# Query Log Analysis

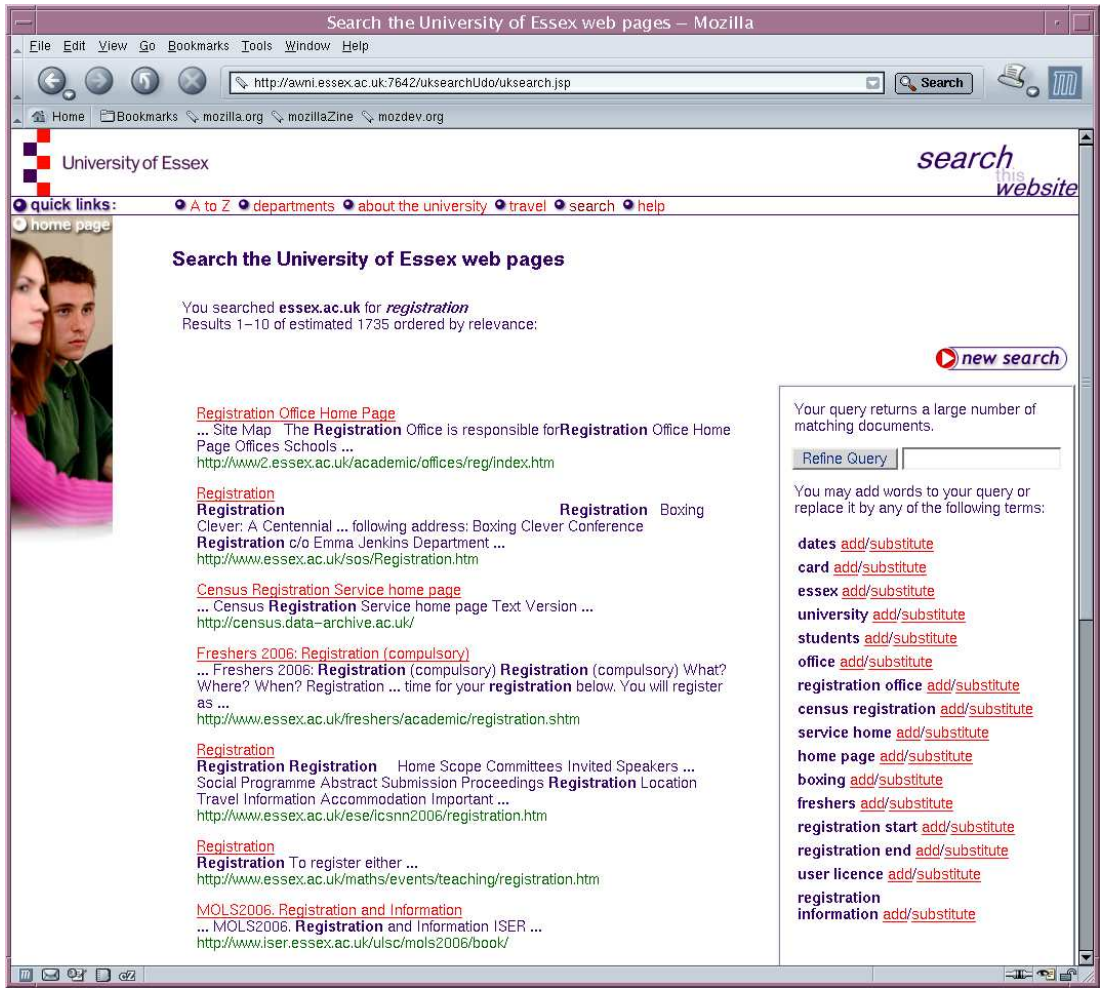
Udo Kruschwitz & Richard Sutcliffe

Department of Computing and Electronic Systems
University of Essex
udo@essex.ac.uk

5th October 2007

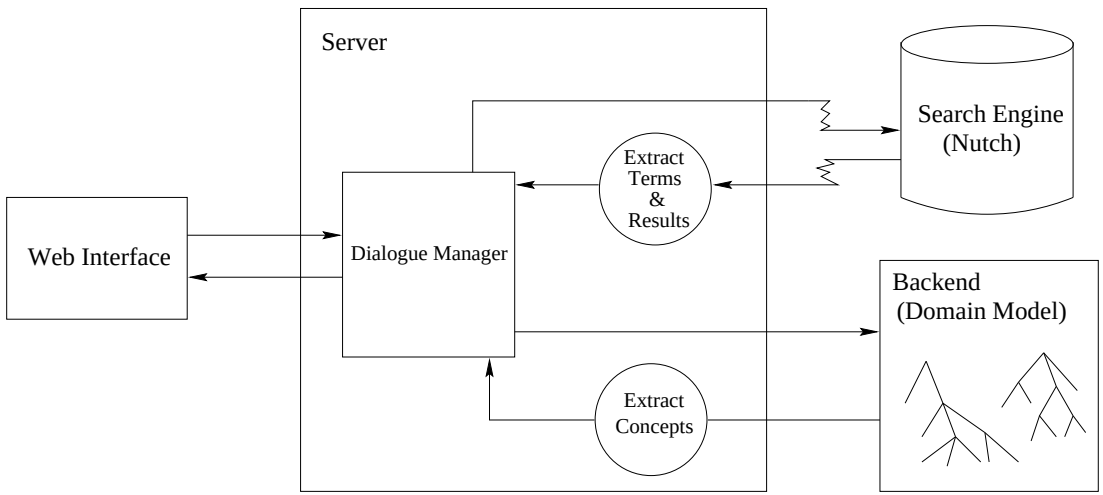
# Overview

- Brief motivation
- Prototype on the Essex intranet
- Preliminary log analysis
- Next steps
- Conclusions

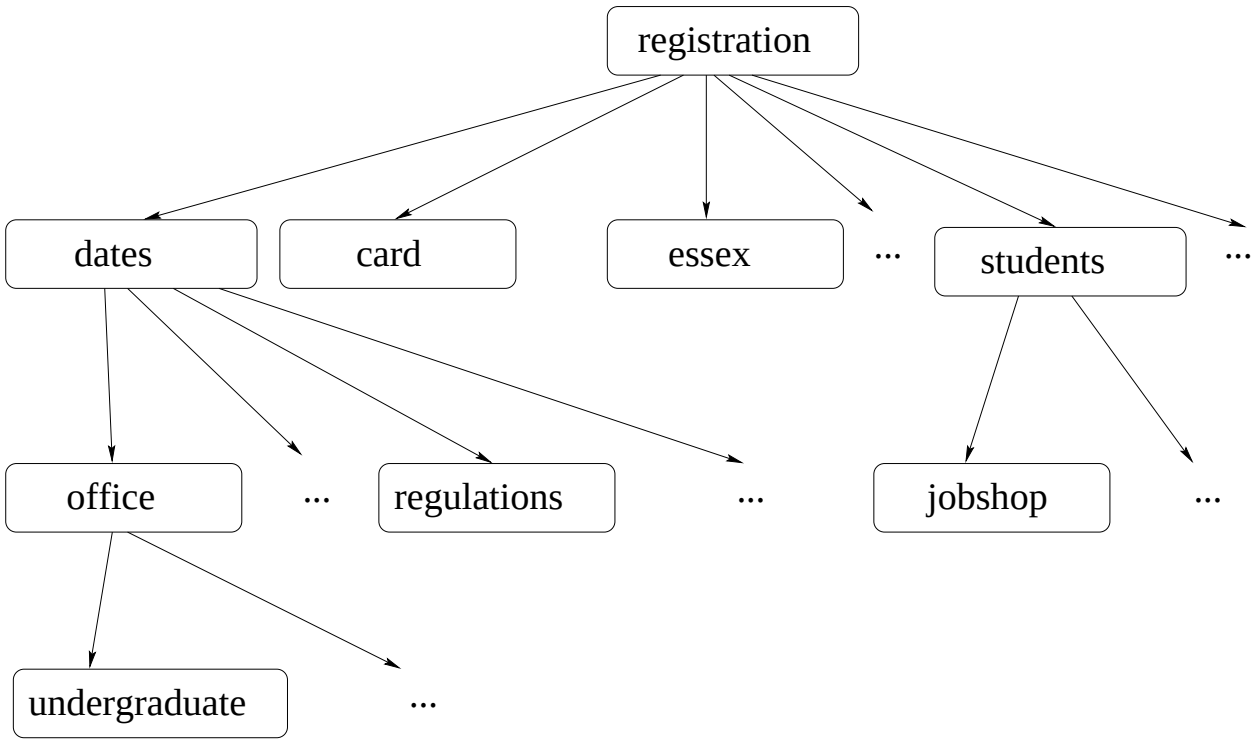

Search the University of Essex web pages – Mozilla
File Edit View Go Bookmarks Tools Window Help
http://awni.essex.ac.uk:7642/uksearchUdo/uksearch.jsp
Search
Home Bookmarks mozilla.org mozillaZine mozdev.org
University of Essex
search this website
quick links: A to Z departments about the university travel search help
home page
Search the University of Essex web pages
You searched essex.ac.uk for registration
Results 1–10 of estimated 1735 ordered by relevance:
new search
Registration Office Home Page
... Site Map The Registration Office is responsible forRegistration Office Home Page Offices Schools ...
http://www2.essex.ac.uk/academic/offices/reg/index.htm
Registration
Registration Registration Boxing Clever: A Centennial ... following address: Boxing Clever Conference Registration c/o Emma Jenkins Department ...
http://www.essex.ac.uk/sos/Registration.htm
Census Registration Service home page
... Census Registration Service home page Text Version ...
http://census.data-archive.ac.uk/
Freshers 2006: Registration (compulsory)
... Freshers 2006: Registration (compulsory) Registration (compulsory) What? Where? When? Registration ... time for your registration below. You will register as ...
http://www.essex.ac.uk/freshers/academic/registration.shtm
Registration
Registration Registration Home Scope Committees Invited Speakers ... Social Programme Abstract Submission Proceedings Registration Location Travel Information Accommodation Important ...
http://www.essex.ac.uk/ese/icsnn2006/registration.htm
Registration
Registration To register either ...
http://www.essex.ac.uk/maths/events/teaching/registration.htm
MOLS2006: Registration and Information
... MOLS2006. Registration and Information ISER ...
http://www.iser.essex.ac.uk/ulsc/mols2006/book/
Your query returns a large number of matching documents.
Refine Query
You may add words to your query or replace it by any of the following terms:
dates add/substitute
card add/substitute
essex add/substitute
university add/substitute
students add/substitute
office add/substitute
registration office add/substitute
census registration add/substitute
service home add/substitute
home page add/substitute
boxing add/substitute
freshers add/substitute
registration start add/substitute
registration end add/substitute
user licence add/substitute
registration information add/substitute

# Overall Architecture

- Dialogue-based search system that makes suggestions using:
  - Bootstrapped domain model
  - Combined with terms extracted on the fly



# Partial Domain Knowledge (Example)



# Applying Domain Knowledge - General Idea

- Combine standard search system with domain model
- Utilize domain model to construct
  - query *refinements*
  - query *relaxations*
- Extract additional query modification suggestions from matching snippets
- Present suggestions alongside matching documents

# First Evaluations

- Several controlled evaluation steps:
  - BBC News domain
  - University of Essex domain
- Task-based evaluations suggest usefulness of system (despite lack of measurable improvements)

... but what about the real world?

# Now what?

- Need to expose system to *real* users
- Find out what happens
- Long-term goals:
  - Learn from the user interactions
  - Improve system over time by adapting to the users' search behaviour
- First step: run system and start collecting log files

# Essex System Setup

- Essex intranet search engine
- Running alongside standard Essex search engine
- Operating since summer 2006
- More than 50,000 queries collected

# Observations I

- We expect that most queries are answered by top-matching documents
- But: $>$10% of queries are modifications
- Huge number of typos ($\approx$6% of queries)

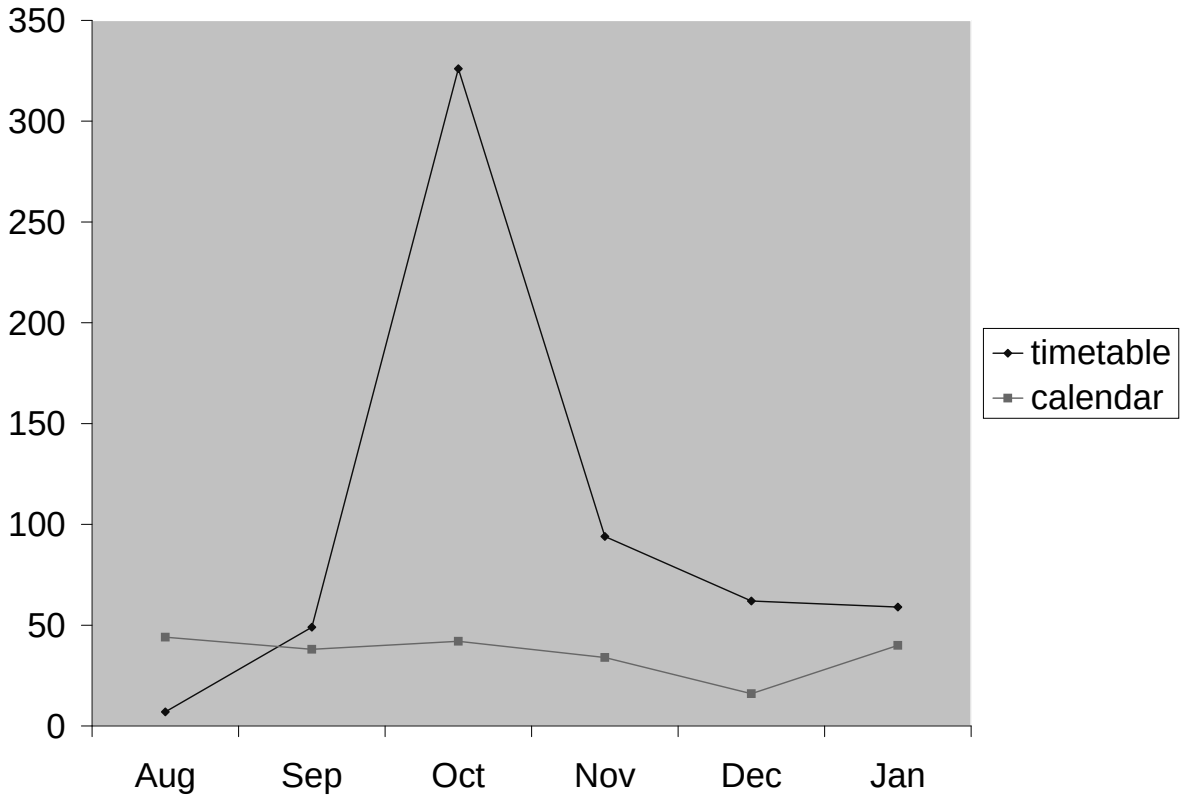
# Most Frequent User Queries

```
305 timetable
230 enrol
168 cmr
132 term dates
122 timetables
100 jobshop
 94 calendar
 83 map
 73 printing credit
 70 change password
 68 spss
 65 printing
 65 ocs
 63 car parking
```

# Observations II

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

# Frequent Queries



# Observations III

- Seasonal variations
- Again different from general Web search

# Query Statistics

| | **Set 1** | **Set 2** |
|---|---|---|
| Number of Queries | 100 | 20,578 |
| Average Query Length (Types) | 1.55 | 1.97 |
| Length of Longest Query | 4 | 17 |
| Queries with Spelling Errors | 3% | ≈6% |
| Fraction of Query Corpus | 21% | 100% |

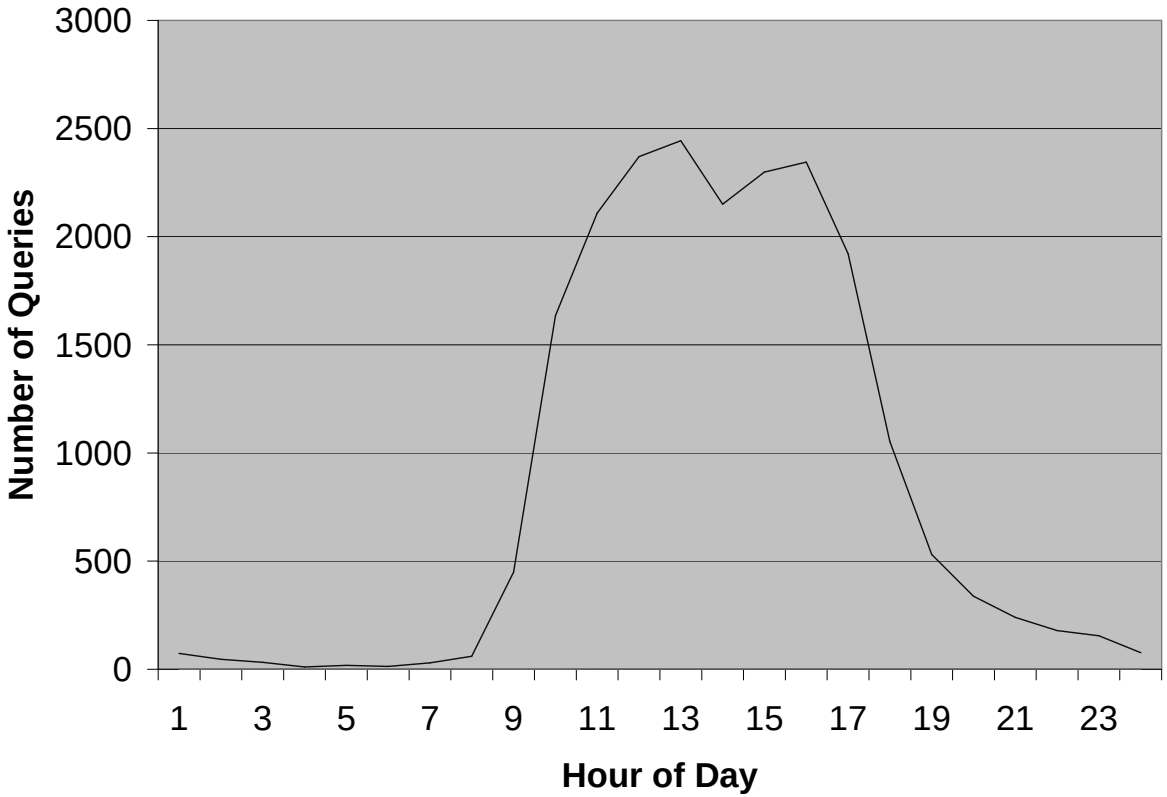
# Observations IV

- Queries are even shorter than on the Web!

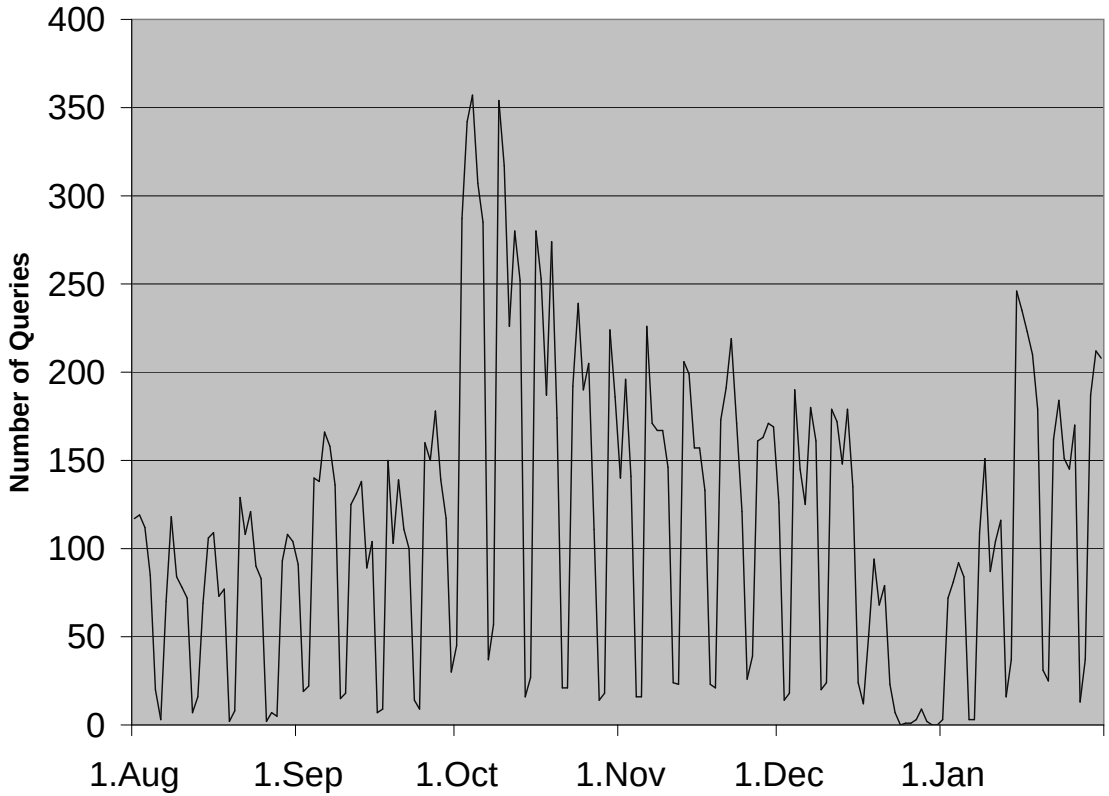
# Spelling Variations (Example)

| Frequency | Query |
|---:|---|
| 48 | plagiarism |
| 13 | plagarism |
| 4 | plagerism |
| 2 | plaigarism |
| 2 | plagirisim |
| 1 | plaigrism |
| 1 | plaigirism |
| 1 | plaigiarism |
| 1 | plagarisim |

# Query Traffic I



# Query Traffic II



# Query Modifications

- About 11% of submitted queries are modifications
- More system suggestions than manual modifications (different to Web study (Anick 2003))
- More additions of terms than replacements
- Long tail of modifications only submitted once

# Selection of Query *Refinement* Steps

| **Query** | **Refinement** | **Frequency** |
|---|---|---|
| parking | car parking | 9 |
| printing | credit | 4 |
| printer | accounting | 4 |
| ... | ... | ... |
| time table | timetable | 3 |
| subscription list | smallads | 2 |
| dog day care | pet care | 2 |

# Query Modifications

| Frequency | Query |
|---|---|
| 22 | car parking |
| 21 | university of essex |
| 15 | timetable |
| 12 | staff |
| 11 | parking of vehicles |
| 10 | small ads |
| 10 | online submission |
| 9 | printing |
| 9 | calendar |
| 8 | subscription list |
| 8 | room |
| 8 | coursework submission |
| 8 | campus shop |

# Manual Query Analysis

- Sample of 1,794 queries analysed in detail
- Classified into topic categories (Academic units, computer use, ...)
- Further classifications (capitalised queries, acronyms, typos etc.)

# Observations V

- Again: very domain-specific
- Lots of named entity queries (23%)
- Lots of capitalised queries (19%)

# User Feedback

- Entry page with link for feedback
- Hardly any feedback submitted

# Problems with Existing System

- Domain knowledge is noisy and incomplete
- System suggestions not always useful/helpful
- Document collection is changing
- Very little work on updating/adjusting/adapting extracted domain knowledge

# Next Steps

- Start by employing initially extracted domain knowledge
- Observe user interaction with the system
- Use this *implicit relevance feedback* to adjust domain knowledge accordingly (Deirdre)
- Aim: evolving domain knowledge that adjusts to the users' search behaviour

# Examples of User Interaction

- System suggestions, e.g.

  *printing* → *credit*
  *gallery* → *university gallery*
  *sports center* → *sports centre*

- User replacements/additions (perhaps most useful), e.g.

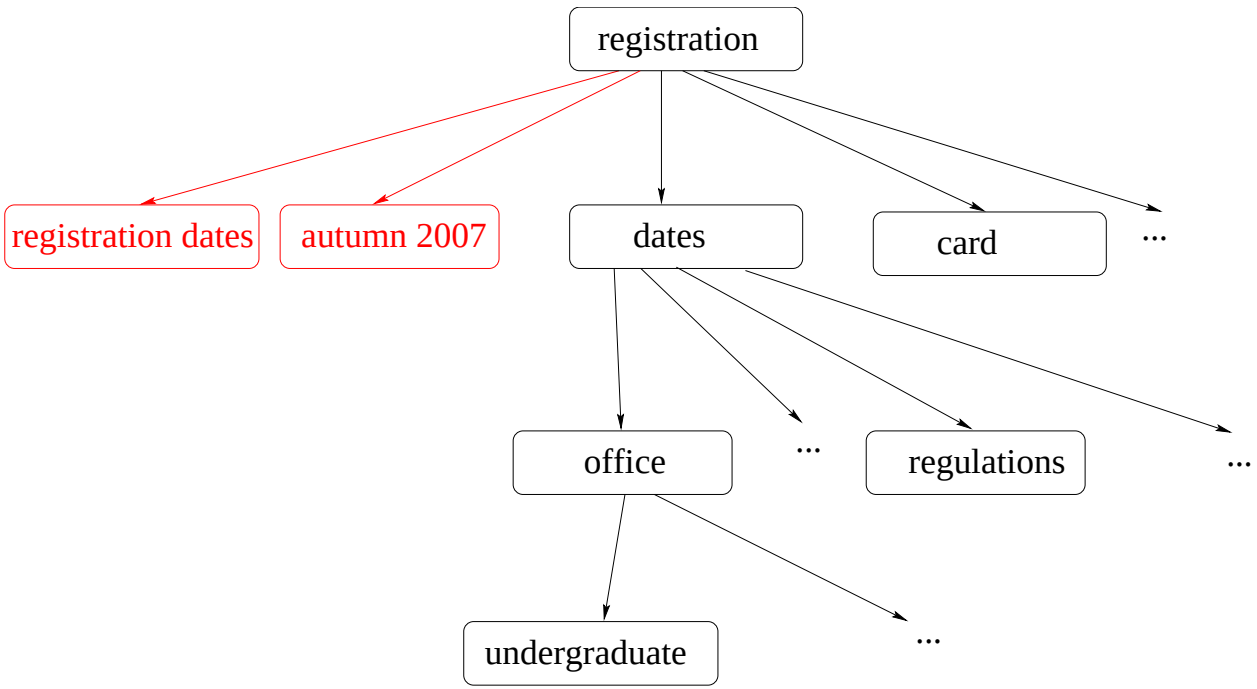
  *ojnline payment* → *online payment*
  *mexsoc* → *Mexican Society*
  *registration* → *autumn 2007*
  *registration* → *registration date*

# Questions & Problems

- How *exactly* do we update existing domain knowledge?
  - Addition/replacement of system suggestions
  - Addition/replacement of terms provided by the user
  - New queries within the same session, e.g.

    $$\textit{switch board} \longrightarrow \textit{telephone operator}$$
- Data sparsity
- Domain-specific data

# Ultimate Goal: Evolved Domain Knowledge



# Conclusions

- Simple dialogue-driven search for intranets / local Web sites
- Prototype suggests usefulness of general setup
- Interesting differences (similarities) with general Web search
- Next step: evolve domain model based on users' search behaviour
- To do that we need to keep collecting real data

# References


- P. Anick. Using Terminological Feedback for Web Search Refinement - A Log-based Study In *Proceedings of the 26$^{th}$ Annual International ACM SIGIR Conference*, pages 88-95, Toronto, 2003.
- Kruschwitz, U., R.F.E. Sutcliffe, and N. Webb (2008) "Query Log Analysis for Adaptive Dialogue-Driven Search". In J. Jansen, I. Taksa and A. Spink (eds.): *Handbook of Web Log Analysis*, IGI Global. Accepted for publication.