The ad-hoc search process

Formulates: Query

Processes: Search / Matching

Analyzes: Ordered results
Why to bother about presentation?

- Looking through the search results is a part of the process of finding relevant documents
- The overall process could be improved if this part is improved
- The standard presentation is the ordered list of matched documents
- What can we improve?
Where we will learn InfoVis?

- Improving search process (L8)
  - Better interfaces for query formulation
  - Graphical interfaces for the search process
- Presentation (L7)
  - Better presentation of search results
- Alternative access techniques (L6)
  - InfoViz as an alternative to search, an alternative approach for information access

What can we do?

- Manage quantity
- Decide what to present for a document
- Provide additional information
- Show context
- Explain relevance to the query
- Group the results
- Present results not in a linear list
- Present results graphically
- Let the user explore the results interactively
Managing quantity

- More is better?
- Quantity and quality
- Let the user choose
- Setting standard cut-off point
- Adaptation to the user’s task and background
  - Adaptive filtering
  - Adaptive cut-off

Documents and surrogates

<table>
<thead>
<tr>
<th>Metadata, Content data</th>
<th>Digitally stored, used for search, presentation, and selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Document</td>
<td>Digitally stored, used for presentation and selection, not used for search</td>
</tr>
<tr>
<td>Externally stored document / object</td>
<td>Externally stored, not used for search</td>
</tr>
</tbody>
</table>
What to present?

- Document ID
- Metadata, content data
  - Metadata: author, title, year, source
  - Keywords
  - Abstract
- An extract of the full document
  - First paragraph
  - Thumbnail
- Full document

Why it is a bad idea?

Two-step / three-step presentation

- Two steps:
  - Level 1 - list with minimal information
  - Level 2 - full information by request
- Three steps:
  - Level 1 - list with minimal information
  - Level 2 - more detailed information by request
  - Level 3 - full information by request
Example: Photo archive

- Photos are stored, but are not searchable
- Searchable are *descriptions*
- Description: what, when, where
  - Content (abstract vs. classifier)
  - Time (granularity!)
  - Location (coding scheme vs words)
- What to present?

Case study: Movie rental store
The case of search engines

- The choice:
  - Header
  - URL
  - Content
- Core elements: Header and URL
  - Why they are important?
  - Why they are not enough?

Providing More Information

- Additional information about document could help the user in making selection
- What kind of information?
  - Document-based information
    - Document alone
  - Search-based
    - Document for this query
  - Personalized
    - Document for this user (and query)
Document-Based Information

- Search engine can get, calculate, or mine additional useful information
- Type of the document (.html, .pdf, .ppt)
- Type of the source/collection (.com, .edu)
- Document size
- Likely dates of creation and update
- Document language
- Presence of images
- Most of that can be also shown visually

Query-Based Information

- Mostly document relevance
- Why relevance?
  - Simple order is misleading
  - Even the first result in the list can have low relevance
  - There could be relevance gaps
- Example: ACM Digital Library
  - [http://portal.acm.org/portal.cfm](http://portal.acm.org/portal.cfm)
**Evolution of Presentation in Search Engines**

- **First Shot: Lycos 1995**
  - Core only

- **Classic design: Excite, 2000**
  - Core+Start of the document

- **Modern design: Lycos, AltaVista 2001**
  - Core+KWAC (keywords and context)

- **Advanced design: 2003**
  - NorthernLight: relevance, category
  - Google: Size, Link to cashed annotated document

- **Visualization: Kartoo, Grokster…**

---

**Personalized Information**

- **Individual history**
  - Have I seen it?
  - How many times?
  - Have I marked/annotated/selected it?
  - Amazon.com case

- **Group history**
  - How my group likes it
  - Have anyone like me marked/annotated…
  - Have anyone like me selected it in the same query

- **Knowledge-based data (relevance, readyness)-> adaptive annotation in AH**
Visual Cues in KS II search

- Traffic-based
  - More group traffic
    - Darker background color
  - More user traffic than others
    - Darker foreground color of the “human” icon
  - 0~9 traffic levels

- Annotation-based
  - More group annotations
    - Darker background color
  - User own annotation
    - Foreground color
    - Type: Sticky notes, Thumbs-up, Question
  - General attitude
    - Page “quality” temperature

Knowledge Sea II search

<table>
<thead>
<tr>
<th>KnowledgeSea Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query: [Dynamically allocated memory]</td>
</tr>
<tr>
<td>Retrieved common words:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Result pages</th>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.62</td>
<td>C. Page 57.html</td>
</tr>
<tr>
<td>2</td>
<td>0.48</td>
<td>S. Summit Chapter 11: Memory a....</td>
</tr>
<tr>
<td>3</td>
<td>0.45</td>
<td>C. Page 57.html</td>
</tr>
<tr>
<td>4</td>
<td>0.42</td>
<td>R. Miles Memory</td>
</tr>
<tr>
<td>5</td>
<td>0.40</td>
<td>R. Miles 57.html</td>
</tr>
<tr>
<td>6</td>
<td>0.37</td>
<td>S. Summit 11.4 Pointer Safety</td>
</tr>
<tr>
<td>7</td>
<td>0.20</td>
<td>R. Miles first</td>
</tr>
<tr>
<td>8</td>
<td>0.34</td>
<td>R. Miles printf.c</td>
</tr>
<tr>
<td>9</td>
<td>0.30</td>
<td>D. Marshall sect02_21_3.html</td>
</tr>
<tr>
<td>10</td>
<td>0.29</td>
<td>C. Page Question 17.23</td>
</tr>
</tbody>
</table>

Similarity score
- General annotation
- Question
- Positive
- Negative

Example of Social Traffic

Example of Social Annotations
Showing the Context

- For structured or networked documents, it makes sense to show the results in a proper context.

- Context in hierarchy (book, three, some sites)
  - Cha-Cha system collects Web pages that satisfy a query, then attempts to climb the context three to show to the user a hierarchy (or a path) that goes to the query results.

- Context in a graph (hypertext, Web site)
  - Mappuchino and VIDAS show relevant documents in the context of the whole site or subset of the site connected to the search results. The idea is that the user will find what he needs in the results pages or in the linked pages.

Show the Context (path)

- Cha-Cha shows the user a path to each query result, helping to see the context.

Can be done with little visualization.
Relevance and Context (links)

- Mappuccino allows the user to search on a specific web site. The pages that satisfy the query are shown together with the other linked pages.

Relevance and Context (structure)

- WIDAS shows relevant documents in the context of a Web site - showing relevance, links, and structure
Show the Relationship Between the Query and the Results

- The motivation: to show how the document relates to the query
  - If there was a year restriction -> show year
  - If there were keywords - show which are found (KWOC - KWIC - KWAC)

- Some efforts to better show keyword relevance between document and query
  - Semantic Highlighting / Google
  - TileBars
  - More on that in graphical presentation of results

TileBars

The system shows the degree of match for each query word in the documents (darker squares represent more frequent matches)
Grouping

- The idea of grouping is to present similar documents together - once one document in a group is relevant, the others will likely too
- How to group?
  - Semantic vs. keyword-level grouping
- How to present?
  - Text-based clustering
  - Graphical clustering
  - Similarity-based visualization (infospace)

Semantic Grouping

- Semantic grouping: the idea is to group documents together by a semantic feature (taken from metadata or mined)
  - Source / Author
  - Media
  - Date / Time
- If no metadata is available the category can be deduced using automatic classification
Hierarchical Classification

- Dumais and Chen approach to present search results
- Uses automatic classification with CVM

Similarity-based Clustering

- If no category for classification is available, documents can be simply grouped by their similarity
- The idea of clustering is to group together documents with similar content
  - Based on keywords-level similarity between documents
  - There are many clustering algorithms that differ in speed, precision, presentation power
  - The problem of cluster naming
Presentation of Clustered Results

- One-level clustering
- Hierarchical clustering with navigation
  - Example: Vivismo’s Clusty engine
    http://vivisimo.com/
- Graphical presentation of results
  - Example: Mooter http://mooter.com/moot
  - Example: Grokker http://www.grokker.com/

Information Visualization for search result presentation

- Present results not in a linear list (2-3D)
  - Table: Envision, SenseMaker
  - 2D or 3D space: VIBE, InfoCrystal, LyberWorld, Lighthouse
- Let the user explore the results by manipulation with visualization
  - VIBE, BIRD, GUIDO, LyberWorld, Envision
Graphical results presentation

- Most graphical presentation approaches are based on the same ideas
  - Group similar documents
  - Show relevance to the query
- In a table similar documents can be shown in the same cell
  - Metadata-based: Envision
  - Similarity-based: SOM (see lecture 6)

Present Results as a Table

- Envision
  - Virginia Tech Digital Libraries project
    http://www.dlib.vt.edu/projects/Envision/
- 2-D Table interface for data exploration
  - This user-controlled system facilitates examining very large data sets, displaying multiple aspects of the data simultaneously and efficiently, and interactive discovery of patterns in the data
  - The user chooses which aspect to present using each dimension, color, and shape
Presenting results using SOM

- SOM is similarity-based visualization technology that is limited to “cells” - can be visualized as a table (see Lecture 6)
- SOM has to be trained in advance to create map of the information space, however, classification of results on trained SOM is fast
- Search results (from one or more searches) can be shown in corresponding SOM cells
Relevance: InfoCrystal

InfoCrystal ( Spoerri, 1993) attempts to show document relevance to specific terms of the query and their combinations. The shape of the icon defines how many of the specified criteria it satisfies (1 – circle, 2 – rectangle, 3 – triangle, 4 – square, and so on). The icons are oriented so that their sides face the matching criteria. The number associated with each icon indicates how many of the retrieved documents satisfy the conditions represented by it.

Relevance + Grouping: Cougar

The Cougar interface [Hearst, 94b] displays at the top of the screen the most frequent categories assigned to the retrieved documents. The user can select up to three of these categories and view via a Venn-diagram how the documents intersect with respect to these categories. In the example shown above, the user chose the categories “government”, “weapons” and “physics”. The chosen categories are highlighted and the same colors are used in the Venn-diagram.
Fully 2D and 3D Spatial Views

- In real spatial views a position of each document is calculated down to its coordinates in 2D or 3D
- Relevance-based 2D/3D visualization
  - VIBE (2D)
  - LyberWorld (3D)
- Spatial visualization
  - Lighthouse
  - Kartoo

VIBE: terms hits between documents

- VIBE system places the query terms at the boundaries of a space and the documents are scattered inside this space
VR-VIBE has two different visualization modes:
(left) the POIs are placed in three dimensions and the relevance is shown by the size and the shade of the documents' icons (figure 10a);
(right) the POIs are positioned in a 2-D plane, and the third dimension is used to indicate overall relevance of the documents.