Search
Exploge
Modify Engine

Arjen P. de Vries
arjen@acm.org

Centrum Wiskunde & Informatica
Delft University of Technology
Spinque B.V.
Library of the “Muntmuseum” in Utrecht (Erik van Hannen)
Search Intermediaries

- Travel agency
- Real estate agents
- Recruiters
- Librarians
- Archivists
- Digital forensics detectives
- Patent information specialists
Trend

- Do-It-Yourself (DIY) information seeking
  - Convenient access to online search engines
  - Perceived time efficiency
“We should recognise that **shallow text operations - select, match, show** - are right for information access. Information is primarily conveyed by natural language and this has to be shown to the user for them to assess.”

Trend

- Do-It-Yourself (DIY) information seeking
  - Convenient access to online search engines
  - Perceived time efficiency

- Let’s face it:
  - Google/Bing/Y! is often best
  - Even Google Enterprise Search ("the Google Box") is far worse than Google Web Search!
Kuhlthau six stages

- **Initiation**: user “becomes aware of a lack of knowledge or understanding”
- **Selection**: user needs to “identify and select the general topic to be investigated”
- **Exploration**: user needs to “investigate information on the general topic in order to extend personal understanding”
- **Formulation**: user forms “a focus from the information encountered”
- **Collection**: user needs “to gather information related to the focused topic”
- **Presentation**: user completes the search and presents findings
Exploration, Formulation

- I want to buy a house in Amsterdam and I want it with ‘sfeer’ but still in good shape
- I can afford about €350K. I need 3 bedrooms, the size should be about 80m². It should have a balcony or a backyard
- The closer to the station and an AH, the better. BUT... I do not want to live in Amsterdam-Noord, unless there is a quick bus connection to the ferry
- I may be willing to drop some of these constraints, but I’m not sure which
Formative Stages of the Information Seeking Process

Seeking Search Intermediary?!
Disclosure: I have been a librarian!
Trend

- Do-It-Yourself (DIY) information seeking
  - Convenient access to online search engines
  - Perceived time efficiency

- Let’s face it:
  - Google/Bing/Y! is often best
  - Even Google Enterprise Search ("the Google Box") is far worse than Google Web Search!

- Lack of tools for the search intermediary to do better than Google?!
Search = IR + DB

- Search tasks in the formative stages of ISP are likely to benefit from
  - a mix of exact (DB) and ranked (IR) searches
  - on structured (DB) and unstructured (IR) data
- Current technical solutions support either/or
- Combining results requires significant effort
  - copy & paste result sets between interfaces, “human (probabilistic) joins”
<table>
<thead>
<tr>
<th>Matching Inference</th>
<th>Data Retrieval (DR)</th>
<th>Information Retrieval (IR)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact match</td>
<td>Partial match, best match</td>
</tr>
<tr>
<td></td>
<td>Deduction</td>
<td>Induction</td>
</tr>
</tbody>
</table>

Van Rijsbergen, 1979
Search = IR on-top-of DB?

- IR on-top-of DB: let exact and ranked operations both be processed by the same engine, so they can be mixed freely
- IR responsible for ranking models, using DB as a data-access layer; no physical details necessary
- DB responsible for reliable, dynamically optimised, data access; no logical details necessary
IR on-top-of DB???

- Traditional, general-purpose DB technology cannot compete with custom IR search tools
  - Working assumption: using column stores should solve the efficiency problem
Cannot we ‘remove’ this IR engineer from the loop, like DBMS software removes the data engineer from the loop?
Search by Strategy

- Visually construct search strategies by connecting building blocks
Need building blocks which correspond to frequently used task (example a synonym builder which potentially community driven). Open Source ontology maybe available.
Search by Strategy

- Visually construct search strategies by connecting building blocks
- Each block describes either data or actions upon that data
Strategy Builder

Keywords
C Topic: Keywords (generated diesel)

alexandria
DATA

SOURCE
stemTERM
F STEMMER = snowball-english

SOURCE
QTERMS
rank_DOC_BM25
F STEMMING = snowball
F LANGUAGE = en

RESULT

Status Panel
1. Correctness
2. Compilation
3. Query Form
4. Results

Response: 10 - 1
patent-document - 302307
patent-document - 5514
patent-document - 108303
patent-document - 354294
patent-document - 46661
patent-document - 306888
patent-document - 87108
Search by Strategy

- Data sources are internally represented as quadruples, triples extended with an additional probability value
- Actions are scripts expressed in (a variant of) Fuhr and Roelleke’s PRA (TOIS 1997)
  - Boolean search: limit probabilities to 0 and 1!
- A search strategy may include multiple data sources
Implementation

- PRA translates into SQL (!)
- Current system setup using CWI’s MonetDB column-store
- Strategies are dynamically transformed into a REST API and a GWT UI
Exploratory Search

- Search & (Faceted) Browsing
  - Help discover schema, ontology, etc.
  - Help discover the relevant sources
    - Within-collection (by year/location, by type, ...)
    - Across multiple collections (by source)

- Tony Russel Rose is likely to tell us more later this afternoon!
Exploratory Search

- PRA enables soft (or “fuzzy”) faceted selections
  - Re-weight based on preferences, no more zero-result-set problem!
From Patent to Inventor
Limitations Search & Browse

- Faceted exploration does not include joins
  - Cannot construct new data sources from existing ones!
  - Only the pre-defined paths through the information space can actually be traversed
Who needs a Join?

- You!!!
  ... whenever ‘relevance cues’ are typed:
    - People (e.g., inventors)
    - Companies (e.g., assignees)
    - Categories (e.g., IPTC)
    - Time (e.g., expiry date)
    - Location (e.g., country)

... or whenever multiple sources are to be combined

- E.g., patents & news, patents & Wikipedia, ...
Patents on X by Y(y)
1. Which universities/colleges hold patents?

2. Who are the inventors named in those patents?

Real-life patent search example:

Which researchers associated to universities and colleges should our Human Resources manager know to hire the right people on time?
How Strategies Help

- Strategies improve communication between search intermediary and user
  - Encapsulate domain expert knowledge
  - Abstract representation of search expert knowledge
  - Analyze information seeking process at any stage
- Strategies facilitate knowledge management
  - Store / share / publish / refine
- Strategies mix exact (DB) and ranked (IR) searches
  - Avoid the need for “human (probabilistic) joins”
Conclusion

- “No idealized one-shot search engine”
- Hand over control to the user (or, most likely, the search intermediary)
  - Patent information specialists
  - Digital forensics detectives
  - Librarians / archivists
  - Real estate agents
  - Travel agency
Interactive Information Access

- **Feedback:**
  - Interaction improves information representation

- **Faceted Browsing:**
  - Interaction can let user take over where machine would fail

- **Search by Strategy:**
  - Interaction can let user take over where system designer would fail
Research Opportunities

- Assist the user make the best out of their increased level of control
  - Integrate usage data from live system to help improve or adapt strategies
- Handle “even larger” scale data
  - Patent demo fine on ~17GB semi-structured data (i.e., Fairview Research’s Green Energy collection), without specific optimizations, even with fairly large strategies
- Formalism
  - Score normalization
- Close the loop!
Current Situation

- index;
- repeat {
  - specify;
  - retrieve
} until 😊

Schema definition

Search & explore
Desirable Situation

- repeat {
  - index ;
  - specify ;
  - retrieve
} until 😊

Mixed Initiative
Schema definition
Search & explore
Acknowledgements

- Wouter Alink  
  Spinque BV
- Roberto Cornacchia  
  (co-founders)
- Martin Kersten & team  
  CWI/MonetDB BV
- Thomas Rölleke  
  QMUL/A Priori
- Henk Tomas  
  IP specialist
- Francisco Webber  
  IRF & Matrixware