Log Analysis at Essex

Udo Kruschwitz

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

- Query log analysis for adaptive intranet search (academic research)
- Query log analysis for learning to match job seekers against best-matching jobs (industry collaboration)

... will primarily focus on the first one.
Overview

- Motivation & context
- Prototype on the Essex intranet
- Preliminary log analysis
- Current research
Context

- Collection of documents, e.g. corporate or academic intranet
- Not Web search in general
- Ad hoc queries
Problems

- Common problem with *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
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
What Sort of Domain Knowledge?

- Online clustering (e.g. Vivisimo)
- *Subsumption hierarchies* (Sanderson & Croft 1999)
- *Lexical modification* approach (Anick & Tipirneni 1999)
- Formal Concept Analysis (e.g. CREDO)
- Term associations using neural networks and fuzzy logic (e.g. Aquabrowser)
- Flat list of terms using simple *tf.idf* applied to top matching documents
- ...

Partial Domain Knowledge (Example)
Applying Domain Knowledge - General Idea

- Combine standard search system with initial domain model
- Utilize domain model to construct
  - query refinements
  - query relaxations
- Present suggestions alongside matching documents
Log Data Collection

- Essex intranet search engine
- Originally running alongside standard Essex search engine
- Operating since summer 2006
- About 40,000 queries collected in 12 months
- November 2007: system replaced old search engine altogether
  ... about 700,000 queries collected since then
Towards Adaptive Intranet 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
- Aim: evolving domain knowledge that adjusts to the users’ search behaviour

... let’s see what the log files tell us so far.
Observations I

- More than 10% of queries are modifications!

... suggests the general system setup makes sense
Most Frequent User Queries (since Nov 2007)

14659 library
14152 search
9291 moodle
6799 timetable
3879 cmr
3757 accommodation
3746 graduation
3376 enrol
3262 accommodation
3012 exam timetable
2826 term dates
2785 fees
2544 courses
2482 psychology
Observations II

- Queries are domain-specific
- This is different from general Web search

... suggests that domain-independent knowledge (e.g. WordNet, Google n-grams) might not be suitable.
Frequent Queries
Query Traffic I
Query Traffic II
Observations III

- All sorts of variations (seasonal, usage patterns, ...)
- Again different from general Web search

... suggests that system should perhaps adapt to “context”.
## Query Statistics

<table>
<thead>
<tr>
<th></th>
<th>Set 1</th>
<th>Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Queries</td>
<td>100</td>
<td>40,006</td>
</tr>
<tr>
<td>Average Query Length</td>
<td>1.54</td>
<td>1.98</td>
</tr>
<tr>
<td>Length of Longest Query</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>Queries with Spelling Errors</td>
<td>2%</td>
<td>≈6%</td>
</tr>
<tr>
<td>Fraction of Query Corpus</td>
<td>20%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Observations IV

- Queries are even shorter than on the Web!
User-System Interaction

- More system suggestions than manual modifications
- More additions of terms than replacements
- Long tail of modifications only submitted once
## Sample Interactions: Most Frequent Query Pairs

<table>
<thead>
<tr>
<th>Frequency</th>
<th>q1</th>
<th>q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>122</td>
<td>fees</td>
<td>tuition fees</td>
</tr>
<tr>
<td>92</td>
<td>accommodation</td>
<td>accommodation</td>
</tr>
<tr>
<td>77</td>
<td>student union</td>
<td>students union</td>
</tr>
<tr>
<td>70</td>
<td>accommodation</td>
<td>accommodation office</td>
</tr>
<tr>
<td>69</td>
<td>post room</td>
<td>postroom</td>
</tr>
<tr>
<td>68</td>
<td>my essex</td>
<td>myessex</td>
</tr>
<tr>
<td>68</td>
<td>mondo</td>
<td>mondo pizza</td>
</tr>
<tr>
<td>62</td>
<td>map</td>
<td>campus map</td>
</tr>
<tr>
<td>61</td>
<td>time table</td>
<td>timetable</td>
</tr>
<tr>
<td>55</td>
<td>foundation</td>
<td>foundation degree</td>
</tr>
<tr>
<td>54</td>
<td>su</td>
<td>student union</td>
</tr>
<tr>
<td>52</td>
<td>library</td>
<td>library</td>
</tr>
</tbody>
</table>
### Sample Interactions: Query Pairs with Highest MLE

<table>
<thead>
<tr>
<th>q1</th>
<th>q2</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>moolde</td>
<td>moodle</td>
<td>1.000</td>
</tr>
<tr>
<td>library</td>
<td>library</td>
<td>1.000</td>
</tr>
<tr>
<td>email on web</td>
<td>email on web</td>
<td>1.000</td>
</tr>
<tr>
<td>psychology</td>
<td>psychology</td>
<td>1.000</td>
</tr>
<tr>
<td>pschology</td>
<td>psychology</td>
<td>1.000</td>
</tr>
<tr>
<td>timetab e</td>
<td>timetable</td>
<td>1.000</td>
</tr>
<tr>
<td>graduatin</td>
<td>graduation</td>
<td>1.000</td>
</tr>
<tr>
<td>sociolgy</td>
<td>sociology</td>
<td>1.000</td>
</tr>
<tr>
<td>registry</td>
<td>registry</td>
<td>1.000</td>
</tr>
<tr>
<td>registry</td>
<td>registry</td>
<td>1.000</td>
</tr>
<tr>
<td>prospective</td>
<td>prospectus</td>
<td>1.000</td>
</tr>
<tr>
<td>myesex</td>
<td>myessex</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Sample Interactions: Query *Refinements* with Highest MLE

<table>
<thead>
<tr>
<th>q1</th>
<th>q2</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>lost registration card</td>
<td>registration card</td>
<td>1.000</td>
</tr>
<tr>
<td>film society</td>
<td>art film society</td>
<td>1.000</td>
</tr>
<tr>
<td>exam timetable 09</td>
<td>exam dates</td>
<td>1.000</td>
</tr>
<tr>
<td>erol</td>
<td>enrol</td>
<td>1.000</td>
</tr>
<tr>
<td>web email</td>
<td>email</td>
<td>1.000</td>
</tr>
<tr>
<td>w800</td>
<td>creative writing</td>
<td>1.000</td>
</tr>
<tr>
<td>stavrakakis</td>
<td>yannis stavrakakis</td>
<td>1.000</td>
</tr>
<tr>
<td>mechanical</td>
<td>mechanical engineering</td>
<td>1.000</td>
</tr>
<tr>
<td>lakeside</td>
<td>lakeside theatre</td>
<td>1.000</td>
</tr>
<tr>
<td>endsleigh</td>
<td>endsleigh insurance</td>
<td>1.000</td>
</tr>
<tr>
<td>study in europe</td>
<td>study abroad</td>
<td>1.000</td>
</tr>
<tr>
<td>switch board</td>
<td>telephone operator</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Observations V

- Users select query modification options on a continuing basis
- Lots of implicit (domain-specific) relationships
- Different “types” of relationships, e.g.
  - Dialogue-based vs. session-based interactions
  - Adding vs. replacing query terms
- But: data sparsity issues
Next Steps

- Automatic Adaptation of Domain Model
- Focus of a new EPSRC project (Essex, Robert Gordon University Aberdeen & Open University): AutoAdapt (November 2008 - November 2011)
- Will look at a variety of adaptation models
- So far we have already experimented with one approach using Formal Concept Analysis (FCA)
AutoAdapt: FCA Approach to Adaptation

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

- 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
- Promising evidence that lattice improves over time (Lungley & Kruschwitz, 2009)
**Alternative: 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>
Alternative: Domain Model derived from Query Logs
Reminder: Original Domain Knowledge

- registration
  - dates
    - office
    - undergraduate
  - card
  - essex
  - students
  - regulations
  - jobshop
Conclusions

- Query logs for adaptive intranet search
- Utilize automatically acquired domain knowledge
- Prototype suggests usefulness of general setup
- Interesting differences (similarities) with general Web search
- Next step: evolve domain model based on users’ search behaviour (domain-specific!)
- Promising direction: clickthrough data, lattice structures, query logs
Limitations

- Data sparsity
- Not easy to apply extracted knowledge to a different domain
- Privacy/ethical constraints
- Difficult to evaluate results
CareerPath Project

▶ Knowledge Transfer Partnership (KTP) with JobServe Ltd. (to start in July 2009)

▶ Data input:
  ▶ CVs
  ▶ Job search queries
  ▶ Clickthrough logs

▶ Expected output:
  ▶ Predict best matching job openings
  ▶ Career path patterns
Acknowledgements

- Deirdre Lungley (FCA)
- Dawei Song, Anne De Roeck, Maria Fasli, Stephen Dignum, Yunhyong Kim (AutoAdapt)
References