Enterprise and Desktop Search

Lecture 1: Introduction

Pavel Dmitriev
Yahoo! Labs
Sunnyvale, CA
USA

Pavel Serdyukov
Delft University of Technology
Netherlands

Sergey Chernov
L3S Research Center
Hannover
Germany
Search Environment of a Company Employee

Corporate Web (Intranet)

Public Web (Internet)

- Social Networks
- CMS
- DVD
- Disk
- FShare
- Web
- Online Shops
- DB
- People
- Online Libraries
- Web Sites
- E-mail
- Web Sites
- Online Shops
- CMS
Search Settings and Goals

• Desktop Search
  – Files & directories I created, files I downloaded, OS and application data, e-mail, data on DVD/USB...

• Enterprise Search
  – Intranet Search: web pages and files from the group/department/company intranet, database contents, corporate email, people...
  – Public Site Search: inventory, web site contents...

• Web Search
  – Publicly accessible documents on the Web
Benefits of Enterprise Search
(Bennet 08)

• Direct benefits for the user
  – Employees/user can find what they want → they are happy

• Financial benefits/ROI
  – Employees are more productive
  – Users are likely to buy more products from your site

• Strategic/Business Intelligence benefits
  – Monitor and detect how users’ interest changes over time
  – See what people are searching for but not finding (documents, products, etc)
  – Track closely important users/products/customers
Information Retrieval System

Crawler
Downloads/collects the data

Indexer
Processes the data and builds Inverted Index

Ranker
Evaluates user queries against the index and computes a list of (ranked) results

Display
Organizes and displays the results to the user, facilitates navigation through the result set
1. **RuSSIR’2009**: III Российская летняя школа по информационному поиску
   11-16 сентября 2009, Петрозаводск, русский.
   Цели Российской летней школы по информационному поиску (RuSSIR) — познакомить слушателей со спектром современных проблем и методов информационного...
   romip.ru/russir2009
   Сохраненная копия · Еще с сайта · Рубрика: Поисковые системы

2. **"Интернет-математика 2009" @ RuSSIR 2009 - Интернет-математика 2009 - я.р**
   5 мая 2009 года, 16:37.
   clubs.ya.ru/imat2009/replies.xml?item_no=45
   Сохраненная копия · Еще с сайта

3. **RuSSIR 2009 - Конференция молодых ученых / CSIN RU**
   III Российская летняя школа по информационному поиску (RuSSIR) пройдет 11-16 сентября 2009 года в Петрозаводске.
   Сохраненная копия · Еще с сайта

4. **Блог PicLab » RuSSIR**
   Конференция молодых ученых по информационному поиску в рамках III Российской летней школы по информационному поиску (RuSSIR) III Российская летняя школа по информационному поиску (RuSSIR) пройдет 11-16 сентября 2009 года в Петрозаводске.
   blog.piclab.ru/tag/russir
   Сохраненная копия · Еще с сайта
Web Search

• Crawl and index data from the public web, present simple result set (plus ads) to the user
• Main challenges are ranking quality and scale

• Huge industry (est. $30B by 2010), very active research community (Cyber 09)
  – Major web search companies employ tens of thousands of people
• Good search quality, ubiquitous usage
• Still evolving, but more or less standard architecture, evaluation, and algorithms
• Example Systems:
  – Yahoo, Yandex, Google, Bing, Baidu, etc.
Enterprise Search

Found 14826 site documents

Tool Shed
Tools Home > Resources > Tool Shed Toolkit Section Overview Approval Center "Vacation" Responder Author Admin Tool Guesthouse User Admin Tool Document Sharing ...
backyard.yahoo.com/resources/toolshed/

Using Thunderbird on IMAP
... for offline use Copying local email to the IMAP server Using the "Vacation" Responder Creating Email Filters Downloading Email for Offline ...
backyard.yahoo.com/resources/isnet/self_support/thunderbird_on_imap.html

Using Outlook on IMAP
... to the IMAP Server Using the "Vacation" Responder Creating Email ... into the folder. Using the "Vacation" Responder This section describes ...
backyard.yahoo.com/resources/isnet/self_support/outlook_on_imap.html

Holiday Tips
... be out of the office during the holiday season. Tip Details I'm going on vacation and I don't know how to set up a vacation auto-responder. (Outlook users on Exchange ...
backyard.yahoo.com/resources/isnet/holiday.htm

Using Outlook on Exchange
... assistant automatically notifies coworkers that you are on vacation when they send you an email. Note: The Out of Office assistant only sends ...
backyard.yahoo.com/resources/isnet/self_support/outlook_on_exchange.html
Enterprise Search

• Crawl and index information from a variety of repositories in variety of formats
• Need to rank different types of entities
• Simple result set not enough, need to provide exploratory interfaces

• Rapidly growing industry (20+% annual growth, est. $2.55B by 2010), relatively little research (Gref 09)
• Rather poor quality, comparing to web search
• Lots of unsolved problems and research opportunities
• Example Systems:
  – Autonomy, Endeca, FAST, Google, IBM OmniFind, Oracle, etc.
Differences between Web Search and Enterprise Search (Crawling)

- Diverse information sources and formats, many are not “crawl-friendly”
  - Web pages, files, databases, etc.
  - “Compound” and “composite” documents
- A “click” may have undesirable side effects
  - Document deleted
  - Charge for accessing a 3rd party’s database
- Many security domains
- Hard to create a research test collection
Differences between Web Search and Enterprise Search (Indexing)

• Data is semi-structured and semantics is often known
  – Search for “objects” (people, rooms, printers, etc)
• Need to incorporate access-control info
• Vocabulary mismatch is a big problem
  – Need to use thesaurus
• Need to index special symbols/punctuation
• Need to efficiently support the kinds of queries generated by exploratory interfaces
Differences between Web Search and Enterprise Search (Ranking)

• Small set of correct answers (often just 1)
• Less hyperlinks and anchortext, and of poorer quality
• Poorer quality of content (pages are not created with a search engine in mind)
• User identity is often known
  – Can use user context
• Often need to retrieve ALL relevant documents
• No (intentional) spam
• Federation and blending is often necessary
Differences between Web Search and Enterprise Search (Display)

• Known identity and user history → personalized results presentation

• Search clients are not just browsers
  – Applications/Advanced search interfaces

• Since ranking is hard, need to provide exploratory interfaces / interactive search & browse experience, give the user more control
How to Measure Success?

• Evaluation metrics used for Web Search are not always suitable for Enterprise Search. Need ways to evaluate
  – Quality of interaction
  – Success of task completion
  – User satisfaction
Desktop Search

• Why *desktop search*?
• Size of data on the desktop is **big** and continuously **growing**
• We are moving towards Social Semantic Desktop
  – **Social** – communication in a social network
  – **Semantic** – metadata descriptions and relations
Desktop Search – Current Status

• Documents on the desktop are not linked to each other in a way comparable to the web

• Simple full text search
  – no personalization
  – no context
  – no ranking possible or too poor

• Metadata enriched search makes use of
  – associations to contexts and activities
  – provenience of information
  – sophisticated classification hierarchies
Differences between Web Search and Desktop Search

• Search on the desktop vs. Search on the Web
  – Re-finding vs. finding
  – Integration across many applications and file formats
  – Users prefer to navigate, not to search
  – Many information types: ephemeral, working, archival
  – Extra sources for ranking improvement:
    • File metadata
    • Usage metadata
    • Folder structure
  – Privacy concerns
Blending of the 3 Search Settings

• Many web collections require login to access
  – I want to search over the collections I have access to (ACM Digital Library)
  – I want to search over the collections I don’t have access to to decide if I want to buy access (Amazon)
• More and more personal data is stored online
  – E-mail, photos, blogs, profiles
• To further improve Web Search need to understand user intent and context
• Would like to search seamlessly over all the three mediums
Course Overview

• Lecture 1: Introduction

• Lecture 2: Searching the Enterprise Web
  – What works and what doesn’t?
  – Using User Feedback in Enterprise Web Search

• Lecture 3: Exploratory Search
  – Look-up search vs Exploratory search, faceted search, result categorization and clustering
Course Overview

• Lecture 4: Expert Finding
  – Profile-based, document-based, and graph based methods

• Lecture 5: Desktop Search
  – Architecture and evaluation, just-in-time retrieval, context detection and usage

• Conclusion
What we would like you to get out of this Course

• Understand the challenges in Desktop and Enterprise search
• Have an idea of what the current state-of-the-art is in the area
• Be inspired to apply techniques developed for one setting to another setting
• Think about creating applications that blend the boundaries among the 3 settings
References

- [Hawking 04] Hawking, D. “Challenges in Enterprise Search”. ADC Conference, Dunedin, NZ.