

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

LIVEJOURNAL™ Вы читаете ленту друзей пользователя: [d_d_k](#)
 Вход, создать журнал в ЖЖ, Подписки

Друзья
 [Свежие записи] [Друзья] [Друзья] [Дневная информация]

Below are the 15 most recent friends journal entries:

Март 13, 2008
 11:27 am **А какие у нас в моде блогсайты?**
 garick Пора уже куда-то переползать с ЖЖко. При этом хочется сохранить возможность комментарить :))
 P.S. После [новости](#), которая быстро расплывается :)
 [ссылка] (1 комментарий)

Март 6, 2008
 03:43 pm **Вы еще продолжаете жрать кактус? Я - нет.**
 Собственно даже и не удивляет, почему-то раз и два
[permalink](#)

WIKIPEDIA

English
The Free Encyclopedia
 2 268 000+ articles

Deutsch
Die freie Enzyklopädie
 718 000+ Artikel

Français
L'encyclopédie libre
 631 000+ articles

Polski
Wolna encyklopedia
 477 000+ haseł

日本語
 フリー百科事典
 474 000+ 記事


Italiano
L'enciclopedia libera
 421 000+ voci

Nederlands
De vrije encyclopedie
 414 000+ artikelen

Português
A enciclopédia livre
 364 000+ artigos

Español
La enciclopedia libre
 339 000+ artículos

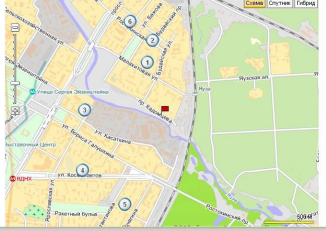
Svenska
Den fria encyklopedin
 277 000+ artiklar



Яндекс
 Что: школа Где: Калужская Область
 Найти все Только в этой категории Вспомогательные и др.
 Везде Новосты Маркет Карты Сайты Блоги Картины Disc...
 Помощь

Искать все на карте
 Калужская область / Москва и Московская область
 Калужь / Смолень / Восточная облужь / Южная облужь

Школы общеобразовательные
 найдено 6 организаций



- # 352 1.01 км Москва, Малаховская ул., 15
+7 495 987 8333 тел.
- # 18 СТЕПАНОВКА 1.01 км Москва, Рязанская ул., 7
+7 495 9879609 тел.бюс.
- # 283 ЦЕНТР ОБРАЗОВАНИЯ 1.40 км Москва, Космодемьян ул., 1А
+7 495 6634570 тел.
- # 277 1.77 км Москва, Космодемьян ул., 5
+7 495 6632048 тел.
- # 1261 1.77 км Москва, Паша Коржанин ул.,

del.icio.us / tag /
 your bookmarks | your network | subscriptions | links for you | post
 popular | recent
 logged in as [dkozlov](#) | settings | logout | help

Popular tags on del.icio.us

This is a tag cloud - a list of tags where size reflects popularity.
 sort: alphabetically | by size


net 2008 3d actionscript advertising agency **ajax** api apple architecture
 art article articles asp.net audio **blog** blogging **blogs** books
 business cms community computer cool **css** culture database **design**
 development diy download drupal **education** english
 environment fashion fic finance **firefox** **flash** fonts **food** forum framework
 free freeware fun funny gallery game **games** **google** graphics hardware
 health history **howto** html humor **illustration** images **imported**
 inspiration **internet** iphone **java** **javascript** jobs jquery learning


Сегодня сеть Интернет - это среда обитания



одноклассники.ru

SPONSORS

 **CiteSeer.IST**
 Scientific Literature Digital Library

Microsoft Research 

Mirrors of CiteSeer are available at the following locations:
[U. of Kansas](#) [MIT](#) [U. of Zurich](#) [National U. of Singapore](#)

Interested in sponsoring CiteSeer?
[Contact](#)

Searching 767,558 documents.

YouTube Broadcast Yourself™

Зарегистрироваться

На главную Видео Каналы

Видео Поиск

Сейчас смотрят...

Рекомендуем посмотреть

В центре внимания Другие видео

В центре внимания | Лидеры просмотров | Лидеры обсуждений | Лидеры в избранном

Советы огорода

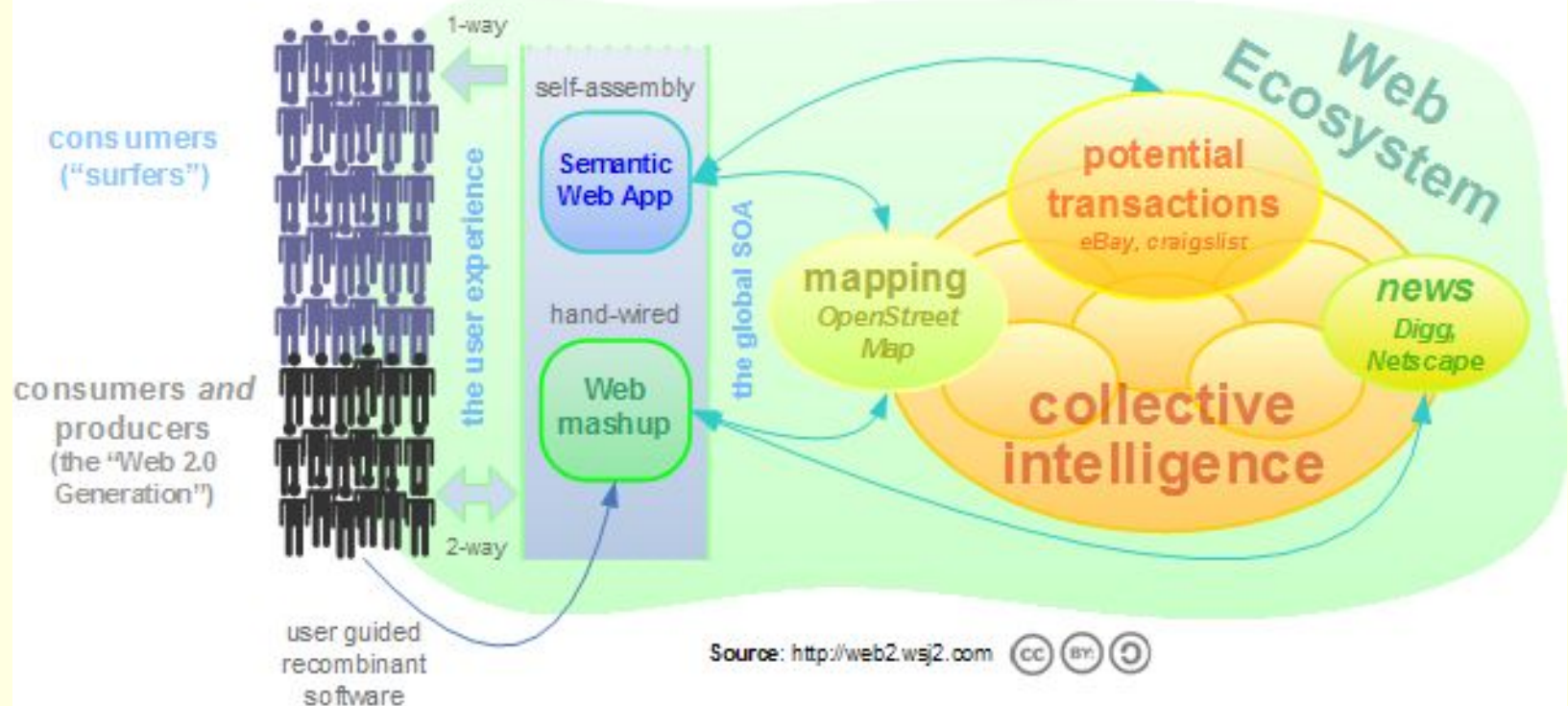
Розы, фиалки и многое другое
 Городские флористы. Доставка в течение дня. Свежие цветы от 300 рублей
[www.proverka.ru](#)

Разместите рекламу на своем сайте

Google AdSense

Тенденция развития

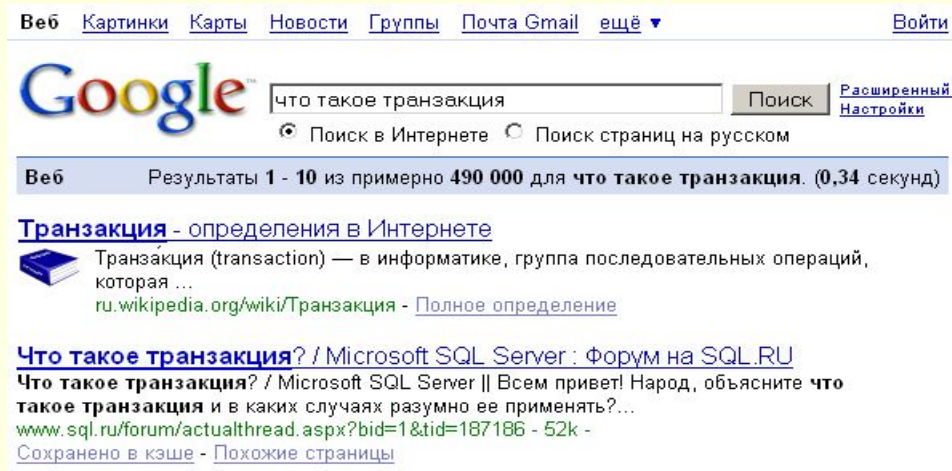
Trends in Web Apps: User Generated and Machine Generated Online Software



Информационный поиск

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

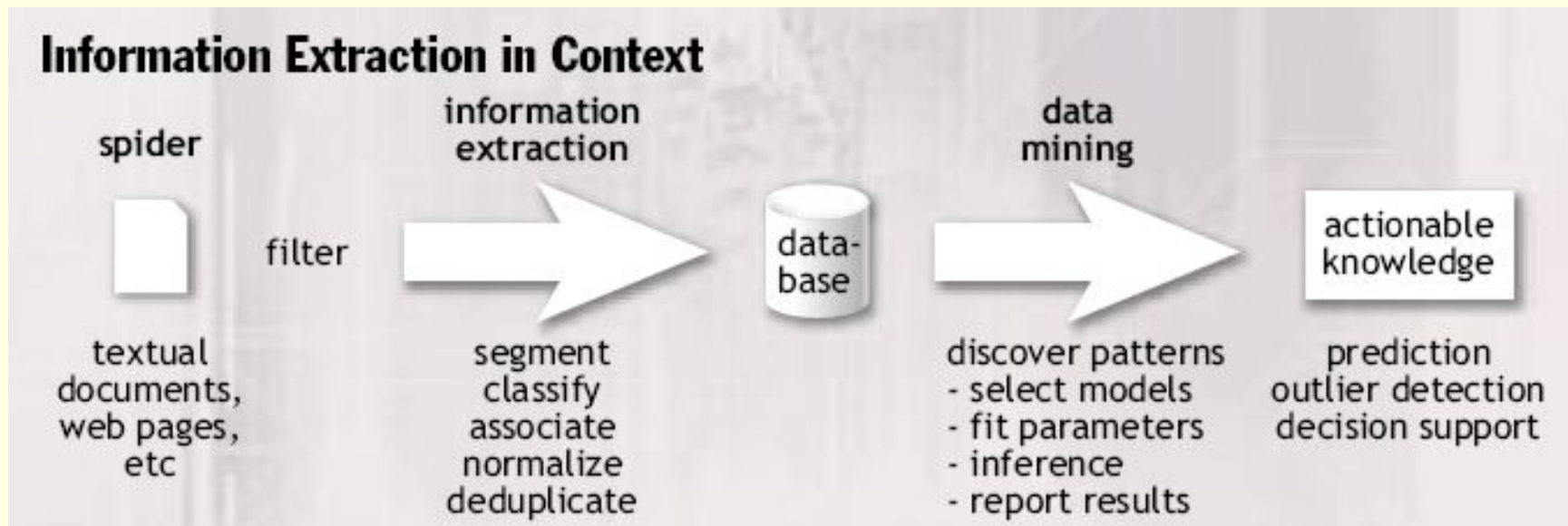
- Сегодня:



The screenshot shows a Google search interface. At the top, there are navigation links: Веб, Картинки, Карты, Новости, Группы, Почта Gmail, ещё ▾, and Войти. The Google logo is on the left, and the search bar contains the text "что такое транзакция". To the right of the search bar is a "Поиск" button and a link to "Расширенный Настройки". Below the search bar, there are radio buttons for "Поиск в Интернете" (selected) and "Поиск страниц на русском". The search results section shows "Веб" and "Результаты 1 - 10 из примерно 490 000 для что такое транзакция. (0,34 секунд)". The first result is titled "Транзакция - определения в Интернете" and includes a small blue book icon. The text of the result says "Транзакция (transaction) — в информатике, группа последовательных операций, которая ..." and provides a link to "ru.wikipedia.org/wiki/Транзакция - Полное определение". The second result is titled "Что такое транзакция? / Microsoft SQL Server : Форум на SQL.RU" and includes the text "Что такое транзакция? / Microsoft SQL Server || Всем привет! Народ, объясните что такое транзакция и в каких случаях разумно ее применять?..." and a link to "www.sql.ru/forum/actualthread.aspx?bid=1&tid=187186 - 52k -". At the bottom of the results, there are links for "Сохранено в кэше" and "Похожие страницы".

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

Извлечение метайнформации - 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

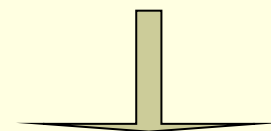
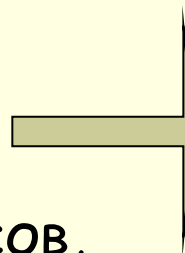
ddk@yandex.ru

Список RSS-каналов,
адреса, которые я ищу

ddk@livejournal.com

Мои друзья, сообщества,
Темы, которые
меня интересуют

...



Персонализированная
реклама



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
- Персонализация 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
 - Персонализированная 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 Scholar.

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](#)

 [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 hirt, 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) [\(Correct\)](#)

Кстати, тут и домашняя страничка автора

В статье есть обзор существующих работ а в нем - библиографические ссылки. Давайте поищем и эти статьи тоже

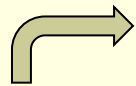
Слишком старая работа? Давайте посмотрим кто на нее ссылается

Может стоит поискать похожие работы?

Поиск научной информации (3)

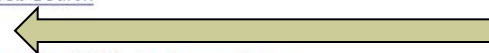
- Нам нужны научные работы - давайте посмотрим в CiteSeer и Google Scholar.

The screenshot shows the Google Scholar interface. At the top, there's the Google Scholar logo and navigation links for Web, Images, Video, News, Maps, and more. A search bar contains the text 'Information Extraction' and a 'Search' button. Below the search bar, there are links for 'Advanced Scholar Search', 'Scholar Preferences', and 'Scholar Help'. The search results are displayed under the heading 'Scholar All articles - Recent articles' and show 'Results 1 - 10 of about 2,540,000 for Information Extraction. (0.16 seconds)'. The results list several articles, including 'Wrapper Induction for Information Extraction' by Nicholas Kushmerick, 'Information Extraction' by JIM COWIE, 'Learning Information Extraction Rules for Semi-Structured and Free Text' by S Soderland, and 'Toward information extraction: identifying protein names from biological papers.' by K Fukuda et al. Each result includes a brief description, the source, and citation information.



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?

The screenshot shows a del.icio.us profile for user 'dkozlov' with the tag 'machineLearning'. The page displays a list of items tagged with 'machineLearning', including links to articles and documents. A search bar is visible at the top right. A 'related tags' sidebar on the right lists various tags associated with the main tag. Two green arrows point from the text on the right to the 'machineLearning' tag in the header and the 'related tags' sidebar.

del.icio.us / dkozlov / machineLearning

popular | recent

your bookmarks | your network | subscriptions | links for you | post

logged in as dkozlov | profile | help

Your items tagged machineLearning (create tag description) → view all, popular

del.icio.us search

« earlier | later » page 1 of 3

Основные методы, применяемые для распознавания рукописного текста edit / delete
to machinelearning ... saved by 6 other people ... on oct 29

CRF Project Page edit / delete
to java machinelearning ... saved by 31 other people ... on oct 12

Conditional Random Fields edit / delete
to machinelearning ... saved by 93 other people ... on oct 04

FRC: Forecasting, Recognition, Classification edit / delete
to machinelearning ... saved by 1 other person ... on sept 13

K.Vorontsov: home page edit / delete
to machinelearning ... saved by 3 other people ... on sept 13

sebastiani02machine.pdf (application/pdf Object) edit / delete
to machinelearning ... on june 28

ABNER: A Biomedical Named Entity Recognizer edit / delete
Linear CRF implementation in Java
to machinelearning ... saved by 8 other people ... on june 14

Machine Learning (Theory) edit / delete
to machinelearning ... saved by 279 other people ... on june 14

▼ related tags + datamining + ir + java

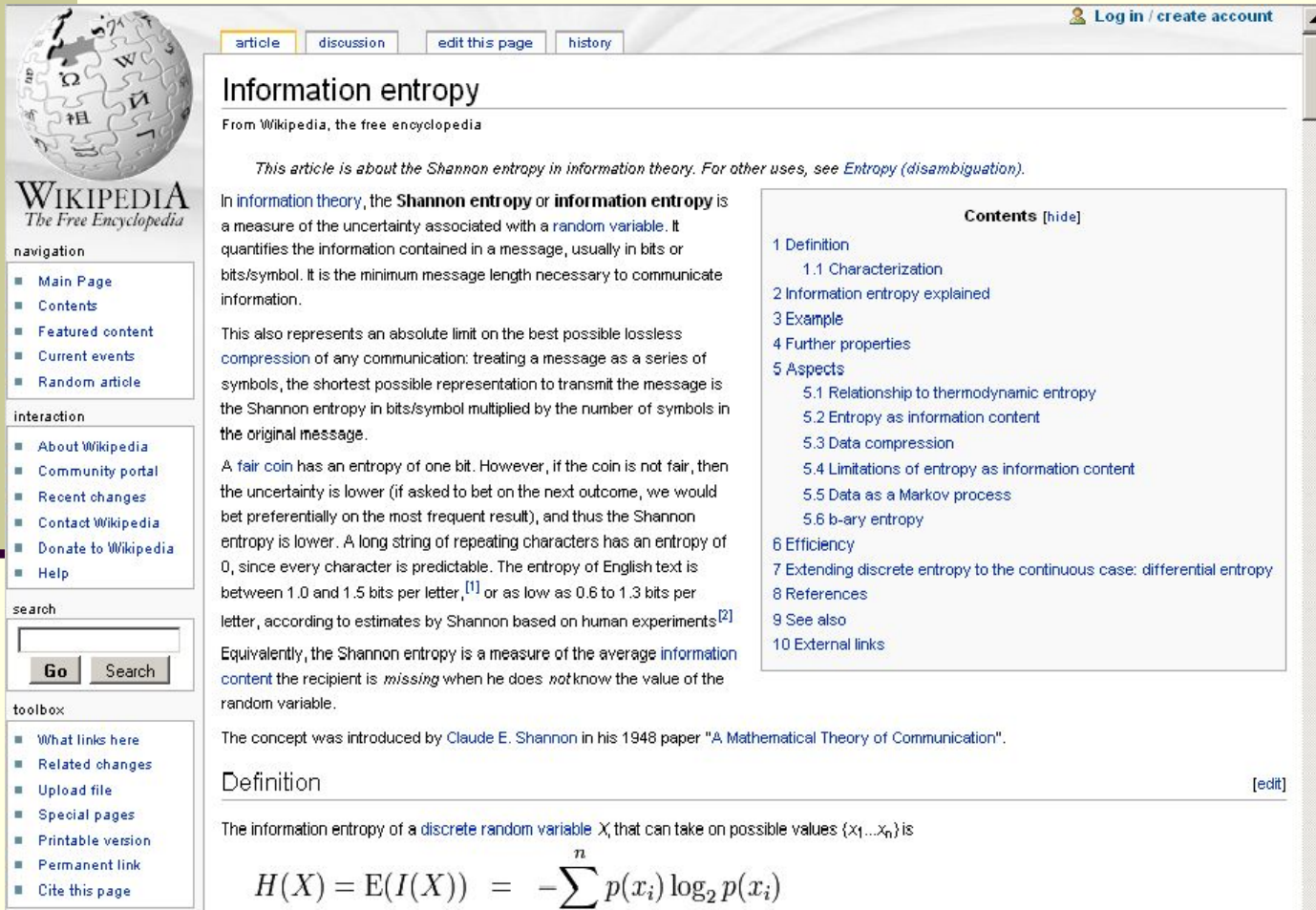
▼ tags -mozilla-information adsl ... de
AJAX anonymousInternet ant apache
architecture auto autoSimulators bar
Barradosfavoritos beanshell beta
biletov blogs bookmark books browser
business ca cas cellphone certification
child cim citeseer client cluster
clustering omni oms codeAnalys
computer continue crawler crm css
customizing dataflow datamining
debugging del.icio.us delicious
desktop deti devel Development dict
dictionary dictionnaire digitallibrary
docflow driving dtrace e2k EAI eclipse
ed2k edonkey edu edu_lunix
education electro Email Emule engine
engineering english Entertainment erp
extensions films fin Firefox
firefox:bookmarks firefox:toolbar
Forum fp7 français fun genetic
Geronimo gifts google googleNews
gost goto grants groupware health

Не вы одни
интересуетесь
этой темой, и
многие уже
нашли

Похожие тэги

Поиск научной информации (5)

- Вы не понимаете этих слов? Не страшно!



The screenshot shows the Wikipedia article for "Information entropy". At the top, there are navigation tabs for "article", "discussion", "edit this page", and "history". The article title is "Information entropy" with a subtitle "From Wikipedia, the free encyclopedia". A note states: "This article is about the Shannon entropy in information theory. For other uses, see Entropy (disambiguation)." The main text explains that in information theory, Shannon entropy or information entropy is a measure of 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. A paragraph notes that 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. Another paragraph states that a fair coin has an entropy of one bit, but if it's not fair, the uncertainty is lower. It also mentions that a long string of repeating characters has an entropy of 0, and that the entropy of English text is between 1.0 and 1.5 bits per letter, or as low as 0.6 to 1.3 bits per letter. A final paragraph explains that 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. It notes that the concept was introduced by Claude E. Shannon in his 1948 paper "A Mathematical Theory of Communication". A "Definition" section is partially visible, starting with "The information entropy of a discrete random variable X , that can take on possible values $\{x_1, \dots, x_n\}$ is". To the right of the main text is a "Contents" table of contents with 10 items: 1 Definition, 1.1 Characterization, 2 Information entropy explained, 3 Example, 4 Further properties, 5 Aspects (with sub-items 5.1 to 5.6), 6 Efficiency, 7 Extending discrete entropy to the continuous case: differential entropy, 8 References, 9 See also, and 10 External links. The left sidebar contains navigation links like "Main Page", "Contents", "Featured content", "Current events", "Random article", "interaction" links like "About Wikipedia", "Community portal", "Recent changes", "Contact Wikipedia", "Donate to Wikipedia", "Help", a search box, and a "toolbox" with links like "What links here", "Related changes", "Upload file", "Special pages", "Printable version", "Permanent link", and "Cite this page".

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 орфографически-ми 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

Как писать хорошую английскую прозу:

- знайте, что хотите сказать,
- подражайте классикам.

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, rnv@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 тематическом содержании 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 *Robot File Request*. 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 Интернет-страниц, индексируя все полученные документы. Очевидно, что такой подход требует значительных сетевых и аппаратных ресурсов. Однако теку-

*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

щий 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 собирающим 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 улучшить 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.

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](mailto:ddk@cs.Msu.su) или очно в к. 764