Расширение запроса при поиске

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

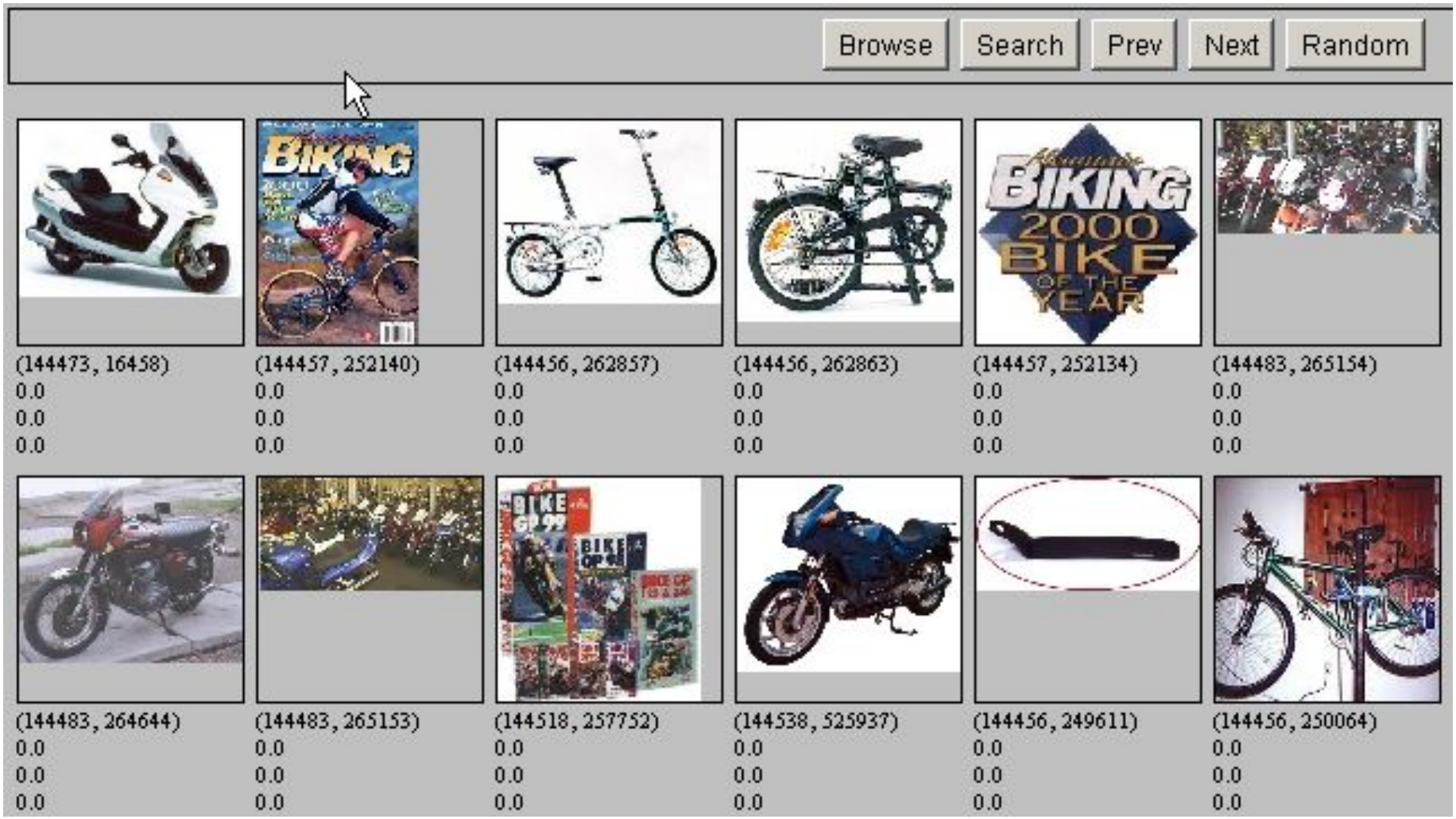
Методы расширения запроса

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

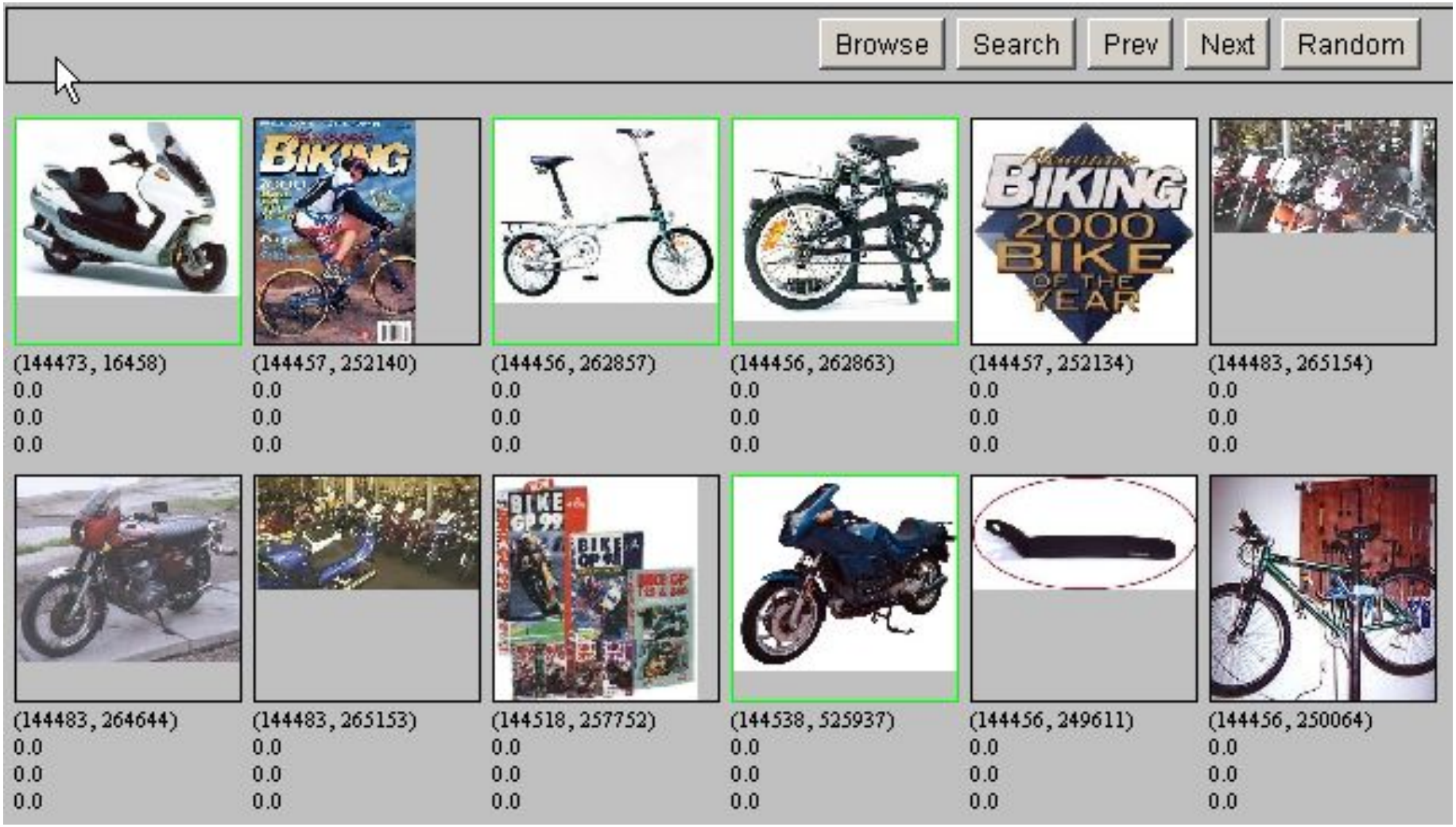
Обратная связь по релевантности

- 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.
- Основная идея: сформулировать хороший запрос трудно, если пользователь не знаком с коллекцией, поэтому – итеративное построение запроса

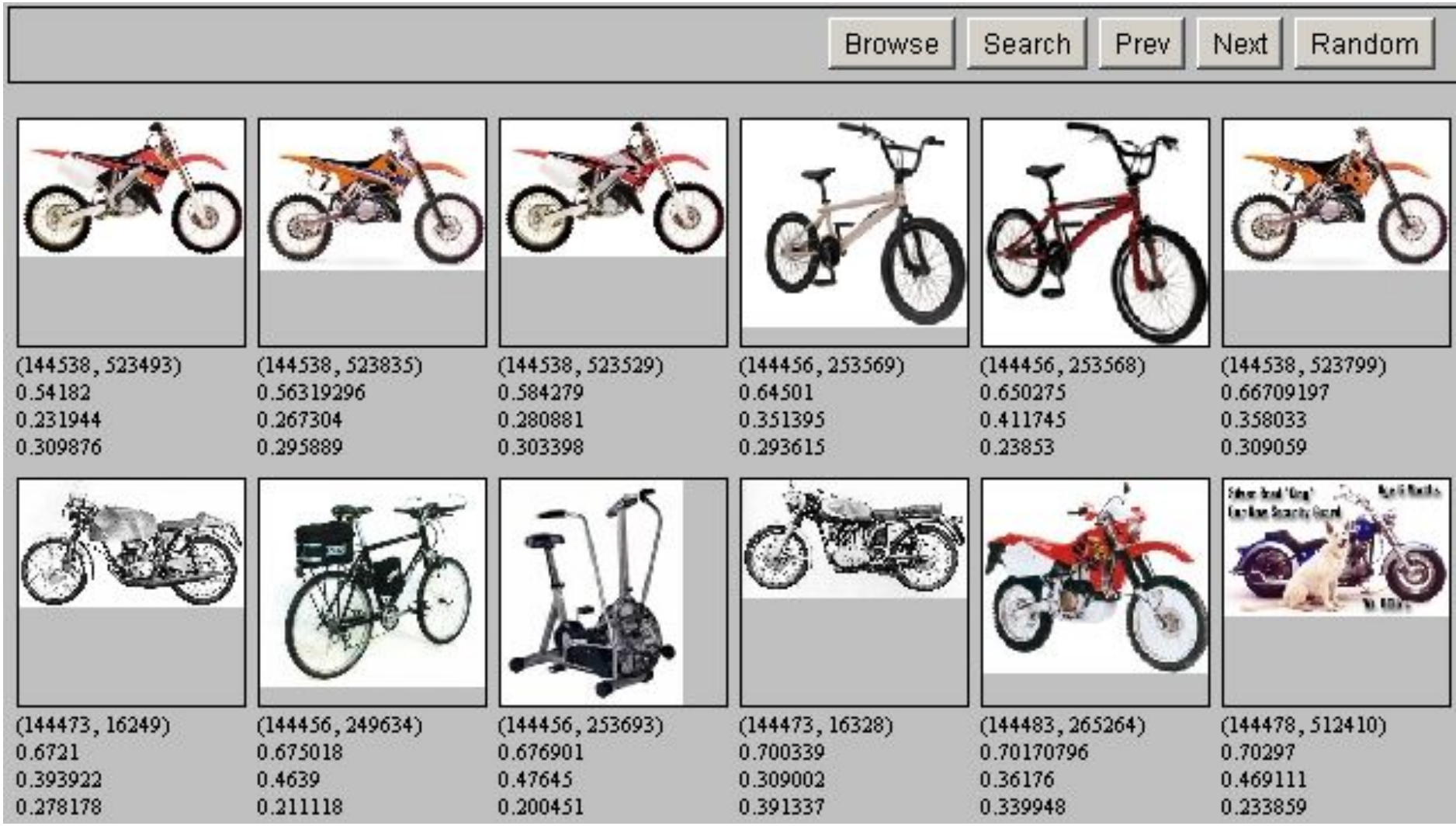
Результаты для начального запроса



Разметка пользователя

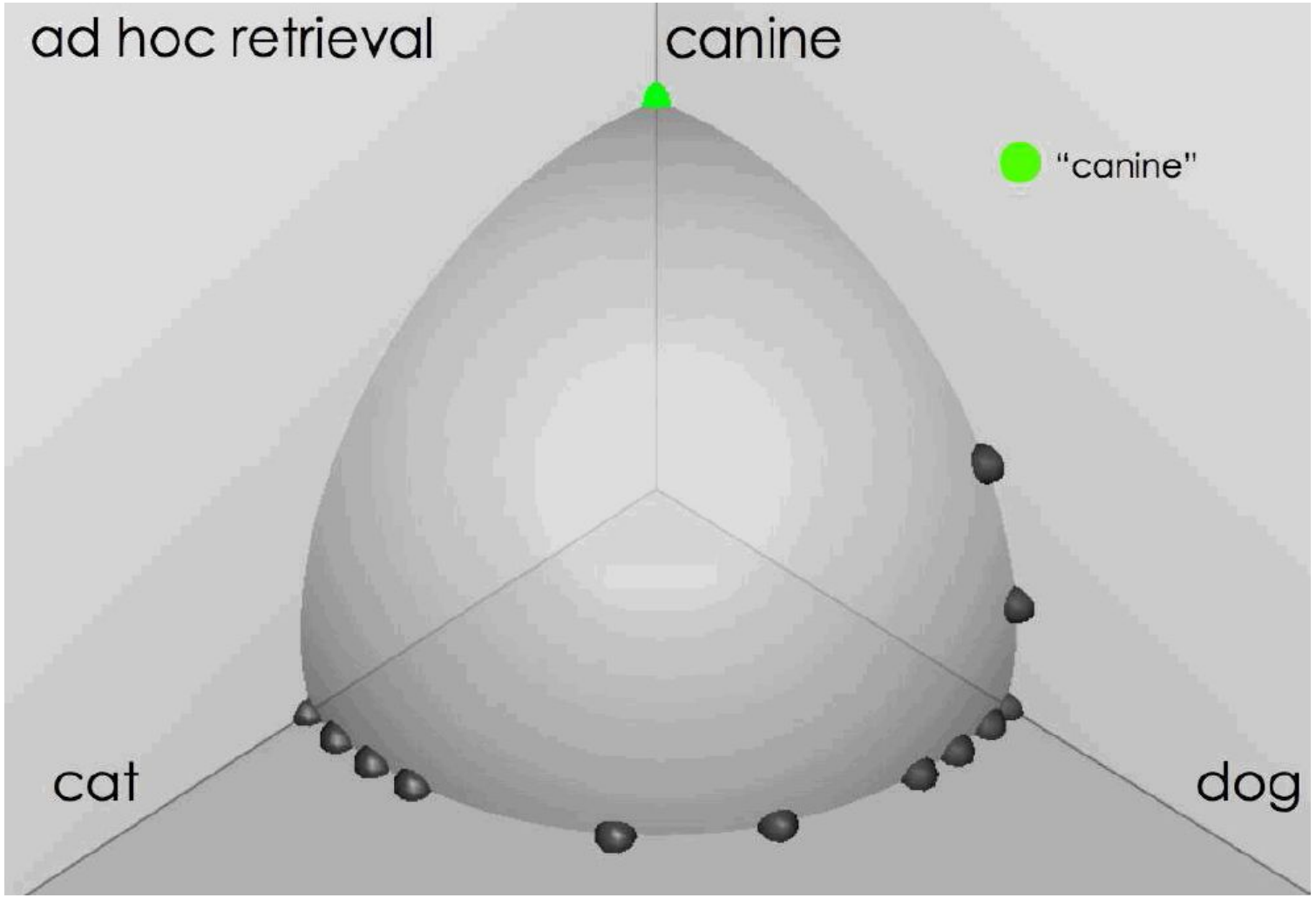


Результаты после разметки



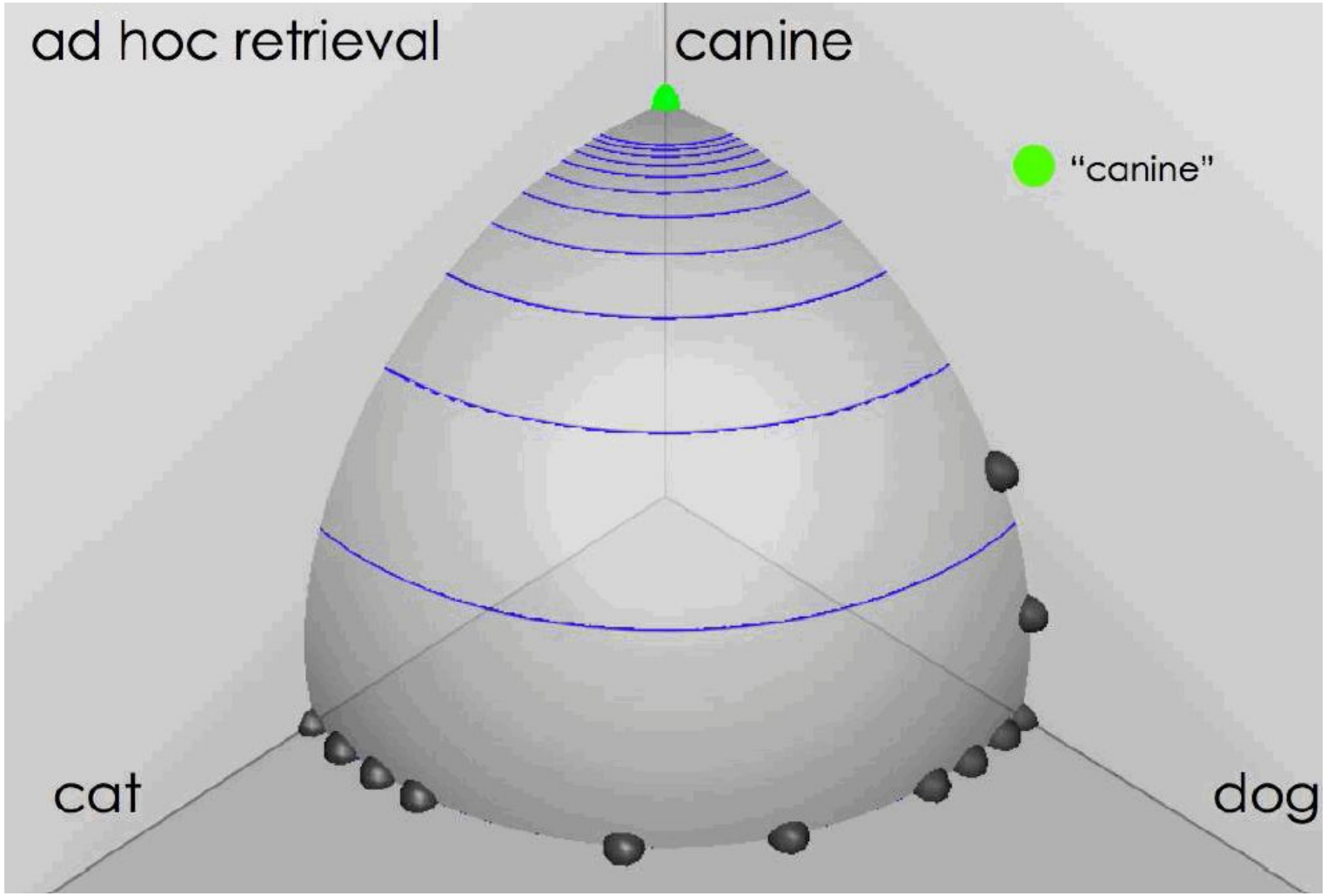
Выдача по запросу *canine*

source: Fernando Diaz



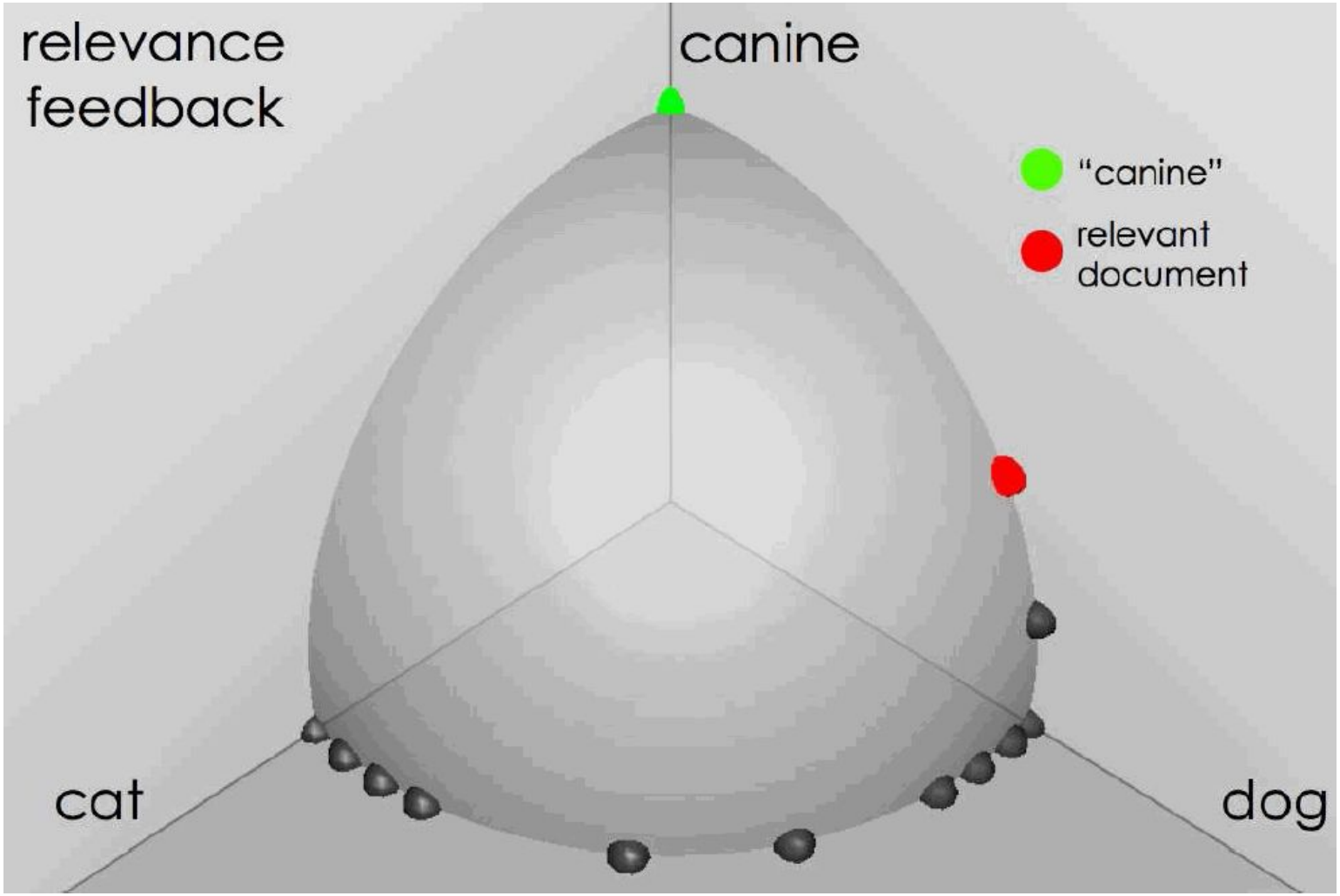
Выдача по запросу *canine-2*

source: Fernando Diaz

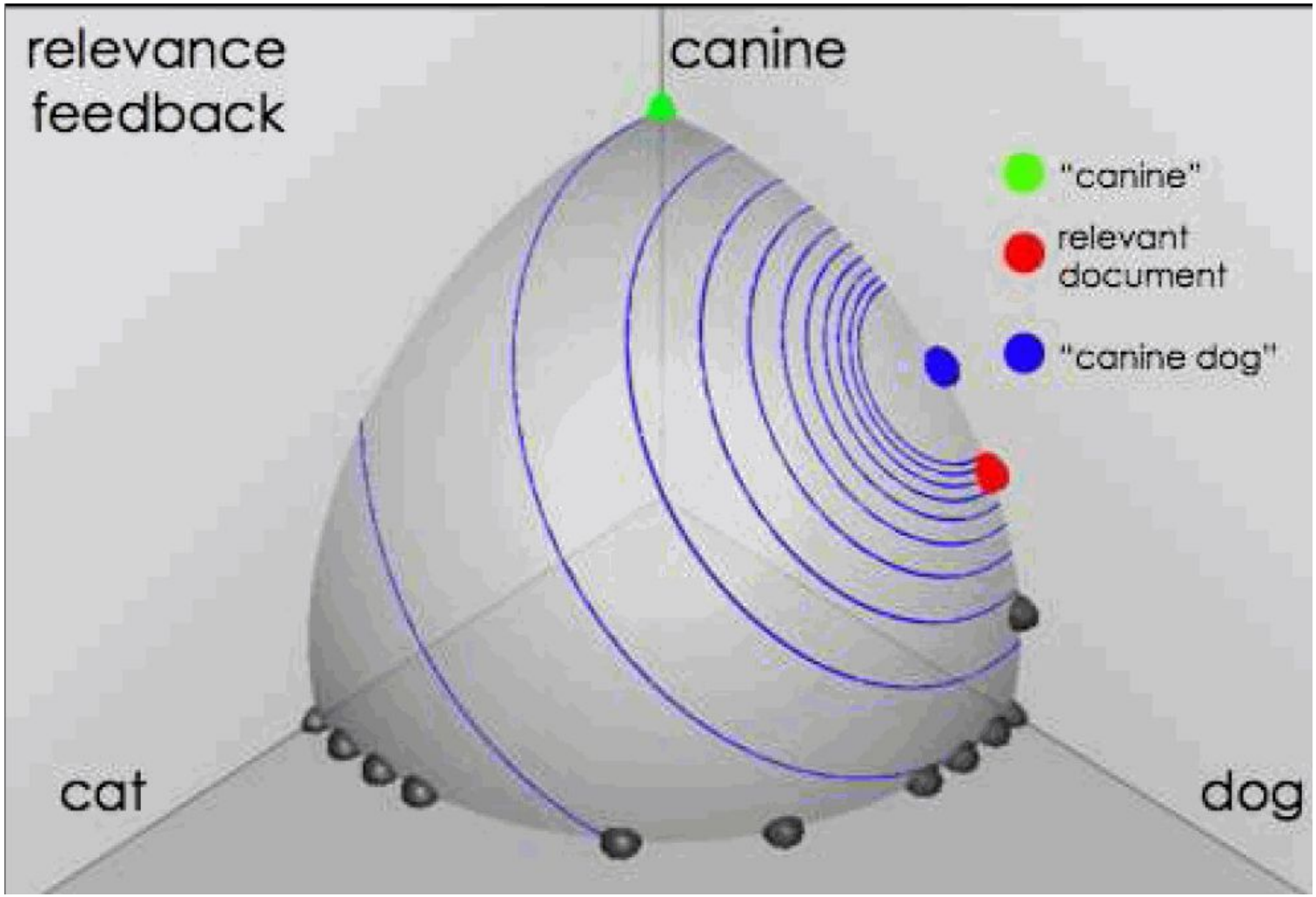


Пользователь выбирает релевантное

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

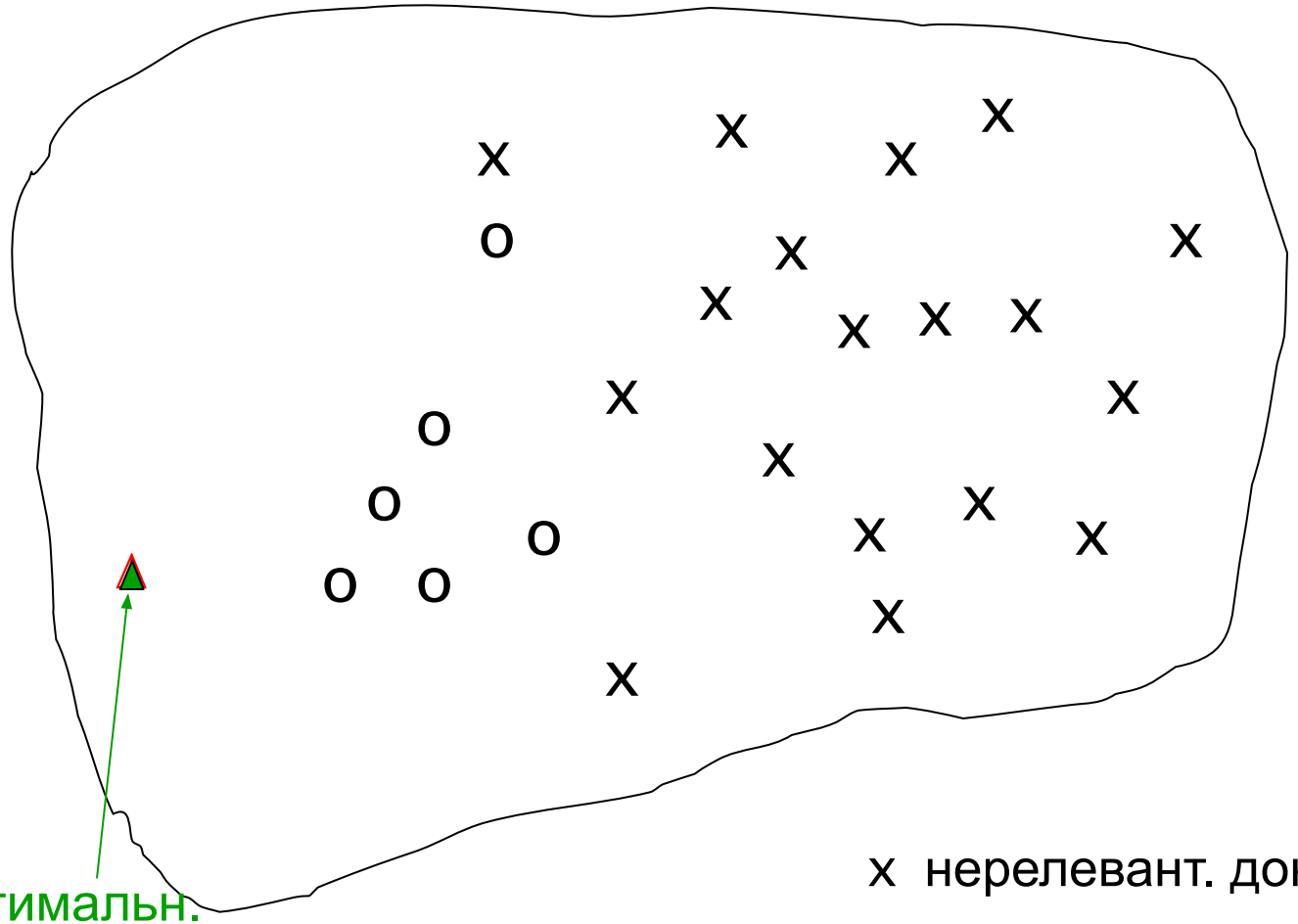
$$\vec{q}_{opt} = \arg\max_{\vec{q}} [\cos(\vec{q}, \vec{\mu}(C_r)) - \cos(\vec{q}, \vec{\mu}(C_{nr}))]$$

- Пытается отделить релевантные и нерелевантные документы

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

- Проблема: мы не знаем все релевантные документы

Лучший запрос



Rocchio 1971 алгоритм (SMART)

- На практике используется:

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

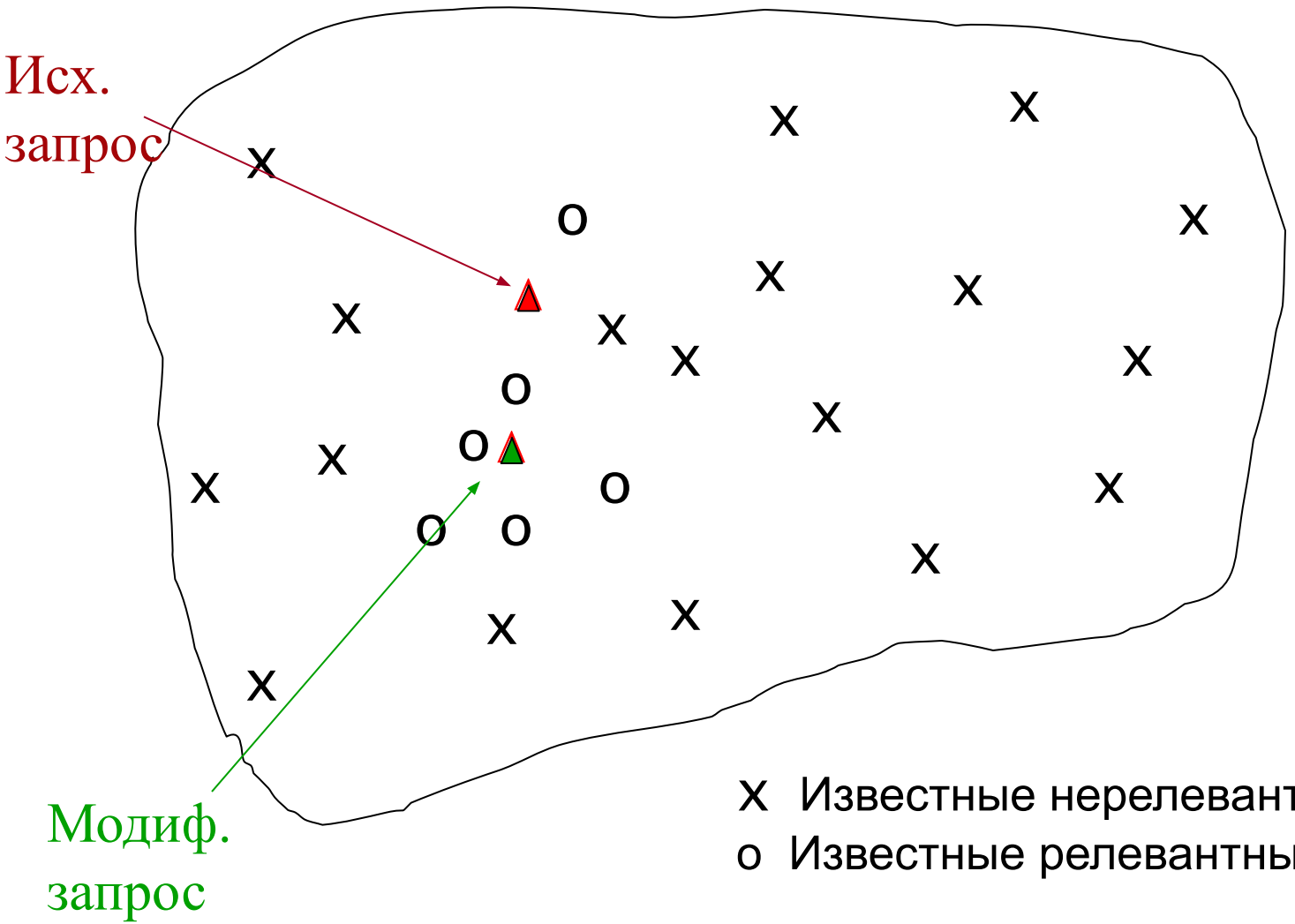
- D_r = множество известных релевантных doc векторов
- D_{nr} = множество известных нерелевантных doc векторов
  - Отличны от C_r *и* C_{nr}
- q_m = модифицированный вектор запроса; q_0 = исходный вектор запроса; α,β,γ: веса
- Новый запрос «сдвигается» по направлению к релевантным документам и «уходит»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

Особенности параметров

- Соотношение α vs. β/γ : Если у нас много оцененных документов, то лучше более высокие β/γ.

- Некоторые веса в модифицированном векторе запроса становятся отрицательными
  - Отрицательные веса слов игнорируются (устанавливаются равными 0)

Relevance feedback по исходному запросу



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

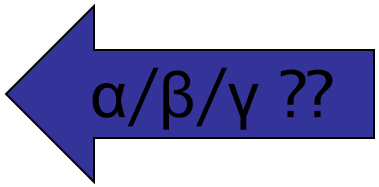
Relevance Feedback: Проблемы

- Длинные запросы – неэффективны для типичной поисковой машины
  - Большее 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 «ручнею» 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»

Методы расширения запроса

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

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

Слайды доклада Расширение поисковых запросов

А. Сокирко
Е. Соловьев (Яндекс)
http://romip.ru/russir2010/slides/yandex_lecture.pdf

Типы синонимов для расширения запроса

- С соответствиями между внутренними элементами:
  - Словообразование: Москва – московский; компиляция - компилирование
  - Аббревиатуры: МГУ - Московский государственный университет
  - Транслиты: Гугл – Google
  - Слитно – раздельно: ватер-поло – ватерполо
  - Орфоварианты: colour - color
- Без поддержки внутренних элементов
  - Переводы (стол – table)
  - Синонимы: бегемот – гипопотам
  - Подвиды: фильм - биопик

Overall design

- One system for all classes? For each word? For each class?

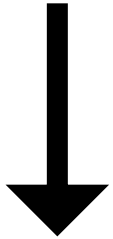
- Our solution is to supply each class with a separate algorithm of expansion.

Алгоритм составления базы

- Получение списка гипотез

~ 200 миллионов гипотез

- Машинное обучение



~ 150 миллионов гипотез

- Отсечение результатов – отсекается первый миллион и объявляется **словарем**

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.

Признаки синонимов: Скобочное написание

Скобочное написание – это набор n-gram, которые встречаются с текстах рунета в контексте скобок:

Московский государственный университет (МГУ)

Владимир Путин (Vladimir Putin)

Признаки синонимов: Открытые словари

Русская Википедия содержит около миллиона жестких перенаправлений, типа:

Абрикос сибирский --- Даурсат

Авачинская бухта --- Авачинская губа

Оценка качества

- 1) Оценка пары синонимов без контекста
- **2) Оценка пары синонимов в контексте запроса**
- 3) Оценка качества поисковой выдачи

Morphological derivation

- The linguistic model consists of the same suffix transformation(=flexia models), like:
  memorize -> memorization: -e,-ation
- There are enough false positives, like sense -> sensibility.
- Generalize models in order to unify the following transformations:
  memorize-> memorization : e->ation
  induce -> induction: e -> tion
  publish -> publication sh->cation

Sense deviation, term boundaries

- F-measure for the dictionary is around 87% (Metric 1).
- F-measure for query expanding by derivation pairs is 65% (Metric 2).

[Australian population] (*Australian*=>*Australia* +)

[Australian gold] (*Australian* => *Australia* -)

[milk diet] (diet => dietary +)

[The Diet of the German Empire] (diet=>dietary -)
(a kind of Parliament)

Лексические расширения с использованием автоматического тезауруса

- Те же проблемы, что и в ручном:
  - Многозначность запроса или расширения
  - Проблемы с расширением устойчивых словосочетаний
  - Влияние контекста запроса и/или документа

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 (сети интернет)
договор поручительства	–	договор поручения
аварийное отключение	–	знак аварийной остановки
замкнутая граница	–	закрыть границу

речное-судно — Яндекс: Нашлось 504 тыс. ответов - Mozilla Firefox

Файл Правка Вид Журнал Закладки Инструменты Справка

http://yandex.ru/yandsearch?text=речное-судно&lr=20674

louk_nat@mail.ru: Ответ отпра... | TROITSK.ORG: Нет воды и теп... | речное-судно — Яндекс: На...

Поиск Почта Карты Маркет Новости Словари Блоги Видео Картинки ещё ▾ Войти

Яндекс

Нашлось 504 тыс. ответов

речное-судно Найти

в найденном в Троицке расширенный поиск

Мои находки Помощь

Настройка

Регион: Троицк

2011 год

1 Как устроены **морские суда** - Заглавная страница

Надстройки и Рубки. Штевни **морского судна**. Кронштейны, выкружки и коридор гребного вала. Люки и шахты на **морском судне**. Фундаменты **судовых** машин. Фальшборт и леерное ограждение.

www.seaships.ru копия ещё

2 Российские **речные суда**

© "Российские **речные суда**", 1998-2005. © "Российский **речной** флот и туризм INFOFLOT.RU"

www.riverships.ru › Российские **речные суда** копия ещё

3 приобретение в собственность **морского/речного судна**

Annuitet Пишет: - дык у меня аренда, а не ипотека, и **речное судно**. > (плавучая гостиница), а не **морское** - поэтому вроде.

forum.garant.ru/read.php?7,1262078 копия ещё

4 оценка **морского судна**, оценка корабля, оценка **речных судов**, оценка...

Все **речные** и **морские суда**, которые необходимо, согласно существующему законодательству регистрировать, причисляются к недвижимому имуществу.

www.profiexpert.ru/ocenka-transporta/morskoi-i-… копия

5 **РЕЧНОЕ СУДНО**-АВТОМОБИЛЕВОЗ, - **СУДА** - ДОСКА ОБЪЯВЛЕНИЙ

Разместить объявление. **Речное судно**-автомобилевоз, **Судно** катамаранного типа для перевозки автомобилей, проект Р-19. для перевозки авто, негабаритов, яхт.

www.korabel.ru › Доска объявлений › detail/34122.html копия ещё

Разместить объявление по запросу «речное-судно»

«речное-судно» в картинках

Все картинки

Видео «речное-судно»

Восстановление Шелкового пути в современном формате

Все видеоролики

расширение-отверстия — Яндекс: Нашлось 207 тыс. ответов - Mozilla Firefox

Файл Правка Вид Журнал Закладки Инструменты Справка

http://yandex.ru/yandsearch?text=расширение-отверстия&lr=20674

расширение-отверстия — Янд...

Поиск Почта Карты Маркет Новости Словари Блоги Видео Картинки ещё ▾

Яндекс

Нашлось 207 тыс. ответов

расширение-отверстия

Найти

в найденном в Троицке расширенный поиск

Мои находки

Настройка

Регион: Троицк

2011г.

1 Новости сети

Сегодня произведено плановое **расширение канала** связи Интернет. Дата размещения: 27.03.2009.

www.oz-web.ru/news/index.php?news=34 Орехово-Зуево копия ещё

2 Роль предварительного **расширения каналов** - Медицинская библиотека...

Определение апикального сужения в изогнутых корнях нижнечелюстных моляров - предварительно **расширенные каналы** в сравнении с нерасширенными.

www.medlinks.ru/article.php?sid=27150 копия ещё

3 Largal ultra – жидкость для химического **расширения каналов**. Septodont...

С помощью пипетки ввести ЛАРГАЛЬ УЛЬТРА в полость зуба, а затем в **каналы**, но уже используя для этого корневую иглу. Сразу после этого можно начать механическое **расширение канала**.

www.konsort.ru/index.php?productID=934 Москва копия ещё

4 ОАО ГОТТЦ Гарант - **Расширение канала** Интернет Гродно-Минск

Для удобства пользователей в ближайшее время **канал** Гродно-Минск для выхода в сеть Интернет будет **расширен** дополнительно на 6 Мегабит/с...

www.garant.by/internet/news/items/**rasshirenie**... копия ещё

5 Новости

Расширение канала до д. Волжский. Уважаемые абоненты, на прошлой неделе был

Разместить объявление по запросу «расширение-отверстия»

«расширение-отверстия» в картинках

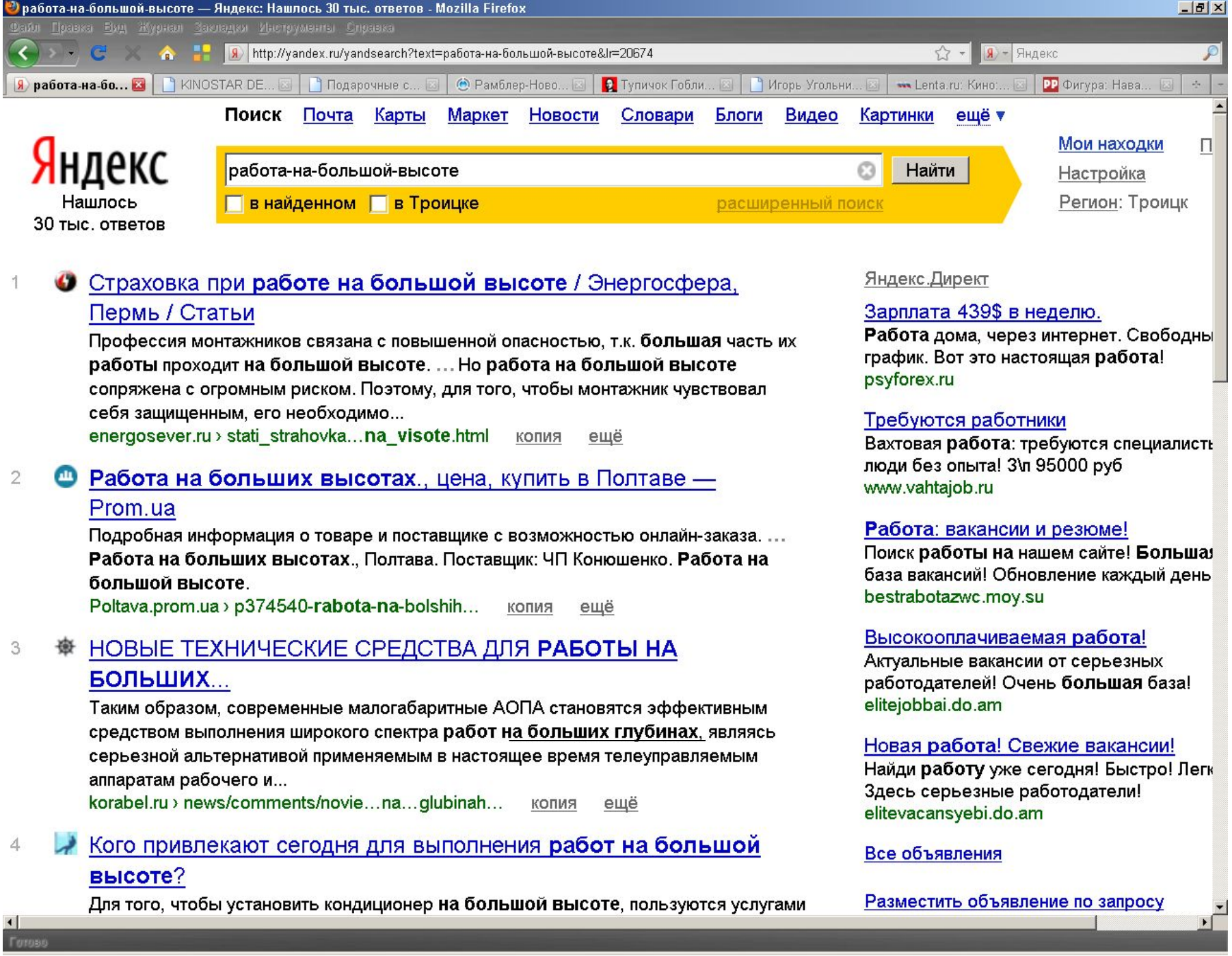
Все картинки

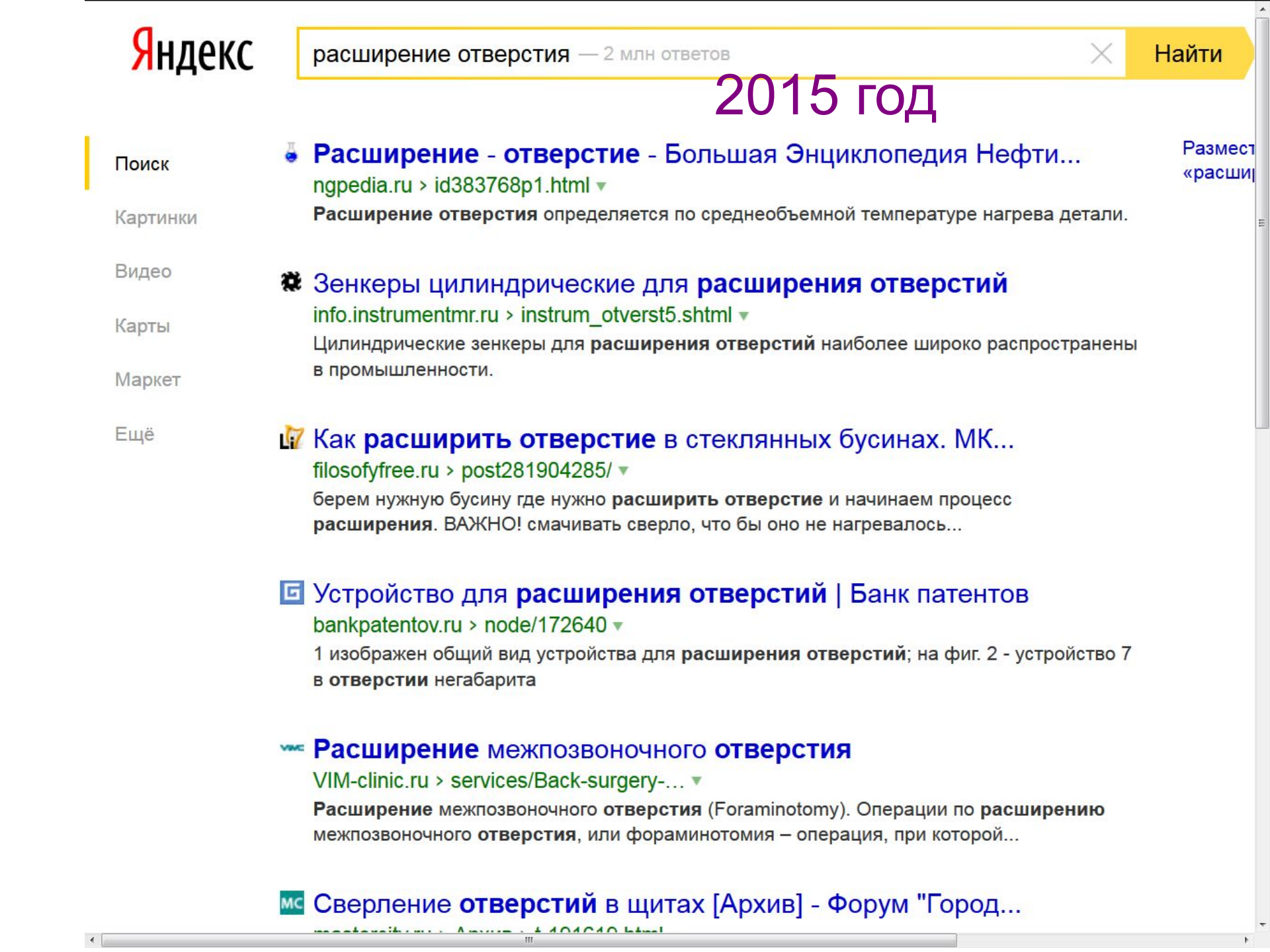
Видео «расширение-отверстия»

БИА микрохирургическое **расширение канала**

Все видеоролики

Готово


работа-на-большой-высоте — Яндекс: Нашлось 30 тыс. ответов - Mozilla Firefox
Файл Правка Вид Журнал Закладки Инструменты Справка
http://yandex.ru/yandsearch?text=работа-на-большой-высоте&lr=20674
Яндекс
работа-на-бо...
KINOSTAR DE...
Подарочные с...
Рамблер-Ново...
Туличок Гобли...
Игорь Угольни...
Lenta.ru: Кино:...
Фигура: Нава...
Поиск Почта Карты Маркет Новости Словари Блоги Видео Картинки ещё
Яндекс
Нашлось
30 тыс. ответов
работа-на-большой-высоте
Найти
в найденном
в Троицке
расширенный поиск
Мои находки
Настройка
Регион: Троицк
1 Страховка при работе на большой высоте / Энергосфера, Пермь / Статьи
Профессия монтажников связана с повышенной опасностью, т.к. большая часть их работы проходит на большой высоте. ... Но работа на большой высоте сопряжена с огромным риском. Поэтому, для того, чтобы монтажник чувствовал себя защищенным, его необходимо...
energosever.ru › stati_strahovka...na_visote.html копия ещё
2 Работа на больших высотах., цена, купить в Полтаве — Prom.ua
Подробная информация о товаре и поставщике с возможностью онлайн-заказа. ... Работа на больших высотах., Полтава. Поставщик: ЧП Конюшенко. Работа на большой высоте.
Poltava.prom.ua › p374540-rabota-na-bolshih... копия ещё
3 НОВЫЕ ТЕХНИЧЕСКИЕ СРЕДСТВА ДЛЯ РАБОТЫ НА БОЛЬШИХ...
Таким образом, современные малогабаритные АОПА становятся эффективным средством выполнения широкого спектра работ на больших глубинах, являясь серьезной альтернативой применяемым в настоящее время телеуправляемым аппаратам рабочего и...
korabel.ru › news/comments/novie...na...glubinah... копия ещё
4 Кого привлекают сегодня для выполнения работ на большой высоте?
Для того, чтобы установить кондиционер на большой высоте, пользуются услугами
Яндекс.Директ
Зарплата 439$ в неделю.
Работа дома, через интернет. Свободны график. Вот это настоящая работа!
psyforex.ru
Требуются работники
Вахтовая работа: требуются специалисты люди без опыта! 3\n 95000 руб
www.vahtajob.ru
Работа: вакансии и резюме!
Поиск работы на нашем сайте! Большая база вакансий! Обновление каждый день
bestrabotazwc.moy.su
Высокооплачиваемая работа!
Актуальные вакансии от серьезных работодателей! Очень большая база!
elitejobbai.do.am
Новая работа! Свежие вакансии!
Найди работу уже сегодня! Быстро! Легк
Здесь серьезные работодатели!
elitevacansyebi.do.am
Все объявления
Разместить объявление по запросу
Готово


Яндекс
расширение отверстия — 2 млн ответов
Найти
2015 год
Поиск
Картинки
Видео
Карты
Маркет
Ещё
Расширение - отверстие - Большая Энциклопедия Нефти...
ngpedia.ru › id383768p1.html
Расширение отверстия определяется по среднеобъемной температуре нагрева детали.
Размест
«расши
Зенкеры цилиндрические для расширения отверстий
info.instrumentmr.ru › instrum_otverst5.shtml
Цилиндрические зенкеры для расширения отверстий наиболее широко распространены в промышленности.
Как расширить отверстие в стеклянных бусинах. МК...
filosofyfree.ru › post281904285/
берем нужную бусину где нужно расширить отверстие и начинаем процесс расширения. ВАЖНО! смачивать сверло, что бы оно не нагревалось...
Устройство для расширения отверстий | Банк патентов
bankpatentov.ru › node/172640
1 изображен общий вид устройства для расширения отверстий; на фиг. 2 - устройство 7 в отверстии негабарита
Расширение межпозвоночного отверстия
VIM-clinic.ru › services/Back-surgery-...
Расширение межпозвоночного отверстия (Foraminotomy). Операции по расширению межпозвоночного отверстия, или фораминотомия – операция, при которой...
Сверление отверстий в щитах [Архив] - Форум "Город...

Заключение: методы расширения запроса

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

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, отобрать – 50000, выбрать 70000
- Как изменится запрос, если
  - alpha=0.7 (коэффициент учета запроса),
  - beta=0.3 (коэффициент учета релевантных документов),
  - Запрос представляется как вектор частот
  - Документ представляется как нормализованный вектор tf.idf