# HCI View of Information Retrieval Evaluation

Zinal, January 24° 2012


SAPIENZA
UNIVERSITÀ DI ROMA

Tiziana Catarci

# The mythical perfect user

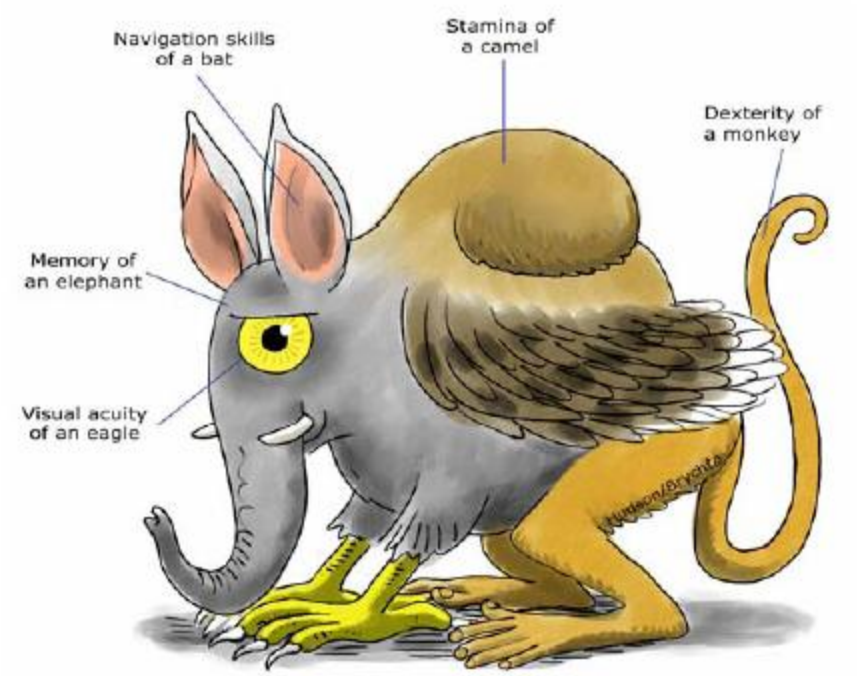



Figure 2, the mythical "perfect user"

# Outline

- HCI and IR
- User interface techniques
- Interaction measures
- Usability models
- Usability evaluation

# HCI and IR

- IR and HCI are related fields with strong traditions that have been challenged and energized by the WWW.
- The type and nature of content have evolved and changed
  - New data types
  - New content relationships
  - More dynamicity
- The type and nature of users have evolved.
  - No expert users
  - High expectations

# HCIR

- “we think of information interaction from the perspective of an active human with information needs, information skills, powerful digital library resources (that include other humans) situated in global and local connected communities – all of which evolve over time.” Marchionini’06

(http://www.asis.org/Bulletin/Jun-06/marchionini.html)

# User-oriented IR system features

- Should be implemented and evaluated in a way that reflects the users' needs
- Cannot just deliver the relevant documents, but must also provide facilities for making meaning with those documents
- Should increase user responsibility and control
- Should have flexible architectures
- Should aim to be part of information ecology of personal and shared memories and tools
- Should support the entire information life cycle
- Should support tuning by end users
- Should be engaging and fun to use

# Key Issues

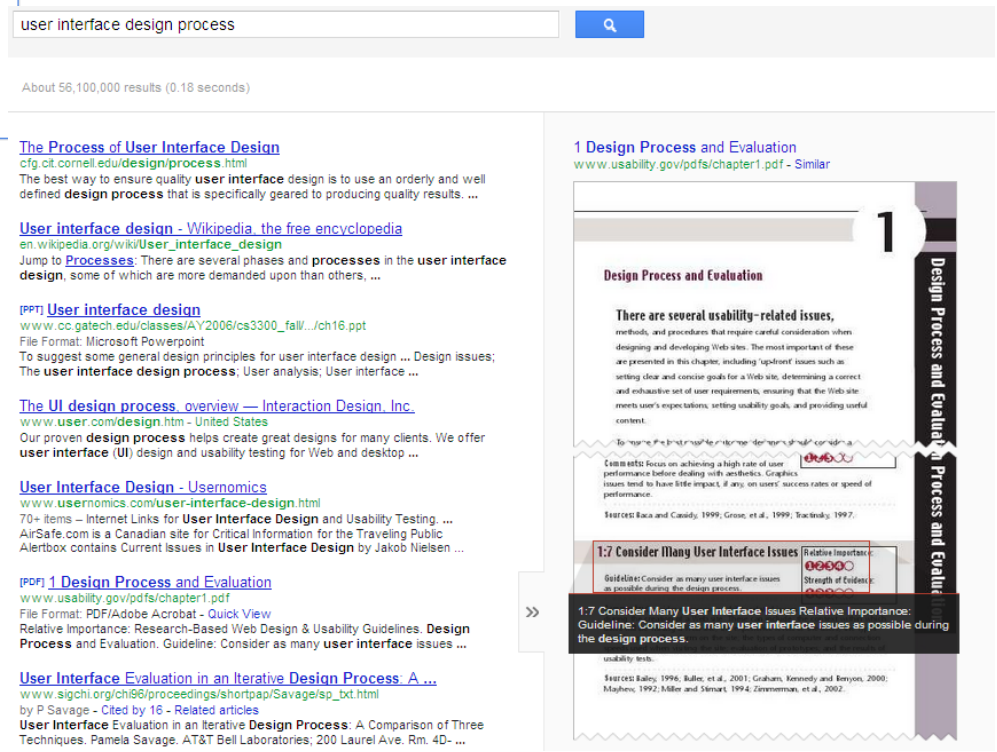
- User interface
  - comprises the elements that the user comes into contact with when using a computing system
  - the interaction part
    - how the user interface works and its behaviour in response to what the user does while performing a task
  - the interface software part
    - the implementation of the interaction component
- User-oriented evaluation
  - traditional methods mainly concerned with system-oriented measurements (precision and recall), but not on usability
  - no well-established evaluation approaches for studying users and their interactions with information retrieval systems

# User interface techniques

- Query reformulation
- Browsing
- Faceted search and navigation
- Lookahead
- Relevance feedback
- Summarization, analytics and visual presentation


user interface de
user interface design
user interface design concepts
user interface design pdf
user interface design process ppt
user interface design lecture notes
user interface design principles
user interface design examples
user interface definition
user interface design patterns
user interface design process
Google Search
I'm Feeling Lucky

# Lookahead


user interface design process
About 56,100,000 results (0.18 seconds)
The Process of User Interface Design
cfg.cit.cornell.edu/design/process.html
The best way to ensure quality user interface design is to use an orderly and well defined design process that is specifically geared to producing quality results. ...
User interface design - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/User_interface_design
Jump to Processes: There are several phases and processes in the user interface design, some of which are more demanded upon than others, ...
[PPT] User interface design
www.cc.gatech.edu/classes/AY2006/cs3300_fall/.../ch16.ppt
File Format: Microsoft Powerpoint
To suggest some general design principles for user interface design ... Design issues; The user interface design process; User analysis; User interface ...
The UI design process, overview — Interaction Design, Inc.
www.user.com/design.htm - United States
Our proven design process helps create great designs for many clients. We offer user interface (UI) design and usability testing for Web and desktop ...
User Interface Design - Usernomics
www.usernomics.com/user-interface-design.html
70+ items – Internet Links for User Interface Design and Usability Testing. ... AirSafe.com is a Canadian site for Critical Information for the Traveling Public Alertbox contains Current Issues in User Interface Design by Jakob Nielsen ...
[PDF] 1 Design Process and Evaluation
www.usability.gov/pdfs/chapter1.pdf
File Format: PDF/Adobe Acrobat - Quick View
Relative Importance: Research-Based Web Design & Usability Guidelines. Design Process and Evaluation. Guideline: Consider as many user interface issues ...
User Interface Evaluation in an Iterative Design Process: A ...
www.sigchi.org/chi96/proceedings/shortpap/Savage/sp_txt.html
by P Savage - Cited by 16 - Related articles
User Interface Evaluation in an Iterative Design Process: A Comparison of Three Techniques. Pamela Savage. AT&T Bell Laboratories; 200 Laurel Ave. Rm. 4D- ...
1 Design Process and Evaluation
www.usability.gov/pdfs/chapter1.pdf - Similar
1
Design Process and Evaluation
There are several usability-related issues,
1:7 Consider Many User Interface Issues
1:7 Consider Many User Interface Issues Relative Importance: Guideline: Consider as many user interface issues as possible during the design process.

# Appropriate Evaluation Metrics and Models

- Performance measures
- Interaction measures
- Usability measures
- Contextual measures

# Evaluation Metrics, Models and Techniques

Some classic IR evaluation measures.

| Measure | Description |
|---|---|
| Recall | The number of retrieved relevant documents divided by the number of relevant documents in corpus. |
| Precision | The number of relevant retrieved documents divided by the number of retrieved documents. |
| $F$-measure | The $F$-measure is a way of combining precision and recall and is equal to their weighted harmonic mean $[F = 2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})]$. The $F$-measure also accommodates weighting of precision or recall, to indicate importance. |
| Average precision (AP) | Individual precision scores are computed for each relevant retrieved document (with 0 assigned to relevant documents that are not retrieved). These values are then summed and divided by the total number of relevant documents in the collection. Thus, AP has a recall component to it and is typically described as the area underneath the precision/recall curve. AP also takes into account the position of relevant documents in the result list. |
| Mean average precision (MAP) | This is a run level measure and consists of taking the average of the average precision values for each topic. |
| Geometric average precision (GMAP) | The geometric mean of $n$ values is the $n$th root of the product of the $n$ values. Robertson [219] recommends taking the logs of the values and then averaging. GMAP was developed for the TREC Robust Track, which explored retrieval for difficult topics and does a better job than MAP of distinguishing performance scores at the low end of the AP scale. |
| Precision at $n$ | The number of relevant documents in the top $n$ results divided by $n$. Typical values for $n$ are 10 and 20, which is thought to better represent the user's experience since research has shown that this is the extent to which users look through Web search results [146]. |
| Mean reciprocal rank (MRR) | This measure was developed for high-precision tasks where only one or a small number of relevant documents are needed. For a single task with one relevant document, reciprocal rank is the inverse of its ranked position. MRR is the average of two or more reciprocal rank scores (used when there is more than one task). |

# Interactive Recall and Precision

Modified versions of recall and precision for interactive IR [284, 285] and relative relevance [33, 36].

| Measure | Description |
|---|---|
| Interactive recall | Number of TREC relevant saved by user/number of TREC relevant documents in the corpus. |
| Interactive TREC precision | Number of TREC relevant documents viewed by the user/total number viewed. |
| Interactive user precision | Number of TREC relevant documents saved by the user/total number saved by the user. |
| Relative relevance (RR) | Cosine similarity measure between two lists of relevance assessments for the same documents. |

# Multi-level relevance and rank measures

- Cumulated gain measures
- Ranked half-life
- Expected search length
- Expected search duration
- Average search length
- Immediate accuracy

# Cumulated gain and ranked half-life

Cumulated gain measures [148, 149] and ranked half-life [33, 36].

| Measure | Description |
|---|---|
| Cumulated gain (CG) | Cumulated gain can be computed at different cut-off values for search result of lists of varying sizes. At the cut-off point, CG is the sum of the relevance values of all documents up to and including the document at the cut-off point. |
| Discounted cumulated gain (DCG) | Discounted cumulated gain discounts the value of relevant documents according to their ranked position. New relevance values are computed by dividing the relevance score of a document by the logarithm of its rank. The discounted relevance scores are then summed to a particular cut-off point. |
| Normalized discounted cumulated gain (nDCG) | The DCG measure is normalized according to the best DCG available for a given results list. This normalization transforms DCG scores, which can take on a large range of numbers, to a 0–1 scale, which is easier to interpret and compare. |
| Ranked half-life (RHL) | The point in the results list at which half of the total relevance value for the entire list of documents has been achieved. If binary assessments are used, this is the point at which half of the relevant documents in the list have been observed. If multi-level assessments are used, this point is when half of the sum total of all of the relevance values are observed. |

# Time-based measures

Time-based measures from Käki and Aula [158].

| Measure | Description |
|---|---|
| Search speed | The proportion of answers that are found per minute. This measure consists of dividing the total number of answers found by the length of time it took to find the answers. All answers are included in this computation regardless of whether they are correct. |
| Qualified search speed | This measure accommodates multi-level relevance and consists of computing search speed for each relevance category, including non-relevant. |

## Other measures

- Informativeness
- Cost and utility measures
- Contextual measures
- User characteristics

# Measures of information needs

- Task-related measures (e.g. task-type, task familiarity, task difficulty and complexity)
- Topic-related measures (e.g. topic familiarity and domain expertise)
- Persistence of information need
- Immediacy of information need
- Information-seeking stage
- Purpose, goals and expected use of the results

# Interaction measures

- Number of queries
- Number of search results viewed
- Number of documents viewed
- Number of documents saved
- Query length

# Usability

- According to ISO 9241-11, usability is defined as

"the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use".

- ISO 9241 also identifies the most useful indicators in measuring the level of usability of a product. They are:
- effectiveness in use (accuracy and completeness through which users achieve certain results);
- efficiency in use (resources utilized in relation to accuracy and completeness through which users achieve certain results);
- satisfaction in use (freedom from inconveniences and positive attitude towards the use of a product).

# Usability and System design

The usability measures the distance between the “designer model” or the computer-system model and its modalities of use in possession of the designer, and the “user model” - the functioning model of the system that the user creates and which regulates its interaction with the system itself. The closer the models, the more the system is to be considered as usable in relation to the customer-user data. In this respect, the usability level of a product is only measurable in the real work environment, during the real use of the product by a specific user.

ISO 9241-11 defines the context of use as an environment containing users, tasks to carry out, hardware and software resources or other materials used and the physical and social conditions in which the product is used. However, the learning capacity, the level of work experience, education, the personal consideration that the user has on how the system works and the purposes of its use are mainly the characteristics to be taken into consideration.

# Some Metrics from ISO 9241

| Usability objective | Effectiveness measures | Efficiency measures | Satisfaction measures |
|---|---|---|---|
| Suitability for the task | Percentage of goals achieved | Time to complete a task | Rating scale for satisfaction |
| Appropriate for trained users | Number of power features used an expert user | Relative efficiency compared with power features | Rating scale for expert satisfaction |
| Learnability | Percentage of functions learned | Time to learn criterion | Rating scale for ease of learning |
| Error tolerance | Percentage of errors corrected successfully | Time spent on correcting errors | Rating scale for error handling |

## Usability measures

- **Effectiveness**
  - Error rate
  - Binary task completion
- **Efficiency**
  - the overall time the subject takes
  - amount of time the subject spends doing specific things
  - amount of time the subject spends in specific or different modes
- **Satisfaction**
- **Preference**
- **Mental effort and cognitive load**
- **Flow and engagement**
- **Learning and cognitive transformation**

# IR system

- A device interposed between a potential user of information and the information collection itself (Harter 1986)
- Three major components:
  - Database
  - Communication channel (interface)
  - User
- Cognitive aspects of the user are getting more consideration

# Models

- Ingwersen's cognitive model, five dimensions for the cognitive viewpoint
  - Information processing takes place in senders and recipients of messages;
  - Processing takes place at different levels;
  - During communication of information any actor is influenced by its past and present experiences (time) and its social, organizational and cultural environment;
  - Individual actors influence the environment or domain;
  - Information is situational and contextual

# Models (2)

- Belkin's episodes model, what happens in interaction as a process
  - Processes of judgement, use, interpretation, etc depending on user's goals, tasks
  - Processes of navigation, comparison, summarization, etc
  - Involving different aspects of information and information objects
- Saracevic's stratified model of IR interaction
  - Users interact with IR systems in order to use information
  - The use of information is connected with cognition and then situational application

# Models(3)

- Ellis' model of information-seeking behaviours
  - Five information seeking characteristics: 1) starting, 2) chaining, 3) browsing, 4) differencing, 5) monitoring, and 6) extracting
- Kuhlthau's model
  - Defines the tasks involved in the information seeking process from a psychological perspective, containing affective/feelings, cognitive/thoughts, and physical/action activities
  - Stages: Initiation, Selection, Exploration, Formulation, Collection and Presentation
- IIR (interactive information retrieval) evaluation model
  - realistic scenarios and the (call for) alternative performance measures

# Models(4)

- Allen's model

| COMPONENT | METHOD | TASK |
|---|---|---|
| Resource Analysis | Description of information system functionality | Describe resources used to complete the tasks. |
| User Needs Analysis | • Questionnaire (qualitative and quantitative data)<br>• Log statistics (quantitative data) | • Users goals, purpose, objectives, actions, individual preferences.<br>• Measures like time, type of actions |
| Task analysis | Hierarchical Task Analysis | Users tasks, goals and activities that they accomplish when meeting their needs. |
| User Modeling | | |
| Designing for usability | Requirement lists (qualitative data) | Requirements for user interface redesign |

# Models(5)

- Ahmed's user-centered approach
  - competitive analysis of an existing IR system to perform usability testing.
  - user task analysis based on activities during usability test.
  - initial prototype design drawn from task analysis.
  - heuristic evaluation of the initial prototype design.
  - interactive prototype design, incorporating input from heuristic evaluation.
  - formative evaluation of the interactive prototype using task scenarios.
  - revised prototype design based on formative evaluations, and finally.
  - summative evaluation of the final prototype design and a comparison of the results with the results of competitive analysis for performing the same tasks

# Usability evaluation

- Goals of evaluation
  - assess extent of system functionality
  - assess effect of interface on user
  - identify specific problems
- Where? Occurs in laboratory, field and/or in collaboration with users
- What? Evaluates both design and implementation
- When? Should be considered at all stages in the design life cycle
- Who? Users and/or experts

# Framework for Usability Evaluation in IR

- Participants
  - HCI Experts (as usual)
  - Users
    - To investigate people information seeking needs
    - Crowdsourcing and mechanical Turk
- Tasks
  - Formulation and submission of a query
  - Examination of the results
  - Possible feedback loop to re-formulate the query
  - Integration of search results and evaluation of the whole search

# Crowdsourcing

- The act of sourcing tasks traditionally performed by specific individuals to a group of people or community (crowd) through an open call.
- The concept of crowdsourcing depends essentially on the fact that because it is an open call to a group of people, it gathers those who are most fit to perform tasks, solve complex problems and contribute with the most relevant and fresh ideas.
- The public may be invited to develop a new technology, carry out a design task, refine or carry out the steps of an algorithm, etc.
- Mass collaboration enabled by Web 2.0 technologies to achieve business goals.
- The **Amazon Mechanical Turk** (**MTurk**) is a crowdsourcing Internet marketplace. It is one of the suites of Amazon Web Services.
- The Requesters are able to post tasks requiring human intelligence.
- *Workers* can then browse among existing tasks and complete them for a monetary payment set by the Requester.
- Requesters can ask that Workers fulfill Qualifications before engaging a task, and they can set up a test in order to verify the Qualification. They can also accept or reject the result sent by the Worker, which reflects on the Worker's reputation

# Framework for Usability Evaluation in IR(2)

- Usage of realistic scenarios
- Simulated  work task situation
- Usability measures
  - Effectiveness
    - *interactive recall*
    - *interactive precision*
    - *interactive TREC precision*
    - informativeness
    - cost
    - Utility
  - Efficiency
    - the overall time the user takes
    - the time the user takes doing specific things
    - the time the user takes in specific or different modes
  - Satisfaction

# Framework for Usability Evaluation in IR(3)

- Interaction measures
  - Number of queries
  - Number of search results viewed
  - Number of documents viewed
  - Number of documents saved
  - Query length
  - Appropriate combinations of the above measures
- User characteristic measures
  - *sex, age, profession, computer experience, search experience*, Internet *perceptions*, *cognitive style,* etc.
  - Preference
  - Mental effort and cognitive load
  - Flow and engagement
  - Learning and cognitive transformation

## Information need measures

- Task-related measures (e.g., task type, task familiarity, task difficulty)
- Topic-related measures (e.g., topic familiarity and domain expertise)
- Peristence of information need
- Immediacy of information need
- Information-seeking stage
- Purpose, goals and expected use of the results

# Information Foraging theory (Pirolli and Card 1999)

- Describes IR behaviour through the similarity with food foraging
- The basis is a cost and benefit assessment of achieving a goal where cost=resources consumed, benefit=what is gained
- Cost-benefit assessment essential for any goal-driven activity (like IR)
- Key concepts: food source, location where to find it, tools available, benefit
- Information: items fulfilling the information need, information patches in which information is clustered, information scent determining the value of the items, information diet determining the decision on which items
- Challenge of IR community is to design interfaces that effectively support these concepts
- ACT-IF cognitive process model for evaluating IR systems

# Framework for Usability Evaluation in IR(4)

- Evaluation methods
  - Expert-based
    - heuristic evaluation
    - cognitive walkthrough
  - User-based
    - usability tests
    - observational methods (e.g. think aloud, stimulated recall/post-task walkthrough, transaction logging)
    - query techniques (e.g., questionnaires and interviews)
    - physiological monitoring methods (e.g., eye tracking, measuring skin conductance, measuring heart rate)

# Heuristic Evaluation

- Proposed by Nielsen and Molich.
- usability criteria (heuristics) are identified
- design examined by experts to see if these are violated
- Example heuristics
  - system behaviour is predictable
  - system behaviour is consistent
  - feedback is provided
- Heuristic evaluation `debugs' design

# Cognitive Walkthrough

Proposed by Polson *et al.*

- evaluates design on how well it supports user in learning task
- usually performed by expert in cognitive psychology
- expert ‘walks though’ design to identify potential problems using psychological principles
- forms used to guide analysis

- For each task walkthrough considers
  - what impact will interaction have on user?
  - what cognitive processes are required?
  - what learning problems may occur?
- Analysis focuses on goals and knowledge: does the design lead the user to generate the correct goals?

# Usability tests (Controlled experiments)

- Controlled evaluation of specific aspects of interactive behaviour
- Evaluator chooses hypothesis to be tested
- A number of experimental conditions are considered which differ only in the value of some controlled variable.
- Changes in behavioural measure are attributed to different conditions

# Usability tests (2)

- Subjects
  - who – representative,  sufficient sample
- Variables
  - things to modify and measure
  - independent variable (IV)

    characteristic changed to produce different conditions - e.g. interface style, number of menu items
  - dependent variable (DV)

    characteristics measured in the experiment - e.g. time taken, number of errors.

# Usability tests (3)

- Hypothesis
  - what you'd like to show
  - prediction of outcome
  - framed in terms of IV and DV - e.g. "error rate will increase as font size decreases"
  - null hypothesis: states no difference between conditions, aim is to disprove this - e.g. null hyp. = "no change with font size"

# Usability tests (4)

- Experimental design
  - how you are going to do it
  - within groups design
    - each subject performs experiment under each condition.
    - transfer of learning possible
    - less costly and less likely to suffer from user variation.
  - between groups design
    - each subject performs under only one condition
    - no transfer of learning
    - more users required
    - variation can bias results.

# Usability tests – Analysis of data

- Before you start to do any statistics:
  - look at data
  - save original data
- Choice of statistical technique depends on
  - type of data
  - information required
- Type of data
  - discrete - finite number of values
  - continuous - any value

# Heuristic evaluation vs. Usability testing

- Doubleday et al. (1997)
  - The expert evaluators identified 86 usability problems whereas 38 problems were identified in the user testing. However, none of the 38 problems found by user testing were identified by the expert evaluators.
- Cogdill (1999)
  - the expert evaluators identified 27 usability problems compared to 21 problems found in the usability test.
  - Using both heuristic evaluation and usability testing resulted in a high degree of comprehensiveness in the study.
- Expert-based and user-based evaluation methods can play a complementary role in evaluating information retrieval systems

# Transaction logging

- Re-popularized method
- Relies on computer and Web monitoring tools in order to collect logs characterizing user's interaction with the system
- Most transaction logging tools can run in the background while the user interacts with the information retrieval system, without causing any distractions or disruption
- Can capture users' natural search behaviours without interrupting them

# Questionnaires (1)

- Set of fixed questions given to users
- Advantages
  - quick and reaches large user group
  - can be analyzed more rigorously
- Disadvantages
  - less flexible
  - less probing
- Need careful design
  - what information is required?
  - how are answers to be analyzed?
- Styles of question
  - General, open-ended, scalar, multi-choice, ranked

# Questionnaires (2)

- Can be used at various points during an evaluation of an IR system
- Various types
  - screening questionnaire
  - pre-study questionnaire
  - post-study questionnaire
- Can be administered electronically or manually
- Studies show that subjects' responses to closed-questions were significantly more positive when elicited electronically, than manually

## Interviews

- Analyst questions user on one-to -one basis usually based on prepared questions
- Informal, subjective and relatively cheap
- Advantages
  - can be varied to suit context
  - issues can be explored more fully
  - can elicit user views and identify unanticipated problems
- Disadvantages
  - very subjective
  - time consuming
- In IR interviews seem to be more appropriate when one is asking complex, abstract questions than when one is asking relatively easy questions
- Can also be useful in information retrieval evaluation during simulated recall/post-task walkthrough

# Think aloud (1)

- User observed performing task
- User asked to describe what he is doing and why, what he thinks is happening etc.

- Advantages
  - simplicity - requires little expertise
  - can provide useful insight
  - can show how system is actually use
- Disadvantages
  - subjective
  - selective
  - act of describing may alter task performance

# Think-aloud (2)

- Users may have a difficult time simultaneously articulating their thoughts and carrying out the information retrieval task that they have been given
- Additional cognitive demands
- Short training task
- Also simulated recall/post-task walkthrough

## Post-task walkthrough

- The researcher records the screen of the computer as the user performs the searching task. After the searching task is complete, the recording is played back to the user who is then asked to articulate his/her thoughts and decision-making as the recording is played
- Transcript played back to participant for comment
  - immediately → fresh in mind
  - delayed → evaluator has time to identify questions
- Useful to identify reasons for actions and alternatives considered
- Necessary in cases where think aloud is not possible

# Protocol analysis

- Paper and pencil – cheap, limited to writing speed
- Audio – good for think aloud, difficult to match with other protocols
- Video – accurate and realistic, needs special equipment, obtrusive
- Computer logging – automatic and unobtrusive, large amounts of data difficult to analyze
- User notebooks – coarse and subjective, useful insights, good for longitudinal studies
- Mixed use in practice.
- audio/video transcription difficult and requires skill.
- Some automatic support tools available

## Choosing an evaluation method

| | |
|---|---|
| when in process: | design vs. implementation |
| style of evaluation: | laboratory vs. field |
| how objective: | subjective vs. objective |
| type of measures: | qualitative vs. quantitative |
| level of information: | high level vs. low level |
| level of interference: | obtrusive vs. unobtrusive |
| resources available: | time, subjects, equipment, expertise |

More work needed to define a complete IR-tailored evaluation framework

# References (1)


- Ahmed, S. M. Z., McKnight, C., Oppenheim, C. (2006). A user-centred design and evaluation of IR interfaces, Journal of Librarianship and Information Science, Vol. 38 No.2, pp.157-72.
- Allen, B. (1996a). From research to design: A user-centered approach. In: Ingwersen, P. and Pors, N. O. (eds.), CoLIS 2. Second International Conference on Conceptions of Library and Information Science: Integration in Perspective, Copenhagen, Denmark. pp. 45-59. Copehagen : The Royal School of Librarianship.
- Allen, B. (1996b), Information tasks. Towards a user-centered approach to information systems. San Diego : Academic Press.
- Alonso, O., Rose, D. E., and Stewart, B. (2008). Crowdsourcing for relevance evaluation. SIGIR Forum, vol. 42, pp. 10–16.
- Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern information retrieval Reading, MA: Addison-Wesley.

# References (2)


- Belkin, N. J., Cool, C., Stein, A. and Thiel, U. (1995). Cases, scripts and information seeking strategies: on the design of interactive information retrieval systems. Expert Systems with Applications. 9: 379-395.
- Belkin, N., J., Cole, M. and, Liu, J. (2009). A Model for Evaluation of Interactive Information Retrieval. SIGIR Workshop on the Future of IR Evaluation.
- Bernal, J. D. (1948). Preliminary Analysis of Pilot Questionnaires on the Use of Scientific literature. The Royal Society Scienti?c Information Conference. pp. 589–637.
- Borgman, C. L., Hirsh, S. G., Walter, V. A., Gallagher, A. L. (1995). Children's searching behaviour on browsing and keyword online catalog: the Science Library Catalog Project, Journal of the American Society for Information Science, Vol. 46 No.9, pp.663-84.
- Borlund, P. (2003a). The concept of relevance in IR, Journal of the American Society for Information Science, vol. 54, pp. 913–925.

# References (3)


- Borlund, P. (2003b). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Information Research, Vol. 8 No. 3, April 2003.
- Borlund, P. and Ingwersen, P. (1998). Measure of relative relevance and ranked halflife: Performance indicators for interactive information retrieval, in Proceedings of the 21st ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR ’98), pp. 324–331, Melbourne, Australia.
- Boyce, B. R., Meadow, C. T., and Kraft, D. H. (1994). Measurement in Information Science. London, UK: Academic Press, Inc.
- Callan, J., Allan, J., Clarke, J. L. A., Dumais, S., Evans, D. A., Sanderson, M., and Zhai, C. (2007). Meeting of the MINDS: An information retrieval research agenda, SIGIR Forum, vol. 41, pp. 25–34.

# References (4)


- Cleverdon, C. W. (1967). The Cranfield tests on index language devices, in Readings in Information Retrieval, (Spark-Jones, K. and Willett, P. eds.), (Reprinted from Aslib Proceedings, pp. 173–192.) San Francisco: Morgan Kaufman Publishers, 1997/1967.
- Cleverdon, C. W., Mills, L., and Keen, M. (1996). Factors Determining the Performance of Indexing Systems, vol. 1 — Design. Cranfield, England: Aslib Cranfield Research Project.
- Cogdill, K. (1999). MEDLINEplus Interface Evaluation: Final Report, University of Maryland, Human-Computer Interaction Lab (HCIL), College Park, MD.
- Cooper, W. S. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems, American Documentation, vol. 19, pp. 30–41.

# References (5)


- Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness, part 1: The “subjective” philosophy of evaluation, Journal of the American Society for Information Science, vol. 24, pp. 87–100.
- Csikszentmihalyi, M. (1997). Finding Flow: The Psychology of Engagement with Everyday Life. New York: Basic Books.
- Dalrymple, P. W. (1990). Retrieval by reformulation in two university library catalogs: Toward a cognitive model of searching behavior. Journal of the American Society.
- Dalrymple, P. W. (1991) User-Centered Evaluation of Information Retrieval. Evaluation ofPublic Services and Public Services Personnel. Bryce Allen, ed. Urbana, IL: University of Illinois, pp. 85-102.
- Doubleday, A.R., Ryan, M., Springett, M. and Sutcliffe, A. (1997). A comparison of usability techniques for evaluating design. Proceedings of the Conference on DesigningInteractive Systems: Processes, Practices, Methods, and Techniques, 18-20 August,Amsterdam, ACM, New York, NY, pp. 101-10.

# References (6)


- Dunlop, M. (1997). Time, relevance and interaction modeling for information retrieval, in Proceedings of the 20th ACM Conference on Research and Development in Information Retrieval (SIGIR ’97), pp. 206–213, Philadelphia, PA.
- Ellis, D. (1987). The derivation of a behavioral model for information system design. Unpublished doctoral dissertation. University of Sheffield, England.
- Ellis, D. (1989). A behavioural approach to information retrieval system design. Journal of Documentation, 45 (3), 171-212.
- Ellis, D., Haugan, M. (1997). Modeling the information seeking patterns of engineers and research scientists in an industrial environment. Journal of Documentation, 53 (4), 384-403.
- Fenichel, C. H. (1981). Online searching: Measures that discriminate among users with different types of experience, Journal of the American Society for Information Science, vol. 32, pp. 23–32.

# References (7)


- Ford, N., Miller, D., and Moss, N. (2001). The role of individual differences in Internet searching: An empirical study, Journal of the American Society for Information Science and Technology, vol. 52, pp. 1049–1066.
- Hansen, P. (1998). Evaluation of IR User Interface. Implications for user interface design. Human IT, 1998, No. 2, pp. 28-41.
- Harter, S. (1986). Online information retrieval. Concepts, principles, and techniques. Orlando: Academic Press.
- Hearst, M. (1999). User Interfaces and Visualization, Chapter 10 of Baeza-Yates, R. and Ribeiro-Neto, B., Modern Information Retrieval.
- Hewett, T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G., and Verplank, W. (1992). ACM SIGCHI Curricula for Human-Computer Interaction.

# References (8)


- Hix, D. and Hartson, H. R. (1993). Developing user interfaces. Ensuring usability through product and process. New York : Wiley.
- Hornbaek, K. (2005). Current practice in measuring usability: Challenges to usability studies and research, International Journal of Human–Computer Studies, vol. 64, pp. 79–102.
- Hutchinson, H. B., Drunin, A., Bederson, B. B. (2007). Supporting elementary-age children's searching and browsing: design and evaluation using the International Children's Digital Library, Journal of the American Society for Information Science, Vol. 58 No.11, pp.1618-30.
- Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory, Journal of Documentation, vol. 52, pp. 3–50.
- Ingwersen, P. and J¨arvelin, K. (2005). The Turn: Integration of Information Seeking and Retrieval in Context. Dordrecht, The Netherlands: Springer.

# References (9)


- ISO (1998). Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part II, Guidance on Usability (ISO 9241-11:1998).
- J¨arvelin, K. and Kek¨al¨ainen, J. (2000). IR evaluation methods for retrieving highly relevant documents, in Proceedings of the 23rd ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR ’00), pp. 41–48, Athens, Greece.
- J¨arvelin, K. and Kek¨al¨ainen, J. (2002). Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems (TOIS), vol. 20, pp. 422–446.
- Johnson, F. C., Griffiths, J. R., and Hartley, R. J. (2003). Task dimensions of user evaluations of information retrieval systems. Information Research, Vol. 8, No. 4.
- K¨aki, M. and Aula, A. (2008). Controlling the complexity in comparing search user interfaces via user studies, Information Processing and Management, vol. 44, pp. 82–91.

# References (10)


- Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval, 3(1-2), 1-224.
- Kelly, D., Harper, J., and Landau, B. (2008). Questionnaire mode effects in interactive information retrieval experiments. Information Processing and Management.
- Kuhlthau, C. C. (1991). Inside the search process: information seeking from the user's perspective. Journal of the American Society for Information Science. 42: 361-371.
- Kuhlthau, C. C. (1994). Seeking meaning: a process approach to library and information services. 1994, Norwood, NJ.: Ablex Publishing.
- Lin, X. (1997). Map displays for information retrieval, Journal of the American Society for Information Science, Vol. 48 No.1, pp.40-54.
- Losee, R. M. (1996). Evaluating retrieval performance given database and query characteristics: Analytical determination of performance surfaces, Journal of the American Society for Information Science, vol. 47, pp. 95–105.

# References (11)


- Marchionini, G. (1987). An invitation to browse: designing full text systems for novice users, Canadian Journal of Information Science, Vol. 12 No.3, pp. 69-79.
- Marchionini, G. (2006). Toward Human-Computer Information Retrieval Bulletin, in June/July 2006 Bulletin of the American Society for Information Science. Available online at http://www.asis.org/Bulletin/Jun-06/marchionini.html.
- Mira working group (1996). Evaluation Frameworks for Interactive Multimedia Information Retrieval Applications. Available online at http://www.dcs.gla.ac.uk/mira
- O'Brien, H. and Toms, E. (2008). What is user engagement? A conceptual framework for defining user engagement with technology, Journal of the American Society for Information Science and Technology, vol. 59, pp. 938–955.
- Pirolli, P. and Card, S. K. (1999). Information foraging. Psychological Review Vol. 106 No. 4, pp. 643-675.

# References (12)


- Rocchio, J. (1971). Relevance feedback in information retrieval. In: Salton, G. (Ed), The SMART Retrieval System.
- Salton, G. (1970). Evaluation problems in interactive information retrieval, Information Storage and Retrieval, vol. 6, pp. 29–44.
- Salton, G. (1989). Automatic text processing: the transformation, analysis and retrieval of information by computer. Reading, MA: Addison-Wesley.
- Salton, G. (1992). The state of retrieval system evaluation, Information Processing and Management, vol. 28, pp. 441–449.
- Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. JASIS 41, 4 (1990), 288–297.
- Saracevic, T. (1996). Modeling interaction in information retrieval (IR): A review and proposal. Proceedings of the 59th ASIS Annual Meeting 1996, 33, 3-9.
- Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. Proceedings of the 60th ASIS Annual Meeting 1997, 34, 313-327.

# References (13)


- Saracevic, T, and Kantor, P. (1988). A study of information seeking and retrieving. Journal of the American Society for Information Science, 39(3), 177-216.
- Siatri, R. (1999). The Evolution of User Studies. Libri, vol. 49, pp. 132–141.
- Su, L. T. (1992). Evaluation measures for interactive information retrieval, Information Processing and Management, vol. 28, pp. 503–516.
- Tague, J. (1987). Informativeness as an ordinal utility function for information retrieval, SIGIR Forum, vol. 21, pp. 10–17.
- Tague-Sutcli?e, J. M. (1992). The pragmatics of information retrieval experimentation, revisted, Information Processing and Management, vol. 28, pp. 467–490.
- Tague-Sutcliffe, J. M. (1995). Measuring Information: An Information Services Perspective. San Diego, California: Academic Press.

# References (14)


- Urquhart, D. J. (1948). The Distribution and Use of Scientific and Technical Information. The Royal Society Scienti?c Information Conference. pp. 408–419.
- Veerasamy, A. and Belkin, N. J. (1996). Evaluation of a tool for visualization of information retrieval results, in Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 85–92, Zurich, Switzerland.
- Veerasamy, A. and Heikes, R. (1997). Effectiveness of a graphical display of retrieval results, SIGIR Forum, vol. 31, pp. 236–245.
- Voorhees, E. M. and Harman, D. K. (2005). TREC: Experiment and Evaluation in Information Retrieval. Cambridge, MA: MIT Press.
- Yuan, W. and Meadow, C. T. (1999). A study of the use of variables in information retrieval user studies, Journal of the American Society for Information Science, vol. 50, pp. 140–150.