Learning from the Users: from Search to Adaptive Search

Udo Kruschwitz

School of Computer Science and Electronic Engineering
University of Essex
udo@essex.ac.uk

18th March 2011
Overview

- Motivation and context
- Exploiting query logs
- Adaptive search
- *AutoEval*: evaluating adaptive search
- Next steps
Context

- Collection of documents, e.g.
  digital library, local Web site, intranet
- Not Web search in general
- Ad hoc queries
Problems

▶ Common problem with \textit{too many} matches
  ▶ General queries
  ▶ Ambiguous queries
  ▶ Short queries
▶ Data sparsity problem
▶ Typical intranet problem: recall can be important (e.g. single matching document)
▶ Express information need as a query
▶ Usable knowledge sources not available
Another Problem

(Source: http://xkcd.com/773)
Our Approach

- Search system that makes suggestions using automatically extracted domain knowledge
- But ...
  - Domain knowledge is noisy and incomplete
  - System suggestions not always useful/helpful
  - Document collection is changing
- Learn from the users’ interactions
- Improve system over time by adapting to the users’ search behaviour
- No single *user profile* but “*community profile*”
Partial Domain Knowledge (Web Site)
Partial Domain Knowledge (Digital Library)
Applying Domain Knowledge - General Idea

- Combine standard search system with initial domain model
- Utilize domain model to construct
  - query *refinements*
  - query *relaxations*
- Visual graph representation for navigation
- Present suggestions alongside matching documents
Log Data Example (Web site)

... 33136 1FEE0F65A1DA07ABE70F497C900D5E7E Wed Jan 02 08:36:02 GMT 2008 \ 0 0 0 posrgarduate application form \ posrgarduate application form posrgarduate application form 33137 1FEE0F65A1DA07ABE70F497C900D5E7E Wed Jan 02 08:36:58 GMT 2008 \ 1 0 1 application application application<\r>

...
Log Data Example (Digital Library)

...  
903779;guest;83.33.xxx.xxx;83et8b7j010eh4vlht3ucj8d1l;en;  
("pomegranate fertilization");search_sim;;0;--;2007-10-05 13:52:30
...  
1889115;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;  
("mozart");search_url;;0;--;2008-06-24 22:02:52
...  
1889118;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;  
("mozart");view_full;;1;--;2008-06-24 22:03:03
...  
1889120;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;  
Klavierkonzerte;search_res_rec_all;;0;--;2008-06-24 22:03:55  
1889121;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;  
("klavierkonzerte");view_full;;1;--;2008-06-24 22:04:10
...
Using Log Data to Acquire a Domain Model

- Queries submitted by users
- Identify sessions
- Associate related queries (many possible ways of doing so)
- Result is a query association graph (of some sort)
Using Log Data to Acquire a Domain Model - Example

... 903779;guest;83.33.xxx.xxx;83et8b7j010eh4vlht3ucj8d11;en;  
("pomegranate fertilization");search_sim;;0;-;;;2007-10-05 13:52:30
...
1889115;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;  
("mozart");search_url;;0;-;;;2008-06-24 22:02:52
...
1889118;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;  
("mozart");view_full;;1;;;;2008-06-24 22:03:03
...
1889120;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;  
Klavierkonzerte;search_res_rec_all;;0;-;;;2008-06-24 22:03:55
1889121;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;  
("klavierkonzerte");view_full;;1;;;;2008-06-24 22:04:10
...
Using Log Data to Acquire a Domain Model - Example

8eb3bdv3odg9jncd71u0s2aff6 xxxx 1889115 xxxx mozart xxxx 2008-06-24 22:02:52
8eb3bdv3odg9jncd71u0s2aff6 xxxx 1889120 xxxx klavierkonzerte xxxx 2008-06-24 22

...
Using Log Data to Acquire a Domain Model - Example
Our Log Data

- We use query logs collected on different collections, e.g.
  - *University of Essex* intranet search engine: almost 2 million queries (since Nov 2007)
  - *The European Library*: 1.8 million interactions (Jan 2007 - Jun 2008)
- Query log analysis (not discussed here)
- Bootstrap (adaptive) domain models
Towards Adaptive Search

- Start by employing initially extracted domain knowledge
- Observe user interaction with the system
- Incorporate clickthrough trails
- Use this *implicit relevance feedback* to adjust domain knowledge accordingly
- Do this fully automatically
- Aim: evolving domain knowledge that adjusts to the users’ search behaviour
- Should learn common patterns over time, e.g. “map” → “campus map”
- Should deal with seasonal terms appropriately, e.g. “registration”

This should improve search ...
... and Navigation
Automatic Domain Model Adaptation

Variety of adaptation models, including:

- Exploiting Maximum Likelihood Estimates (MLE)
- Formal Concept Analysis (FCA)
- Ant Colony Optimization analogy (ACO)
- Adaptive Intranet Navigation

... let’s quickly look at the first three approaches.
## MLE: Domain Model derived from Query Logs

<table>
<thead>
<tr>
<th>q1</th>
<th>q2</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>registration</td>
<td>online registration</td>
<td>0.045</td>
</tr>
<tr>
<td>registration</td>
<td>registration office</td>
<td>0.035</td>
</tr>
<tr>
<td>registration</td>
<td>timetable</td>
<td>0.025</td>
</tr>
<tr>
<td>registration</td>
<td>enrol</td>
<td>0.020</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>online registration</td>
<td>registration</td>
<td>0.211</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>registration office</td>
<td>careers centre</td>
<td>0.053</td>
</tr>
<tr>
<td>registration office</td>
<td>albert sloman library</td>
<td>0.053</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>enrol</td>
<td>course enrolment</td>
<td>0.050</td>
</tr>
</tbody>
</table>
MLE: Domain Model derived from Query Logs II
MLE: Reminder - Original Domain Knowledge

- registration
  - dates
  - card
  - essex
  - students
    - jobshop
  - undergraduate

- regulations
MLE: Results

- Studied 18-months log file of Essex search engine (about 670,000 queries)
- Sampled frequent and less frequent queries
- User study to assess quality of derived term suggestions
- Compared these terms to alternatives, e.g. Google suggestions, snippet approach, association rules approach
- Log analysis beats alternatives
- Session-based approach very powerful, dialogue-based even better
FCA Approach to Adaptation

- Lattice structure representing terms and corresponding documents
- Concept in lattice defined by objects (URLs) and attributes (terms)
FCA Approach to Adaptation II

- Learn from past user queries (implicit relevance judgements) using relative judgements (Radlinski & Joachims, 2005)
- Train a classifier (SVM) that associates terms with documents
- Rerun lattice construction
FCA: Architecture
FCA: Screenshot
FCA: Results

- User study similar to MLE evaluation
- Sampled frequent and less frequent queries
- User study to assess quality of derived term suggestions
- Compared these terms to alternatives, e.g. association rules approach, and unadapted FCA lattice
- FCA adaptation beats both alternatives
ACO: Adaptive Domain Model

- Simple approach using the idea of ant colony optimization:
  - Update domain model in daily batches
  - Add weight to query pairs observed that day, e.g. “libraary” → “library”
  - Normalise the weights so that all outgoing graphs in a node sum up to 1
  - And so on ...

- Idea: learn associations as they become popular, allow for forgetting relations as well!
ACO: Results

- Again: user studies to assess quality of derived term suggestions
- Two studies: Essex university logs (Essex), European Library logs (TEL)
- Compared these terms to alternatives, e.g. Google suggestions, association rule approach, snippet processing
- Essex: ACO beats all alternatives and suggestions improve over time
- TEL: ACO better than association rule approach but not snippet baseline
- Suggestions derived using different methods can be complementary (TEL)

... see (Kruschwitz et al., 2011) and (Dignum et al., 2010)
**AutoEval: Evaluate Adaptive Search**

- Limitations of user studies
- Evaluate suggestions *without* recruiting subjects
- Compare different models automatically
- Idea: use log files and exploit past user interactions

![Diagram showing the process of evaluating adaptive search with log data.](image)

\[
score = \frac{1/2 + 1/4 + 1/1}{3} = 0.583
\]

... come along to our ECIR’11 talk
AutoEval Results (Web Site)
AutoEval Results (Digital Library)

![Graph showing MRR scores over months Jan 2007 - Jun 2008.
ACO, FONSECA(MinSup=2), FONSECA(MinSup=3).]
Next Steps

- Prototypes to go live at the University of Essex
- Work with industrial partners
- TREC 2011 and CLEF 2011
- Ongoing EPSRC project (Essex, Robert Gordon University Aberdeen & Open University): AutoAdapt (November 2008 - November 2011)

... any collaboration welcome!
Conclusions

- Adaptive search by exploiting query logs
- Adaptive domain models can be learned, experiments with different approaches demonstrate this
- Have to deal with noisy data
- Data sparsity
- Navigation support as a suitable alternative to query suggestions (Saad & Kruschwitz, 2011)
Acknowledgements

- Dyaa Albakour, Nikolaos Nanas, Jinzhong Niu, Dawei Song, Anne De Roeck, Maria Fasli, Stephen Dignum (EPSRC-funded AutoAdapt research project: http://autoadaptproject.org/)
- Deirdre Lungley, Sharhida Saad (Language and Computation Group at Essex)
- Johannes Leveling (DCU)
References


References II


ECIR 2011 Industry Day in Dublin

- Co-located with last day of ECIR 2011
- Dedicated to the latest innovations in web and enterprise search
- Mix of research and application
- Excellent list of speakers
- Highly interactive and collegial, with attendance limited to 60-80 delegates
- Date: 21st April
- http://ecir2011.dcu.ie/program/industry-day/
That’s it ...