

# 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

University of Essex :: Search results – Mozilla

File Edit View Go Bookmarks Tools Window Help

http://search.essex.ac.uk/uksearch.jsp Search

Home Bookmarks mozilla.org mozillaZine mozdev.org

# University of Essex

home  
a to z  
contact  
help

prospective students new and current students staff alumni visiting

you are here: home > search

about the university  
virtual tours  
maps  
job vacancies  
departments  
research and expertise  
business  
key dates

**Targeted results.**  
Were you looking for any of these: [International Academy](#); [Modern Languages](#); [Department of Language and Linguistics](#)?

**Search results**

- [Results for web pages and other online documents](#)
- [Results from the phonebook](#)

**Results for web pages and other online documents**

You searched [essex.ac.uk](#) for *language*  
Results 1–10 of estimated 5150 ordered by relevance:

[Department of Language and Linguistics at the University of Essex, UK](#)  
... Department of **Language** and Linguistics at ...  
<http://www.essex.ac.uk/linguistics/>

[University of Essex - International students - English language requirements](#)  
... of Essex :: International students :: English **language** requirements skip to content ...  
<http://www.essex.ac.uk/international/language.aspx>

[mySkills: Academic Skills at Essex – Skills – Language](#)  
... Academic Skills at Essex – Skills – **Language** Skip to: site ...  
<http://www.essex.ac.uk/myskills/skills/language/default.asp>

[Department of Language and Linguistics at the University of Essex, UK](#)  
... Who to contact Modern **Language** Courses BA **Language** Studies BA Modern Languages  
BA ... Department of **Language** ...

**find out more...**

Your query returns a large number of matching documents.

You may add words to your query or replace it by any of the following terms:

[courses](#) [add/substitute](#)

[centre](#) [add/substitute](#)

[studies](#) [add/substitute](#)

[skills](#) [add/substitute](#)

[university of essex](#) [add/substitute](#)

[department of language](#) [add/substitute](#)

[computation day](#) [add/substitute](#)

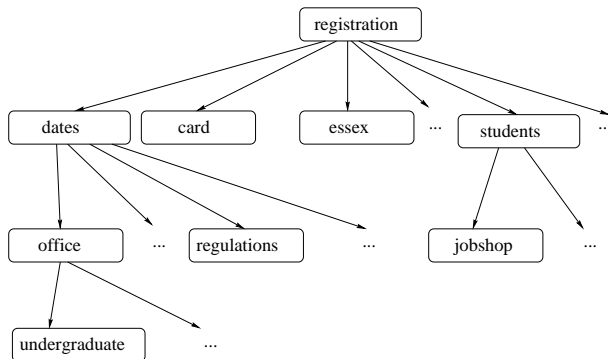
[information](#) [add/substitute](#)

[linguistics](#) [add/substitute](#)

## 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)

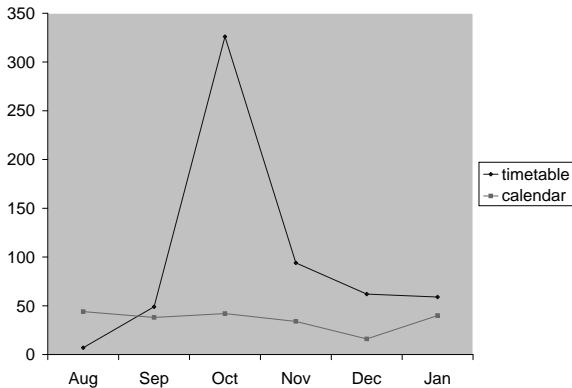
|       |                |
|-------|----------------|
| 14659 | library        |
| 14152 | search         |
| 9291  | moodle         |
| 6799  | timetable      |
| 3879  | cmr            |
| 3757  | accomodation   |
| 3746  | graduation     |
| 3376  | enrol          |
| 3262  | accommodation  |
| 3012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

## Observations II

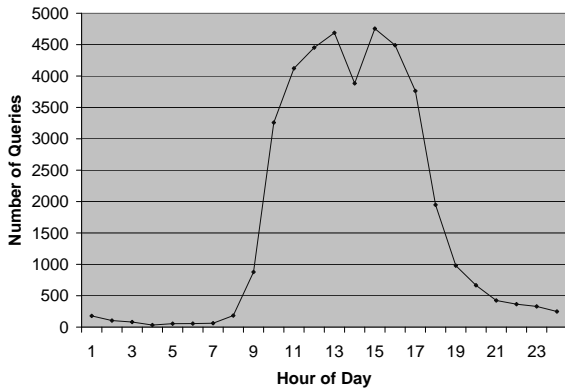
- ▶ Queries are domain-specific
- ▶ This is different from general Web search

... suggests that domain-independent knowledge (e.g. WordNet, Google n-grams) might not be suitable.

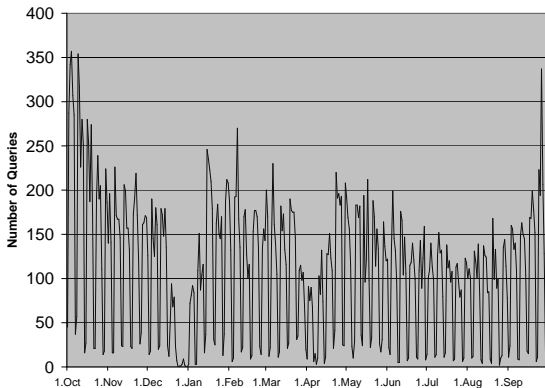
# Frequent Queries



# 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

|                              | <b>Set 1</b> | <b>Set 2</b> |
|------------------------------|--------------|--------------|
| Number of Queries            | 100          | 40,006       |
| Average Query Length         | 1.54         | 1.98         |
| Length of Longest Query      | 3            | 17           |
| Queries with Spelling Errors | 2%           | ≈6%          |
| Fraction of Query Corpus     | 20%          | 100%         |

## 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

| <b>Frequency</b> | <b>q1</b>     | <b>q2</b>           |
|------------------|---------------|---------------------|
| 122              | fees          | tuition fees        |
| 92               | accomodation  | accommodation       |
| 77               | student union | students union      |
| 70               | accomodation  | accomodation office |
| 69               | post room     | postroom            |
| 68               | my essex      | myessex             |
| 68               | mondo         | mondo pizza         |
| 62               | map           | campus map          |
| 61               | time table    | timetable           |
| 55               | foundation    | foundation degree   |
| 54               | su            | student union       |
| 52               | libary        | library             |

## Sample Interactions: Query Pairs with Highest MLE

| <b>q1</b>   | <b>q2</b>    | <b>MLE</b> |
|-------------|--------------|------------|
| moolde      | moodle       | 1.000      |
| lbrary      | library      | 1.000      |
| email on we | email on web | 1.000      |
| psycholgy   | psychology   | 1.000      |
| pschology   | psychology   | 1.000      |
| timetab e   | timetable    | 1.000      |
| graduatin   | graduation   | 1.000      |
| sociolgy    | sociology    | 1.000      |
| registry    | registry     | 1.000      |
| regisrty    | registry     | 1.000      |
| prospective | prospectus   | 1.000      |
| myesex      | myessex      | 1.000      |

## Sample Interactions: Query *Refinements* with Highest MLE

| <b>q1</b>              | <b>q2</b>              | <b>MLE</b> |
|------------------------|------------------------|------------|
| lost registration card | registration card      | 1.000      |
| film society           | art film society       | 1.000      |
| exam timetable 09      | exam dates             | 1.000      |
| erol                   | enrol                  | 1.000      |
| web email              | email                  | 1.000      |
| w800                   | creative writing       | 1.000      |
| stavrakakis            | yannis stavrakakis     | 1.000      |
| mechanical             | mechanical engineering | 1.000      |
| lakeside               | lakeside theatre       | 1.000      |
| endsleigh              | endsleigh insurance    | 1.000      |
| study in europe        | study abroad           | 1.000      |
| switch board           | telephone operator     | 1.000      |

## 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)

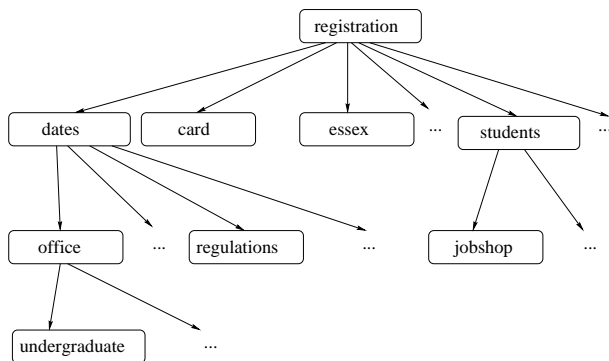
## Alternative: Domain Model derived from Query Logs

| <b>q1</b>           | <b>q2</b>             | <b>MLE</b> |
|---------------------|-----------------------|------------|
| registration        | online registration   | 0.045      |
| registration        | registration office   | 0.035      |
| registration        | timetable             | 0.025      |
| registration        | enrol                 | 0.020      |
| ...                 | ...                   | ...        |
| online registration | registration          | 0.211      |
| ...                 | ...                   | ...        |
| registration office | careers centre        | 0.053      |
| registration office | albert sloman library | 0.053      |
| ...                 | ...                   | ...        |
| enrol               | course enrolment      | 0.050      |

## Alternative: Domain Model derived from Query Logs



## Reminder: Original Domain Knowledge



## 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

- ▶ M. Sanderson & B. Croft. Deriving concept hierarchies from text. SIGIR: 206–213, 1999.
- ▶ P. Anick & S. Tipirneni. The paraphrase search assistant: terminological feedback for iterative information seeking. SIGIR: 153–159, 1999.
- ▶ F. Radlinski & T. Joachims. Query Chains: learning to rank from implicit feedback. SIGKDD: 239–248. 2003.
- ▶ U. Kruschwitz, 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. 2008.
- ▶ D. Lungley & U. Kruschwitz. Automatically Maintained Domain Knowledge: Initial Findings. ECIR: 739-743, 2009.