IR in Context of the User: Interactive IR Evaluation

Peter Ingwersen
Royal School of LIS
Denmark
pi@iva.dk – http://www.iva.dk/pi

Oslo University College, Norway
Agenda - 1

- **Introduction** (20 min)
  - Research Frameworks vs. Models
  - Central components of Interactive IR (IIR)
  - The Integrated Cognitive Research Framework for IR

- **From Simulation to ‘Ultra-light’ IIR** (20 min)
  - Short-term IR interaction experiments
  - Sample study – Diane Kelly (2005/2007)
Agenda - 2

- Experimental Research Designs with Test persons (25 min)
  - Interactive-light session-based IR studies
  - Request types
  - Test persons
  - Design of task-based simulated search situations
  - Relevance and evaluation measures in IIR
  - Sample study – Pia Borlund (2000; 2003b)
Agenda - 3

- **Naturalistic Field Investigations of IIR** (20 min)
  - Integrating context variables
  - Live systems & (simulated) work tasks

- **Wrapping up** (5 min)

Questions are welcome during the sessions

Frameworks & Models – difference?

- **Frameworks** describe
  - Essential **objects** to study
  - The **relationships** of objects
  - The **changes** in the objects / relationships that affect the functioning of the system
  - Promising **goals** and **methods** of research

- **Frameworks contain** (tacit) shared assumptions
  - ontological, conceptual, factual, epistemological, and methodological

- **The concept model**
  - A precise (often formal) representation of objects and relationships (or processes) **within** a framework
  - Modeling may also in principle encompass human actors and organizations

- **Frameworks may lead to**
  - Research Designs, incl.
  - Research Questions; Experimental Setting; Methodology
The Lab. Research Framework – cave with central variables (*The Turn*, 2005)
User-centered (contextual) MODELS

- **Examples** *(in display order)*
  - Wilson, 1999 *(conceptual: Info. Behavior; Seek; IR)*
  - Byström & Järvelin, 1995 *(flow chart: Info. Seek)*
  - Saracevic, 1996 *(conceptual, stratified: IR)*
  - Vakkari, 2000 *(flow chart, Online Search; Relevance)*
Information behaviour and IR

T. Wilson’s Onion Model, 1999 - extended:

Non-job-related
Tasks and Interests
Daily-life behavior

Job-related
Work Tasks
Interests

Information behaviour

Seeking

Interactive IR

IR

Behaviour
Perceived Task

Personal Factors

Situational Factors

Information

Need analysis

Choice of Action

- identification of alternatives
- ranking them
- choosing an action

Implementation

Evaluation
a) needs satisfied, task may be completed
b) needs cannot be satisfied
c) further information is needed

(From: The Turn, p. 69)
Saracevic’ stratified model for IIR (1996)
Wang & Soergel 1998

Decision Rules
- Elimination
- Multiple criteria
- Dominance
- Scarcity
- Satisfice
- Chain

Knowledge of
- topic
- person
- organization
- journal
- document type

DIEs: Document Information Elements
- Title
- Author
- Abstract
- Journal
- Series
- Date
- Type

Criteria
- Topicality
- Orientation
- Quality
- Novelty
- Availability
- Authority
- Relation

Values
- Epistemic
- Functional
- Conditional
- Social
- Emotional

Decision
- Acceptance
- Maybe
- Rejection

DIEs: Document Information Elements
Values: Document Values/Worth

(From: The Turn, p. 201)

Peter Ingwersen
IR and relevance in Seeking context – Seeking into IS&R: Vakkari 2000
Research Setting Types

- Laboratory experiments – no test persons, but
  - Simulations – Log analyses (not treated in presentation)

- Laboratory study – with test persons:
  - ‘Ultra-light’ (short-term interaction: 1-2 retrieval runs)
    – or ‘Interactive light’ (session-based multi-run interaction)

- Field experiment – experimental (artificial) situation in natural setting with test persons

- Field study – study of natural performance or behavior in natural setting with test persons
  - Longitudinal studies

- Case study – (qualitative) study with few test persons
Variables involved in a test:

- **Independent** *(the ‘cause’), e.g.,*
  - Interface functionalities; Different IR models; Searcher knowledge

- **Dependent** *(the ‘effect’), e.g.,*
  - Performance measures of output (recall/prec.; CumGain; usability)

- **Controlled** *(held constant; statistically neutralized; randomized):*
  - Database; Information objects
  - Search algorithms
  - Simulated work task situations – Assigned TREC topics
  - Test persons

- **Hidden variables (Moderating or Intervening), e.g.,**
  - Variation of test persons’ levels of experience …!!! – see the Integrated Research Framework for IR
Agenda - 1

- **Introduction to Tutorial** (20 min)
  - Research Frameworks vs. Models
    - Central components of Interactive IR (IIR)
    - The Integrated Cognitive Research Framework for IR
  - From Simulation to ‘Ultra-light’ IIR (20 min)
    - Short-term IR interaction experiments
    - Sample study – Diane Kelly (2005/2007)
Central Components of Interactive IR – the basic integrated framework

- Information objects
- IT: Engines
- Logics
- Algorithms
- Interface
- Cognitive Actor(s) (team)
- Org.
- Social Context
- Cultural

Essir2011
Ingwersen
Central Components of Interactive IR – the basis of the integrated framework

- Information objects
- IT: Engines Logics Algorithms
- Interface
- Cognitive Actor(s) (team)
- Org.
- Social Context
- Cultural

= Cognitive transformation and influence
= Interactive communication of cognitive structures
= Cognitive transformation and influence over time

The Lab. Framework

Ingwersen
Dimensions and Range of Variables in the Integrated IIR framework:

**9 dimensions from 6 components**

- Information objects
- IT: Engines
- Logics
- Algorithms
- Interface
- Cognitive Actor(s) (team)
- Org.
- Social Context
- Cultural

Interaction
Categories of Dimensions in the Cognitive Research Framework

1. Natural work task dimension
2. Natural search task dimension
3. Actor characteristics dimension
4. Perceived work task dimension
5. Perceived search task
6. Document dimension
7. Algorithmic search engine dimension
8. Algorithmic interface dimension
9. Access and interaction dimension

Each containing multiple variables

Socio-org. task dimensions

Actor dimensions

Algorithmic dimensions
<table>
<thead>
<tr>
<th>Natural Work Tasks (WT) &amp; Org</th>
<th>Natural Search Tasks (ST)</th>
<th>Actor</th>
<th>Perceived Work Tasks</th>
<th>Perceived Search Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT Structure</td>
<td>ST Structure</td>
<td>Domain Knowledge</td>
<td>Perceived WT Structure</td>
<td>Perceived Information Need Content</td>
</tr>
<tr>
<td>WT Dependencies</td>
<td>ST Dependencies</td>
<td>Experience on Search Task</td>
<td>Perceived WT Dependencies</td>
<td>Perceived ST Specificity &amp; Complexity</td>
</tr>
<tr>
<td>WT Requirements</td>
<td>ST Requirements</td>
<td>Stage in Work Task Execution</td>
<td>Perceived WT Requirements</td>
<td>Perceived ST Dependencies</td>
</tr>
</tbody>
</table>

Variables with values:
<table>
<thead>
<tr>
<th>Natural Work Tasks (WT) &amp; Org</th>
<th>Natural Search Tasks (ST)</th>
<th>Actor</th>
<th>Perceived Work Tasks</th>
<th>Perceived Search Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT Structure</td>
<td>ST Structure</td>
<td>Domain Knowledge</td>
<td>Perceived WT Structure</td>
<td>Perceived Information Need Content</td>
</tr>
<tr>
<td>WT Dependencies</td>
<td>ST Dependencies</td>
<td>Experience on Search Task</td>
<td>Perceived WT Dependencies</td>
<td>Perceived ST Specificity &amp; Complexity</td>
</tr>
<tr>
<td>WT Requirements</td>
<td>ST Requirements</td>
<td>Stage in Work Task Execution</td>
<td>Perceived WT Requirements</td>
<td>Perceived ST Dependencies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sources of Difficulty</td>
<td></td>
<td>Perceived ST Domain &amp; Context</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Motivation &amp; Emotional State</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Document and Source</strong></td>
<td><strong>IR Engines IT Component</strong></td>
<td><strong>IR Interfaces</strong></td>
<td><strong>Access and Interaction</strong></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>------------------------------------------</td>
<td>------------------------------------</td>
<td>----------------------------</td>
<td></td>
</tr>
<tr>
<td>Document Structure</td>
<td>Exact Match Models</td>
<td>Domain Model</td>
<td>Interaction Duration</td>
<td></td>
</tr>
<tr>
<td>Document Types</td>
<td>Best Match Models</td>
<td>System Model Features</td>
<td>Actors or Components</td>
<td></td>
</tr>
<tr>
<td>Document Genres</td>
<td>Degree of Doc. Structure and Content Used</td>
<td>User Model Features</td>
<td>Kind of Interaction and Access</td>
<td></td>
</tr>
<tr>
<td>Information Type in Document</td>
<td>Use of NLP to Document Indexing</td>
<td>System Model Adaption</td>
<td>Strategies and Tactics</td>
<td></td>
</tr>
<tr>
<td>Communication Function</td>
<td>Doc. Metadata Representation</td>
<td>User Model Building</td>
<td>Purpose of Human Communication</td>
<td></td>
</tr>
<tr>
<td>Temporal Aspects</td>
<td>Use of Weights in Doc. Indexing</td>
<td>Request Model Builder</td>
<td>Purpose of System Communication</td>
<td></td>
</tr>
<tr>
<td>Document Sign Language</td>
<td>Degree of Req. Structure and Content Used</td>
<td>Retrieval Strategy</td>
<td>Interaction Mode</td>
<td></td>
</tr>
<tr>
<td>Layout and Style</td>
<td>Use of NLP to Request Indexing</td>
<td>Response Generation</td>
<td>Least effort Factors</td>
<td></td>
</tr>
<tr>
<td>Document Isness</td>
<td>Req. Metadata Representation</td>
<td>Feedback Generation</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Document Content</td>
<td>Use of Weights in Requests</td>
<td>Mapping ST History</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contextual Hyperlink Structure</td>
<td></td>
<td>Explanation Features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Source (see Actor)</td>
<td></td>
<td>Transformation of Messages</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scheduler</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Number of variables using the Framework

- Maximum application of **three independent** variables simultaneously!
- Can be done **in pairs** – and by total control of binary values of variables, e.g.
  1. **Interface function X**, value a/b
  2. **Personal IR expertise**, values none/much
  3. **In domain Z, work task type**: routine – but Rich/Poorly defined

There are many relevant combinations made from the Framework!
Agenda - 1

✓ Introduction to Tutorial (20 min)
  ✓ Research Frameworks vs. Models
  ✓ Central components of Interactive IR (IIR)
  ✓ The Integrated Cognitive Research Framework for IR

■ From Simulation to ‘Ultra-light’ IIR (20 min)
  ■ Short-term IR interaction experiments
  ■ Sample study – Diane Kelly (2005/2007)
IR interaction ‘Ultra-light’ – short-term IIR
Lab IR - ’ultra light’ interaction

In this 1-2 run setting we have two ways of measuring performance:

1. By **assessor in pre-existing test collection** (as in TREC with unrealistically long delay between run and assessment – but equal to all).

   Assessments may be applied to judging **second run results** (made by the test persons) – like using pseudo RF after first run.
Lab IR - ‘ultra-light’ interaction

2. By all test searchers of the first run results of the same query session (Secondary run assessments of results used if first run done by pseudo RF).

- One may pool performance scores across the set of assigned requests (topics) or simulated tasks given in experiment – because of the max. two runs:

- **Good:** No learning effects can influence the experiment
  - or the same effects appear as when using TREC assessors
  - That is why this setting is ‘interactive ultra-light’
  - Graded relevance assessments possible
  - Can be used OUTSIDE traditional test collections!

- **Bad:** Quite few documents are commonly assessed for relevance per test searcher (or vary much)!
  - The setting is limited in realism (only 2 runs)
Interactive ‘Ultra-light’ experiment. Research question concerned with variables from actor & IT.
IIR Interactive ‘Ultra-light’ sample


- **RQ**: Does multi-evidence of users’ information need situation improve retrieval performance through query expansion, compared to initial request and pseudo relevance feedback?
Research setting

- 13 test persons supplied …
  - 45 natural ‘topics’ to HARD TREC (title and description) and
  - Relevance assessments

- HARD TREC collection; Lemur system (BM25)
  1. Topic title+description run by Lemur (bag-of-words) one run; 
     serves as baseline (BL).
  2. Pseudo RF modes (top-5; top-10;…) run on top of BL
  3. Each test person asked 4 questions via a form:
Research setting 2

- (Q1) state the times in the past he/she had searched that topic;
- (Q2) describe what he/she already knows about the topic (knowledge state);
- (Q3) state why he/she wants to know about the topic; and
- (Q4) add any keywords that further describe the topic.
Research setting 3

- Controlled variables: BM 25; 45 topics; HARD coll.
- Independent variables:
  1. Pseudo RF variations – on top of baseline (BL)
  2. Q2-Q4 combinations (term weights) – on top of BL
- Dependent var.: MAP – statistical significance test

RESULTS, yield of different words (mean):
- Single Q-forms outperform BL
- Q2-Q3 (and Q2-Q4) combined outperform BL plus any pseudo RF
- Performance increases with query length.
Summary: IIR ‘Ultra-Light’

- **Strength:**
  - Easy to apply existing test collections, with …
  - Relevance assessments existing \textit{a priori} (as in TREC or INEX)
  - New relevance assessments possible – with graded assessments and over many assessors (test persons): weighted assessments
  - Can lead to more solid interactive investigations in later studies

- **Weakness:**
  - Are all variable values known?? (people means hidden ones!)
  - ‘Ultra-light’ IIR is limited in realism (1-2 iterations; context features hardly in play)
  - Limited number of documents assessed (per test person)
Agenda - 1

✓ Introduction to Tutorial (20 min)
  ✓ Research Frameworks vs. Models
  ✓ Central components of Interactive IR (IIR)
  ✓ The Integrated Cognitive Research Framework for IR

✓ From Simulation to ‘Ultra-light’ IIR (20 min)
  ✓ Short-term IR interaction experiments
  ✓ Sample study – Diane Kelly (2005/2007)
Agenda - 2

- Experimental Set-ups with Test Persons (25 min)
  - Interactive-light session-based IR studies
  - Request types
  - Test persons
  - Design of task-based simulated search situations
  - Relevance and evaluation measures in IIR
  - Sample study – Pia Borlund (2000; 2003b)
IR interaction ‘Light’

- **Documents**
- **Search request**
- **Query**
- **Matching**
- **Evaluation**
- **Evaluation Result**
- **Recall base**
- **Searcher**
  - MUST DO
  - Posteriori
  - Relevance Assessments
- **Context**
  - MANY
  - Relevance Feedback
  - Runs Allowed

Relevance assessment

Essli2011 Ingwersen
Data Collection Means

- Observation
- Thinking (talking) aloud - Introspection
- Eye-tracking
- Critical incidence
- Questionnaires
- Interviews (structured; open-ended; closed)
  - Post or/and pre-interviews
- Focus groups
- Diaries – Self reporting
- Logging and recording of behavior (system/client logs)
  - Assessments of relevance
Request Types in (‘ultra’) ‘light’ IIR

- **Natural request**/ real need of test person - or
- **Assigned** to test person
- ‘**Query**’ is the retrieval mechanism’s internal translation of the REQUEST

- **Topical** (as TREC ‘topics')
- **Factual**
- ‘**Known Item**’
- **Other** metadata

- **Simulated Task Situation** (Cover Story)
- ‘**Sample**’ as request
- **Simplistic request formulation** (context free)
**Number of Test Persons**

- Number depends on goal of research & no. of variables:

- **Behavioral** field study/experiment: many persons (>30 test persons required – and some (2-3) search jobs, to be statistically valid)

- **Performance-like** field experiment: many search jobs per person (4-10) – but less (~ 15) test persons required.
  - Note: Sanderson et al. paper: IIIX 2005 on no. of topics necessary for statistical validity:
    > 60!! (if applying MAP on top-15)

- The best design: always > 25 persons
Test Persons …

- In order to be statistically significant, or really indicative, each cell in the cross tabulation result matrix should contain **25-30 units** (rule of thump).

- Example (performance/evaluation goal with 3 independent (binary) variables, done in pairs: 2x2x2 = 8 cells x 30 units = 240 units in total):
  - You have 2 x 10 test persons (doctors & med. stud.)
  - They need to do **12 search jobs per person** = 120 units per group over 2x2 additional variable values, for reasons of cross tabulations = 120 x 2 = 240 jobs!
  - or 2 x 20 persons doing 6 jobs each.
Latin Square research design – *The Turn*, p. 253-254

| 2: C, B, A | 5: F, E, D | 2: E, F, D | 5: B, C, A |

6 test persons (1-6);
6 real / simulated work tasks/ or assigned topics (A-F)
Agenda - 2

- Experimental Set-ups with Test Persons (25 min)
  - Interactive-light session-based IR studies
  - Request types
  - Test persons
  - Design of task-based simulated search situations
  - Relevance and evaluation measures in IIR
  - Sample study – Pia Borlund (2000; 2003a)
Simulated Work Task Situations
– or ‘cover stories’ – to trigger natural information needs (& requests)

Example from study on relevance assessments on the Web (See ‘Ultra-Light’ in Bibliography: Papaeconomou, 2008):

Beijing is hosting in 2008 (8th-24th August) the Olympic Games. A friend of yours, who is a big fan of the Olympic Games, wants to attend the events and asks you to join in this trip. You find this invitation interesting. You are not a big fan of the games but you always wanted to visit China, therefore you want to find information about the sightseeing in the city and the activities that the Chinese will offer during the games. Find for instance places you could visit, activities you could do in relation to the Chinese culture or in the spirit of the games.
Simulated situation: sim A

**Simulated work task situation:** After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

**Indicative request:** Find for instance something about future employment trends in industry, i.e. areas of growth and decline.
Integrated Framework and Relevance Criteria

A: Recall, precision, efficiency, (quality of information/process)

B: Usability, Graded rel., CumGain; Quality of information/process

C: Quality of info & work task result; Graded R.

D: Socio-cognitive relevance; Social utility: rating; citations; inlinks;

Evaluation Criteria:

Essir2011

Ingwersen
Relevance & Evaluation Measures

The measurement of performance by use of non-binary (graded, scaled or gliding) based performance measures (generalized by Järvelin & Kekäläinen, 2002)

- Realistic assessment behaviour
- Indication of users’ subjective impression of system performance and satisfaction of information need: usability (Hornbæk, 2006; Nielsen, 2006)

- Other measurements to be used on Interaction Process:
  - Display time; No. of requests/queries; Visits & Downloads
  - Selection patterns; Views & clicks; Social utility assessments;
  - No. of documents assessed; Perceived ease of process; …
Agenda - 2

✓ Experimental Set-ups with Test Persons (40 min)
  ✓ Interactive-light session-based IR studies
  ✓ Request types
  ✓ Test persons
  ✓ Design of task-based simulated search situations
  ✓ Relevance and evaluation measures in IIR

■ Sample study – Pia Borlund (2000; 2003a)
The Borlund Case (2000; 2003b)

Research questions

1) Can simulated information needs substitute real information needs?
   - Hypothesis is: YES!

2) What makes a ‘good’ simulated situation with reference to semantic openness and types of topics of the simulated situations?
Experimental Setting:

**Data collection:** Financial Times (TREC data) and The Herald (current)

**Test system:** Full-text online system
Probabilistic based retrieval engine

**Test persons:** 24 university students
(undergraduates and graduates)

**From:** Computing, engineering, psychology, geography, English history, etc.

**Info. needs:**
- **24 real needs** (1 real need per test person)
- **96 simulated information needs**
  (4 simulated task situations per test person)

**Location of tests:** IR Laboratory at Glasgow University
<table>
<thead>
<tr>
<th>Natural Work Tasks (WT) &amp; Org</th>
<th>Natural Search Tasks (ST)</th>
<th>Actor</th>
<th>Perceived Work Tasks</th>
<th>Perceived Search Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT Structure</td>
<td>ST Structure</td>
<td>Domain Knowledge</td>
<td>Perceived WT Structure</td>
<td>Perceived Information Need Content</td>
</tr>
<tr>
<td>WT Dependencies</td>
<td>ST Dependencies</td>
<td>Experience on Search Task</td>
<td>Perceived WT Dependencies</td>
<td>Perceived ST Specificity &amp; Complexity</td>
</tr>
<tr>
<td>WT Requirements</td>
<td>ST Requirements</td>
<td>Stage in Work Task Execution</td>
<td>Perceived WT Requirements</td>
<td>Perceived ST Dependencies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sources of Difficulty</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Motivation &amp; Emotional State</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Independent Variables**
<table>
<thead>
<tr>
<th>Document and Source</th>
<th>IR Engines IT Component</th>
<th>IR Inter-faces</th>
<th>Access and Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document Structure</td>
<td>Exact Match Models</td>
<td>Domain Model Attributes</td>
<td>Interaction Duration</td>
</tr>
<tr>
<td>Document Types</td>
<td>Best Match Models</td>
<td>System Model Features</td>
<td>Actors or Components</td>
</tr>
<tr>
<td>Document Genres</td>
<td>Degree of Doc. Structure and Content Used</td>
<td>User Model Features</td>
<td>Kind of Interaction and Access</td>
</tr>
<tr>
<td>Information Type in Document</td>
<td>Use of NLP to Document Indexing</td>
<td>System Model Adaption</td>
<td>Strategies and Tactics</td>
</tr>
<tr>
<td>Communication Function</td>
<td>Doc. Metadata Representation</td>
<td>User Model Building</td>
<td>Purpose of Human Communication</td>
</tr>
<tr>
<td>Temporal Aspects</td>
<td>Use of Weights in Doc. indexing</td>
<td>Request Model Builder</td>
<td>Purpose of System Communication</td>
</tr>
<tr>
<td>Document Sign Language</td>
<td>Degree of Req. Structure and Content Used</td>
<td>Retrieval Strategy</td>
<td>Interaction Mode</td>
</tr>
<tr>
<td>Layout and Style</td>
<td>Use of NLP to Request Indexing</td>
<td>Response Generation</td>
<td>Least effort Factors</td>
</tr>
<tr>
<td>Document Isness</td>
<td>Req. Metadata Representation</td>
<td>Feedback Generation</td>
<td>-</td>
</tr>
<tr>
<td>Document Content</td>
<td>Use of Weights in Requests</td>
<td>Mapping ST History</td>
<td></td>
</tr>
<tr>
<td>Contextual Hyperlink Structure</td>
<td></td>
<td>Explanation Features</td>
<td></td>
</tr>
<tr>
<td>Human Source (see Actor)</td>
<td></td>
<td>Transformation of Messages</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Controller</td>
<td></td>
</tr>
</tbody>
</table>

**Controlled Variables**
Agenda - 3

- **Naturalistic Field Investigations of IIR** (20 min)
  - Integrating context variables
  - Live systems & (simulated/real) work tasks

- **Wrapping up of Tutorial** (5 min)

Questions are welcome during the tutorial sessions
Keep things simple!

- If you can isolate one (or two) variables as independent – then stick to that.
- Real-life studies are much more uncertain and complex than laboratory tests.
- **A robust research setting is crucial**
- **Natural search jobs** (e.g. exploratory) mixed with simulated ones (but must be realistic!)
- Test persons do relevance assessments!
Agenda - 3

✓ Naturalistic Field Investigations of IIR (20 min)
  ✓ Integrating context variables
  ✓ Live systems & (simulated) work tasks

■ Wrapping up (5 min)

Questions are welcome during the tutorial sessions
Natural IR Interaction KMS Sample


- **Research setting:** Danish Pharmaceutical Company

- **Goal:** To observe if a company thesaurus (ontology) based on *human conceptual associations* affects searching behavior, retrieval performance and searcher satisfaction different from a *domain-based thesaurus*. 

Made from several association tests with 35 employees from the company, supplying synonyms, narrow and broader concepts, based on the "company vocabulary" (task/product-based).

This thesaurus was larger in number of entries (379 more) and associative terms than the "control thesaurus" – made by domain expert and based on the "scientific vocabulary".
Research Design - 1

- **20 test persons** from the basic and clinical researchers, including marketing employees (also with scientific background)

- **3 simulated search task situations** (next slide) per test person, all having same structure and based on real work tasks observed by recently logged requests to company retrieval system.

- **“Blind testing”** of the two thesaurus types: test persons were told that the investigation was part of the system design process. Only the research team knew who searched which thesaurus type!
Search Job A

You are Product Manager working for Lundbeck Pharma.

A physician who wants to know if the combination of Citalipram and Lithium leads to approve therapeutic effect on Bipolar Disorders, has consulted you.

You need to find reports or articles investigating interaction and effect of the two drugs.
Research Design - 2

- **Steps in the field study** *(2 hours per test person):*
  1. Capture search skills (e-mail questionnaire)
  2. Explanation session
  3. **Pre-search interview** of searcher’s mental model concerning each search job / expectations
  4. **Search session with relevance assessments** (logging and structured observation of each job)
  5. **Post-search interview** of motivation & satisfaction for each search job.
Research Design - 3

- **Latin square execution** (slide 59) to avoid learning effects & all search jobs are tried out on both thesauri:
  - 10 persons x 3 search jobs in ASSO = 30 units
  - 10 persons x 3 search jobs in DOMAIN = 30 units (in reality there were only 2 x 29, due to error)
- **Relevance assessments**: three-graded: Highly relevant; Partially relevant; Not relevant.
- **Measures**: Recall/Precision; Behavior; Satisfaction
Research Design - 4

- **Independent Variable:**
  - Document Metadata Representation (two values)

- **Controlled Variables:**
  - Natural Work/Search Task Org. setting
  - Perceived Work Task Structure; Complexity (high)
  - Perceived Information Need
  - Database; Retrieval Engine; Interface

- **Hidden Variables:** Test person characteristics
Naturalistic Field Study (M.L. Nielsen) - variables

Information Objects

Interface

Social

Task Org. Context

Cultural

IT

Metadata Struc.

Thesauri

Actor Char.

Work task perception

Search task perception

Task

Org. Context
<table>
<thead>
<tr>
<th>Document and Source</th>
<th>IR Engines IT Component</th>
<th>IR Inter-faces</th>
<th>Access and Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document Structure</td>
<td>Exact Match Models</td>
<td>Domain Model</td>
<td>Interaction Duration</td>
</tr>
<tr>
<td>Document Types</td>
<td>Best Match Models</td>
<td>Attributes</td>
<td></td>
</tr>
<tr>
<td>Document Genres</td>
<td>Degree of Doc. Structure and Content Used</td>
<td>System Model Features</td>
<td>Actors or Components</td>
</tr>
<tr>
<td>Information Type in Document</td>
<td>Use of NLP to Document Indexing</td>
<td>User Model Features</td>
<td>Kind of Interaction and Access</td>
</tr>
<tr>
<td>Communication Function</td>
<td>Doc. Metadata Representation</td>
<td>System Model Adaption Strategy</td>
<td>Strategies and Tactics</td>
</tr>
<tr>
<td>Temporal Aspects</td>
<td>Use of Weights in Doc. indexing</td>
<td>User Model Building</td>
<td>Purpose of Human Communication</td>
</tr>
<tr>
<td>Document Sign Language</td>
<td>Degree of Req. Structure and Content Used</td>
<td>Request Model Builder</td>
<td>Purpose of System Communication</td>
</tr>
<tr>
<td>Layout and Style</td>
<td>Use of NLP to Request Indexing</td>
<td>Retrieval Strategy</td>
<td>Interaction Mode</td>
</tr>
<tr>
<td>Document Isness</td>
<td>Req. Metadata Representation</td>
<td>Response Generation</td>
<td>Least effort Factors</td>
</tr>
<tr>
<td>Document Content</td>
<td>Use of Weights in Requests</td>
<td>Feedback Generation</td>
<td>-</td>
</tr>
<tr>
<td>Contextual Hyperlink Structure</td>
<td>Mapping ST History</td>
<td>Explanation Features</td>
<td></td>
</tr>
<tr>
<td>Human Source (see Actor)</td>
<td>Transformation of Messages</td>
<td>Transformation of Messages</td>
<td></td>
</tr>
</tbody>
</table>

Scheduler
<table>
<thead>
<tr>
<th>Natural Work Tasks (WT) &amp; Org</th>
<th>Natural Search Tasks (ST)</th>
<th>Actor</th>
<th>Perceived Work Tasks</th>
<th>Perceived Search Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT Structure</td>
<td>ST Structure</td>
<td>Domain Knowledge</td>
<td>Perceived WT Structure</td>
<td>Perceived Information Need Content</td>
</tr>
<tr>
<td>WT Dependencies</td>
<td>ST Dependencies</td>
<td>Experience on Search Task</td>
<td>Perceived WT Dependencies</td>
<td>Perceived ST Specificity &amp; Complexity</td>
</tr>
<tr>
<td>WT Requirements</td>
<td>ST Requirements</td>
<td>Stage in Work Task Execution</td>
<td>Perceived WT Requirements</td>
<td>Perceived ST Dependencies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sources of Difficulty</td>
<td>Perceived WT Domain &amp; Context</td>
<td>Perceived ST Domain &amp; Context</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Motivation &amp; Emotional State</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sources of Difficulty

Motivation & Emotional State
Selected Results

- Both thesauri show **same IR performance level**
- Both thesauri applied to **Query Formulation & Modification** or as **Lead-in Terms**:
  - Finding **synonyms** and/or more **specific** terms
  - Clarifying **meaning** (in task perspective) of terms
- ASSO applied slightly more time (used for Narrow Terms capture)
- **DOMAIN** applied more in pre-search stage
Selected Results 2

- **Recall / Precision:**
  - **ASSO:** .14 / .32 – **DOMAIN:** .11 / .37

- **Note:** Test persons assessed same documents quite differently!

- **This was due to two fundamentally different groups of test persons** (hidden variable!):
  - Basic researchers (exploring new drugs)
  - Clinical researchers (clinical drug tests)

  This also concerns the satisfaction of the use of the thesauri for **IR** (which was quite high)
Agenda - 3

- Naturalistic Field Investigations of IIR (20 min)
  - Integrating context variables
  - Live systems & (simulated) work tasks

- Wrapping up of Tutorial (5 min)

Questions are welcome during the tutorial sessions
Step-by-Step into Light!

- In pure ‘laboratory experiments’ only simulations of searcher behavior can be done;

- If one wishes to stick to existing test collections, with existing sets of relevance assessments and ‘topics’, only IR interaction ‘ultra-light’ can be done (in order to avoid learning effects by test persons):
  - Requires short-term IR interaction;
  - In the form of ‘laboratory studies’.
  - Number of test persons, search jobs and research setting follow same line as Interactive ‘light’ IR.
Step-by-Step into Context - Light!

- **IR interaction ‘light’** entails session-based IR, with test persons’ relevance assessments and more intensive monitoring (logs; interviews; observation);
  - Can be carried out as laboratory study or field experiment
- Like in ‘ultra-light’ and ‘naturalistic’ IR, **number of test persons and search jobs** must assure that ‘statistically enough’ data is present in the result matrices when cross tabulating independent variables (see slides 40-41).
- **IR interaction ‘light’**: assigned realistic requests, simulated task situations and **searcher relevance assessments**
- **Naturalistic IR interaction** assumes natural tasks (mixed with simulated ones) in natural environments
‘Ultra-light’ and ‘Light’ IIR

- Assessments can be 4-graded (Vakkari & Sormunen, 2004);
- Realistic but few relevance assessments per person;
- Assessments can be pooled for same search job over all test persons – weighted doc. assessments
- Common recall/precision, MAP, CumGain, P@n, etc. feasible
- You require min. 30 responses per result cell
- Ultra-light lab. studies are effective for tightly controlled IIR experiments (like Kelly et al.)
The Cognitive Research Framework informs about …

- Central variables to combine as independent ones
- Major variables kept controlled/neutralized in a setting
- What kind of variables that are hidden!
- Dependent variables depend on the research goals (the independent variables!)

Novel possible research designs, settings and measures … there is a lot to do - really!
THANK YOU!

http://www.springeronline.com/1-4020-3850-X