Complex Search
From Lookup to Understanding

Ulrich Norbisrath
University of Tartu, Estonia

Keynote at Datasyys and Softnet 2011, Barcelona
2011/10/24
Where am I from?

USA (live)

Germany (born)

Estonia (tele-work)

Kazakhstan (future)

maps taken from wikipedia
Where am I from?

Tartu (research)

Aachen (PhD)
Research Background

- PhD in Software Engineering
- Still teaching in SE – currently writing textbook on story driven modeling
- Leader of two research groups at UT:
  - Ubiquitous Systems (p2p systems, mobile computing, biometry)
  - Information Management
    - Measurement of complex search
    - Graph-based storage of personal information
    - Graph-based information visualization
My research groups

• Information Management
  - 3 PhD students, 2 Master students
  - 1 PhD Student here:
    • Tonu Tamme (IMMM4)
    • “Impr. Email Management”

• Ubiquitous Systems
  - 1 PhD student, 3 Master, 4 Bachelor students
  - 1 of them here:
    • Artjom Lind (Mobility4) “Mobile F2F”
    → go to their talks!
Content

- Search today – complex search identified
- Current Search research
- Measuring Complex Search (Search-Logger)
- Ways to better understanding?
Search Means – What do you use? (in class)

- Internet
  - Google, Google Scholar, webcrawler
  - baidu
  - wikipedia
  - facebook
- Other means
  - library, newspapers
Search Means – what is used

• Internet
  - Google (search, maps, etc.)/Yahoo/Bing
  - Craigslist/Amazon/Ebay/Kayak/Facebook/Ask
  - PubMed/IEEE
  - Wikipedia
  - Wolfram Alpha

• Other means
  - Library (also Internet)
  - Newspaper (news and classifieds)
  - Bulletin Board (in supermarket)
  - Ask human/expert
Market Share of the Search Titans

search engines market share 2008 – source: hitslink.com

search engines market October 2011 – source: hitslink.com
Do you search? How and why?

- Who of you searches a lot?
- Who of you uses for search the Internet? (Who not? Who still uses anything else?)
- Who is sometimes unhappy with their search experience?
- Why, in which cases?
Search Cases
Personal Motivation

- Find a list of still open shops, where you can buy dance shoes (ballroom dance shoes, sneakers) near Courthouse or Clarendon, Arlington, Virginia, USA
- Is there a video pcmcia expresscard with Nvidia chipset? (List of them)
- Document, how you can make "Immutable Pages" in the moin moin wiki editable?
Your daughter Anna is to graduate from high-school and wants to study abroad in the field of either political sciences or architecture. You can support her with 2000 Euro per month. Which universities can you recommend that are affordable and will offer the best career perspectives after graduating? Please compile a list of 10 universities.
Search Cases

You have been offered the job as rector of a Goethe institute (coordinating regional German language education abroad). It is very likely that you will be sent to Astana, Kazakhstan. Collect one page of comprehensive facts about the political situation and quality of life in Astana.
Problems in Today’s Search

- Information outdated
- Information need not specifiable
- Search only query/answer
- Context important (and not specifiable)
- Choice of websites too diverse
- All parts of pages indexed → unreasonable combinations
- Information scattered via multiple web-pages, cannot be directly found
- Information need hard to aggregate
- Information from query is dropped (try to google: broadcom BCM4727 linux module)
What’s wrong?

• These cases are special.

• Why?
  - Much more than query/answer
  - Some kind of iteration/interaction, multiple search windows/tabs
  - Take a long time (complex)
  - Require critical thinking
  - Part of the search is happening outside of the browser (maybe in users head)
How do we store information?

picture source: wikipedia
In Documents/Files and Folders

![File folders with handwritten labels](attachment:image)

- (root)
  - Logfile
  - Widgets
  - Foobar
  - Payroll
    - Employees
    - Timecards
    - Salaries
      - Managers
      - Leads
      - Clerks

picture source: wikipedia
What Information is Retrieved?

• Still documents
• Today’s search research:
  - Retrieving single documents
  - Optimize precision
  - Optimize ranking in first 10 hits
  - Improve display of the 10 results
Search Measurement

• Precision
  - Ratio of retrieved relevant documents to total number retrieved in a search.

• Recall
  - Ratio of retrieved relevant documents to total number of relevant documents.
Fundamental Problem

- Only query/answer paradigm
- No intuitive translations
  - Need → query
  - Document list → need
- Only looking up documents supported (lookup)

source: Traditional IR Model, Bates, 1989
Current Trends
Personalized Ranking

Dance Supplies Directory of Virginia
Dance Supplies Directory of Virginia, Find or Locate Dance Supplies, shoes, dance, capezio, bloch, ballet, tap, wear, pointe, jazz, bags, tights, leotards, ...
virginia.uscity.net/Dance/Supplies/index.html - Cached - Similar - ☞ ▼

Costumes, Ballet shoes, Dance wear, Chantilly, Va - Prima Dance ...
Pointe Shoes; Ballet Shoes; Jazz Shoes; Dance Sneakers; Tap Shoes; Character Shoes; Lyrical Shoes ... Classes are located in Chantilly, Virginia. ...
www.primadancesuppliesandcostume.com/ - Cached - Similar - ☞ ▼

Athletic & Dance Shoes Retail in VA Yellow Pages by SuperPages
Select a Virginia City to view Athletic & Dance Shoes Retail Listing. Athletic & Dance Shoes Retail · Aldie Athletic & Dance Shoes Retail ...
www.superpages.com/yellowpages/...Dance+Shoes.../S-VA/ - Cached - Similar - ☞ ▼

Ulrich Norbisrath (http://ulno.net)
University of Tartu
Complex Search
http://im.ulno.net
Current Trends
Suggestions/Discovery

Google

<table>
<thead>
<tr>
<th>horoscope love</th>
<th>compatibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>horoscope love</td>
<td>match</td>
</tr>
<tr>
<td>horoscope love</td>
<td>matches</td>
</tr>
<tr>
<td>horoscope love</td>
<td>compatibility chart</td>
</tr>
<tr>
<td>horoscope love</td>
<td>calculator</td>
</tr>
<tr>
<td>horoscope love</td>
<td>match compatibility</td>
</tr>
<tr>
<td>horoscope love</td>
<td>signs</td>
</tr>
<tr>
<td>horoscope love</td>
<td>2009</td>
</tr>
<tr>
<td>horoscope love</td>
<td>test</td>
</tr>
<tr>
<td>horoscope love</td>
<td>meter</td>
</tr>
</tbody>
</table>

Google Search  I'm Feeling Lucky
Wolfram Alpha

Input interpretation:
weather

Latest recorded weather for Washington, District of Columbia:

- temperature: 64 °F
- conditions: cloudy
- relative humidity: 34% (dew point: 36 °F)
- wind speed: 10 mph

(25 minutes ago)

Weather history & forecast:

- Temperature:
  - Minimum: 45 °F (Tuesday, November 3, 6:45 AM)
  - Average: 58 °F
  - Maximum: 72 °F (Wednesday, October 28, 2:15 PM)

Cloud cover:

- 100% coverage on Wednesday, November 2

Exchange history for $859 (US dollars):

- 1-year minimum: €571.69 (October 23, 2009, 12 days ago)
- 1-year maximum: €686.52 (November 11, 2008, 12 months ago)

Additional currency conversions for $859 (US dollars):

- JPY: ¥77,740 (Japanese yen)
- GBP: £520.33 (British pounds)
- CNY: ¥5864 (Chinese yuan)
- CAD: CA$915.09 (Canadian dollars)
- MXN: $11,410 (Mexican pesos)
Classic Query Types
(Broder, 2002)

• Informational
  - Looking for information on a certain topic
  - User wants to view a few relevant pages

• Navigational
  - Looking for a (known) homepage
  - User wants to navigate to this homepage, only one relevant result

• Transactional
  - Looking for a website to complete a transaction
  - One or more relevant results
  - Transaction can be purchasing a product or downloading a file
Facts from User Tests (Lewandowski, 2006)

• More than 40% of searches are navigational

• Queries
  - Average length: < 2 words
  - ~50% of queries just one word

• Search engine results pages (answer)
  - 80% of users no more than first results page
  - Users normally only view the first few results
  - Users only view up to five results per session
  - Session length is less than 15 minutes

• Users are usually satisfied with the results
More Innovation in Search (especially fulfilling transact. lookups)

- Vertical (domain specific) search, search 3.0:
  - Channeled information for particular area
  - Spiders instead of crawlers
  - Spiders scrape the content of pages

- Examples:
  - Travel search sites
  - Professional hardware search sites (for doctors, machine engineers)
  - Media related sites (video, music, books)
  - Domain libraries (CS library, Math library)
Our Definition of Search

- *Search* is the **process** to satisfy an information need.
- Search is *successful*, if need is satisfied.
- *Search task* is the **description** of the Search. If the search takes long, it is possible that the search tasks changes.
Information Need

- Why do you search the Internet (reasons, what do you want to achieve)?
- Pair up with your neighbor
- Write down 5 points (abstract terms, why you search)
- Think,
  - which of these can be easily fulfilled by the traditional lookup offered in the net,
  - which require thinking and iteration.
- Time: 5 minutes
Information Need (in class)

- relation between different fields, mapping
- routing, navigation, a way
- accuracy, ensuring it, credibility
- reputation about projects, people, products
Exploratory Search

Today’s Web Search

- Lookup
  - Fact retrieval
  - Known item search
  - Navigation
  - Transaction
  - Verification
  - Question answering

Information Need

Exploreatory Search

- Learn
  - Knowledge acquisition
  - Comprehension/Interpretation
  - Comparison
  - Aggregation/Integration
  - Socialization

- Investigate
  - Analysis
  - Exclusion/Negation
  - Synthesis
  - Evaluation
  - Discovery
  - Planning/Forecasting
  - Transformation

source Marchionini, 2006
Complex Search

- Involvement of multiple steps
- Time consuming
- At least one of Discovery, Aggregation, Synthesis strongly present
How could we measure complex search?

- Classic measures for single queries
- User behavior
- Demographics
- Whole search spanning multiple queries and information aggregation and synthesis
- Evaluate aggregation information (clipboard, bookmarks)
- User feedback
Research Project: Search-Logger

- **Firefox (3.6) Plug-In**
  - Browser
  - Plug-in
  - Internet
  - PHP Front-end
  - DB
  - Analyzer

- **Proxy-based**
  - Browser
  - Proxy
  - Internet
  - DB
  - Analyzer
## Search-Logger User Study Data

<table>
<thead>
<tr>
<th>User</th>
<th>User Action</th>
<th>Action Description</th>
<th>Timestamp</th>
<th>Task number</th>
<th>Google Searches</th>
</tr>
</thead>
<tbody>
<tr>
<td>231</td>
<td>User started a new SC</td>
<td>Displaying pre-SC form for SC index 9::</td>
<td>09/08/2011 12:06:04</td>
<td>10</td>
<td>Religion (KERNAUP)</td>
</tr>
<tr>
<td>232</td>
<td>User opened a tab</td>
<td></td>
<td>09/08/2011 12:06:41</td>
<td>10</td>
<td>Religion (KERNAUP)</td>
</tr>
<tr>
<td>233</td>
<td>User submitted pre-SC form</td>
<td>(a=5)</td>
<td></td>
<td>(l=(1)</td>
<td></td>
</tr>
<tr>
<td>234</td>
<td>User closed a tab</td>
<td></td>
<td>09/08/2011 12:06:19</td>
<td>10</td>
<td>Religion (KERNAUP)</td>
</tr>
<tr>
<td>241</td>
<td>User closed a tab</td>
<td></td>
<td>09/08/2011 12:10:30</td>
<td>10</td>
<td>Religion (KERNAUP)</td>
</tr>
</tbody>
</table>

### Data Columns:

- **User**: User identifier.
- **User Action**: Action performed by the user.
- **Action Description**: Description of the action.
- **Timestamp**: Date and time of the action.
- **Task number**: Task number associated with the action.
- **Google Searches**: Search queries made using Google.
Search-Logger User Study Data Timeline

- User opened a tab
- User started a new SC
- User opened a tab
- User performed a Google search entered at Google web search homepage
- User opened Google Web Search Homepage
- User viewed a related link
- User performed a Google search entered at Google web search homepage
- User opened a tab
- User viewed a related link
- User opened a tab
- User viewed a related link
- User performed a Google search entered at Google web search homepage

Timeline: 15hr, 16hr, 17hr, 18hr
How to overcome or extend query response?

• User interface
  - Different result visualization (not only list of documents)
  - New interaction patterns (not only keywords/tags or hierarchies)

• Avoid information duplication (blogs do the opposite)

• Use categorization techniques to relate documents or part of documents

• Move from document to graph of objects
Ted Nelson’s vision (1970)

Xanadu
Examples of Information Visualization

Mufin Player
Examples of Information Visualization
TheBrain (thebrain.com)

- Nice (but slow) user interface
- No kick start
3D Discovery Tool

- Camera to root
- Turn path traveler on/off
- Move orbits
- Turn mouse Orientation On/Off
2.5D layout
A better future?
Summary

• What is search and an information need?
• Has today’s search reached its limits? - No!
• Google, Bing & CO great for lookup.
• Support of discovery, aggregations, synthesis seems possible
• Documents → graph?
• Still a lot to do!
  → better understanding?
Interested? Contact us!

- http://contact.ulno.net
- http://search-logger.com
References


Information Overload

- Neither information nor overload easily definable
- Bound to the belief in the “information age”
- Conflicting and vague definitions available

- So does it even exists?
How much time do you spend on Facebook?
• Automatic categorization, tagging
• Manual categorization, tagging
• Summarization
Personal Information Management

Personal information management (PIM) refers to the practice and the study of the activities a person performs in order to acquire or **create, store, organize, maintain, retrieve, use, and distribute** the information needed to **meet life’s many goals** (everyday and long-term, work-related and not) and to fulfill life’s many roles and responsibilities (as parent, spouse, friend, employee, member of community, etc.). PIM activities are an effort to establish, use, and **maintain** a mapping between information and need.

Personal Information Management (PIM) – William Jones