Search engines
Поисковые машины, информационно-поисковые системы, ИПС

В прошлый раз: основные понятия в информационном поиске

- Users and Information Needs – потребность пользователя, информационная потребность

- Relevance – релевантность

- Evaluation - оценка качества

Информационный поиск и поисковые машины-2

Информационный поиск

Релевантность

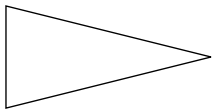
-Эффективное ранжирование

Оценка качество

-Тестирование и измерение

Потребности пользователя

-Взаимодействие с пользователем



Поисковые машины

Исполнение запроса

-Эффективный поиск и индексирование

Включение новых данных

-Покрытие и свежесть

Масштабируемость

-Рост с данными и пользователями

Адаптивность

-Настройка на приложения

Специфические проблемы

-например, спам

Особенности работы поисковых машин

- Выполнение запроса (performance)
  - Измерение и улучшение эффективности поиска
    - Уменьшение времени ответа, увеличение скорости индексирования
  - *Индексы – это структуры данных, которые необходимы, чтобы уменьшить время ответа системы*
    - Важнейший вопрос для поисковых систем

Особенности работы поисковых машин - 2

- Динамические данные
  - «Коллекции» данных для наиболее востребованных приложений постоянно меняются: обновляются, удаляются, пополняются
    - Например, веб-страницы
    - Типичные меры: покрытие (сколько проиндексировано) и новизна (*freshness*) (насколько недавно проиндексировано)
  - Необходимо одновременно менять индексы и обрабатывать запросы

Особенности работы поисковых машин-3

- Масштабируемость
  - Миллионы пользователей и терабайты документов
  - Используется распределенная обработка
- Адаптивность
  - Изменение и настройка компонентов поисковой машины, таких как алгоритм ранжирования, методы индексирования, интерфейсы для различных приложений

Поисковый спам

- Для веб поиска одним из важных направлений работы является поисковый спам
- Важно для качества поисковых результатов
- Много видов спама
  - Порождение текстов похожих на естественные
  - Ссылочный спам и др.

- Новая область информационного поиска - *adversarial IR*,
  - Спамеры - противники с различными целями

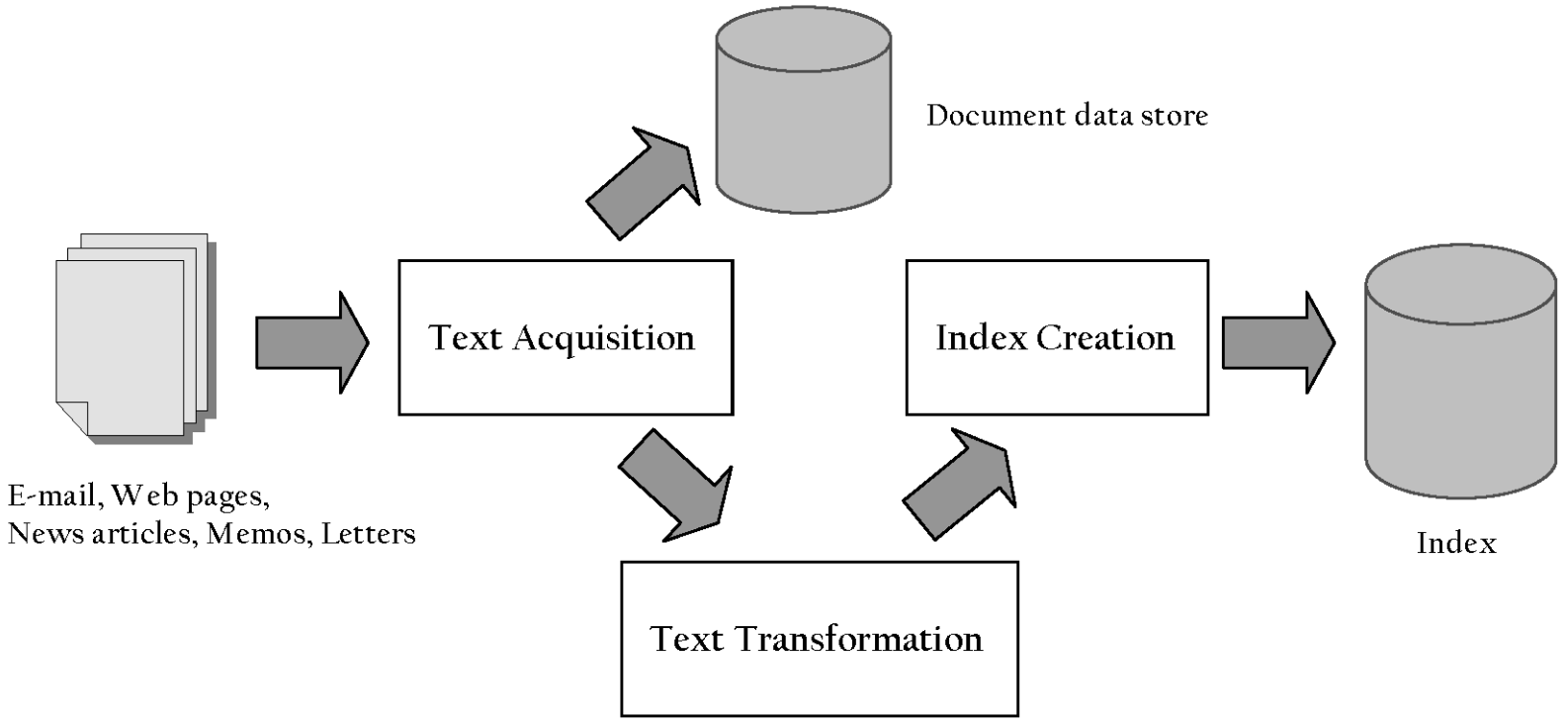
Архитектура поисковых машин

Основные компоненты поисковых машин

Основные подсистемы поисковой машины

- Подсистема сбора и индексирования документов
- Подсистема взаимодействия с пользователем для выполнения его запросов.

Подсистема1. Процесс индексирования



Процесс индексирования-2

- Извлечение текстов
  - Идентифицирует и сохраняет тексты для индексирования
- Трансформация текстов
  - Трансформирует документы в индексные термы
- Создание индексов
  - Берет индексные термы и создает индексы для быстрого поиска

Извлечение текстов. Краулер

- Идентифицирует и извлекает документы для поисковой машины
- Много типов – интернет, предприятие, компьютер
- Интернет-краулеры используют ссылки, чтобы найти документы
    - Должны найти огромное количество веб-страниц (покрытие) и сохранять их в актуальном состоянии
    - Краулеры сайтов
    - Тематические краулеры для вертикального поиска
- Краулеры документов для поиска по документам предприятия или компьютера
    - Используют ссылки и сканируют директории

Получение текстов-2

- Фиды
  - Потоки документов в реальном потоке времени
    - Новости, блоги, видео, радио, tv
  - RSS - стандарт
    - RSS читалка обеспечивает новые XML документы поисковой машине

- Конвертация
  - Конвертирует форматы в текст плюс мета-данные
    - HTML, XML, Word, PDF, и др. → XML
  - Конвертирует кодировки для различных языков
    - Например, в кодировку UTF-8

Получение текстов-3

- Хранилище документов
  - Хранит тексты, метаданные и другое содержание документов
    - Метаданные: тип, дата создания
    - Ссылки, текст ссылки
  - Обеспечивает быстрый доступ к содержанию документов
    - Порождение списка результатов
  - Эффективное хранение
    - не реляционная база данных

Преобразование текстов

- Анализатор (Parser)
  - Обрабатывает последовательность токенов в документе, распознает структурные элементы
    - Заголовки, ссылки, подзаголовки и др.
  - *Токенизатор распознает «слова» в тексте*
    - Обработка капитализации, кавычек, дефисов ..
  - *Обработка структуры, задаваемой* HTML, XML
    - *Теги:* <h2> Overview </h2>
    - Парсер использует синтаксис языка разметки идентифицировать структуру документа

Преобразование текстов-2

- Стоп-слова
  - Удаление наиболее частотных слов
    - Предлоги, союзы, артикли..
  - Может быть проблемой для некоторых запросов

- Стемминг (морф. анализ)
  - “computer”, “computers”, “computing”, “compute”
  - Обычно эффективен, но не для всех запросов
  - Разное действие для разных языков

Преобразование текстов-3

- Анализ ссылок
  - Ссылки и тексты ссылок (анкор ссылки – якори)
  - Анализ ссылок важен для определения популярности сайта и сообщества, связанного с сайтом
    - Например, PageRank
  - Текст ссылки может значительно уточнить содержание связанных страниц,
  - Значительное влияние на интернет-поиск
    - Меньше значимость в других поисковых приложениях

Преобразование текстов-4

- Извлечение информации
  - Идентифицирует семантические классы индексных термов, которые важны для конкретных приложений индекс
  - Например, распознавание имен людей, географических мест, компаний, дат…

- Классификатор
  - Отнесение текста к категориям
    - Тематика, тональность, жанры и др.
    - Зависит от приложения

Создание индекса

- Статистика по документам
  - Собирает частоты и позиции слов и других признаков
  - Используется в ранжирующем алгоритме
- Определение весов
  - Вычисляет веса для индексных термов
  - Используется в алгоритме ранжирования
  - например, вес *tf.idf*
    - Комбинирование частоты слова в документе и инверсной подокументной частоты слова в коллекции

Создание индекса-2

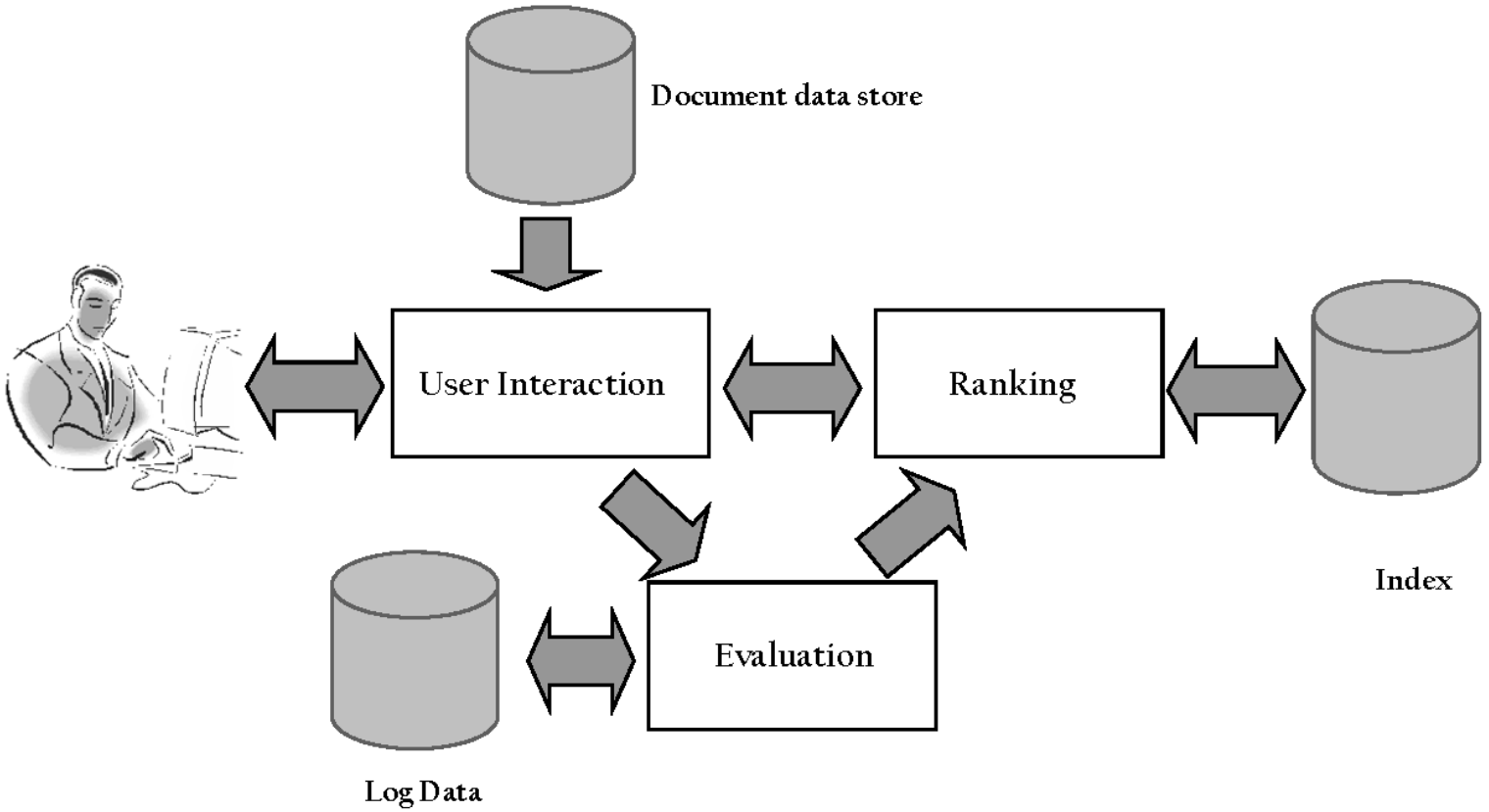
- Инвертирование
  - Преобразует матрицу документ-терм в данные терм-документ, необходимые для индексирования
    - Сложно для большого числа документов
  - Формат инвертированного файла – для быстрой обработки запросов
    - Должен обрабатывать изменения индекса
    - Сжатие данных

Создание индекса-3

- Распределенное хранение индекса
  - Распределяет индексы по многим компьютерам и/или многим дата-центрам
  - Необходимо для быстрой обработки запросов
  - Много вариантов
    - Подокументное распределенное хранение, распределенное хранение термов, повтор данных

  Особое направление исследований: Distributed IR – распределенный информационный поиск

Подсистема 2. Обработка запроса



Обработка запроса

- Взаимодействие с пользователем
  - Поддерживает создание и уточнение запроса, показ результатов
- Ранжирование
  - Использует запрос и индексы породить ранжированный лист документов
- Оценка качества
  - Мониторит и измеряет качество поиска

Взаимодействие с пользователем

- Ввод запроса
  - Интерфейс и парсер для языка запросов
  - Большинство интернет запросов – простые

  - Язык запросов нужен для описания сложных запросов и результатов трансформации запросов (работа т.н. колдунщиков запросов)
    - Булевские запросы
    - Специализированные языки запросов для информационно-поисковых систем (Indri, Galago)
    - Сходны с SQL языками, используемыми в базах данных

Взаимодействие с пользователем-2

- Трансформация запросов
  - Улучшает исходный запрос
    - Спеллчекинг
    - Подсказка запроса
  - Автоматическое расширение запроса – пополнение его дополнительными словами
  - *Relevance feedback* - автоматизированная технология с участием пользователя
    - Пользователь размечает релевантные документы

Взаимодействие с пользователем-3

- Выдача результатов
  - Строит поисковую выдачу (SERP)
  - Порождает сниппеты, чтобы отразить соответствие документа запросу
  - *Подсвечивает важные слова*
  - Показывает релевантную рекламу – основной источник прибыли Интернет-поисковых систем
  - Может обеспечивать кластеризацию результатов и другие виды визуализации

Ранжирование

- Присваивание веса соответствия документа запросам
  - Веса использует алгоритм ранжирования
  - Базовый компонент поисковой машины
  - Базовое вычисление веса - на основе векторных представлений
    - Классические представления документов - вектора слов
    - Позже: машинное обучение и большое количество признако
  - Много вариантов алгоритмов вычисления весов и ранжирования

Ранжирование-2

- Оптимизация выполнения запроса
  - Ранжирующие алгоритмы должны позволять эффективное исполнение
- Распределенное выполнение
  - Обработка запросов в распределенной среде
  - *Брокер запросов рассылает запросы и собирает результаты*
  - *Кэширование*

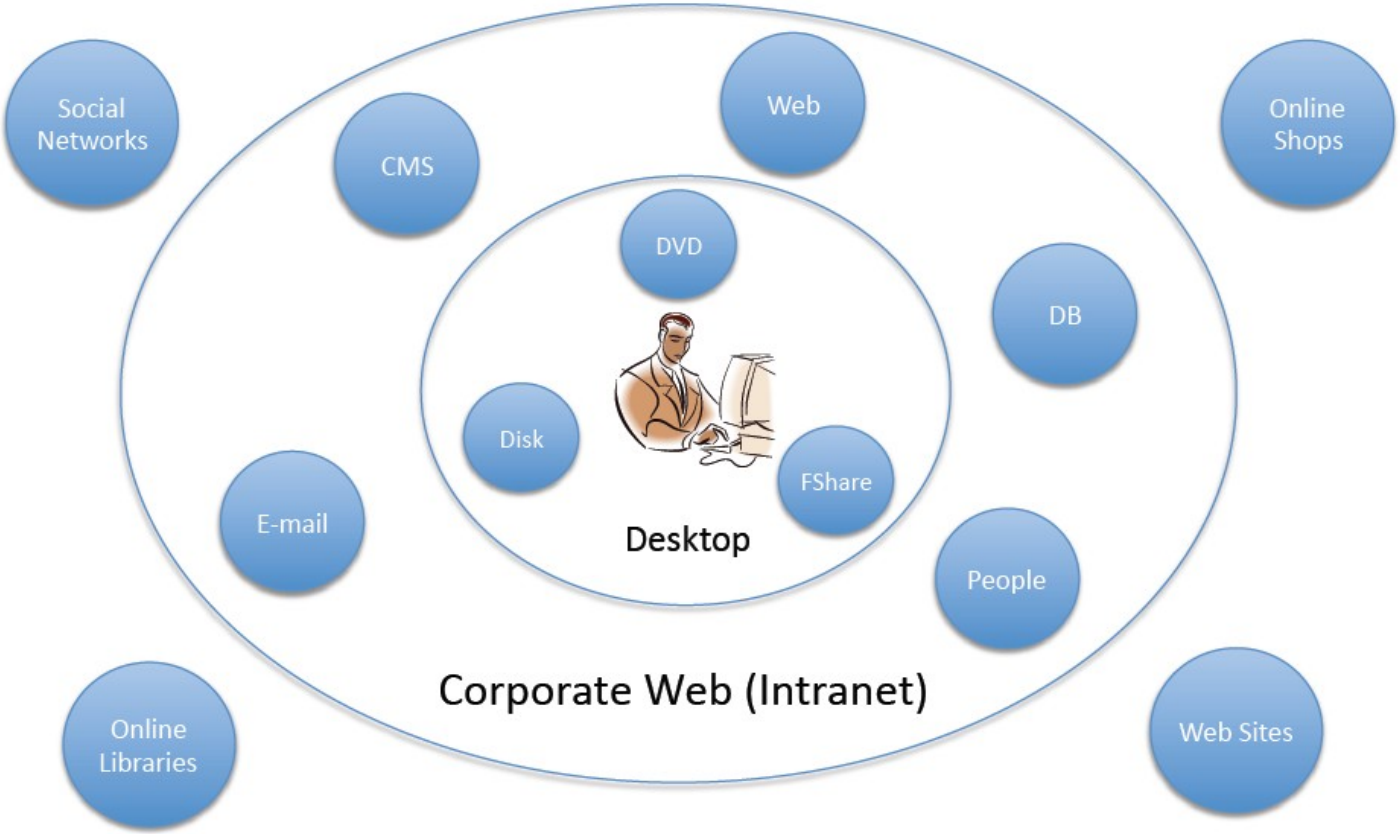
Поисковые системы разного уровня

Russir, 2009

Курс “Enterprise and Desktop Search”

(Дмитриев и др.)

Search Environment of a Company Employee



Интернет-поиск vs. Корпоративный поиск

- Интернет-поиск
  - Собирает результаты по общедоступному Интернету
  - Проблема ранжирования результатов
  - Большие объемы
  - Громадная индустрия – Интернет-реклама
  - Активные исследования: хорошее качество
- Корпоративный поиск
  - Собирает информацию разных форматов из совокупности хранилищ
  - Ранжирование документов разного типа
  - Относительно малый объем исследований
  - Хуже качество поиска – сложнее проводить сравнительные исследования
  - Активная сфера исследований

Различия между интернет-поиском и корпоративным поиском

- Краулинг, т.е. сбор информации
  - Много различных форматов, многие могут быть не для удобного просмотра в поиске
  - Сложные и длинные документы
  - Проблемы безопасности: многочисленные ограничения на просмотр и скачивание
  - Исследования: как создать исследовательскую коллекцию

Различия между интернет-поиском и корпоративным поиском

- Индексирование
  - Больше информации о документах, возможно полуструктурированные данные – эту информацию можно использовать для эффективного поиска
  - Релевантных документов может быть очень мало ->
    - Vocabulary mismatch - несовпадение между словами запроса и документов
    - Нужны специальные интерфейсы, подсказки для поиска нужного документа

Различия между интернет-поиском и корпоративным поиском

- Ранжирование
  - Может быть единственный правильный ответ, который нужно найти
  - Нет или мало ссылок, гипертекста
  - Часто нужно найти все релевантные документы
  - Возможно нужно собирать в единую выдачу документы из разных хранилищ (federation and blending)

Корпоративный поиск: Проблема измерения качества поиска

- Меры качества поиска, которые применяются к интернет-поиску, могут быть неприменимы к корпоративному поиску
- Поскольку нужно измерять:
  - Качество взаимодействия с пользователем, например удобные дополнительные инструменты поиска
  - Удачное завершение поиска
  - Удовлетворенность пользователя

Предметно-ориентированный поиск

(Domain-Specific Search)

Предметные области поиска

- Медицинский поиск
- Патентный поиск
- Поиск научных публикаций
- Поиск по законодательству
- Поиск по химическим документам
- …

Поиск научной литературы

- Рост публикаций
- Цитаты, которые можно использовать как ссылки в Интернет-поиске
  - Большое количество ссылок на работу – фактор значимости работы
  - Поиск близких работ
  - наукометрия

Анализ ссылок в литературе

- Количество ссылок на работу
  - Импакт-фактор работы
- Сходство работ на основе ссылок
  - Если работа A цитирует работы B и C, то, возможно, имеется сходство между работами B и C
  - Если работы B и C цитируют одну и ту же работу A, то возможно есть сходство между этими работами
  - Если таких совпадений много, то связь между работами

Системы поиска научной литературы

- CiteCeer
- ScienceDirect
- Google Scholar
- IEEE Xplore
- Scopus
- Microsoft Academic Search

Сервисы поиска научной информации

- Собирают научные публикации из различных источников: важно покрытие источников
- Обеспечивают индексирование и ранжированную выдачу
- Извлекают ссылки статей на другие статьи
  - Ссылки используются как дополнительный фактор ранжирования: авторитетные статьи обычно в начале списка
- Возможно искать по авторам, датам и т.п.
- Извлекается суммированная информация об авторах
  - Сколько статей, наукометрические индексы, кто ссылается на статьи
- Может работать сервис рекомендаций

Поиск научных публикаций в Google.Scholar

Google Академия

"domain specific information retrieval"

Статьи

Результатов: примерно 707 (0,09 сек.)

Мой профиль

Моя библ

а все время
2021
2020
2017
ыбрать даты

о релевантности
о дате

включая патенты
показать цитаты

Создать оповещение

Toward a semantic granularity model for domain-specific information retrieval [PDF] open.ac.uk

X Yan, RYK Lau, D Song, X Li, J Ma - ACM Transactions on Information ..., 2011 - dl.acm.org

Both similarity-based and popularity-based document ranking functions have been successfully applied to information retrieval (IR) in general. However, the dimension of semantic granularity also should be considered for effective retrieval. In this article, we ...

☆ Цитируется: 65 Похожие статьи Все версии статьи (11)

Concept-based document readability in domain specific information retrieval [PDF] worktribe.com

X Yan, D Song, X Li - Proceedings of the 15th ACM international ..., 2006 - dl.acm.org

Domain specific information retrieval has become in demand. Not only domain experts, but also average non-expert users are interested in searching domain specific (eg, medical and health) information from online resources. However, a typical problem to average users is ...

☆ Цитируется: 70 Похожие статьи Все версии статьи (7)

Using wikipedia and wiktionary in domain-specific information retrieval [PDF] psu.edu

C Müller, I Gurevych - Workshop of the Cross-Language Evaluation Forum ..., 2008 - Springer

The main objective of our experiments in the domain-specific track at CLEF 2008 is utilizing semantic knowledge from collaborative knowledge bases such as Wikipedia and Wiktionary to improve the effectiveness of information retrieval. While Wikipedia has already been used ...

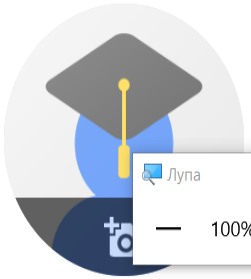
☆ Цитируется: 87 Похожие статьи Все версии статьи (19)

[HTML] Design and development of semantic web-based system for computer science **domain-specific information retrieval** [HTML] sciencedirect.com

R Bansal, S Chawla - Perspectives in Science, 2016 - Elsevier

In semantic web-based system, the concept of ontology is used to search results by

Наукометрические показатели



Наталья Лукашевич, Natalia Loukachevitch, Natalia Loukashevich, Natalia Lukashevich, Natalia

ПОДПИСАТЬСЯ

Лупа 100%

nosov Moscow State University

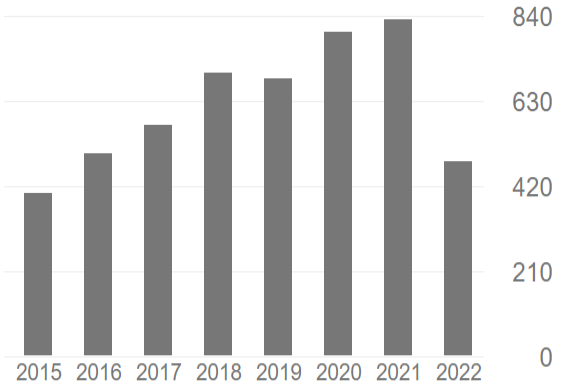
нной почты в домене mail.cir.ru

NLP

НАЗВАНИЕ	ПРОЦИТИРОВАНО	ГОД
Semeval-2016 task 5: Aspect based sentiment analysis M Pontiki, D Galanis, H Papageorgiou, I Androutsopoulos, S Manandhar, ... International workshop on semantic evaluation, 19-30	1915 *	2016
Онтологии и тезаурусы. Учебное пособие ВД Соловьев, БВ Добров, ВВ Иванов, НВ Лукашевич Казань, Москва	486 *	2006
Автоматическая обработка текстов на естественном языке и анализ данных Е Большакова, К Воронцов, Н Ефремова, Э Клышинский, ...	431	2017
Автоматическая обработка текстов на естественном языке и компьютерная лингвистика ЕИ Большакова, ОВ Пескова, ЭС Клышинский, АА Носков, ДВ Ландэ, ...	~~431~~	2015
Тезаурусы в задачах информационного поиска НВ Лукашевич М.: Издательство МГУ, 2011	411 *	2010
Sentiment Analysis Track at ROMIP 2011//Компьютерная	123 *	2012

Процитировано ПРОСМОТРЕТЬ ВСЕ

	Все	Начиная с 2017 г.
Статистика цитирования	6549	4075
h-индекс	30	23
i10-индекс	104	54



Общий доступ ПРОСМОТРЕТЬ ВСЕ

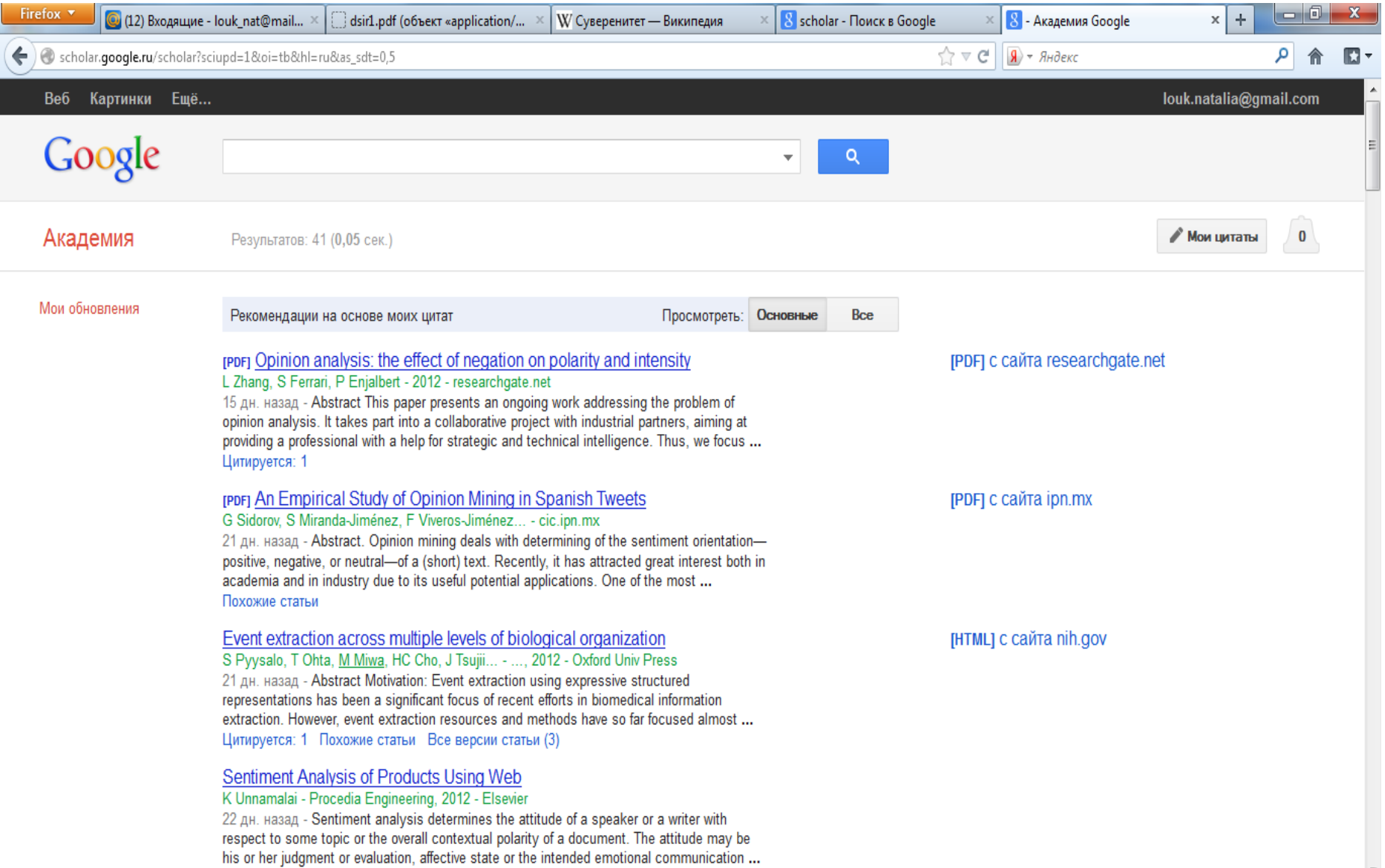
0 статей 4 статьи

недоступно доступно

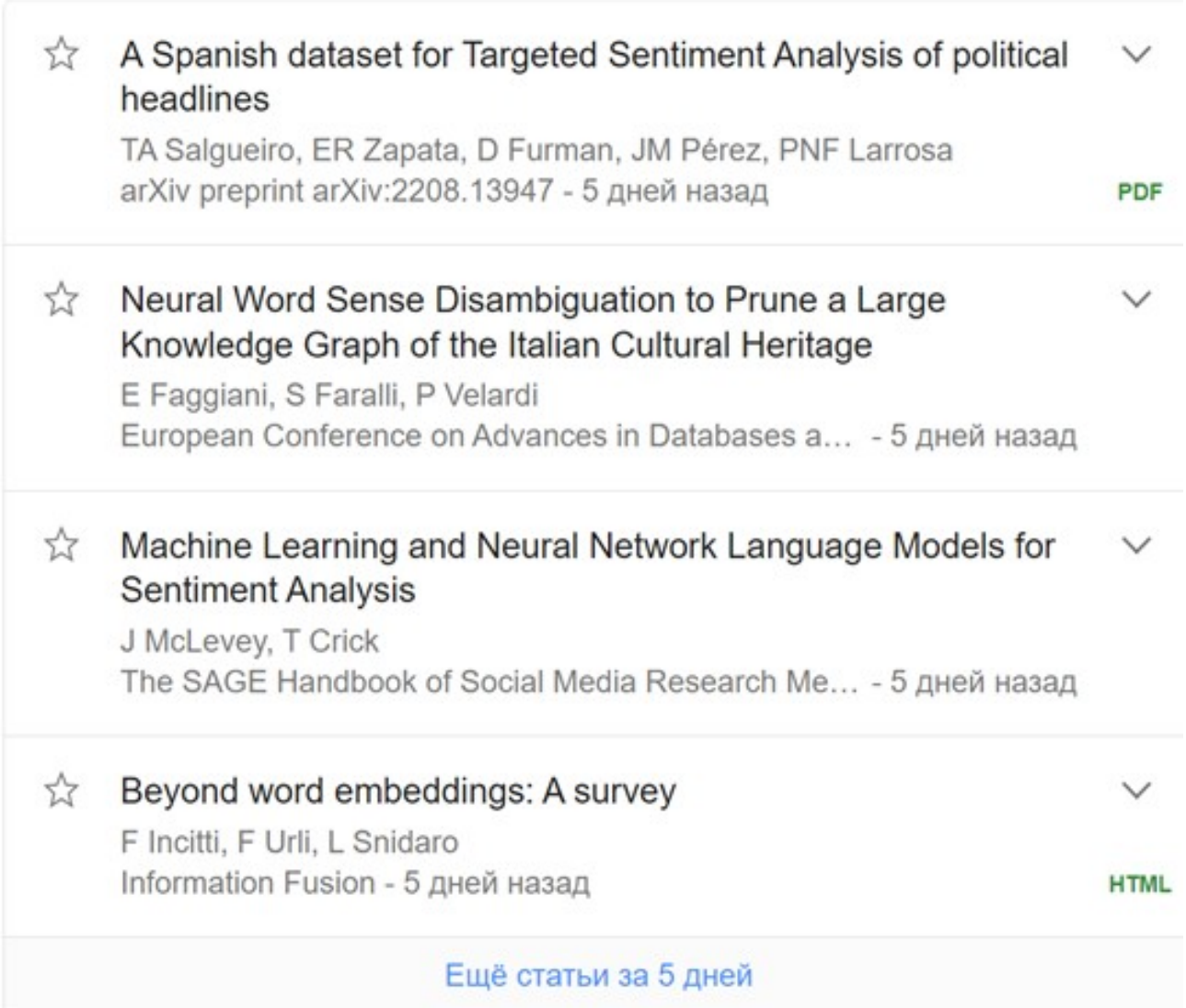
На основе финансирования

Соавторы ИЗМЕНИТЬ

Рекомендуемые исследователю работы, на основе его последних публикаций



Рекомендуемые исследователю работы, на основе его последних публикаций



Заключение

- Архитектура поисковой машины
  - Подсистема индексирования документов
  - Подсистема взаимодействия с пользователем и поиска документов на запрос
- Поисковые системы разного уровня и предметной направленности
  - Интернет-поиск
  - Корпоративный поиск
  - Поиск в специальной предметной области