Institute of Applied Information Technology
Zurich University of Applied Sciences
Information Engineering Group

FROM EVALUATION TO APPLICATIONS
(Beyond evaluating retrieval effectiveness of IR systems)

MARTIN BRASCHLER
(with work by Melanie Imhof & Stefan Rietberger)

08.02.2013
Content

- Retrieval problem/IR cycle
- IR applications
- Application-centric evaluation
- Guerilla campaign
- Criteria
  - Index
  - Query/Document matching
  - User interaction
  - Search results
- Results
- Conclusion
The Retrieval Problem

• Robertson et al. (1982):
  “..... the function of a document retrieval system is to retrieve all and only those documents that the inquiring patron wants (or would want).”

=> this is a difficult problem, due to a number of factors, as will be demonstrated with the following example:
Information Retrieval Cycle

- Example taken from Peters et al. (2012)
Challenges for IR

• Information needs are not verbalised or formulated well enough
• Language is ambiguous: Synonyms, Homonyms, Paraphrasings, Metaphors, Word forms, Typos
• Documents are unstructured and inhomogenous
• Misleading content (spam etc.)
• Authority, source, actuality, copyright.
• Conflicting goals: recall vs. precision
• Relevance is subjective, context dependent, personal preferences, completeness of answer
Cranfield-style evaluations

• "Back-bone" of evaluation campaigns such as TREC, CLEF, NTCIR, …
• Extensively studied, well understood (including limitations!)
• Needs test collections, needs (costly) relevance assessments
• Is an evaluation of IR *systems*
IR applications

• An information retrieval application covers the whole information retrieval cycle
• IR applications support knowledge-intensive business processes (e.g. recruiting, compliance, search for innovation etc.)
• IR applications often have a communication and integration role
• Consider extranets: communication with the customer (user can abort communication at any time!)
• Consider intranets: integrating multiple data sources at search time

• IR application = IR system + data collection + configuration/application layer
Application-centric evaluation

- Cranfield = evaluation of retrieval effectiveness of IR systems
- How to evaluate IR applications? What to evaluate?
Motivation

• Very few system developers

• Many more IR application deployers, implementors and operators

• The latter are actually interested in how well their IR applications perform (this is not just a system development question!)

• We believe IR evaluation can successfully be conducted from the "front" (application) or the "back" (system)
Why not Cranfield?

• IR applications support a *specific* knowledge-intensive business process.
• The performance on the *actual* underlying data collection is important

• There are different motivations for evaluation:
• Not just tuning, but also monitoring, continuous evaluation etc.

• There are different requirements for evaluation:
• Not comparative evaluation. Absolute performance is important!
Interesting problem

• How to measure?
• What to measure?
• Alternative methodologies
• There is some prior work, but there is a lot to do
• Additional, related problems (→ validation, certification, …)
What to measure?

- Operators of IR applications are typically not interested in retrieval effectiveness per se.
- It is more important that the application is "good enough" than if the retrieval effectiveness is (theoretically or practically) optimal.

- What to measure is still a research question.
- We attempt to measure some form of "user perception" of the application.
Evaluation Methodology

• We used a "guerrilla campaign" setup to develop/verify the methodology
• Part of the work in work package 4 of the Promise NoE

• Sites are treated as black box applications
• No ranking of sites
Evaluation Methodology (cont.)

- Measurement is based on a wide range of criteria (large number of tests)
- Scoring of individual tests is simple by design (simple counts etc.)
- Weighting the different tests, and aggregation, is still subject of research
- There is a "script" guiding the testers ("creativity" of testers needs to be minimized as much as possible!)
- Testers are also asked about their subjective experience
Evaluation Methodology
Instantiating the grid of tests

- The methodology can be adapted to different use cases by use-case specific pruning
- This is still work in progress
Example of a single test: Completeness

Assumption
Users expect to potentially find all documents that can be publicly accessed in any way on the site (namely, through browsing the site) when using the search functionality.

Irregularity
Publicly accessible documents (known through browsing or obtaining a direct link) cannot be found using the search functionality.

Root causes
The index is incomplete – documents/sets of documents are missing
The index is incomplete – the index is out of date (→ Freshness)
The index is incomplete – documents of certain types are missing (→ Format support)
Example of a single test: Completeness (cont.)

Test

The content scope of the test is decidedly narrow. No linked resources in the application are expected to be accessible through the search. E.g. if a retailer owns another shopping outlet, the latter's products need not be found.

- Locate 3 documents which match the following criteria:
  - Takes at least 5 clicks to locate document
  - Document is at least 3 levels from root as determined by URL (optional if URL-rewriting is used in the application)
  - URL is at least 100 characters long (optional as above)
- If no documents matching these criteria are found, abort
- For each document
  - Extract a characteristic phrase (2-4 words) from a central location within the document
  - Search for the document using that phrase
  - Score according to the number of documents which can be located within the top 10 search results (0, 1, 2, 3)
Example of a single test: Completeness (cont.)

Use Case Domain Adaptations

Search for Innovation: the notion of completeness is defined by a user's expectation of the presence of specific content rather than the inclusion of all browsable content. Therefore, there is no way to objectively and easily test this feature.
## Tested Sites for Guerilla Campaign

### Enterprise Search
- www.abb.ch/
- www.about.hm.com
- www.axa-winterthur.ch
- www.credit-suisse.com/
- www.e-coop.it/
- www.efginternational.com/
- www.generali.it/
- www.hm.com/se
- www.hug-ge.ch
- www.ikea.com/se/sv/
- www.insel.ch/
- www.nestle.ch/
- www.novartis.ch/
- www.poste.it/
- www.rai.it/
- www.sf.tv/
- www.swissre.com/
- www.telecomitalia.com/
- www.trentitalia.com/
- www.ubs.com
- www.unicreditgroup.eu/
- www.vattenfall.de/
- www.volkswagen.de

### Cultural Heritage
- www.beniculturali.it/
- www.bundesarchiv.de/
- www.culture.fr/collections/
- www.dhm.de/
- www.europeana.eu/
- www.europeana.eu/portal/
- www.gallica.bnf.fr/
- www.helveticarchives.ch/
- www.ige.ch/
- www.pinakothek.de/
- www.smb.museum/
- www.staedelmuseum.de/
- www.unesco.org/

### Patent
- www.boliven.com/patents/
- www.epo.org/
- www.freepatentsonline.com/
- www.google.com/patents/
- www.patft.uspto.gov/
- www.wipo.int/

### Other Domains
- www.aphp.fr/
- www.archives.elysee.fr/
- www.bag.admin.ch/aktuell/
- www.bium.univ-paris5.fr/histmed/medica.htm
- www.bsip.com/
- www.centerpartiet.se/
- www.elysee.fr/
- www.health.ch/
- www.hon.ch/
- www.hon.ch/HONmedia/index_f.html
- www.hu-berlin.de
- www.inpi.fr/
- www.kth.se/
- www.lara.inist.fr/
- www.medicalinfo.ch/
- www.moderat.se/
- www.policlinicogemelli.it/area/
- www.santeromande.ch/
- www.spotify.com/
- www.unilabs.ch/
Criteria

Index
  How are documents processed and stored for retrieval?

Query/Document Matching
  How are queries matched to documents?

User Interaction
  What are the presentation options and what is the usability of the search functionality?

Search Results
  What is the quality of the search result?
Criteria

1. INDEX

- **Completeness**
  - Is everything indexed?
  - 80% of the tested sites are **complete**
    - BP: Make sure that all documents are reachable and processable by the indexer

- **Freshness**
  - Are new pages indexed?
  - 50% of the tested sites are **not fresh**
    - BP: Update the index at least daily

- **Binary Document Handling**
  - Can binary documents (e.g. PDF, Office documents,...) be found?
  - No clear tendency ($\mu = 0.45$, $\sigma = 0.44$)
Criteria

1. INDEX

• Separation of Actual Content and Representation
  • Have structural elements (header, footer, etc.) been indexed?
  • No clear tendency (μ = 0.56, σ = 0.5)
    – BP: Detect and remove structural document parts (e.g. header, footer)

• Special Characters
  • Are diacritics (e.g. ö, ä, é, ...) processed?
  • No clear tendency (μ = 0.42, σ = 0.49)

• Synonyms or Domain Specific Terminology
  • Is there functionality to help users select technical terms?
  • Nearly no domain specific terminology (μ = 0.11)
### Chercher dans plus de 6'800 images et vidéos médicales

<table>
<thead>
<tr>
<th>Index</th>
<th>V</th>
<th>W</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>HONcode</td>
<td>HONselect</td>
<td>News (en)</td>
<td>Conférences</td>
<td>Images</td>
<td></td>
</tr>
<tr>
<td>Williams, syndrome (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wolfram, syndrome (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beckwith Wiedemann, syndrome (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Granulomatose Wegener (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syndrome de Waardenburg (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Von Willebrand, maladie (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fièvre à virus West Nile (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polygone artériel Willis (2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syndrome de Werner (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**All screenshots are © by their respective owners**
Criteria

1. INDEX

- **Duplicate Documents**
  - Are duplicates shown in the result list?
  - No clear tendency ($\mu = 0.61$, $\sigma = 0.49$)
    - BP: Detect and remove duplicates documents using checksum or full vector comparison
- **Meta-Data Quality**
  - Is searching in meta-data fields available?
  - No clear tendency ($\mu = 0.47$, $\sigma = 0.46$)
- **Tokenization**
  - Is domain specific tokenization (e.g. F/A-18, Hans-Peter O'Brian) implemented?
  - No clear tendency ($\mu = 0.61$, $\sigma = 0.43$)
    - BP: Treat domain specific terms with tokenization characters (-, /, .) separately
Criteria

1. INDEX

• Core Business Entities
  • Can clearly defined entities (persons, products) be found?
  • No clear tendency
    – BP: Index business entities as single features

• Stemming
  • Can different word forms (plurals, verbal forms) be found?
  • 63% of the tested sites **don’t support stemming**
    – BP: Use stemming to overcome mismatches between word forms
Résultats de recherche pour « Hôpitaux »

Les hôpitaux de l’AP-HP

Contacter un hôpital de l’AP-HP

Vous pouvez contacter directement un des hôpitaux ou un des services généraux de l’AP-HP, des services auxiliaires et des services de santé publique. Les hôpitaux (classés par ordre alphabétique) ALBERT CHENEVIER : direction.chercheur@ap-hop.fr ANTOINE BECERE : contact.becere@ap-hop.fr ARMAND TROUSSEAU : info.trousseau@ap-hop.fr AVICENNE : contact.avicenne@ap-hop.fr BEAUJON : info.beaujon@bjs.ap-hop.fr BICHAT-CLAUDE BERNARD : contact.bc...
Criteria

2. QUERY/DOCUMENT MATCHING

- Phrasal Queries
  - Are phrasal queries (“man bites dog”) supported?
  - 67% support Phrasal Queries

- Query Syntax
  - Are Boolean queries supported? (AND, OR, NOT)
  - No clear tendency

- Over- and Under-Specified Queries
  - Are there results if we over-specify queries?
  - Are there suggestions if we under-specify queries?
  - Sites with a good overall score can deal with over- and under-specified queries
Criteria

2. QUERY/DOCUMENT MATCHING

• Feedback
  • Can user give feedback about search results?
  • Most of the sites do not allow users to give feedback about the results ($\mu = 0.22$)

• Multimedia
  • Is it possible to retrieve multimedia (e.g. images, videos, music)?
  • Multimedia retrieval is rarely supported ($\mu = 0.26$)
    – BP: Use text-based multimedia retrieval (i.e. captions, filenames, subtitles)
Criteria

2. QUERY/DOCUMENT MATCHING

• Cross-Language Information Retrieval
  • Can a user retrieve documents in different languages? (language-neutral terms, main language)
  • No clear tendency
    – BP: Use combinations of translation resources, translate metadata, maximize vocabulary coverage of translation resource

2.6 Cross-Language Information Retrieval
Criteria

3. USER INTERACTION

• Performance/Responsiveness
  • How responsive are the sites?
  • 86% of the tested sites had a response time smaller than 5 sec

• User Guidance
  • Is there any guidance for the user? E.g. Term suggestion, spell checker,…
  • Most of the pages do not offer user guidance features (μ = 0.16)

• Browsing
  • Is there usable browsing functionality?
  • 82% of the tested sites offered browsing functionality
### Produkte & Leistungen für W

<table>
<thead>
<tr>
<th>ABB W - Cabinets (Sub Distribution Boards)</th>
<th>ABB Rozłączniki i produkty bezpiecznikowe w obudowach...</th>
<th>ABB FCU200-W - Flow Computer Units (Flow Measurement...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Voltage Products and Systems &gt; Enclosures &gt; Sub Distribution Boards &gt; Cabinets &gt; W ... availability... Condition: Product ... Global alias ... Price ... Availability: 0. Quantity: The W wall-mounting and floor-standing cabinet...</td>
<td>Produkty niskiego napędu &gt; Rozłączniki i produkty bezpiecznikowe w obudowach... Collection of plastic and metal enclosed switches. Rozłączniki w obudowach i produkty bezpiecznikowe. Skontaktuj się z ABB w Inne.</td>
<td>Measurement Products &gt; Flow Measurement &gt; Flow Computer Units &gt; FCU200-W... Checking price and availability... Condition: Product ... Global alias ... Price ... Availability: 0. Quantity: FCU 200-W (SensyCal-W) ...</td>
</tr>
</tbody>
</table>

### Produktkatalog für W

<table>
<thead>
<tr>
<th>ABB15FGR643W2LA140</th>
<th>2240 W-54 / 6600-0.0792</th>
<th>2240 W-53 / 6600-0.0791</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC, Dynics Industrial w/Enet</td>
<td>Product ID: 6600-0.0792</td>
<td>Product ID: 6600-0.0791</td>
</tr>
<tr>
<td></td>
<td>Product Main Type: Alarm- and control devices</td>
<td>Product Main Type: Alarm- and control devices</td>
</tr>
<tr>
<td></td>
<td>Twilight switch</td>
<td>Twilight switch</td>
</tr>
</tbody>
</table>

Weitere Ergebnisse in Produktpflanze ansehen

Weitere Ergebnisse im Download Center aufrufen

Weitere Ergebnisse in Produkte & Leistungen ansehen
3. USER INTERACTION

• Personalization
  • Is it possible to personalize the search? E.g. user profiles, user preference, search history,…
  • Only rarely there were personalization features available

• Social Aspects
  • Can search result be shared with other user? Is content suggested based on searches by other users?
  • In general no social aspects are available ($\mu = 0.17$)

• Exception Handling
  • What happens if invalid input is entered in the search field?
  • No clear tendency ($\mu = 0.43$, $\sigma = 0.44$)
All screenshots are © by their respective owners.
3. USER INTERACTION

• Result List Presentation
  • How are the results presented? Snippets, highlighting, arrangement
  • No clear tendency
    – BP: Offer document snippets in the result list

• Entertainment
  • Is the search functionality entertaining?
  • Most of the sites have **decent usability and aesthetics**

• Localization
  • Are different languages selectable? Is the localization consistent?
  • No clear tendency ($\mu = 0.48$, $\sigma = 0.44$)
Criteria

3. USER INTERACTION

- Facets
  - Is it possible to filter?
  - 56% provide filters, 44% don’t

- Result List Import/Export
  - Can the result list be imported or exported?
  - 84% of the tested sites *don’t offer import or export*

- Sorting of Result List
  - Can the result list be sorted?
  - 65% provide sorting, 35% don’t
  -  BP: Offer multiple ways to sort the result list; adapt to requirements of use case domain
3. USER INTERACTION

- **Justification of Results**
  - Can the result be justified based on snippets, highlighting and the number of results?
  - Some of the sites already provide it, but still capacity to improve ($\mu = 0.68$)
- **Monitoring**
  - Can the queries be monitored over time?
  - 85% of the sites **don’t provide monitoring** functionality
- **System Override**
  - Can user input modifying features be overridden or disabled?
  - 81% of the sites **don’t allow system overrides**
Criteria

3. USER INTERACTION  6/7

• Related Content
  • Does the application suggest other content that might be relevant?
  • 78% of the tested sites *don’t suggest other content*
    – BP: Link additional sources and resources that provide context information

• Context Information
  • Does the application present contextual information? E.g. #views, #citations
  • 78% of the tested sites *don’t provide contextual information*
<table>
<thead>
<tr>
<th>Produkt</th>
<th>Artikelnummer</th>
<th>Pris</th>
</tr>
</thead>
<tbody>
<tr>
<td>STEFAN Stol</td>
<td>002.110.68</td>
<td>149 kr</td>
</tr>
<tr>
<td>BIRGIT gardiner</td>
<td></td>
<td>249 kr</td>
</tr>
<tr>
<td>BJÖRKUDDEN matbord</td>
<td></td>
<td>799 kr</td>
</tr>
<tr>
<td>FOTO tändstake</td>
<td></td>
<td>199 kr</td>
</tr>
<tr>
<td>NORDEN skåp</td>
<td></td>
<td>1 795 kr</td>
</tr>
</tbody>
</table>

All screenshots are © by their respective owners.
3. USER INTERACTION

- Navigational Aids
  - Is it possible to navigate back and forth in the query history?
  - 82% of the tested sites *don’t offer a navigation possibility*

- Mobile Access
  - Is there a mobile version of the site and the search functionality?
  - 82% of the tested sites *don’t have a mobile version* of the search functionality
Criteria

4. SEARCH RESULTS

- Navigational Queries
  - Can navigational entry points be found by queries?
  - 50% success, 50% don’t

- Factual Queries
  - Can facts be found through the search functionality?
  - No clear tendency
Criteria

4. SEARCH RESULTS

• Known Item Retrieval
  • Can known items be re-found?
  • 77% of the tested sites succeeded in finding a known item

• Diversity
  • Are the retrieved results diverse?
  • 80% of the tested sites are felt to deliver diverse results

• Geo-Location
  • Is the user’s geo-location part of the retrieval functionality?
  • 87% of the tested sites don’t consider the geo-location
Criteria Results Summary

GOOD RESULTS

• Completeness
• Phrasal Queries
• Performance/Responsiveness
• Browsing
• Known Item Retrieval
• Diversity

POOR RESULTS

• Office Document Handling
• Separation of Actual Content and Representations
• Special Characters
• Duplicate Documents
• Meta Data Quality
• Tokenization
• Named Entities
• Query Syntax
• Over- and Under-Specified Queries
• Cross-Language IR
• Exception Handling
• Result List Presentation
• Entertainment
• Localization
• Facets
• Sorting of Result List
• Justification of Results
• Navigational Queries

• Freshness
• Synonyms
• Stemming
• Feedback
• Multimedia
• User Guidance
• Personalization
• Social Aspects
• Result List Import/Export
• Monitoring
• System Override
• Related Content
• Context Information
• Navigational Aids
• Mobile Access
• Geo-Location
Best Practice

• Even though best practices exist (e.g. Promise NoE work package 2), most of the sites achieve poor results for the following tests

  • Freshness
  • Stemming
  • Multimedia
  • Related Content

  • Separation of Actual Content and Representation
  • Tokenization
  • Cross-Language Information Retrieval
  • Result List Presentation
  • Sorting of Result List
Overall Result

The graph illustrates the overall result for various components, including Index, Query/Document Matching, User Interface, and Search Result. The values range from 0.0 to 1.0, with Search Result showing the highest performance.
Overall Result Interpretation

- High scatter (different use cases)
- User interface has lowest mean and smallest standard deviation
  - most of the sites have a poor performance in this category
- Search results is the only category where the maximal score is reached
Enterprise Search Result Interpretation

- Smaller scatter
- Higher mean in search result category
- Lower mean in document/query matching
Use Case Domains

- Due to too little data these plots are not representative (→ work in progress)
Ranking

- Correlation between rankings based on the four categories is low
- The four best category scores were achieved by four different sites (-> room for improvement!)

![Graph showing Top 4 Sites rankings](image)
User Experience vs. Overall Score

Correlation of User Experience and Overall Score: 0.53
Conclusion

• Index
  • Keep index fresh
  • Apply stemming

• Query/Document Matching
  • Allow the user to give feedback about the search result
  • Include multimedia objects in index

• User Interface
  • User guidance (term suggestions, spell checking, …)
  • Benefit from other users (social aspects, related content, context information)
  • Advanced search functionality (result list import/export, monitoring, system override)

• Search results
  • Consider geo-location in the search
Future work

• Comparing to Google baseline (fixed system!, shows variation from application)
• Already explored in earlier study
• "What would happen, if one used Google web search instead of a specific site search?"
• Evaluation of a search with Google web search and a site restriction (www.google.de site:...)
• Mediocre results
Future work

• Explore automation of tests (such as completeness, freshness, responsiveness...)
• Explore more deeply the actual use of IR applications inside organizations
References

- Cleverdon, CW The Cranfield tests on index language devices. 1967
- Promise NoE, deliverable 2.3 and 4.2, www.promise-noe.eu