Семинар Информационные системы в Интернет
и базы данных

Дмитрий Дмитриевич Козлов
к.ф.-м.н., м.н.с ЛВК
ddk@cs.msu.su

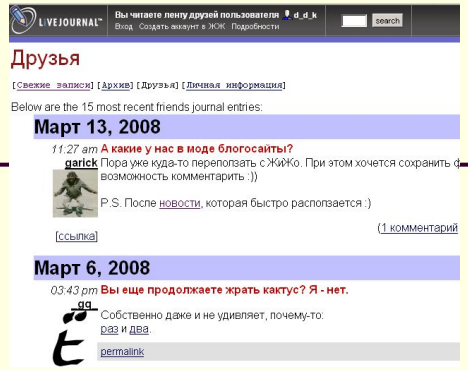
Чем мы занимаемся

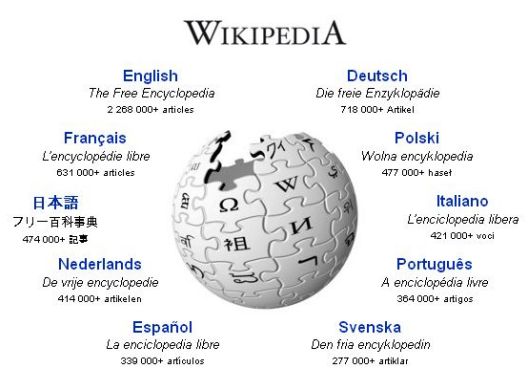
- Информационный поиск в сети Интернет
  - Тематический поиск в сети Интернет.
  - Периодический тематический поиск.
  - Автоматический поиск научных статей в сети Интернет.
- Извлечение информации из слабоструктурированных текстов
  - Применение методов машинного обучения для извлечения информации из текстов русскоязычных научных статей.
  - Извлечение библиографических ссылок из текстов web-страниц.
- Вопросы безопасности web-приложений
  - Обнаружение уязвимостей в web-приложениях, написанных скриптовых языках.
  - Статический анализ безопасности программ на скриптовых языках.
- Базы данных (Александр Чупров)

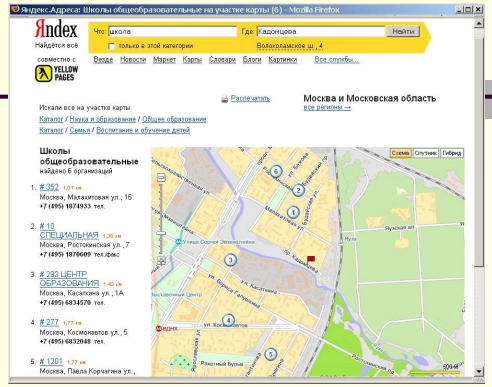
Часть первая

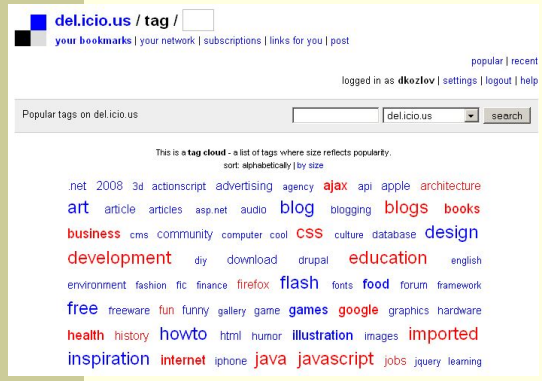
Информационный поиск
в сети Интернет и

извлечение информации из
слабоструктурированных текстов

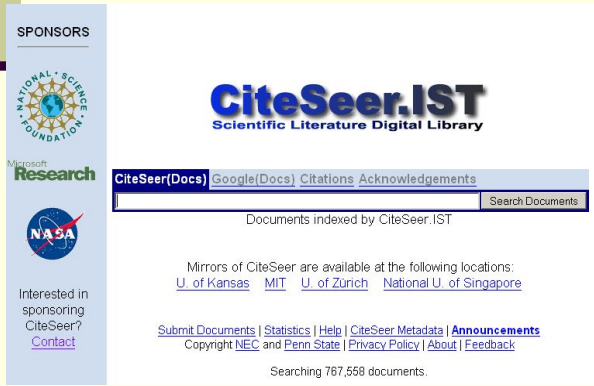


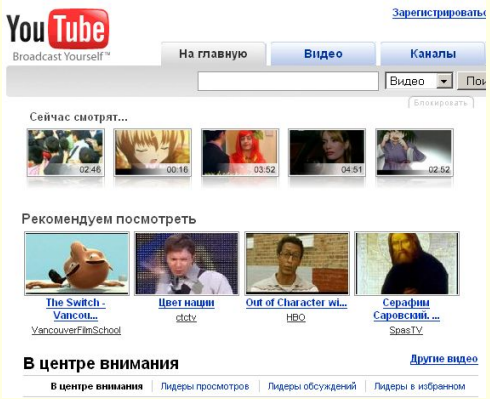


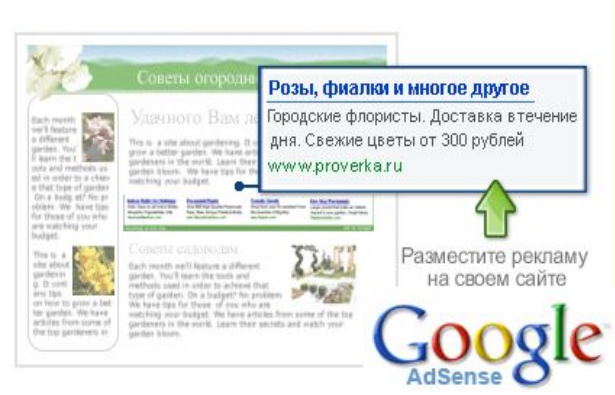




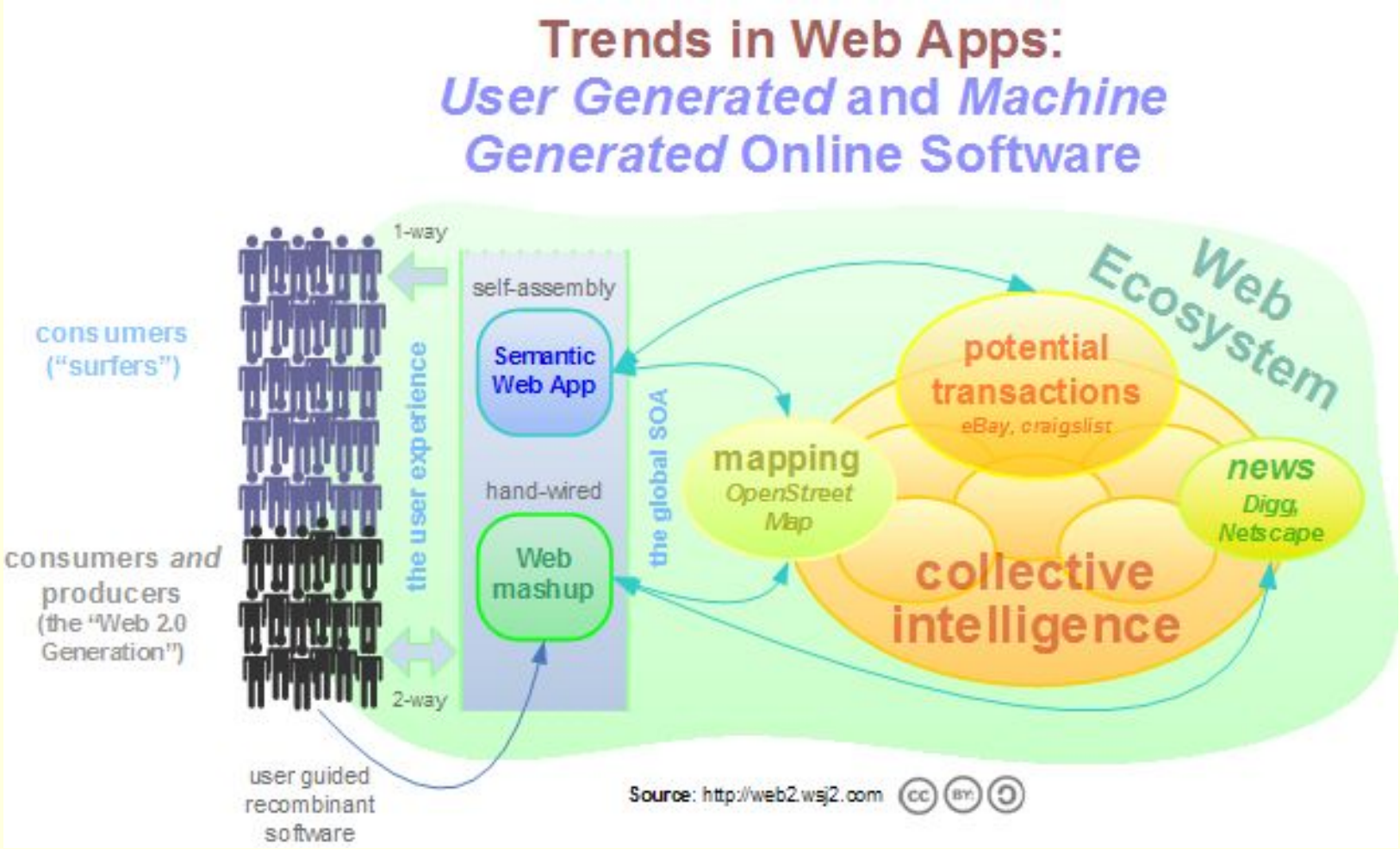
Сегодня сеть Интернет – это среда обитания





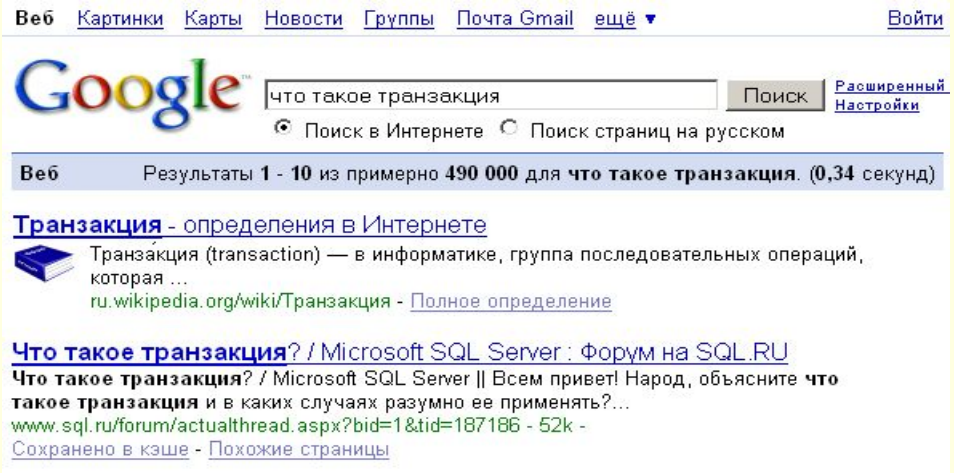


Тенденция развития



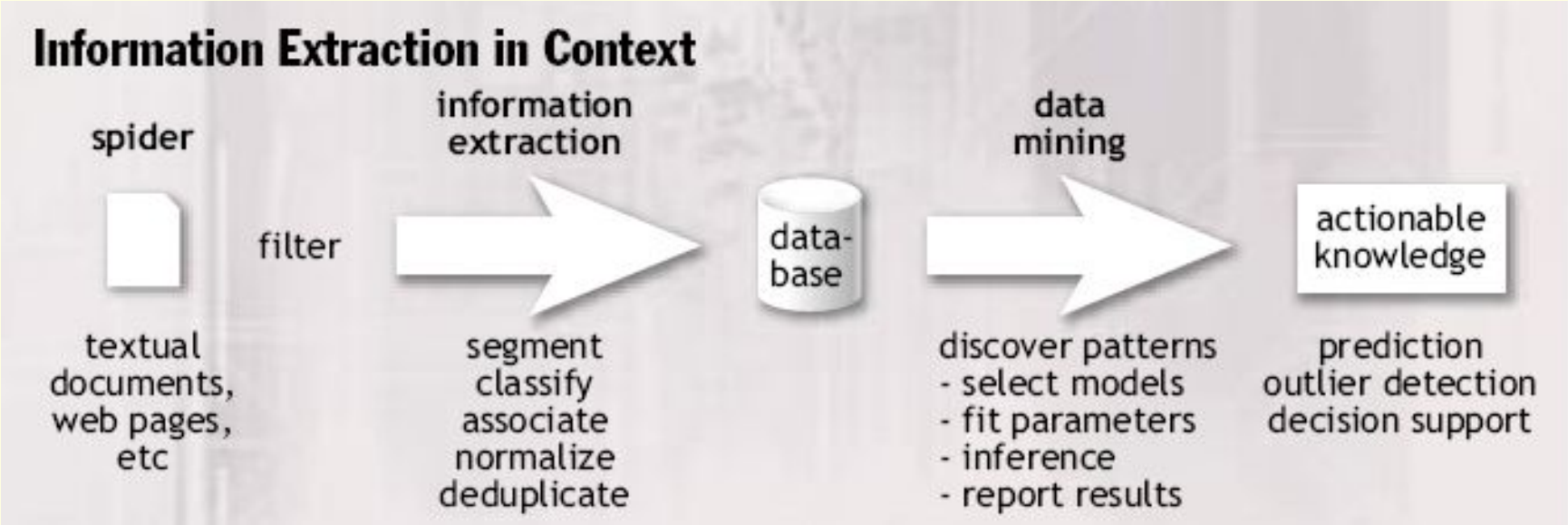
Информационный поиск

- **Вчера:** select * from documents where doc_title contains «сингулярный» or «сингулярное»
- **Сегодня:**



- **Завтра:** Find a suitable wine for every item in this menu. If possible, choose French

Извлечение метаинформации – неотъемлемая часть поиска



Накопление информации о пользователях

ddk@icq.com
История общения,
список контактов
ddk@gmail.com
История поисковых запросов,
Web-страницы, которые я читаю
ddk@yandex.ru
Список RSS-каналов,
адреса, которые я ищу
ddk@livejournal.com
Мои друзья, сообщества,
Темы, которые
меня интересуют
. . .

Персонифицированная реклама

Google AdSense

И не только она …

Актуальные задачи

- Развитие технологий поиска
  - Тематический поиск, как помощь классическим системам поиска по ключевым словам
  - Извлечение информации из накопленных web-страниц
  - Вопросно-ответные системы (фактографический поиск)
  - Семантический поиск
- Персонификация поиска
  - Создание информационного портрета пользователя
  - Поиск с учетом особенностей пользователя
  - Периодический тематический поиск (персональная газета, Push-технологии в блогосфере)

Актуальные задачи (2)

- Целенаправленная реклама
  - Создание информационного портрета пользователя
  - Реклама в блогах и социальных сетях
  - Персонализованная реклама
- «Коллективный разум»
  - Фолксономии (folksonomies)
  - Автоматическая классификация ресурсов Интернет на основе фолксономий, automatic labelling.

Часть вторая - познавательная

Как пользоваться поисковыми системами на примере поиска научных статей

Поиск научной информации

- Информационная потребность пользователя: хочу обзор исследований по методам извлечения метаданных из web-страниц.
  - Автоматическое выполнение (а вдруг вы придумаете такой метод в своей курсовой): семантический поиск
  - Выполнение вручную: тематический поиск (пока только так, вручную)

Поиск научной информации (2)

- Нам нужны научные работы - давайте посмотрим в CiteSeer и Google Scolar.

Кстати, тут и домашняя страничка автора

Слишком старая работа? Давайте посмотрим кто на нее ссылается

Wrapper Induction for Information Extraction (1997) (Make Corrections) (228 citations)
Nicholas Kushmerick, Daniel S. Weld, Robert Doorenbos
Intl. Joint Conference on Artificial Intelligence (IJCAI)

Bookmark in CiteULike

CiteSeer Home/Search Context Related

View or download:
cs.ucd.ie/staff/ni...rickijcai97.ps.gz
washington.edu/pub...erickijcai97.ps.Z
washington.edu/hom...merickijcai97.pdf
Cached: PS.gz PS PDF
Image Update Help

From: cs.ucd.ie/staff/nick/research/... (more)
From: washington.edu/homes/weld/pubs
(Enter author homepages)

(Enter summary)

Rate this article: 1 2 3 4 5 (best)
Comment on this article

Abstract: Many Internet information resources present relational data---telephone directories, product catalogs, etc. Because these sites are formatted for people, mechanically extracting their content is difficult. Systems using such resources typically use hand-coded wrappers, procedures to extract data from information resources. We introduce wrapper induction, a method for automatically constructing wrappers, and identify hlrt, a wrapper class that is efficiently learnable, yet expressive enough to... (Update)

Cited by: More
Hierarchical Wrapper Induction for Semistructured.. - Ion Muslea Steven (Correct)
Extraction Techniques for Mining Services from Web Sources - Hasan Davulcu Saikat (Correct)
Thresher: Automating the Unwrapping of Semantic - Content From The (2005) (Correct)

Active bibliography (related documents): More All
0.4: Programming by Demonstration: a Machine Learning Approach - Lau (2001) (Correct)
0.4: Programming By Demonstration Using Version Space Algebra - Lau, Wolfman, Domingos, Weld (2000) (Correct)
0.3: Wrapper Induction: Efficiency and Expressiveness - Kushmerick (2000) (Correct)

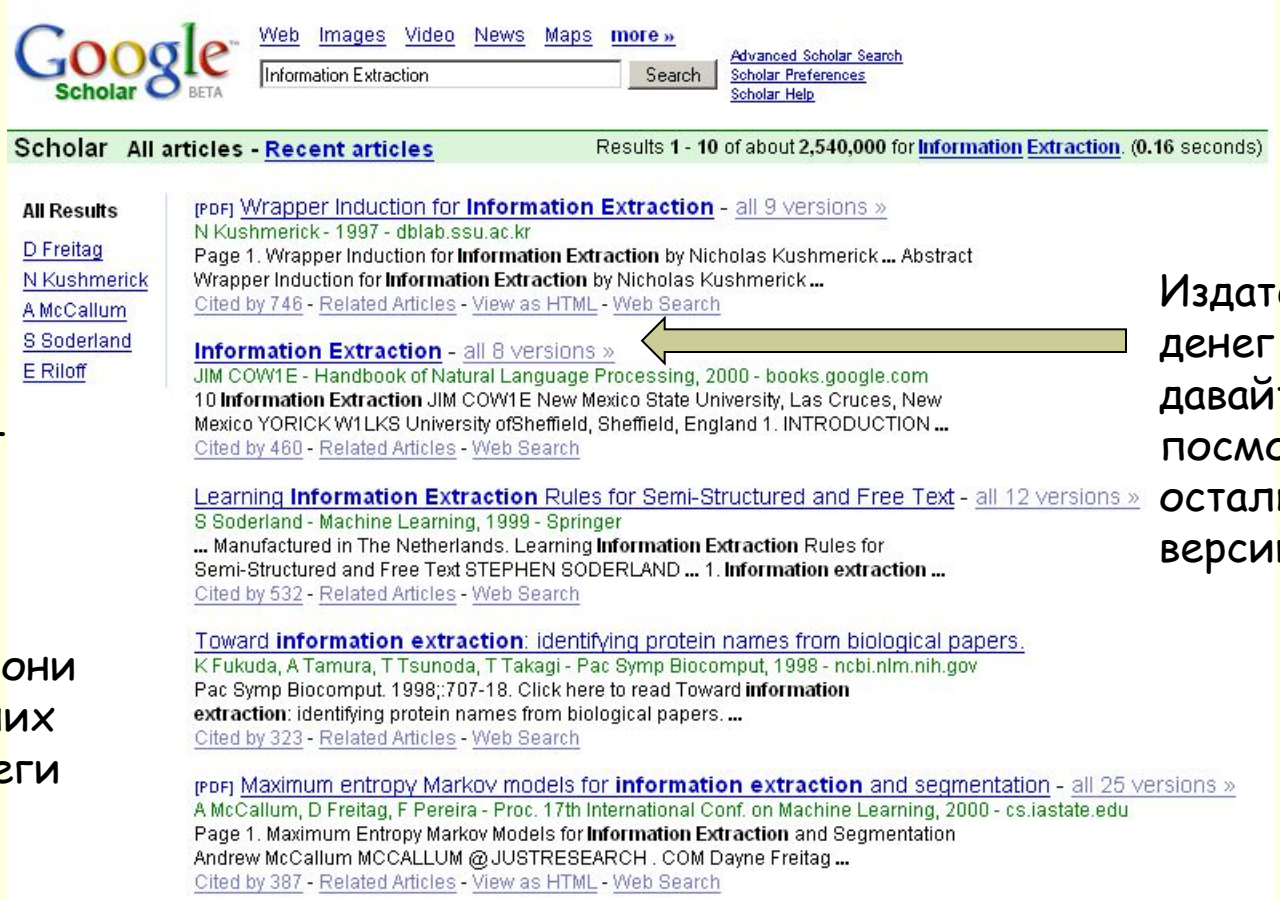
Similar documents based on text: More All
0.4: Gleaning Answers From the Web - Kushmerick (Correct)
0.3: Wrapper induction: Efficiency and expressiveness (Extended.. - Kushmerick (1998) (Correct)
0.3: Wrapper Induction for Information Extraction - Kushmerick (1997) (Correct)

Может стоит поискать похожие работы?

В статье есть обзор существу-ющих работ а в нем – библиографи-ческие ссылки. Давайте поищем и эти статьи тоже

Поиск научной информации (3)

- Нам нужны научные работы - давайте посмотрим в CiteSeer и Google Scolar.



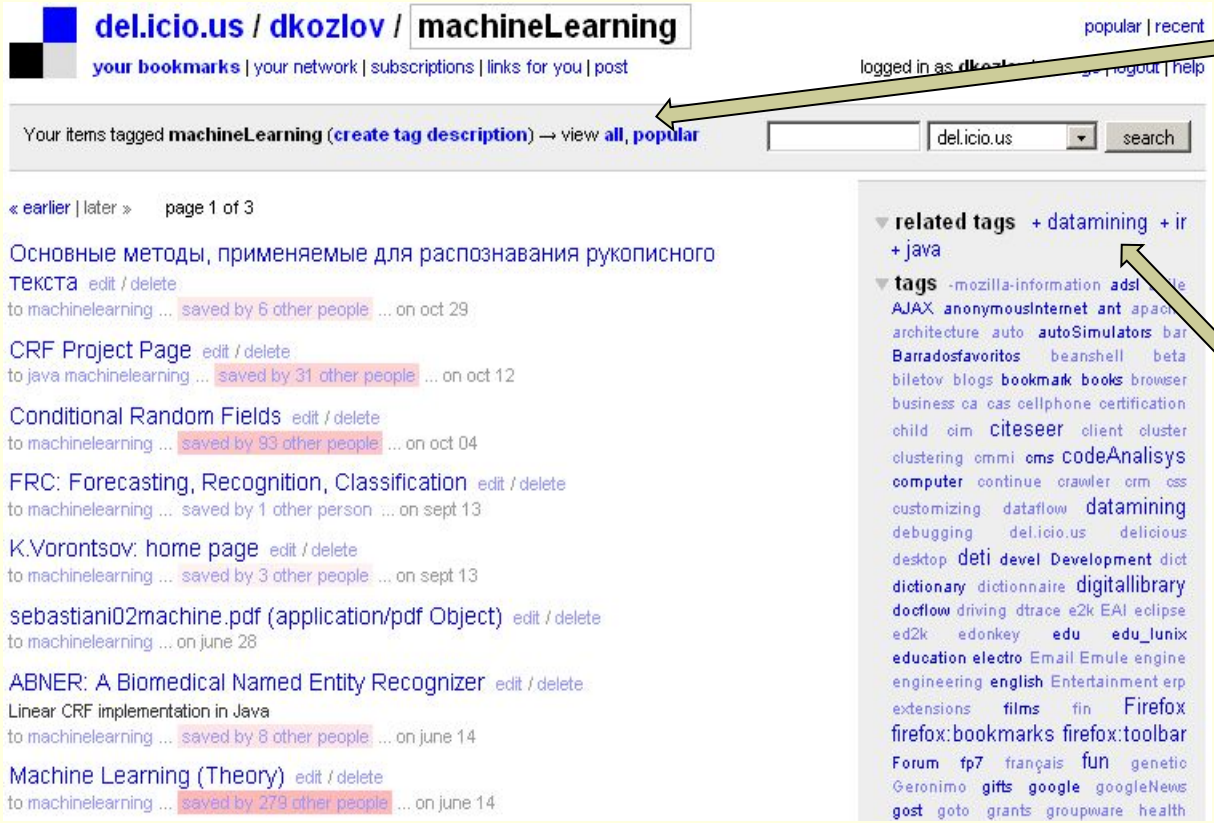
Наиболее популярные авторы, может поищем их домашние страницы.

А, кстати, где они работают – у них может и коллеги есть

Издатель хочет денег за статью, давайте посмотрим остальные 8 версий

Поиск научной информации (4)

- Вы думаете, вы одиноки?

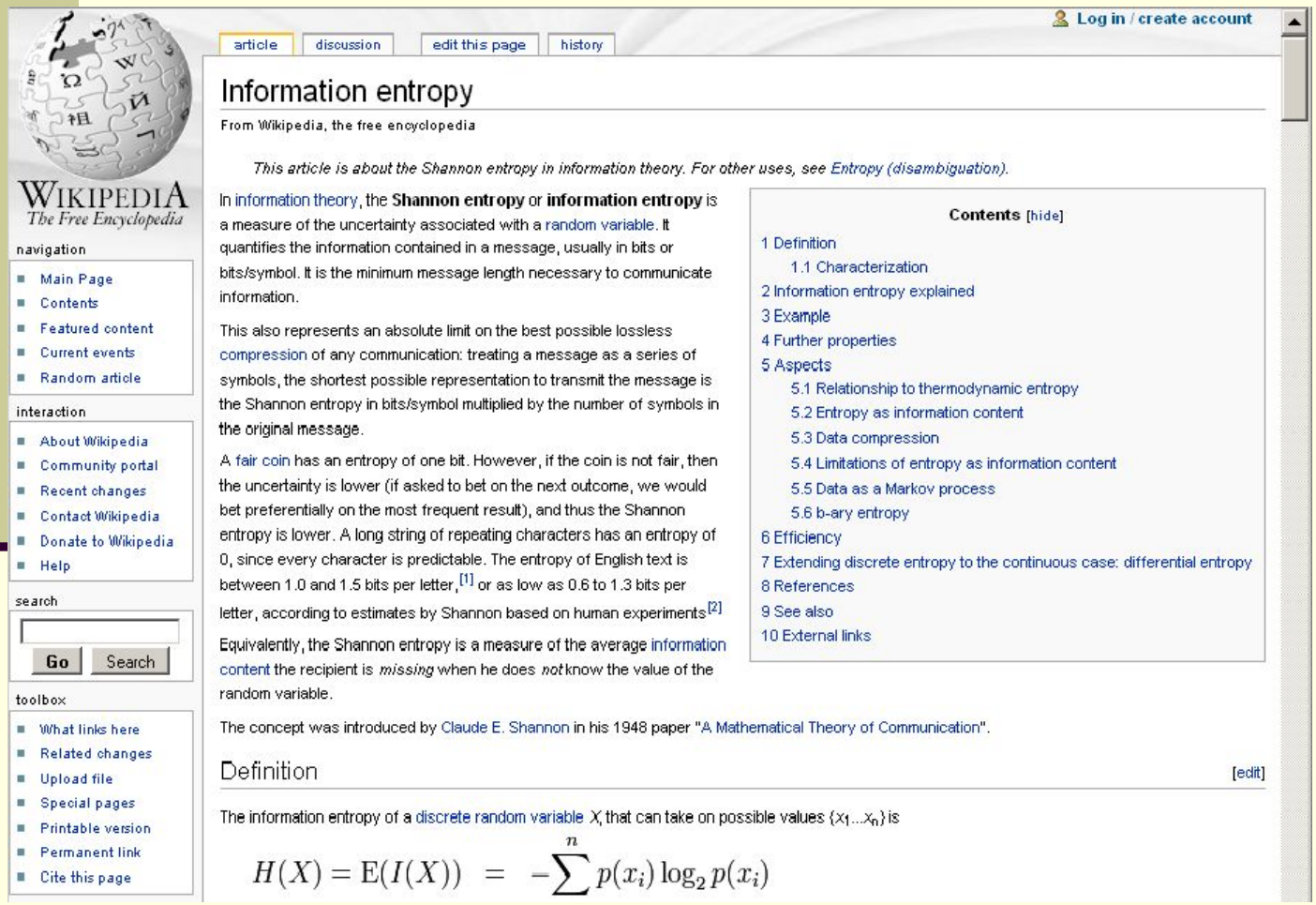


Не вы одни интересуетесь этой темой, и многие уже нашли

Похожие тэги

Поиск научной информации (5)

- Вы не понимаете этих слов? Не страшно!



article | discussion | edit this page | history | Log in / create account

WIKIPEDIA
The Free Encyclopedia

navigation
- Main Page
- Contents
- Featured content
- Current events
- Random article

interaction
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

search
Go | Search

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

Information entropy

From Wikipedia, the free encyclopedia

This article is about the Shannon entropy in information theory. For other uses, see Entropy (disambiguation).

In information theory, the **Shannon entropy** or **information entropy** is a measure of the uncertainty associated with a random variable. It quantifies the information contained in a message, usually in bits or bits/symbol. It is the minimum message length necessary to communicate information.

This also represents an absolute limit on the best possible lossless compression of any communication: treating a message as a series of symbols, the shortest possible representation to transmit the message is the Shannon entropy in bits/symbol multiplied by the number of symbols in the original message.

A fair coin has an entropy of one bit. However, if the coin is not fair, then the uncertainty is lower (if asked to bet on the next outcome, we would bet preferentially on the most frequent result), and thus the Shannon entropy is lower. A long string of repeating characters has an entropy of 0, since every character is predictable. The entropy of English text is between 1.0 and 1.5 bits per letter,[1] or as low as 0.6 to 1.3 bits per letter, according to estimates by Shannon based on human experiments[2]

Equivalently, the Shannon entropy is a measure of the average information content the recipient is *missing* when he does *not* know the value of the random variable.

Contents [hide]

1 Definition
1.1 Characterization
2 Information entropy explained
3 Example
4 Further properties
5 Aspects
5.1 Relationship to thermodynamic entropy
5.2 Entropy as information content
5.3 Data compression
5.4 Limitations of entropy as information content
5.5 Data as a Markov process
5.6 b-ary entropy
6 Efficiency
7 Extending discrete entropy to the continuous case: differential entropy
8 References
9 See also
10 External links

The concept was introduced by Claude E. Shannon in his 1948 paper "A Mathematical Theory of Communication".

Definition [edit]

The information entropy of a discrete random variable X, that can take on possible values $\{x_1 \ldots x_n\}$ is

$$H(X) = \mathrm{E}(I(X)) \quad = \quad -\sum^{n} p(x_i) \log_2 p(x_i)$$

В энциклопедии Wikipedia.org можно об этом прочитать.

Э-э-э, как это будет по-русски ?

Gramota.ru,
Multitran.ru
Slovari.yandex.ru

помогут вам не пугать коллег орфографическими ошибками

Поиск научной информации (6)

- Я уже все нашел и понял, вот как бы написать…

В хорошей научной работе всегда есть ЦЕЛЬ, а также почти всегда:

- аннотация
- введение
- постановка задачи
- обзор существующих методов
- изложение результатов, полученных авторами
- исследование/обоснование результатов
- заключение в результатами работы и выводами
- список литературы

Использование интеллектуальных сетевых роботов для построения тематических коллекций*

Романова Е.В., Романов М.В., Некрестьянов И.С.
Санкт-Петербургский Государственный Университет, Санкт-Петербург.
emails: katya@tepkom.ru, rmv@sparc.spb.su, igor@meta.math.spbu.ru

Abstract

В работе рассматривается задача создания интеллектуального сетевого робота для сбора тематических коллекций. Для повышения производительности обнаружения тематических ресурсов используется специализированный алгоритм обхода сети, учитывающий информацию о тематической релевантности уже посещенных страниц. Робот также производит грубый отсев "мусора" среди посещенных документов, для того чтобы повысить качество рекомендаций.

1 Введение

В течение ряда лет вопросы создания и применения сетевых роботов привлекают все больше внимания [8, 10, 13, 11]. Сетевой робот или *Crawler* — это программа, которая, начиная с некоторой Интернет-страницы, рекурсивно обходит ресурсы Интернет, извлекая ссылки на новые ресурсы из получаемых документов.

Классической областью применения сетевых роботов является построение индексов Интернет-ресурсов для поисковых систем [14, 3, 5, 15]. Однако в последнее время сетевые роботы используются для выполнения множества других задач — сбора статистики, поиска определенных ресурсов сети (например, домашних страниц), проверки целостности существующих гипертекстовых ссылок, и т.п. Разработаны даже соответствующие правила "вежливого" поведения для сетевых роботов — Standard for Robot Exclusion и Rapid Fire Requests. Текущий вариант списка добровольно зарегистрированных роботов на странице info.webcrawler.com содержит более сотни позиций, а общее число существующих сетевых роботов по некоторым оценкам превышает десятки тысяч.

Большинство сетевых роботов посещают огромное количество Internet страниц, индексируя все полученные документы. Очевидно, что такой подход требует значительных сетевых и аппаратных ресурсов. Однако текущий объем доступной информации в Интернет оценивается в 6 терабайт и быстро растет, поэтому даже самый мощный сетевой робот не может посетить все Интернет-страницы.

Поскольку посещение всех Интернет-страниц не представляется возможным, то разумно посещать в первую очередь наиболее важные из них. Простейший критерий важности, используемый многими из современных сетевых роботов собирающими информацию для популярных поисковых систем, является глубина URL, т.е. количество промежуточных каталогов упоминающихся в URL между именем Интернет-узла и именем самого ресурса. *Чем больше глубина, тем ниже важность соответствующего ресурса. Подобный подход позволяет быстро посетить стартовые и близкие к ним страницы на большом числе Интернет-узлов.

7 Заключение

В работе рассматривается задача создания интеллектуального сетевого робота для сбора тематических коллекций.

Описана базовая архитектура системы, структура тематического фильтра и методы оценки тематической релевантности документа. Использование дополнительной информации от клиента робота во время работы для уточнения тематического фильтра позволяет улучшать качество оценок в процессе работы. Описываемая стратегия обхода сети учитывает тематические оценки уже посещенных документов, что позволяет посетить тематически релевантные документы в первую очередь.

Предварительные результаты экспериментов показывают преимущество тематически-ориентированной стратегии обхода над другими стратегиями для сбора тематических коллекций. Все это подтверждает перспективность предлагаемого подхода.

Отметим, что проблема построения тематических коллекций не является специфичной для проекта OASIS и актуальна во многих других задачах информационного рынка, например, таких как построение тематических

Библиография

[1] I.J. Aalbersberg. Incremental relevance feedback. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–22, 1992.

*Эта работа была выполнена в рамках проекта Open Architecture Server for Information Search and Delivery (OASIS) и поддержана грантом Европейской комиссии (INCO Copernicus Programme Project PL 961116).

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

Как писать хорошую английскую прозу:

- знайте, что хотите сказать,
- подражайте классикам.

Часть третья

Безопасность информационных систем в сети Интернет

Актуальность

- Интернет – среда обитания
  - Торговля – можно номер вашей кредитки
  - Интернет-банк – а вы мне не переведете пару (сотен, тысяч …) долларов
  - Оплата услуг – оплатите и мой водопровод тоже
  - Электронные медиа-издания – а я хочу бесплатно посмотреть «Терминатор 6»
  - Мобильный офис – у конкурента в почте интересный финансовый отчет, отнесу-ка я его в налоговую
  - Privacy – сегодня начался новый призыв, а у вас в ЖЖ написано встречаемся в кафе в 7. Вот и прокатимся… в военкомат.

Чем занимаемся мы: предыстория

- Сегодня большинство информационных систем работают в сети Интернет. Они и их пользователи могут быть атакованы.
- 80% создаваемых web-приложений уязвимы.
- Один из способов предотвращения атак – обнаружить и исправить уязвимости.
- Самый эффективный способ обнаружения уязвимостей – code review. Но человек может хотеть спать, плохо себя чувствовать, работать медленно. «Очень хотелось спать, когда я вычитывал код управления ядерным реактором».

Чем занимаемся мы

- Разрабатываем методы и средства автоматизированного обнаружения уязвимостей.
- Тестирование на проникновение (исследование работающего web-приложения, без его исходных кодов)
- Динамический анализ исходных кодов программ (исследование работающего web-приложения с учетом доступности исходных кодов)
- Статический анализ исходных кодов программ (исследование исходных кодов еще не дописанного web-приложения)

Часть четвертая

Примерные темы курсовых работ на следующий год

Примерные темы работ

- Извлечение метаинформации и библиографических ссылок, находящихся внутри текста статьи
- Кластеризация результатов информационного поиска, поиск тематических сообществ
- Идентификация личности в социальных сетях
- Обход web-приложений (автозаполнение форм) поисковым роботом

это еще не все

Заключение: об учебе

«Западная» модель научной работы студентов:

- На каждый из трех курсов дается своя задача.
- Вы учитесь в процессе выполнения научной работы совместно с научным руководителем. Он – старший товарищ, он не заставляет, а может лишь помочь.
- По итогам каждого года вы должны сделать научную статью и выступить с докладом на конференции.
- С каждым годом все больше самостоятельности и ответственности.

До встречи на собеседовании

Вопросы Козлову Дмитрию Дмитриевичу можно задавать по электронной почте ddk @ cs . Msu . su или очно в к. 764