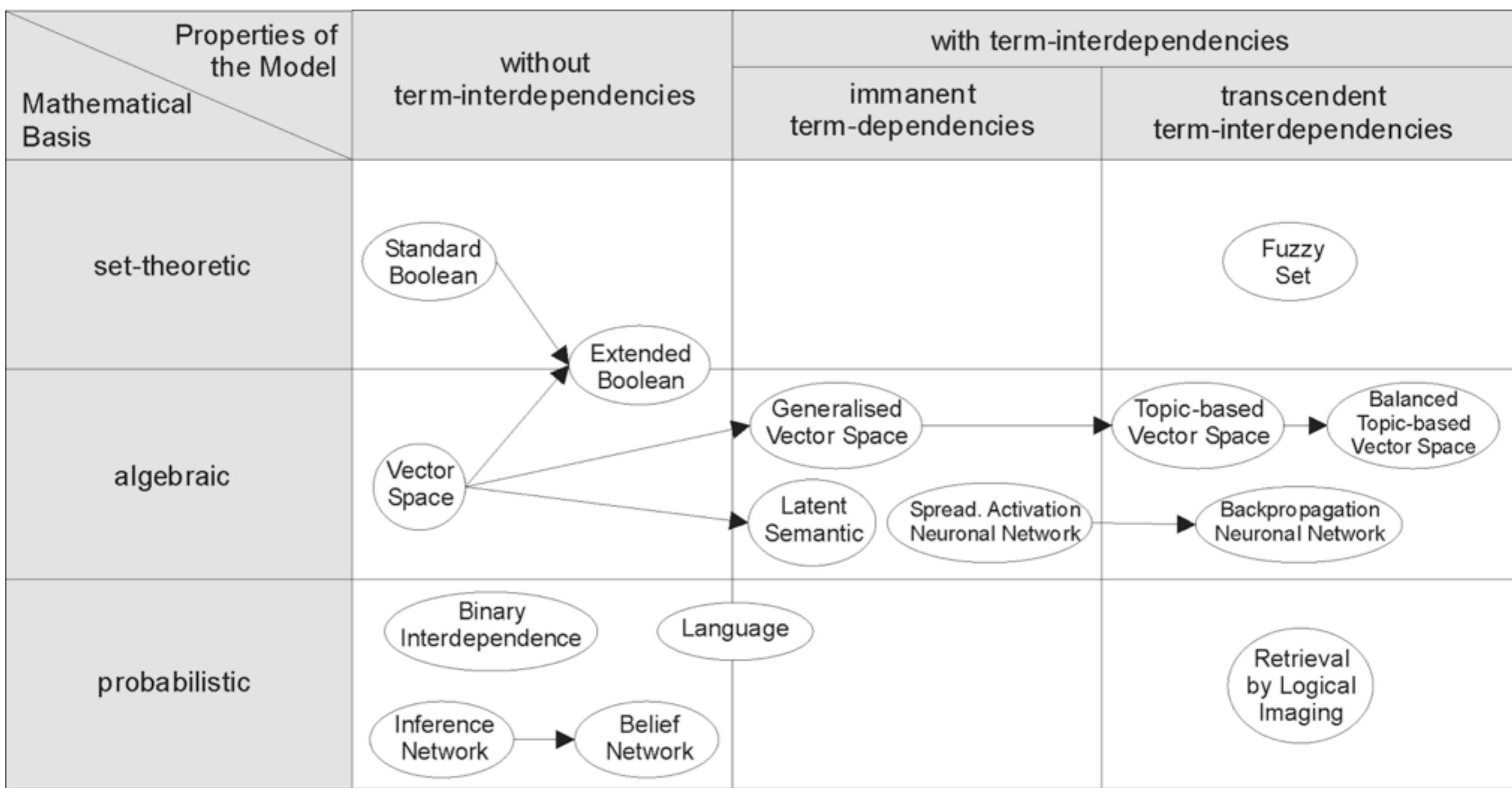
Мера Дайса

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

Мера Жаккара

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

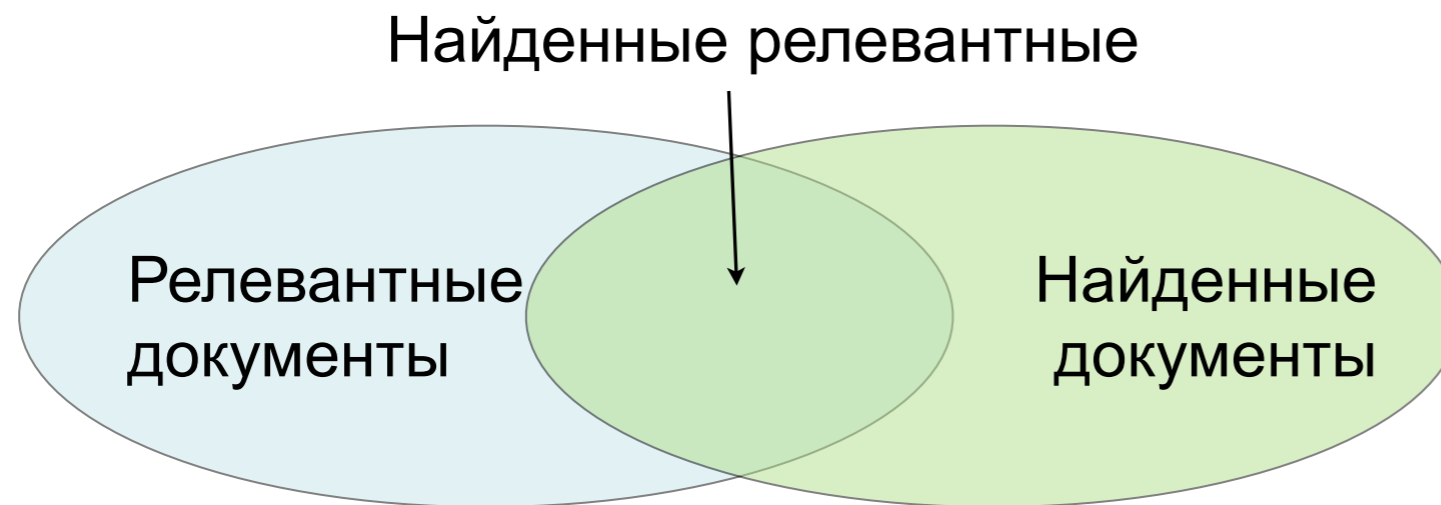
Какие еще бывают модели



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

Оценка систем

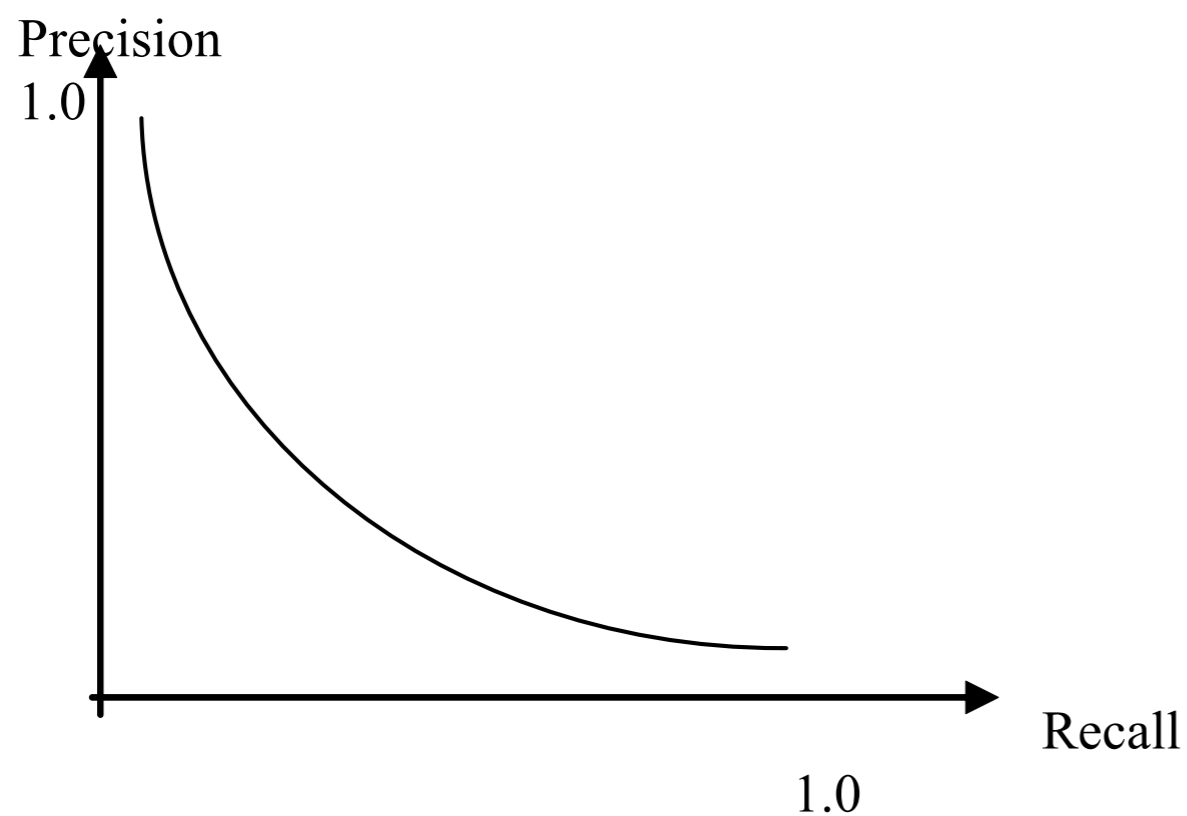
- Точность и полнота



- 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)



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

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

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

$\text{rel}(k) \in \{0, 1\} = 1$, 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	1 ^{ый} рел. док.
	5	8	2 ^{ой} рел. док.
	10		3 ^{ий} 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)

AT&T 7:36 AM

“Today do I need an umbrella Ella Ella a a a a”

Yes, it's likely to rain today:

57° H: 57° L: 36°

8:00 AM	70%	57°
9:00 AM	70%	57°
10:00 AM	80%	55°



Типы вопросов

О фактах

Какая обычная высота жирафа?
Где расположен главный офис Google ?

Списки

Какие страны экспортируют нефть?
Какие названия имеют штаты США?

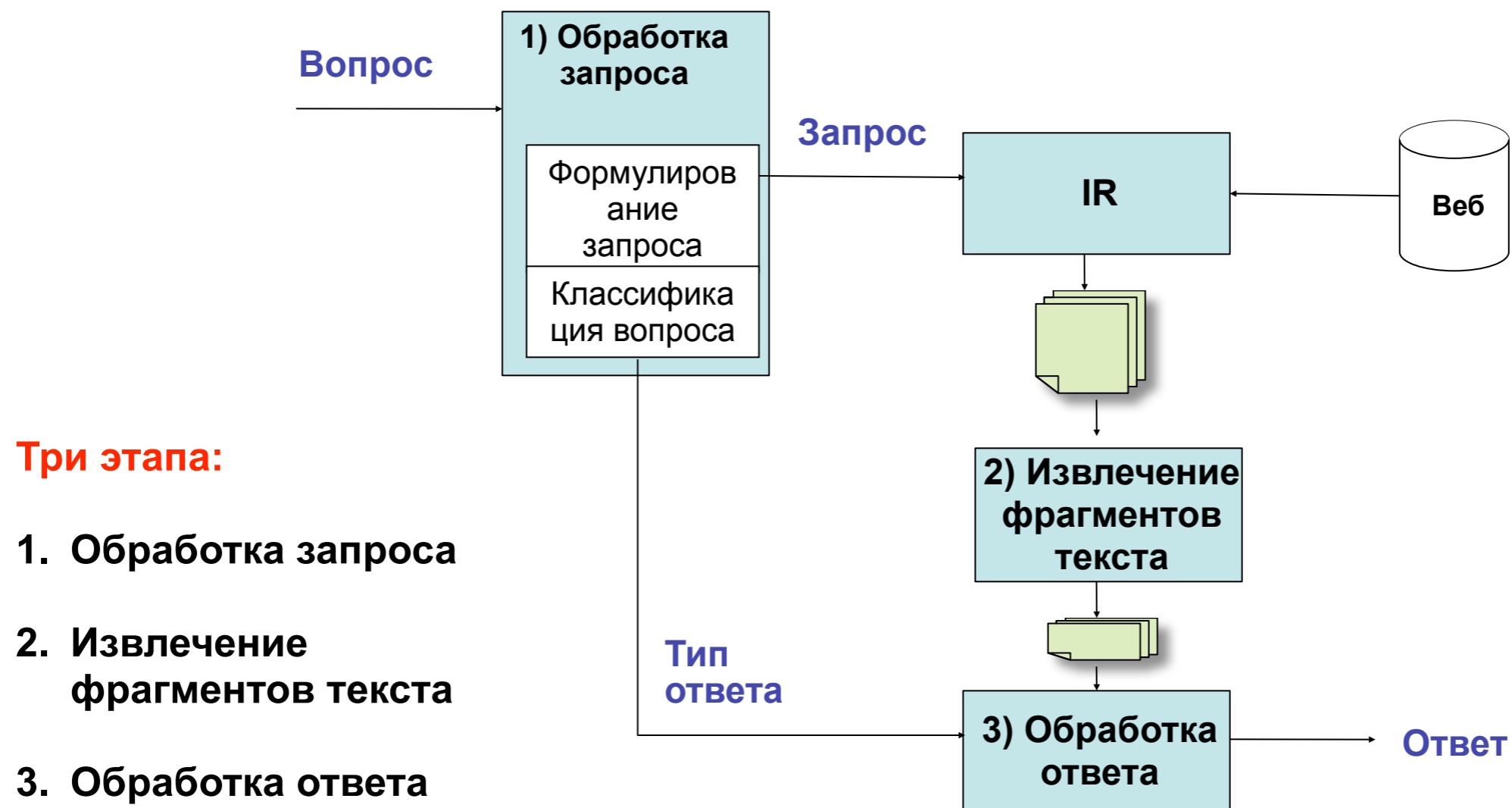
Определения

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



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: <i>What was the monetary value of the Nobel Peace Prize in 1989?</i>	monetary, value, Nobel, Peace, Prize, 1989
Q003: <i>What does the Peugeot company manufacture?</i>	Peugeot, company, manufacture
Q004: <i>How much did Mercury spend on advertising in 1993?</i>	Mercury, spend, advertising, 1993
Q005: <i>What is the name of the managing director of Apricot Computer?</i>	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

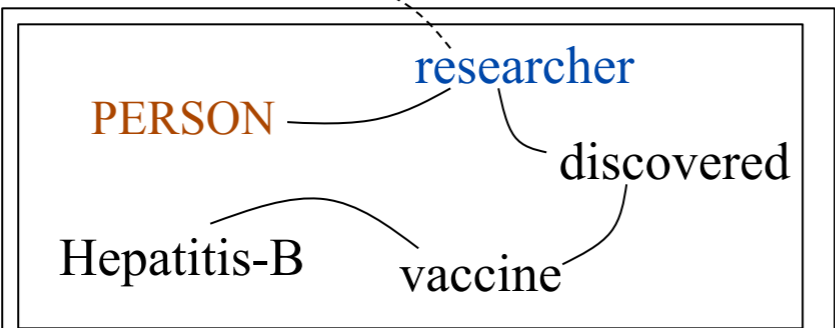
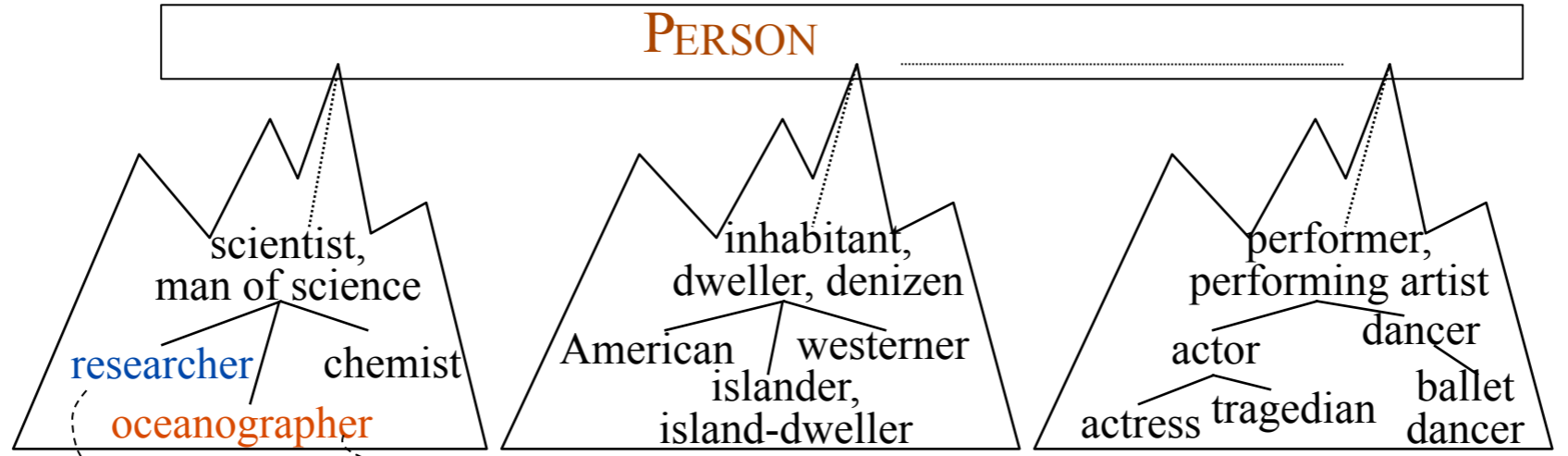
Вопрос	Основа вопроса	Тип ответа
Q555: <i>What was the name of Titanic's captain?</i>	What	Person
Q654: <i>What U.S. Government agency registers trademarks?</i>	What	Organization
Q162: <i>What is the capital of Kosovo?</i>	What	City
Q661: <i>How much does one ton of cement cost?</i>	How much	Quantity

Определение типа ответа

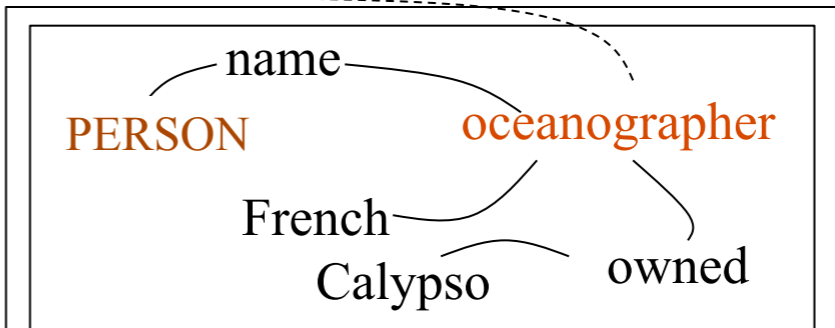
- В некоторых случаях тип ответа можно определить по вопросу
 - Почему → Причина
 - Когда → Дата
- Для многозначных вопросов использовать дополнительные понятия в вопросе
 - *What* was the name of Titanic's *captain*?
 - *What* U.S. Government *agency* registers trademarks?
 - *What* is the *capital*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

что такое сниппет?

в найденном в Москве расширенный поиск

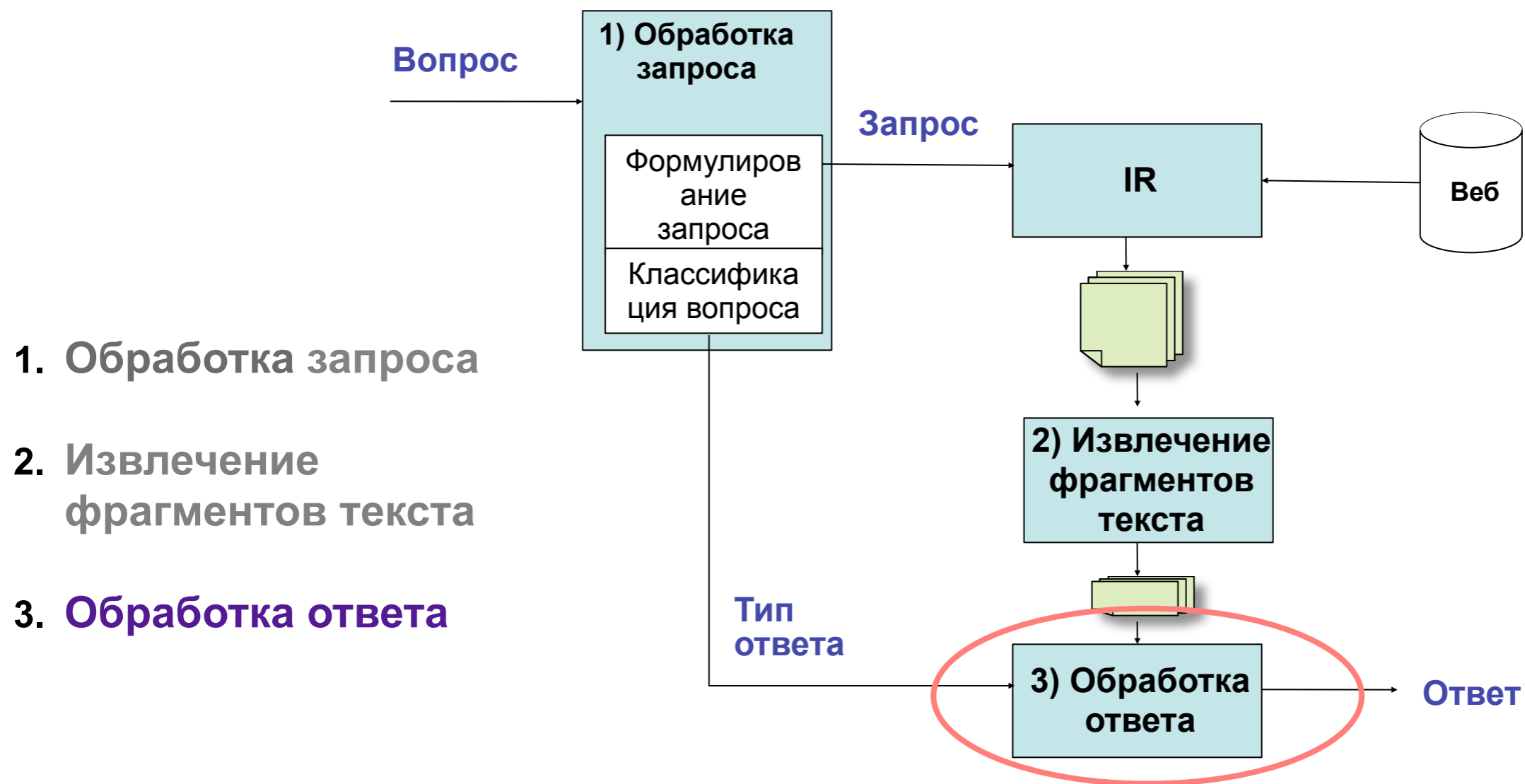
[Описание сайта - Что такое сниппет?](#)

Что представляют из себя навигационные цепочки? Для каких страниц в **сниппетах** показываются даты? Какие специальные данные могут быть показаны в **сниппетах**? **Что такое сниппет?**
[help.yandex.ru](#) > [Помощь](#) > [Вебмастер](#) [копия](#) [ещё](#)

[Что такое сниппет и как его использовать](#)

Сниппет (англ. **snippet** - лоскут, отрывок или фрагмент) - это та короткая текстовая информация по сайту, которая появляется в результатах поиска, сразу же под вылезшим адресом.
[bigfozzy.com](#) > [Articles/Promotion...snippet.php](#) [копия](#) [ещё](#)

Типичная архитектура QA-систем



1. Обработка запроса

2. Извлечение фрагментов текста

3. Обработка ответа

Обработка ответа

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

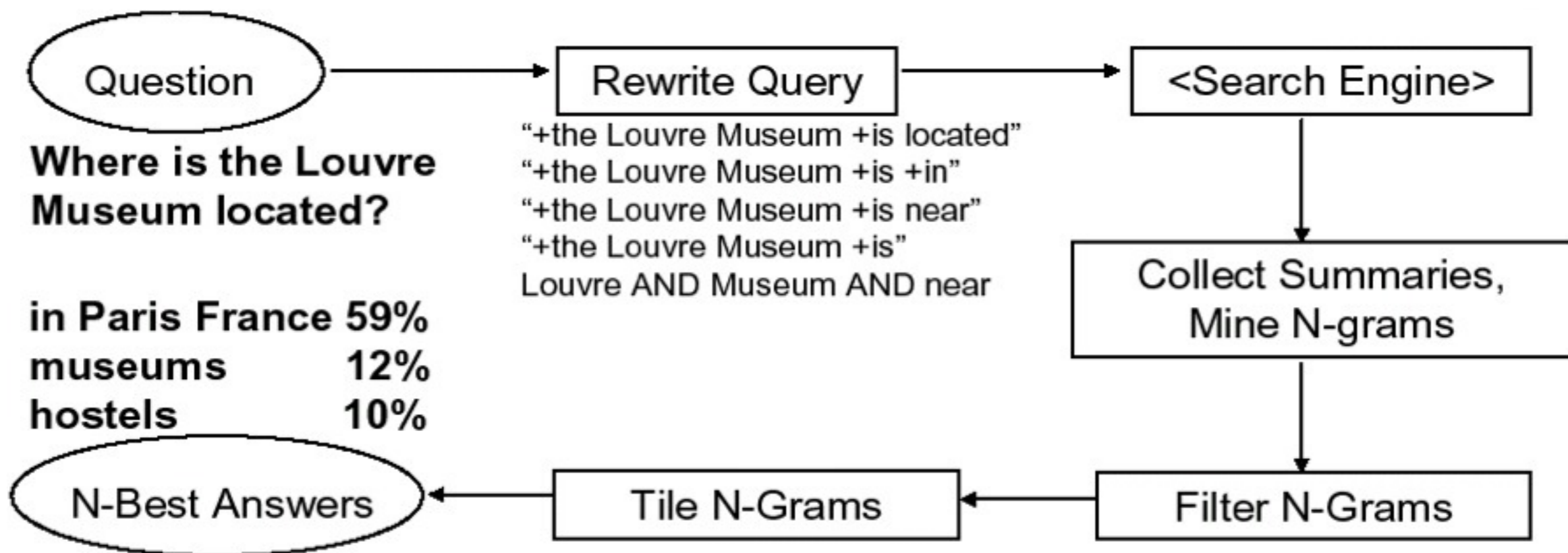
Алгоритмы на основе шаблонов

- 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>	<i>What is autism?</i>	<i>developmental disorders such as autism</i>

Сбор ответа из N-грамм

Архитектура AskMSR

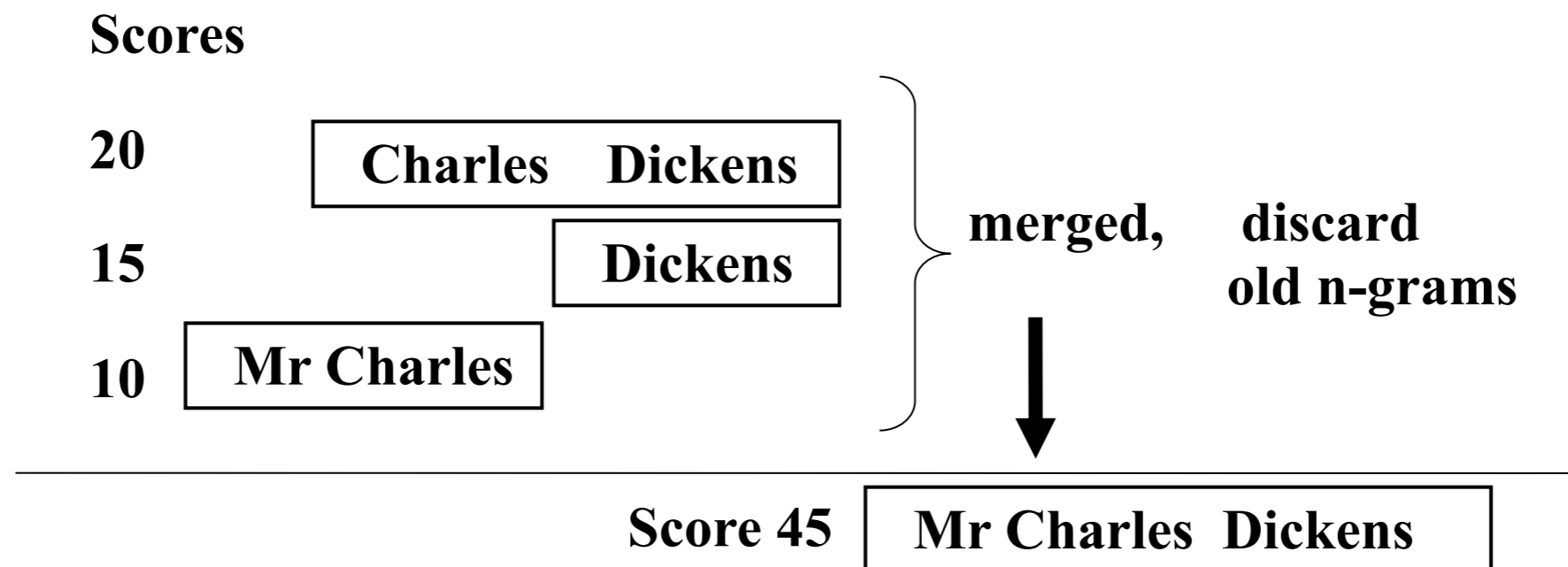


Сбор N-грамм

- Назначить вес N-грамме равный количеству снippetов, 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

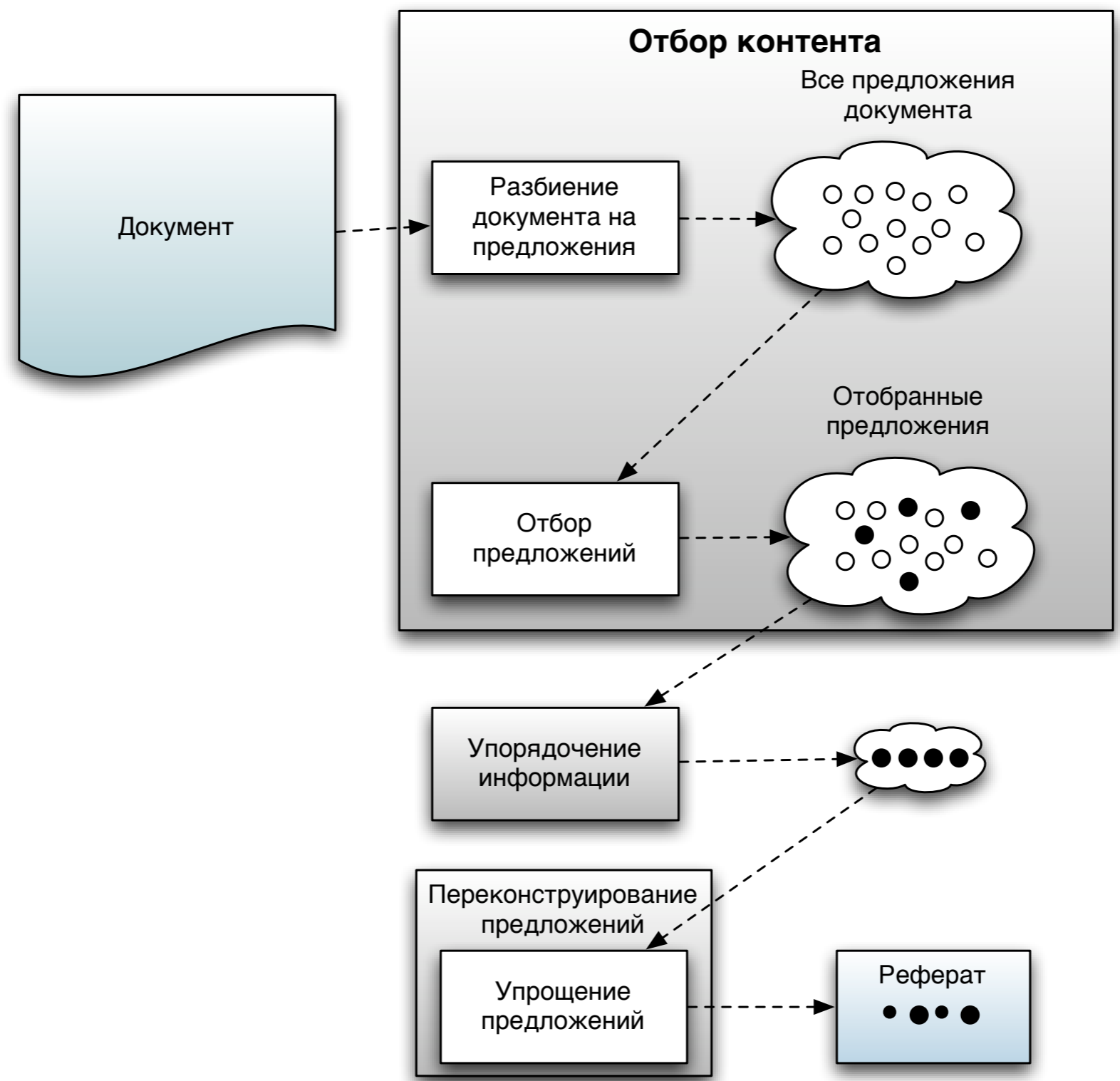
- **Реферат** состоит из частей оригинального текста
- **Аннотация** - главная мысль документа, сформулированная своими словами
 - Более компактная
 - Предполагает генерацию текста

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