Using Interactive Systems

Lecture 7
Finding Your Way Around: Retrieval
- 2009/2010

Herre van Oostendorp
- Master Content and Knowledge Engineering
Content

1. Multimedia representations: feedback and demo’s

2. Modeling information-seeking behavior
   - Specificity of search questions

3. Retrieval by queries
   - Types of required knowledge
   - Strategies of search term queries
   - Spatial visualization abilities
   - User evaluation of search engines

4. Evaluation of information found on internet

5. Retrieval performance evaluation

6. Assignment WP4 retrieval performance and usability of search engines

7. Individual assignment
1. Feedback multimedia representations

- No final judgment yet. First impressions:
  - Nice papers
  - Length is OK.
  - You have tried to come up with **new hypotheses, derived from theory**.
  - Also the experimental setups are OK, but a little bit brief.

- But let us first see how the **demo’s** are.
2. Modeling Information-Seeking Behavior

On basis of Ellis (1987), information-seeking model by Meho & Tibbo (2002), comprising activities such as:

- Starting
- Chaining
- Browsing
- Monitoring
- Accessing*
- Differentiating
- Extracting
- Verifying* (web resources!)
- Networking*
- Information managing*

* means new compared to Ellis (1987)
Information-seeking behavior model by Meho & Tibbo (2002)

- Activities not necessarily sequential.
- Can be grouped in 4 interrelated stages:
  - Searching
  - accessing
  - processing
  - ending
- Implications:
  e.g. accessing → collaboration
Stages in information-seeking behavior

Starting, Chaining, Browsing, Monitoring, Differentiating, Extracting, Networking

 Searching

 New Information Needs

 Direct Sources

 Indirect Sources

 No

 Processing

 Yes

 Accessing

 Decision-Making

 Ending

FIG. 1. Stages in the information-seeking behavior of academic social scientists.
Cognitive model of document search

- Goal formation
- Category selection
- Information extraction
- Integrating new information with previously extracted information
- Recycling through these 4 steps until satisfactory answer

Guthrie (1988)
Factors influencing search effectiveness

- **A. Task complexity (example)**
  - E.g. specificity of search questions → Rouet (2003)

- **B. Document structure**
  - E.g. simple hierarchical structure can be effective (e.g. Van Nimwegen, Pouw & Van Oostendorp, 1999)

- **C. Display features**
  - E.g. (overview) content representation can be effective
A. Specificity of search questions

Task specificity (hypertext about ‘History of anorexia’, ca. 2400 words) (Rouet, 2003):

- Specific questions (answer at one location)
  - “Which authors have provided the first clinical descriptions of anorexia?”
- General questions (answer distributed at several locations)
  - “What treatments may be suggested, and what are their effects?”

Results: General questions take longer, longer planning, more look backs, and retrieval (opened nodes) was less precise.

Why more difficult?
- partial decay of goal representation
- evaluation process more complex because less fit between goal statement and menu items
3. Querying search engines

FIG. 1. The iterative search process.
a. Types of knowledge required when querying search engines

In the context of online catalog search

3 types of knowledge involved (Borgman, 1996):

- **Conceptual** knowledge (translating an information need into searchable query)
- **Semantic** knowledge (how to implement a query in a given system):
  - ACTIONS (e.g. find), ACCESS POINTS or FIELD TAGS (author, title, subject), SEARCH TERMS, BOOLEAN OPERATORS (and, or, not etc)
  - e.g. knowledge of file organization
- **Technical** knowledge (basic computing skills and **syntax**: exact commands to manipulate the system)
b. Strategies of search term queries

- Typology of 9 strategies on basis of the semantic relationships between search terms (Hembrooke, Granka, Gay & Liddy, 2005) in cycles of queries.

- Tasks to find information regarding topics on which one had to give a lecture on, e.g. ‘lecturing on history of computers’.

- Subjects had to generate search queries on the same topic over 10 trials.

- (9) strategies distinguished
  - Differences based on expertise (novices: plural making, redundancy, poke&hope, backtracking; experts: elaboration strategies)
### Strategies of search term queries

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elaboration</td>
<td>The global level of detail and sophistication intrinsic to user search attempts</td>
</tr>
<tr>
<td>Example:</td>
<td>Computers: <em>Microsoft, inventions, technological advances, personal computers, history of Microsoft, history of computers, inventions 20th century, technological advances 20th century, computers now and then, machine language development</em></td>
</tr>
<tr>
<td>Redundancy</td>
<td>An overall index of the extent to which search terms are used repeatedly on successive queries</td>
</tr>
<tr>
<td>Example:</td>
<td>Antiques: <em>18th century antiques, 18th century, antiques, antiques 18th century ...</em></td>
</tr>
<tr>
<td>Broadening</td>
<td>The extent to which a user begins with a specific query and expands the scope of the search phrase over successive trials</td>
</tr>
<tr>
<td>Example:</td>
<td>Butterflies: <em>Monarch butterflies, butterfly migration, insect migration patterns, yellow swallowtail, painted lady, Mexico and monarchs, migration pattern and butterflies, butterfly, migration routes butterflies</em></td>
</tr>
<tr>
<td>Refining</td>
<td>The extent to which a subject begins broadly and narrows the search with increasing specificity</td>
</tr>
<tr>
<td>Example:</td>
<td>Computers: <em>Computer history, Babbage, history of DARPA, history of IBM, mechanical computer, computer history, military computing history, eniac, ưuią, early computing software</em></td>
</tr>
<tr>
<td>Backtracking</td>
<td>The frequency with which a searcher reuses prior search terms over successive trials</td>
</tr>
<tr>
<td>Example:</td>
<td>Sundance; <em>Indian ceremonies, Native American Ceremonies, Native American Ceremonies, Lakota Sundance ceremony, Lakota, Lakota ceremony, Native American Traditions, Native American ceremonies, ceremonial dances, dance ceremonies</em></td>
</tr>
<tr>
<td>Plural Making/Taking</td>
<td>Reflects instances when a user repeatedly incorporates similar nouns into their search attempt, with the slight modification of making the word plural or singular</td>
</tr>
<tr>
<td>Example:</td>
<td>Gardening: <em>ant, ants, spider, spiders, pests, pest, etc ...</em></td>
</tr>
<tr>
<td>Kitchen Sink</td>
<td>The extent to which a searcher incorporates search terms related to the subject, but not specific to the query task</td>
</tr>
<tr>
<td>Example:</td>
<td>Asian Cooking: <em>dumplings, kung pao chicken, rice, eating on the floor, wok, chopsticks, curry, haan, Korean dishes, French cooking</em></td>
</tr>
<tr>
<td>Poké-n-Hope</td>
<td>The extent to which a searcher retains the same basic structure throughout all search queries, changing only a single word within each trial</td>
</tr>
<tr>
<td>Example:</td>
<td>Gardening: <em>garden pest elimination, garden pest kill, garden threats, garden pests, gardening tips, garden rid pests, garden pest wipe, garden pest eliminate, wood eliminate</em></td>
</tr>
<tr>
<td>Topic Terms</td>
<td>The extent to which the user incorporates the given query terms as their search terms. For example, “you are lecturing on play strategies in basketball”</td>
</tr>
<tr>
<td>Example:</td>
<td>Basketball strategies, basketball plays, strategies in basketball playing, strategies for basketball, coaching basketball, learning basketball strategies, play strategies basketball, etc ...</td>
</tr>
</tbody>
</table>
c. Individual characteristics (example)

- Effects of spatial visualization and domain expertise on information seeking
  - E.g. Spatial Visualization Ability (SVA) and domain knowledge of user (Downing, Moore & Brown, 2005):
    - 2x2 factorial design (high vs. low SVA; high vs. low domain expertise)
    - Dependent variables
      - Time to find first article
      - Number of articles found
        - Tasks: with search tool (FirstSearch) 5 questions
        - Subjects: n=44 (students)
Information seeking (1)

Fig. 3. Relationship between SVA and Domain Expertise based on time-to-first-article.
Information seeking (2)

Fig. 4. Relationship between SVA and Domain Expertise based on total-number-of-relevant-articles.
d. User evaluation of search engines

On basis of 4 steps (Salton, 1989)
- Formulation and submission of query
- Examination of results
- Possible feedback loop to reformulate
- Integration of results and evaluation of the whole search

→ **user** evaluation of search engines is a **multidimensional construct** (Johnson, Griffiths & Hartley, 2003) (in contrast to system retrieval performance)

**Dimensions:**

1. **Measures of (perceived) search results**
   - Effectiveness (relevance of retrieved results to a user)
   - Utility (actual usefulness of retrieved results)
   - Efficiency (effectiveness in relation to time or effort)
   - Satisfaction with these three criteria
Relationships dimensions and overall judgment

II. Measures of (perceived) user-system interaction

- Query formulation (perceived ease of expressing a query)
- Query reformulation (feedback received for (re)formulating the query)
- Examine results (process of interpreting results in the given frame of information need)
- Satisfaction on each of these criteria
Relationships dimensions and overall judgment

Results (n=23, Johnson et al, 2003): relationship dimensions and overall judgment (!) of system success (3 search engines: Excite, NorthernLight, Hotbot):

- Results-criteria (effectiveness, efficiency and utility) correlate highly with overall (perceived) judgment of system success, except the interaction criterion
- Efficiency (ease of use) and effectiveness highest correlation
- One search engine behaves somewhat differently (Excite)

Criticism: still all high correlations; strange definition of variables; no relationships established between usability measures and system retrieval performance! (see Assignment 4).
4. Evaluating information found on the internet

How is the **quality of information** evaluated in practice (internet information: lack of quality control, and context deficit). What **kind of criteria**?

- **Technical** criteria (authorship, affiliation, publishing body, currency (date), etc)
- **Design** criteria (layout, etc)
- **Readability** (writing style, etc)
- **Accuracy** (concordance with best evidence or general practice, verifiability of details)
- **Completeness/coverage/scope** (point of view or bias, referral to/knowledge of the literature)


5. Retrieval performance evaluation

How to measure?

- Build a Test Reference Collection
  - Static document collection
    - TREC Web Track: 1.25 million documents
    - TREC Terabyte Track: 1 terabyte documents.
  - Set of topics
  - Relevance judgements
    - ‘Blind pooling review method’:
      - Gather top $N$ results from participating search engines for each topic. ($N > 500$)
      - Let experts judge the relevance of each retrieved document.

[num] 409
<title> legal, Pan Am, 103
<desc> What legal actions have resulted from the destruction of Pan Am Flight 103 over Lockerbie, Scotland, on December 21, 1988?
<narr> Documents describing any charges, claims, or fines presented to or imposed by any court or tribunal are relevant, but documents that discuss charges made in diplomatic jousting are not relevant.
Retrieval performance evaluation

What to measure? [Baeza - Yates 1999]

- Information Request
  - Document Collection $C$
  - Answer Set $A$
  - Relevant documents $R$
  - Relevant documents in Answer Set $R_A$

Recall: $| R_a | / | R |$

Precision: $| R_a | / | A |$
Retrieval performance evaluation

- **Advanced measures, used in TREC (Text retrieval Evaluation Conference)?**
  - **Summary table statistics**
    - Summarizes statistics relative to given task.
  - **Recall-precision averages**
    - Computes the interpolated average precision (over all topics) at 11 standard recall levels.
  - **Document level averages**
    - Computes the average precision at specified cutoff values.
    - Provides detailed insight in the ranking performance of a retrieval strategy.
  - **Average precision histogram**
    - The average precision histogram plots the average precision of a retrieval strategy per topic.
    - Examines the performance of the retrieval strategy for the individual topics, and reveals possible abnormalities.
## Retrieval performance evaluation

- **Summary table statistics:**

<table>
<thead>
<tr>
<th>Total number of documents over all queries</th>
<th>Google</th>
<th>Vivisimo</th>
<th>AlltheWeb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer set (A)</td>
<td>160</td>
<td>160</td>
<td>160</td>
</tr>
<tr>
<td>Relevant Documents (R)</td>
<td>186</td>
<td>186</td>
<td>186</td>
</tr>
<tr>
<td>Rel. docs. in answer set (R&lt;sub&gt;A&lt;/sub&gt;)</td>
<td>106</td>
<td>95</td>
<td>45</td>
</tr>
</tbody>
</table>
## Recall-precision averages:

<table>
<thead>
<tr>
<th>Recall level</th>
<th>All the Web</th>
<th>Google</th>
<th>Vivisimo</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.6855</td>
<td>0.9062</td>
<td>0.8184</td>
</tr>
<tr>
<td>0.1</td>
<td>0.4703</td>
<td>0.8285</td>
<td>0.7316</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3557</td>
<td>0.7749</td>
<td>0.6903</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1228</td>
<td>0.7369</td>
<td>0.6417</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0833</td>
<td>0.704</td>
<td>0.6255</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0812</td>
<td>0.6211</td>
<td>0.4033</td>
</tr>
<tr>
<td>0.6</td>
<td>0</td>
<td>0.207</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0.2062</td>
<td>0</td>
</tr>
<tr>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Retrieval performance evaluation:

- Recall-precision averages (example):

![Graph showing recall-precision averages for different search engines (All the Web, Google, Vivisimo).]
Retrieval performance evaluation:

- Document level averages (example):

<table>
<thead>
<tr>
<th>Doc. Cutoff</th>
<th>All the Web</th>
<th>Google</th>
<th>Vivisimo</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.325</td>
<td>0.7</td>
<td>0.575</td>
</tr>
<tr>
<td>10</td>
<td>0.325</td>
<td>0.6875</td>
<td>0.5875</td>
</tr>
<tr>
<td>15</td>
<td>0.3333</td>
<td>0.65</td>
<td>0.5833</td>
</tr>
<tr>
<td>20</td>
<td>0.2812</td>
<td>0.6625</td>
<td>0.5938</td>
</tr>
<tr>
<td>30</td>
<td>0.1875</td>
<td>0.4417</td>
<td>0.3958</td>
</tr>
<tr>
<td>100</td>
<td>0.0562</td>
<td>0.1325</td>
<td>0.1188</td>
</tr>
<tr>
<td>200</td>
<td>0.0281</td>
<td>0.0663</td>
<td>0.0594</td>
</tr>
</tbody>
</table>
Retrieval performance evaluation

- Document level averages (example):

![Graph showing precision vs. document cutoff for Fastsearch, Google, and Vivissimo]
Retrieval performance evaluation:

- Average precision histogram (example):

![Average precision histogram](image)
6. Assignment WP4: retrieval performance and usability of search engines

- **Evaluation of 2 search engines** in terms of
  - **Usability from user’s perspective** (usability experiment).
    General indicators, e.g. ISO:
    - Effectiveness (correctness)
    - Efficiency (correctness/time needed in sec.)
    - Satisfaction (7-point rating scales—concerning input phase (interaction), output phase (relevance and presentation), and overall)

- **Retrieval performance of the search engine**
  - Recall
  - Precision
  - Etc.  *See Van Zwol & Van Oostendorp (2004)(read it before Friday)*
**Issues to be addressed**

**Report:**
- related work (max 800 words)
- setup of both the retrieval performance experiment *and* usability experiment (test bed TERS)
- describe and evaluate the results
- integrate the results of both experiments into a recommendation

- Max. 2500 words all-in (please also a digital version)

- **Details:** see Assignment WP4 on the UIS website
7. Individual assignment

**Paper** (max 4000 words all-in) and **powerpoint presentation** plus evt. **demo**:

- choose a website – where navigation, personalization, multimedia representation, retrieval is/can be made relevant
- identify, justify and fundament methods to evaluate each of these 4 domains
- **Apply, concretely, at least one** of these methods on a specific domain
- report results
- draw conclusions

**Warning**: individual assignment demands knowledge of all **four** topics (all mandatory articles should be referred to in the paper).
What next?

- **Request:**
  - Course Usability for Interactive Systems (USY), please sign in, if you want to participate

- **Tuesday, April 6, 9-11.00 (BBL109):**
  - Introduction case study on retrieval by Herre van Oostendorp;
  - Introduction TERS (Assignment WP4) by Sandor Spruit, plus practicum

- **Friday, April 9,**
  - 13-14.00 consult WP4 (BBL165)
  - 14-15.00 lecture 6b serious games (BBL165)
  - 15-17.00 practicum retrieval (BBL 452)

- **Deadline Problem Description Individual Assignment:**
  - Friday 9 April (send per email 1 A4)

- **Consult Individual Assignment**
  - 16 April 13.00