

# Расширение запроса при поиске

Маннинг и др. Введение в  
информационный поиск, гл.9

# Методы расширения запроса













- Несовпадение слова запроса:
  - самолет – лайнер
- Методы расширения запроса:
  - Глобальные методы
    - Информационно-поисковый тезаурус
    - Автоматически порождаемый тезаурус
  - Локальные методы (по конкретному запросу)
    - Relevance feedback (обратная связь по релевантности)
    - Pseudo Relevance feedback (обратная связь по псевдорелевантности)

# Обратная связь по релевантности


- Пользователь оценивает документы в поисковой выдаче
  - Пользователь задает относительно простой, короткий запрос
  - Затем пользователь размечает часть результатов как релевантные и нерелевантные
  - Система вычисляет улучшает соответствие документов запросу на основе пользовательской разметки
  - Процедура может выполняться итеративно.
- Основанная идея: сформулировать хороший запрос трудно, если пользователь не знаком с коллекцией, поэтому – итеративное построение запроса

# Результаты для начального запроса













[Browse](#) [Search](#) [Prev](#) [Next](#) [Random](#)

					
(144473, 16458) 0.0 0.0 0.0	(144457, 252140) 0.0 0.0 0.0	(144456, 262857) 0.0 0.0 0.0	(144456, 262863) 0.0 0.0 0.0	(144457, 252134) 0.0 0.0 0.0	(144483, 265154) 0.0 0.0 0.0
					
(144483, 264644) 0.0 0.0 0.0	(144483, 265153) 0.0 0.0 0.0	(144518, 257752) 0.0 0.0 0.0	(144538, 525937) 0.0 0.0 0.0	(144456, 249611) 0.0 0.0 0.0	(144456, 250064) 0.0 0.0 0.0

# Разметка пользователя



[Browse](#) [Search](#) [Prev](#) [Next](#) [Random](#)

					
(144473, 16458)	(144457, 252140)	(144456, 262857)	(144456, 262863)	(144457, 252134)	(144483, 265154)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
					
(144483, 264644)	(144483, 265153)	(144518, 257752)	(144538, 525937)	(144456, 249611)	(144456, 250064)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0



# Результаты после разметки

[Browse](#)[Search](#)[Prev](#)[Next](#)[Random](#)

(144538, 523493)  
0.54182  
0.231944  
0.309876



(144538, 523835)  
0.56319296  
0.267304  
0.295889



(144538, 523529)  
0.584279  
0.280881  
0.303398



(144456, 253569)  
0.64501  
0.351395  
0.293615



(144456, 253568)  
0.650275  
0.411745  
0.23853



(144538, 523799)  
0.66709197  
0.358033  
0.309059



(144473, 16249)  
0.6721  
0.393922  
0.278178



(144456, 249634)  
0.675018  
0.4639  
0.211118



(144456, 253693)  
0.676901  
0.47645  
0.200451



(144473, 16328)  
0.700339  
0.309002  
0.391337



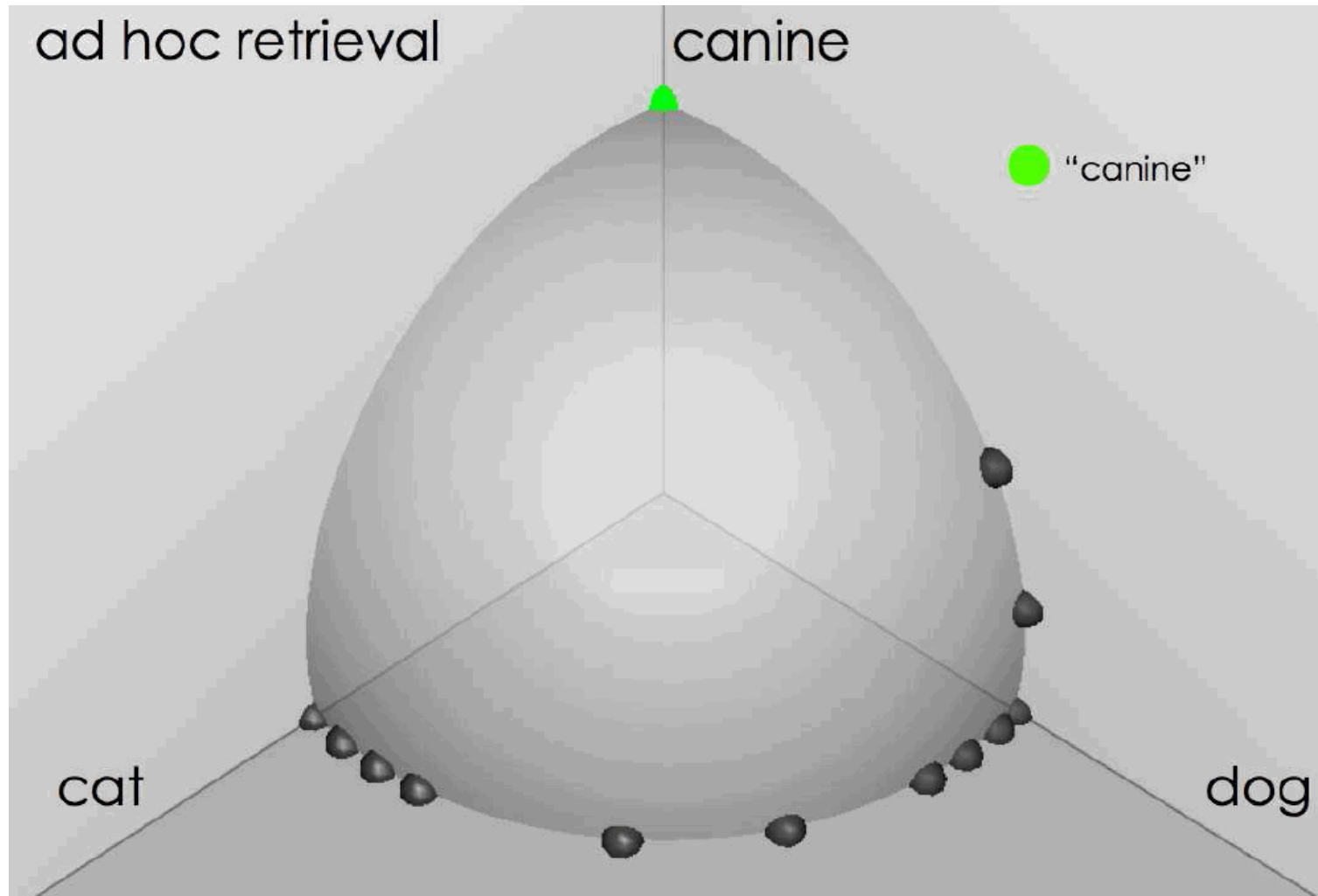
(144483, 265264)  
0.70170796  
0.36176  
0.339948



(144478, 512410)  
0.70297  
0.469111  
0.233859

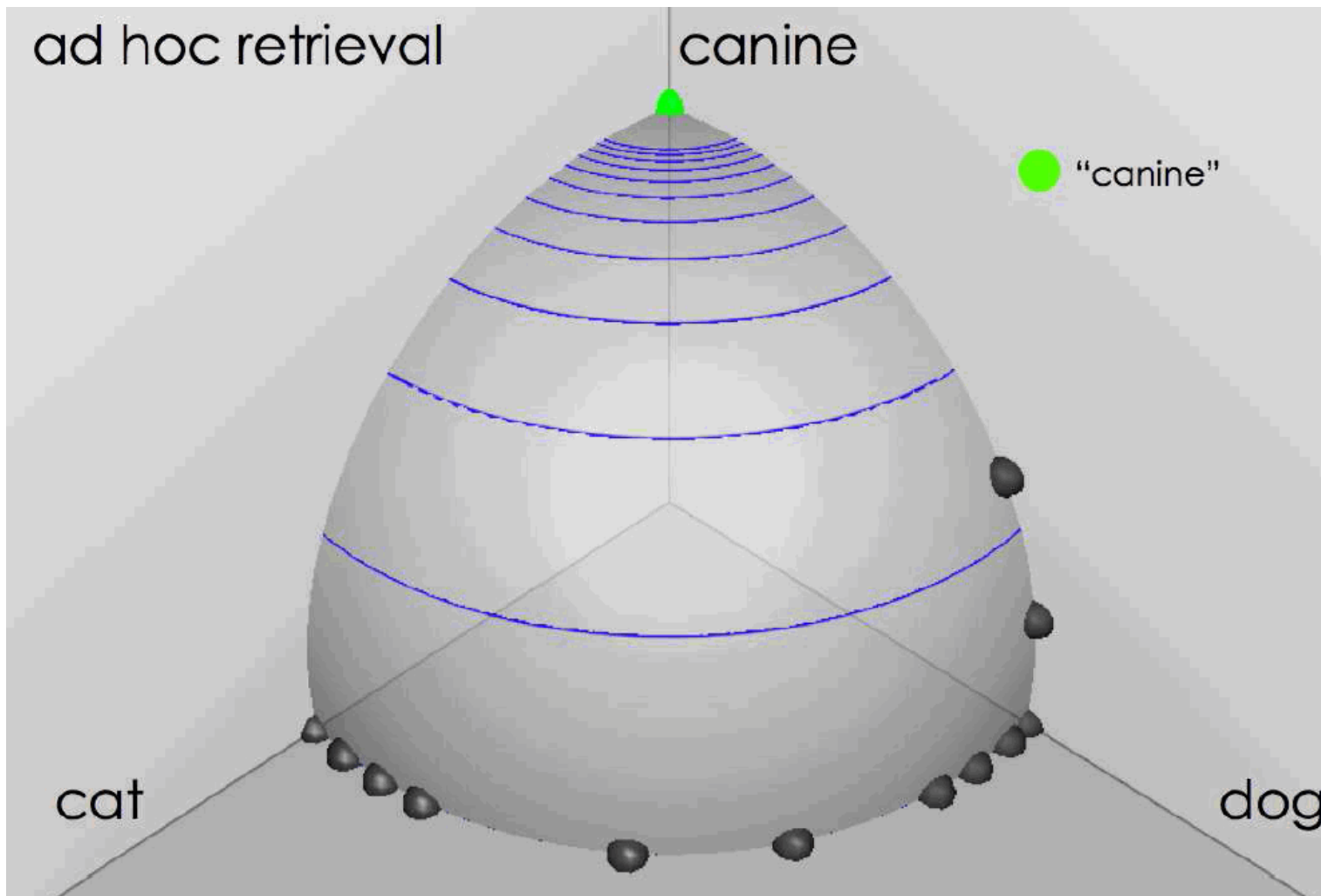
# Выдача по запросу *canine*

source: Fernando Diaz



# Выдача по запросу *canine-2*

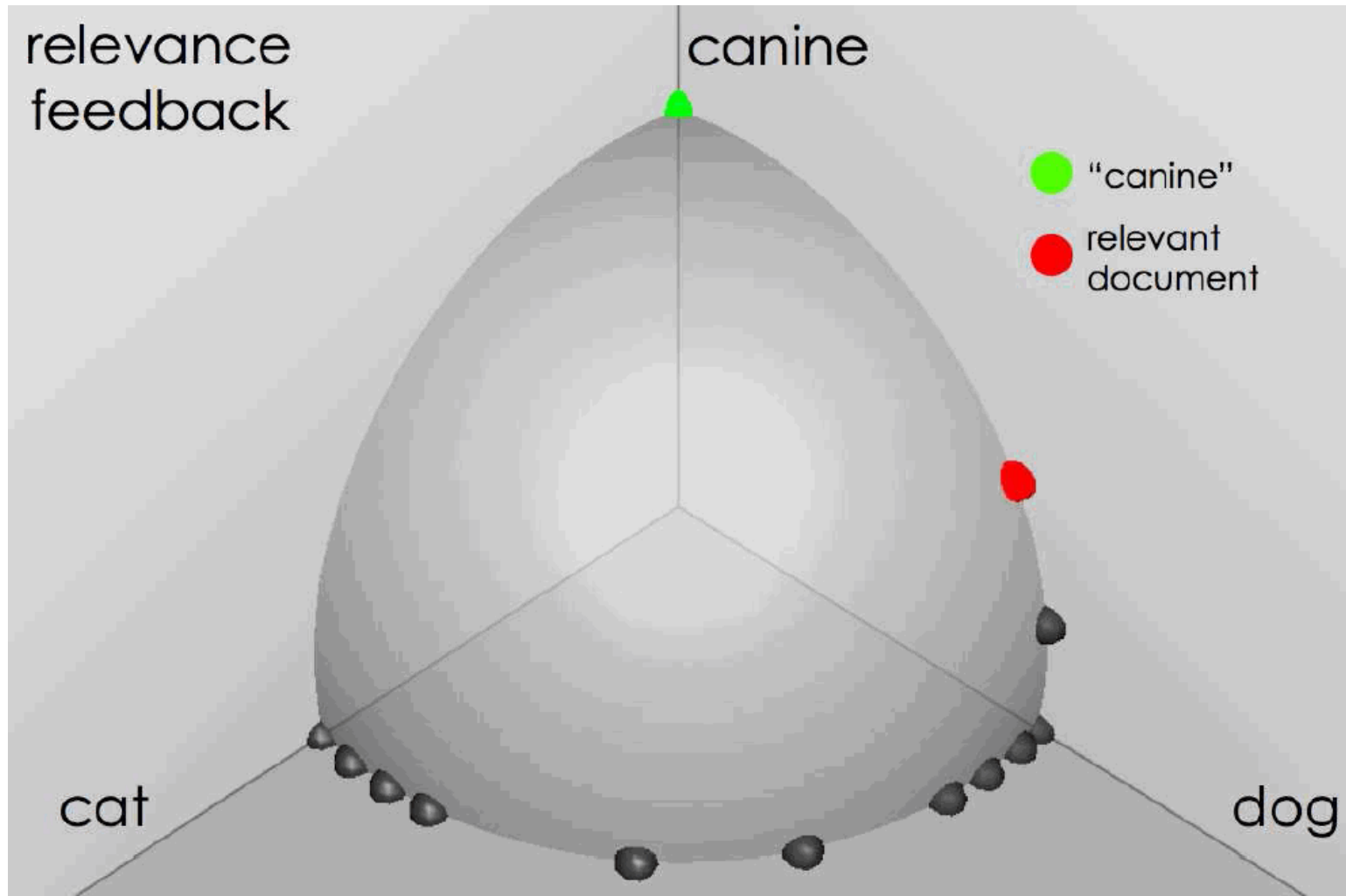
source: Fernando Diaz





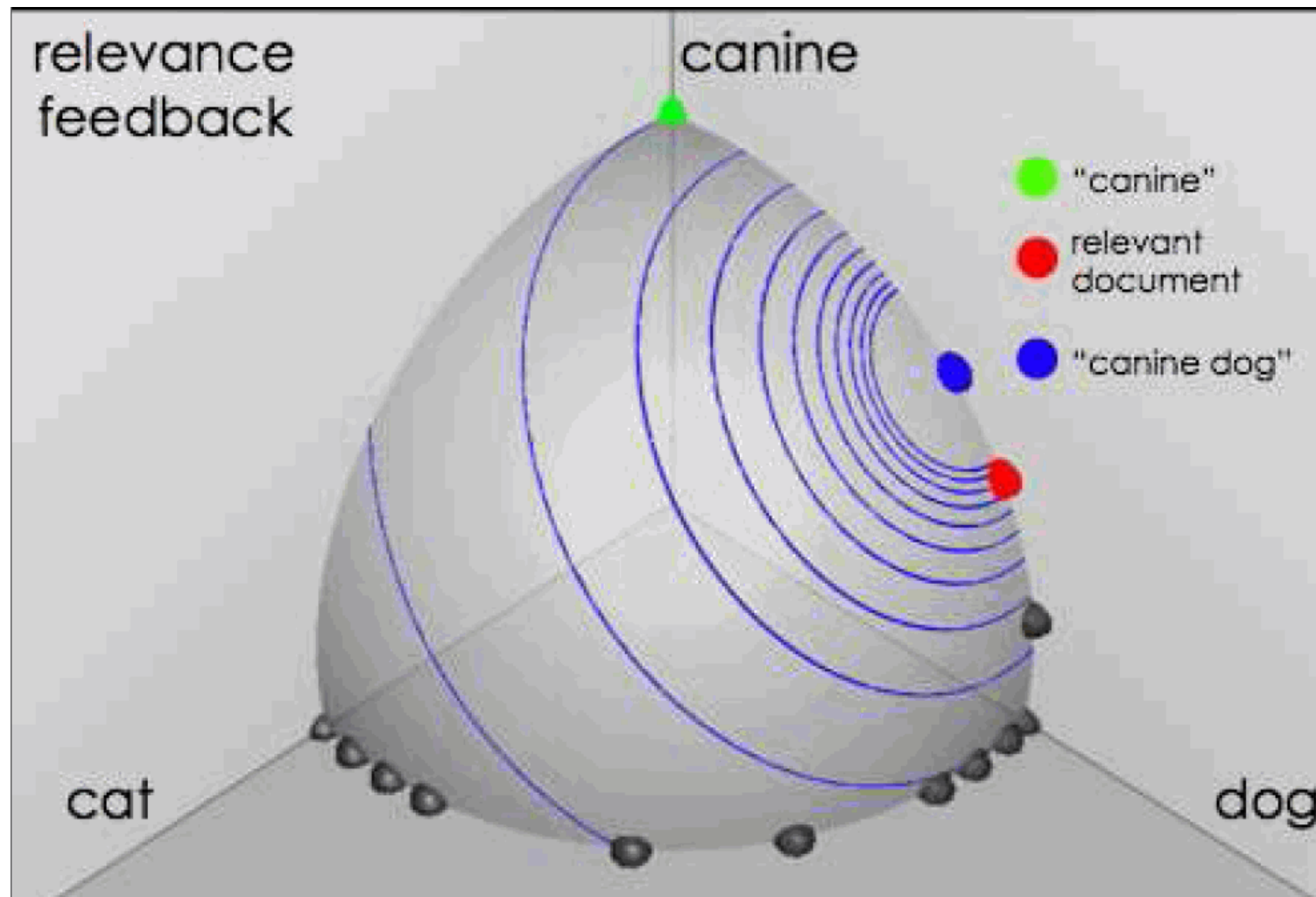
# Пользователь выбирает релевантное

source: Fernando Diaz



# Результаты (relevance feedback)

source: Fernando Diaz



# Начальный запрос и результаты

- Запрос: *New space satellite applications*
  1. 0.539, 08/13/91, [NASA Hasn't Scrapped Imaging Spectrometer](#)
  2. 0.533, 07/09/91, [NASA Scratches Environment Gear From Satellite Plan](#)
  3. 0.528, 04/04/90, [Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes](#)
  4. 0.526, 09/09/91, [A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget](#)
  5. 0.525, 07/24/90, [Scientist Who Exposed Global Warming Proposes Satellites for Climate Research](#)
  6. 0.524, 08/22/90, [Report Provides Support for the Critics Of Using Big Satellites to Study Climate](#)
  7. 0.516, 04/13/87, [Arianespace Receives Satellite Launch Pact From Telesat Canada](#)
  8. 0.509, 12/02/87, [Telecommunications Tale of Two Companies](#)
- Пользователь отмечает релевантные результаты отметкой “+”.

# Расширенные запрос после relevance feedback

- 2.074 new
- 30.816 satellite
- 5.991 nasa
- 4.196 launch
- 3.516 instrument
- 3.004 bundespost
- 2.790 rocket
- 2.003 broadcast
- 0.836 oil
- 15.106 space
- 5.660 application
- 5.196 eos
- 3.972 aster
- 3.446 arianespace
- 2.806 ss
- 2.053 scientist
- 1.172 earth
- 0.646 measure

# Ключевое понятие: центроид

- Центроид – это центр масс совокупности точек
- Документы – это точки в многомерном пространстве
- Определение: Центроид

$$\vec{\mu}(C) = \frac{1}{|C|} \sum_{d \in C} \vec{d}$$

где  $C$  – множество документов.



# Алгоритм Роккьо (Rocchio)

- Алгоритм Rocchio использует векторное пространства найти наилучший запрос на основе пользовательской разметки
- Rocchio ищет запрос  $q_{opt}$ , который максимизирует

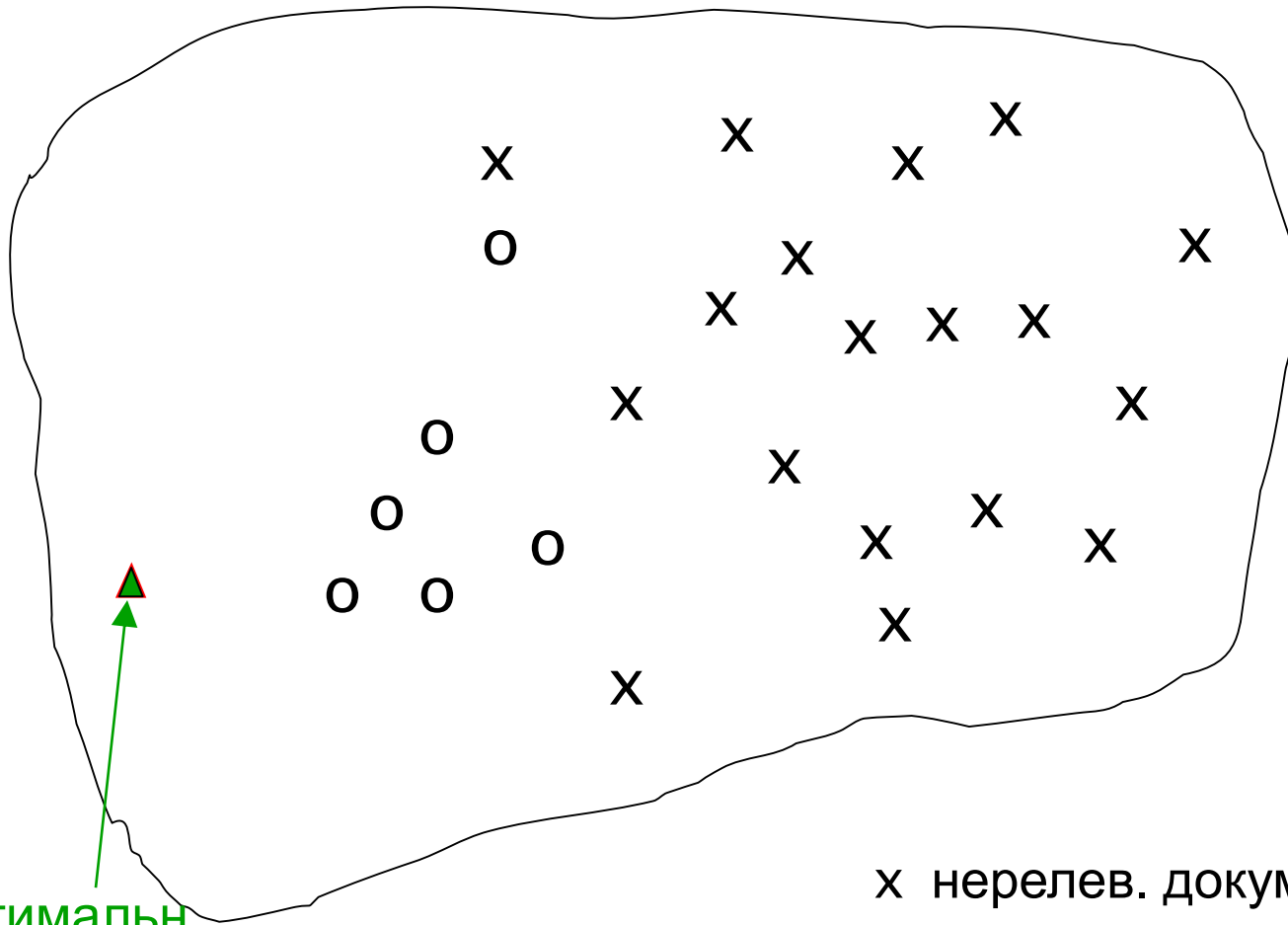
$$\vec{q}_{opt} = \arg \max_{\vec{q}} [\cos(\vec{q}, \vec{\mu}(C_r)) - \cos(\vec{q}, \vec{\mu}(C_{nr}))]$$

- Пытается разделить релевантные и нерелевантные документы

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\vec{d}_j \in C_r} \vec{d}_j - \frac{1}{|C_{nr}|} \sum_{\vec{d}_j \notin C_r} \vec{d}_j$$

- Проблема: мы не знаем все релевантные документы

# Лучший запрос



Оптимальн.  
запрос

x нерелев. документы  
o релевантные документы

# Rocchio 1971 алгоритм (SMART)

- На практике используется:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

- $D_r$  = множество известных релевантных doc векторов
- $D_{nr}$  = множество известных нерелевантных doc векторов
  - Отличны от  $C_r$  и  $C_{nr}$
- $q_m$  = модифицированный вектор запроса;  $q_0$  = исходный вектор запроса;  $\alpha, \beta, \gamma$ : веса
- Новый запрос «сдвигается» по направлению к релевантным документам и «уходит» от нерелевантных документов

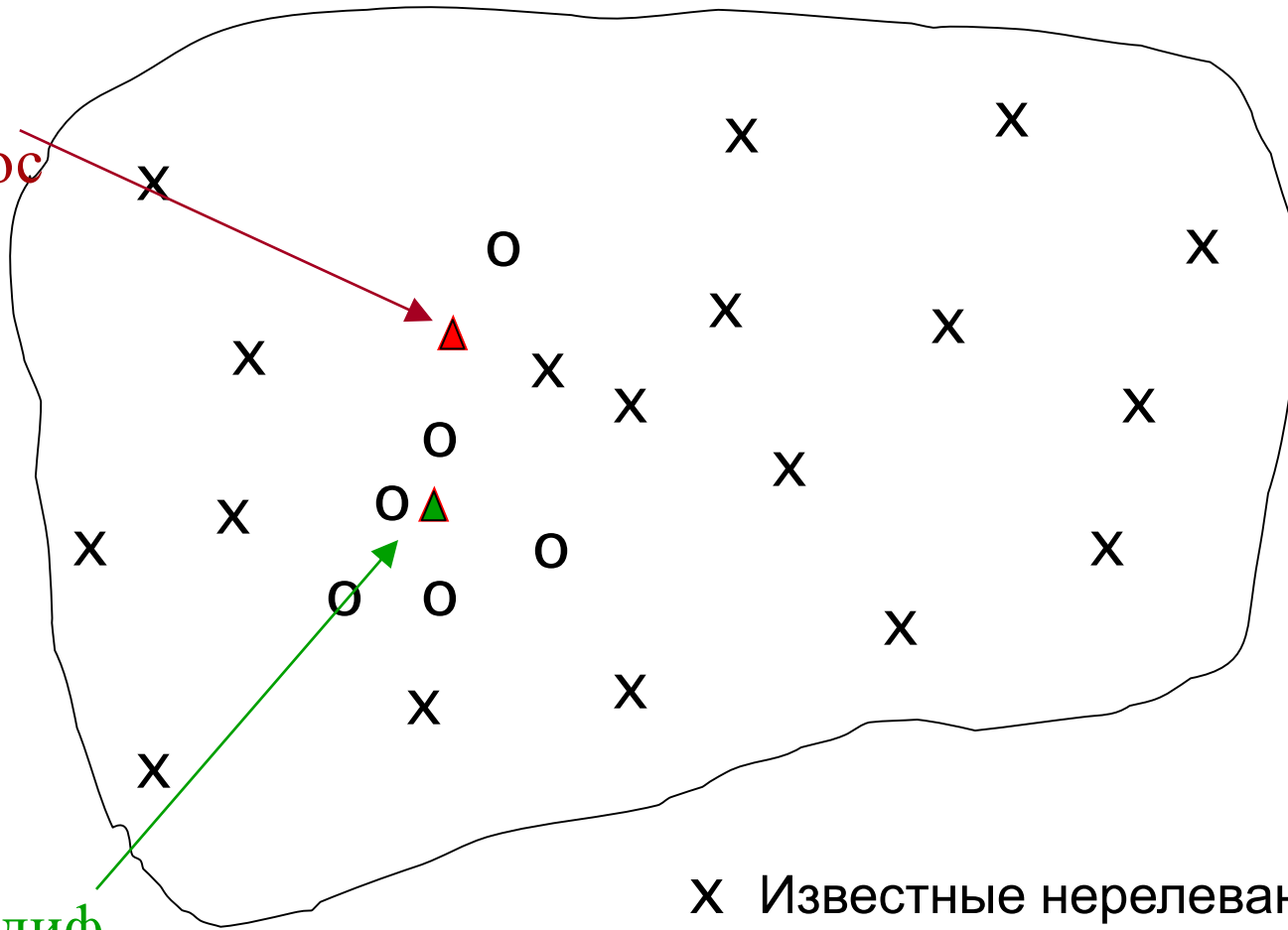


# Особенности параметров

- Соотношение  $\alpha$  vs.  $\beta/\gamma$  : Если у нас много оцененных документов, то лучше более высокие  $\beta/\gamma$ .
- Некоторые веса в модифицированном векторе запроса становятся отрицательными
  - Отрицательные веса слов игнорируются (устанавливаются равными 0)

# Relevance feedback по исходному запросу

Исх.  
запрос



Модиф.  
запрос

X Известные нерелевантн. док-ты  
O Известные релевантные док-ты



# Relevance Feedback

## в векторных пространствах

- Можно модифицировать запрос на основе разметки пользователя и применить стандартную векторную модель.
- Используются только документы, которые размечены.
- Relevance feedback может улучшить и полноту и точность
- Relevance feedback наиболее полезен в увеличении полноты в тех ситуациях, когда полнота важна
  - Пользователи должны просматривать и размечать результаты
    - Несколько итераций

# Позитивный vs Негативный Feedback

- Позитивный feedback более ценен, чем негативный feedback (обычно  $\gamma < \beta$ ; например,  $\gamma = 0.25$ ,  $\beta = 0.75$ ).
- Многие системы позволяют только позитивный feedback ( $\gamma=0$ ).

# Relevance Feedback: предположения

- A1: Пользователь имеет достаточно знаний для исходного запроса
- A2: Прототипы релевантных/нерелевантных документов “ведут себя хорошо”
  - Распределение слов в релевантных документах сходно
  - Распределение слов в нерелевантных документах отлично от распределения слов в релевантных документах
    - 1) Все релевантные документы похожи на один прототип
    - 2) Имеется несколько прототипов, но у них значительное пересечение по составу
    - Сходство между релевантными и нерелевантными документами относительно небольшое

# Нарушение A1

- У пользователя нет достаточного начального знания
- Примеры:
  - Неправильное написание: Brittany Speers.
  - Многоязыковой информационный поиск (hígado).
  - Несоответствие словаря пользователя и словаря коллекции
    - Cosmonaut/astronaut

# Нарушение A2

- Имеется несколько прототипов
- Примеры:
  - Сейчас: Украина – две точки зрения
  - Pop stars that worked at Burger King
- Часто: примеры более общего понятия

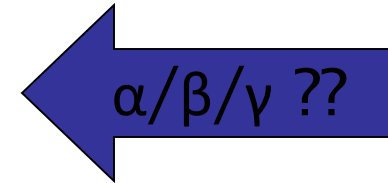


# Relevance Feedback: Проблемы

- Длинные запросы – неэффективны для типичной поисковой машины
  - Большое ожидание для пользователя
  - Высокая стоимость для поисковой системы
  - Частичное решение:
    - Использование только слов с наиболее высоким весом
      - Например, 20 первых по весу
- Пользователи часто не хотят размечать документы
- Трудно понять, почему данный документ был выдан после relevance feedback

# Relevance Feedback в вебе

- Некоторые поисковые машины предлагают возможность просмотра похожих страниц
  - Тривиальная форма relevance feedback
  - Google (link-based)
  - Altavista
  - Stanford WebBase
- Но результаты трудно объяснить среднему пользователю
- Excite
  - вводил настоящий relevance feedback,
  - затем убрал – никто не пользовался



# Pseudo relevance feedback

- Pseudo-relevance feedback автоматизирует «ручную» часть реального relevance feedback.
- Pseudo-relevance алгоритм:
  - Строит поисковую выдачу по запросу
  - Предполагает, что первые k документов - релевантны
  - Выполняет relevance feedback
- В среднем хорошо работает
- Но может получить очень плохие результаты для некоторых запросов
- Несколько итераций могут вызвать «искажение запроса»

# Задача

- Запрос: *отбор кандидатов*
- Пользователь отметил релевантными два документа
  - *Кандидат отобрать претендент*
  - *Отбор выбрать претендент*
- Объем коллекции – 1 млн. документов
- Df:
  - отбор 70000, кандидат – 70000,
  - Претендент - 30000, отбирать – 50000, выбрать 70000
- Как изменится запрос, если
  - $\alpha=0.7$ ,  $\beta=0.3$
  - Вектора запроса и документов нужно составить по формуле  $ntc.nnn$  (см. лекцию про векторные модели)

# Методы расширения запроса

- Несовпадение слова запроса:
  - самолет – лайнер
- Методы расширения запроса:
  - Глобальные методы
    - Информационно-поисковый тезаурус
    - Автоматически порождаемый тезаурус
  - Локальные методы
    - Relevance feedback (обратная связь по релевантности)
    - Pseudo Relevance feedback (обратная связь по псевдорелевантности)



# Информационно-поисковые тезаурусы

- Информационно-поисковый тезаурус – нормативный словарь терминов предметной области, создаваемый для улучшения качества информационного поиска в данной предметной области
- Национальные и международные стандарты
- Особенно популярны в 60-80х годах 20 века

# Цели разработки ИПТ

- Перевод языка авторов на нормативный язык, используемый для индексации и поиска
- Обеспечение последовательности в присваивании индексных терминов
- Обозначение отношений между терминами
- Облегчение информационного поиска

# Примеры тезаурусов

- Тезаурус ООН – UNBIS Thesaurus
- Тезаурус Европейского союза – EuroVoc
- Тезаурус Исследовательской службы Конгресса США – LIV
- СССР
  - Правовой тезаурус
  - ИНИОН
  - Шемакин «Технический тезаурус»
- Стандарты ISO, ГОСТы

# Традиционные информационно-поисковые тезаурусы для ручного индексирования: структура

- Основные понятия ПО – дескрипторы
- Условные синонимы – аскрипторы –
  - Отношения эквивалентности
  - аскриптор – дескриптор
- Отношения между дескрипторами

# Отношения в ИПТ

- Родовидовые отношения (выше – ниже)
  - BT (broader term) – NT (narrower term)
- Отношение ассоциации
  - RT (related term)
- Часть –целое: части должны всегда относиться к своему целому (рекомендация стандарта Z39.19)
  - Органы тела
  - Географические объекты
  - Дисциплины
  - Иерархические структуры (*полк – батальон – рота*)

•

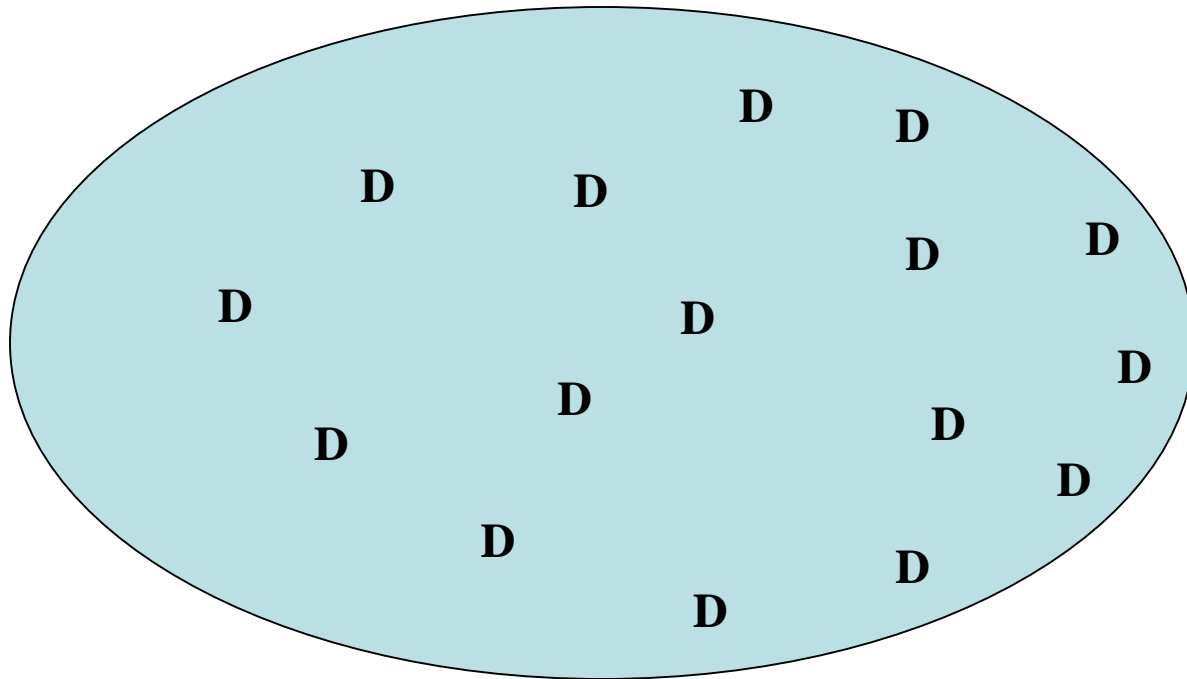
# Ассоциативные отношения (Стандарт Z39.19)

- Сфера деятельности – действующее лицо
  - Математика – математик
- Дисциплина – объект изучения
  - Неврология – нервная система
- Действие – агент или инструмент
  - Охота – охотник
- Действие – результат действия
  - Ткачество – ткань
- Действие – цель
  - Переплетные работы - книга
- Причина-следствие
  - Смерть – похороны
- Величина – единица измерения
  - Сила тока - ампер
- Действие - контрагент
  - Аллерген – антиаллергический препарат и т.п.

# Информационно-поисковые тезаурусы: использование для индексирования

- Процедура индексирования по тезаурусу
  - Индексатор читает документ, формулирует его основное содержание
  - Затем ему нужно подобрать наиболее подходящий набор дескрипторов для описания содержания документа
  - Отношения используются для уточнения набора дескрипторов
- Контекст использования
  - 60-80е годы – в информационных системах не было полных документов
  - Вид библиотечной обработки в парламентах, международных организациях

# Дескрипторы в предметной области



Дескрипторы похожи на теги, но тщательно отобраны профессионалами для обеспечения равномерного покрытия предметной области



# Тезаурус исследовательской службы Конгресса США (LIV)

## Transportation

Scope Note *For writings on transportation in a specific city or metropolitan area, use  
Urban transportation.*

Narrower Term [Choice of transportation](#)

[Commercial aviation](#)

[Energy transportation](#)

[Highspeed ground transportation](#)

[Inland water transportation](#)

[Intermodal transportation](#)

[Marine transportation](#)

[Military transportation](#)

[Motor transportation](#)

[Railroads](#)

[Student transportation](#)

[Terminals \(Transportation\)](#)

[Transportation and the aged](#)

[Transportation and the disabled](#)

# UNBIS Thesaurus: ТРАНСПОРТ

- **Более узкие термины:**

Воздушный транспорт

Городской транспорт

Каботаж

Контейнерные перевозки

Морской транспорт

Немоторный транспорт

Пассажирские перевозки

Перевозка грузов

Перевозки по внутренним путям

Смешанные перевозки

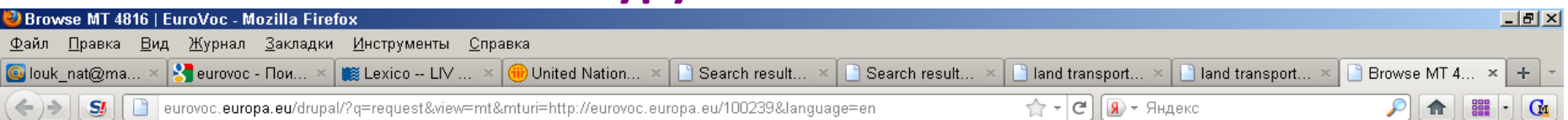
Транспорт для оказания помощи

Устойчивый транспорт

# UNBIS: ТРАНСПОРТ-2

- **Смежные термины (RT):**
  - Космические перевозки
  - Перевозочные документы
  - Правила перевозок
  - Проекты в области транспорта
  - Путешествия
  - Работники транспорта
  - Технология транспортных перевозок
  - Транспортная инфраструктура
  - Транспортная политика
  - Транспортная статистика
  - Транспортное оборудование
  - Транспортное планирование
  - Транспортные исследования
  - Транспортные коридоры
  - Транспортные расходы
  - Транспортные системы
  - Транспортные терминалы
  - Упрощение перевозок
  - Экономика транспорта
  - Экспресс-службы

# Тезаурус EuroVoc



## 4816 land transport

### land transport

RT energy transport [ 6606 ]

RT European Conference of Ministers of Transport [ 7621 ]

#### NT1 rail transport

RT air-cushion vehicle [ 4811 ]

RT European Railway Agency [ 1006 ]

RT high-speed transport [ 4811 ]

RT railway industry [ 6821 ]

RT railway tariff [ 4806 ]

RT transport staff [ 4811 ]

#### NT2 CIV Convention

RT carriage of passengers [ 4811 ]

#### NT2 rail network

RT transport network [ 4811 ]

#### NT3 railway station

#### NT2 rolling stock

#### NT2 vehicle on rails

RT underground railway [ 4811 ]

#### NT1 road transport

RT axle tax [ 4806 ]

RT driving licence [ 4806 ]

RT motor vehicle industry [ 6821 ]

# Информационно-поисковый тезаурус как книга



# Расширение запроса, основанное на тезаурусных знаниях

- Для каждого термина  $t$  в запросе происходит расширение синонимичными словами или близкими по смыслу (связанными отношениями с исходным словом)
  - из тезауруса
    - *feline* → *feline cat*
- Как расширять:
  - Можно добавлять в вектор запроса (с более низкими весами и в зависимости от типа отношения к слову запроса)
  - Можно вставлять в булевское выражение
    - *Налог* → ( *НАЛОГ* или *НАЛОГОВЫЙ* )
- Используется в предметно-ориентированных системах
  - Современные тезаурусы, встроенные в ПО поисковые системы, могут иметь другие формы, чем описано в стандартах, например, только список синонимов и вариантов

# Расширение запроса, основанное на тезаурусных знаниях-2

- Увеличивает полноту поиска
- Обычно снижает точность поиска, обычно для многозначных слов
  - “interest rate” → “interest rate fascinate evaluate”
  - Можно вводить в тезаурус многословные термины «interest rate», но запросы все равно разнообразнее
- Сложность создания и обновления тезаурусов
- Поэтому в интернет-поиске
  - Долгое время не было расширения запросов
  - Затем стали расширять на однокоренные слова
  - Сейчас для расширения запроса используются статистически насчитанные «синонимы»

# Тезаурусные отношения при автоматич. расширении запросов

- Синонимы
  - хорошо работает для однозначных слов (выражений)
- Родовидовые отношения (выше-ниже)
  - Хорошо работает, если запрос совпадает с термином тезауруса
  - В длинном запросе может приводить к снижению точности
  - Города Сибири -> город столица Сибири



# Тезаурусные отношения при автоматич. расширении запросов-2

- Традиционные информационно-поисковые тезаурусы
  - Отношение ассоциации
    - Считается симметричным, но фактически часто не симметрично
    - Принципы установления
  - EvroVoc: Монографии – асц - Типографии
- Предложения:
  - ввести большую градацию отношений (причина, объект, место ...)
  - ввести числовые оценки на отношения
  - Но: в любом случае контекст длинного запроса может сильно влиять на направление расширения

# Методы расширения запроса

- Несовпадение слова запроса:
  - самолет – лайнер
- Методы расширения запроса:
  - Глобальные методы
    - Информационно-поисковый тезаурус
    - Автоматически порождаемый тезаурус
  - Локальные методы
    - Relevance feedback (обратная связь по релевантности)
    - Pseudo Relevance feedback (обратная связь по псевдорелевантности)

# Извлечение синонимов для автоматического расширения запросов

- Компания Яндекс: доклад Russir-2010
- Признаки для извлечения синонимов
  - Совместная встречаемость в одном документе (странице)
  - Совместная встречаемость в тексте ссылки (анкор)
  - Встречаемость в документе и в тексте ссылки
  - Как часто пользователь в запросах заменяет одно на другое
  - Клики пользователя на страницу, содержащую S2, при запросе, содержащем S1
  - сходство контекстов употребления S1 и S2 (запросы, документы) и др.

# Примеры расширения (декабрь 2010 – февраль 2011)

Запрос — документ

- |                        |   |                                      |
|------------------------|---|--------------------------------------|
| речное судно           | — | морское судно                        |
| речной порт            | — | морской порт (Находка)               |
| присуждение имущества  | — | передача имущества                   |
| ледяная горка          | — | холодная гора                        |
| расширение отверстия   | — | расширение канала<br>(сети интернет) |
| договор поручительства | — | договор поручения                    |
| аварийное отключение   | — | знак аварийной остановки             |
| замкнутая граница      | — | закрыть границу                      |

# Современное состояние расширения запросов (яндекс)

- По однокоренным частям речи
  - Решить – решение
  - Дорога - дорожный
- Многие из перечисленного исчезло
- Сохраняется:  
Договор поручительства – договор поручения  
Проблема – задача - вопрос
  - *Решение проблемы – решение задачи*
  - *Постановка проблемы – постановка вопроса (на английском языке)*
  - *Обсудить - рассмотреть*

# Заключение: методы расширения запроса

- Глобальные методы
  - Информационно-поисковый тезаурус
  - Автоматически порождаемый тезаурус
- Локальные методы (по конкретному запросу)
  - Relevance feedback (обратная связь по релевантности)
  - Pseudo Relevance feedback (обратная связь по псевдорелевантности)