Utilizing User Access Patterns in Enterprise Search

Udo Kruschwitz

School of Computer Science and Electronic Engineering
University of Essex
United Kingdom

udo@essex.ac.uk

10th October 2014

Overview

- Motivation and context
- Exploiting query logs
- Adaptive search
- *AutoEval*: evaluating adaptive search
- Current work

Background

- Natural language processing (NLP) at Essex goes back a very long time
- Information retrieval (IR) emerged later
- Essex combines both
- Essex: particular focus on practical applications
- Funded research projects (EPSRC, TSB, BT, EU ...)
- About 10 PhD students in the wider area of IR + NLP

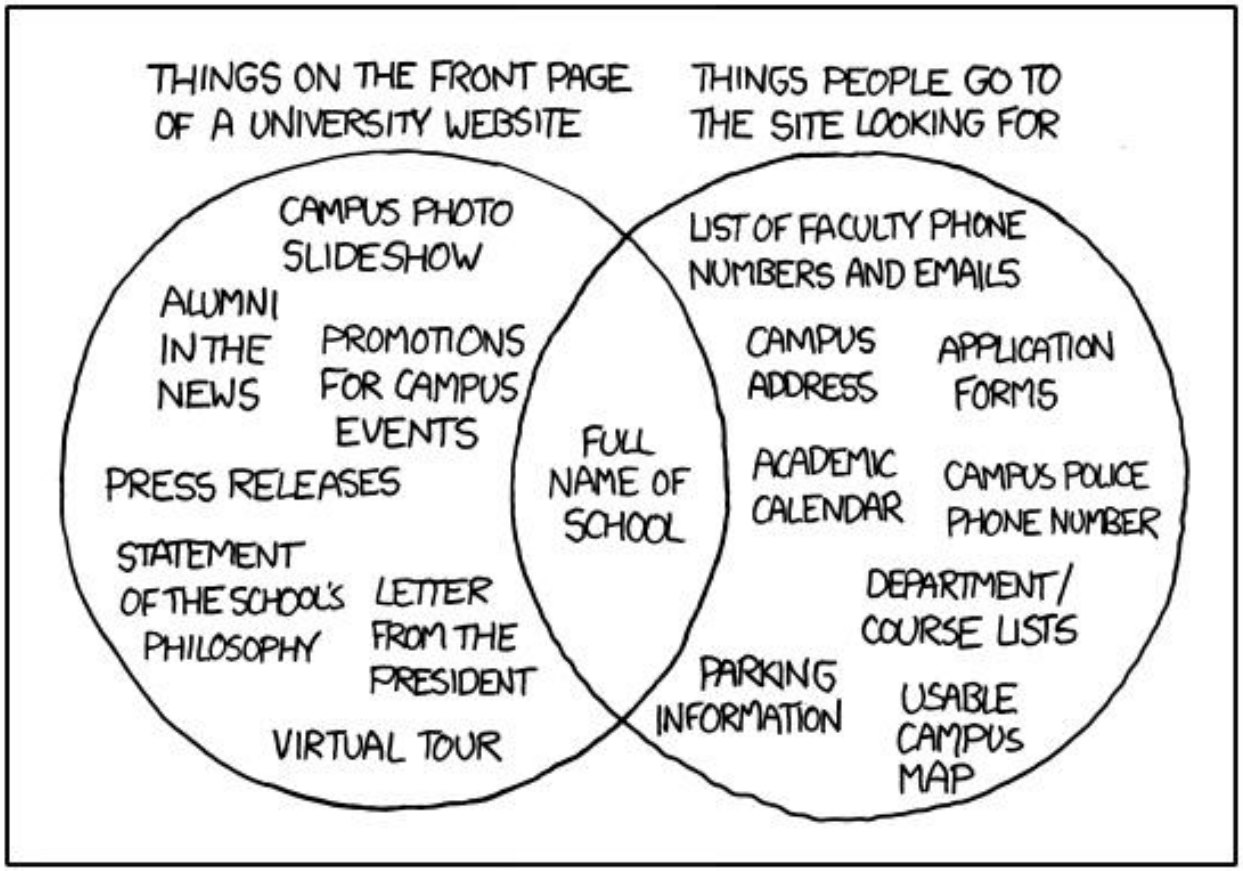
Context

- Collection of documents, e.g.
  digital library, local Web site, 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

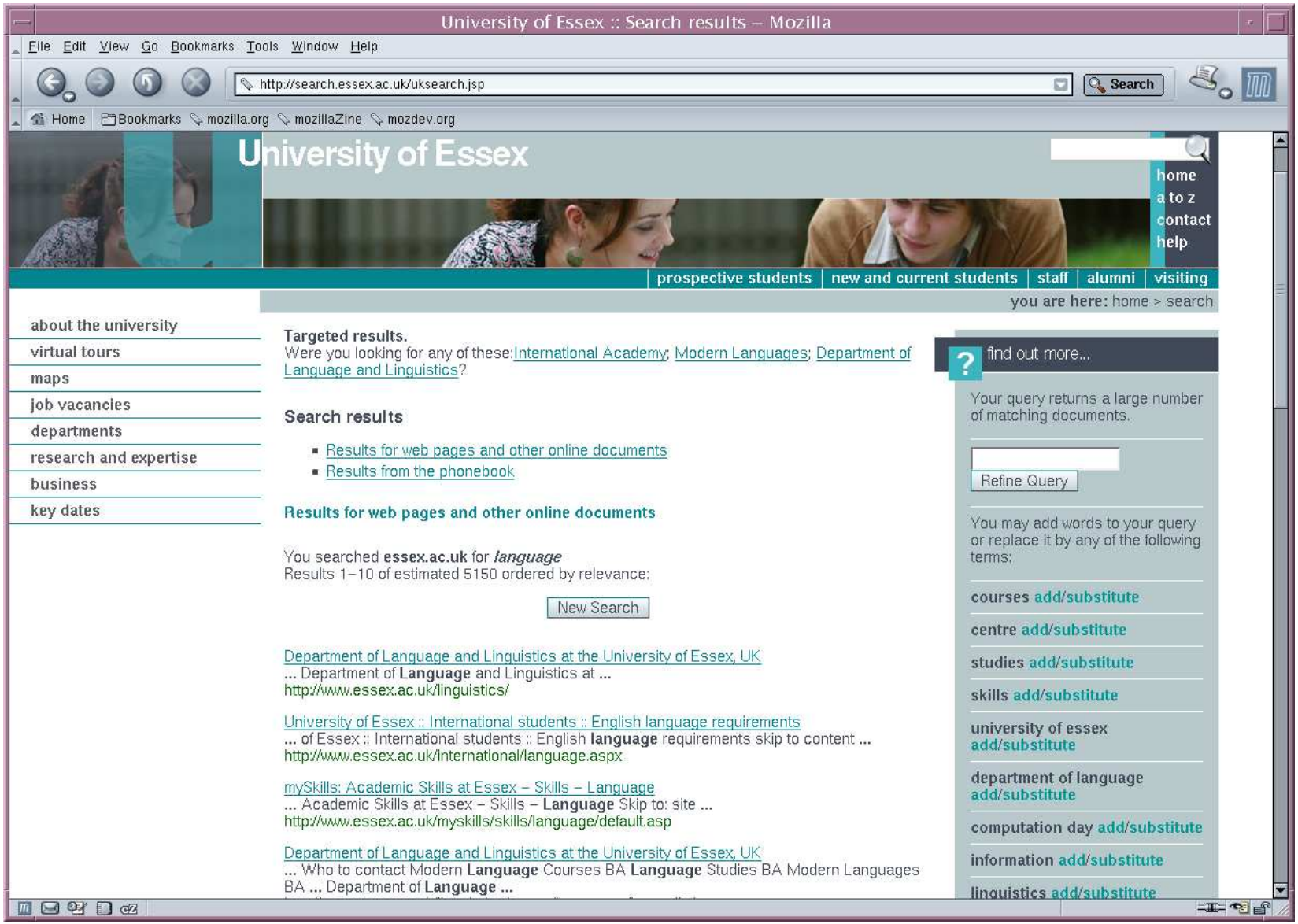
Another Problem



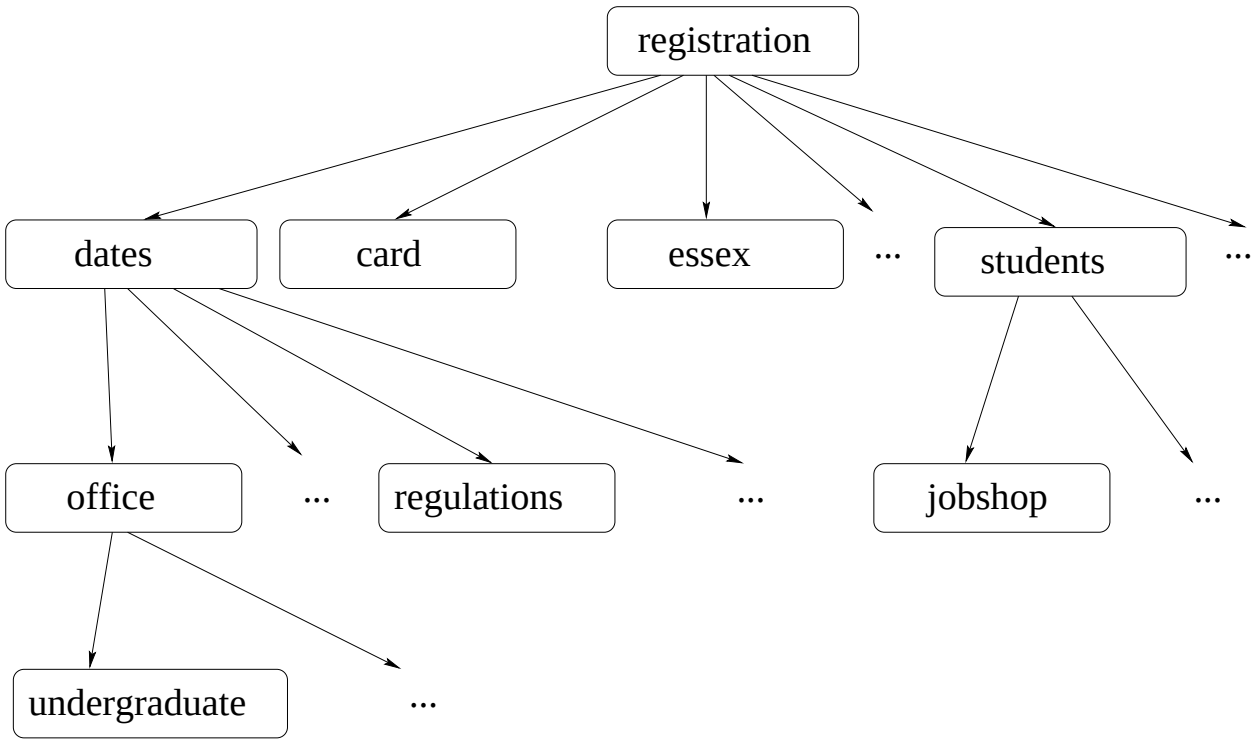
(Source: http://xkcd.com/773)

Our Approach

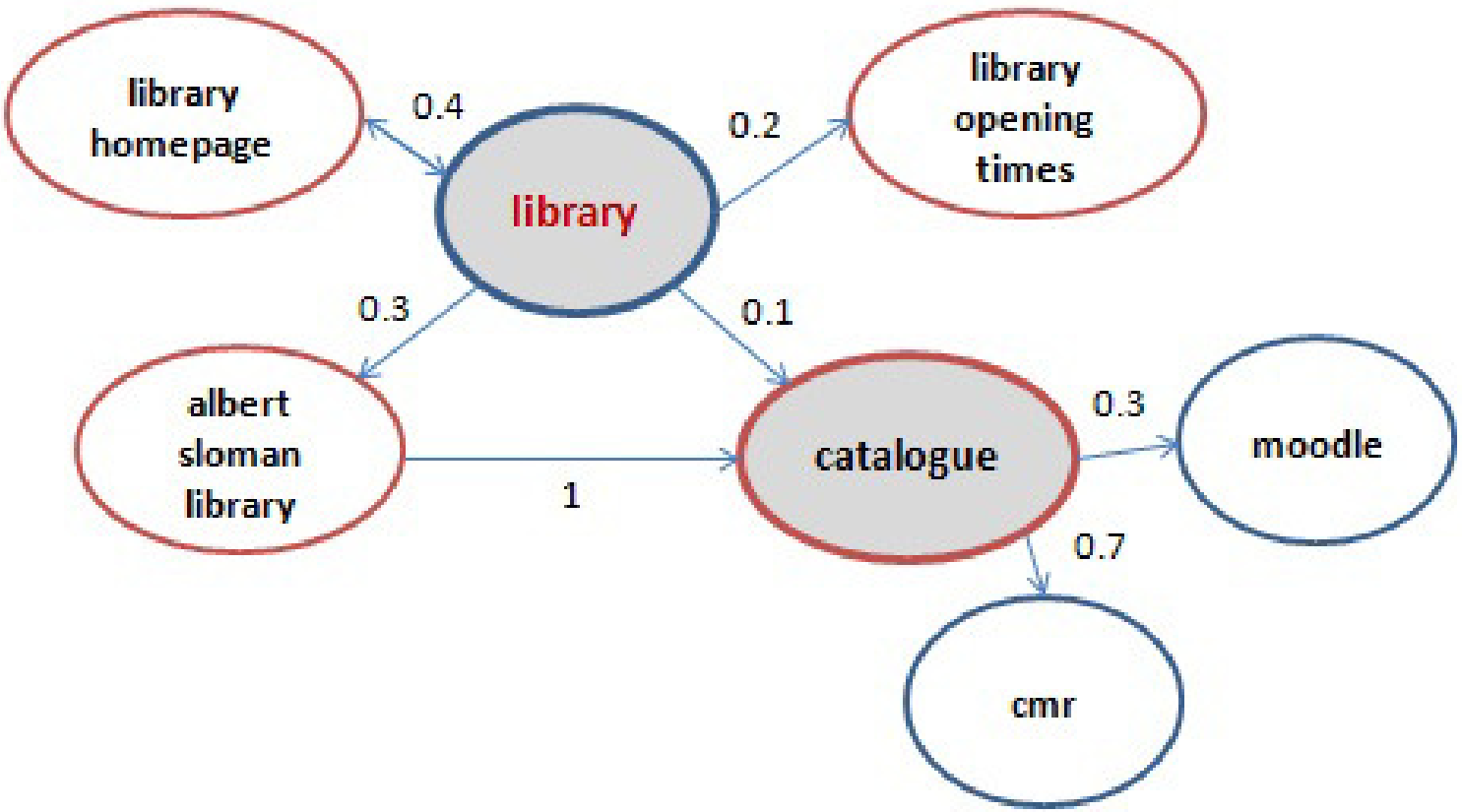
- Search system that makes suggestions using automatically extracted domain knowledge
- But ...
  - Domain knowledge is noisy and incomplete
  - System suggestions not always useful/helpful
  - Document collection is changing
- Learn from the users' interactions
- Improve system over time by adapting to the users' search behaviour
- No single *user profile* but *"community profile"*


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:
New Search
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.
Refine Query
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

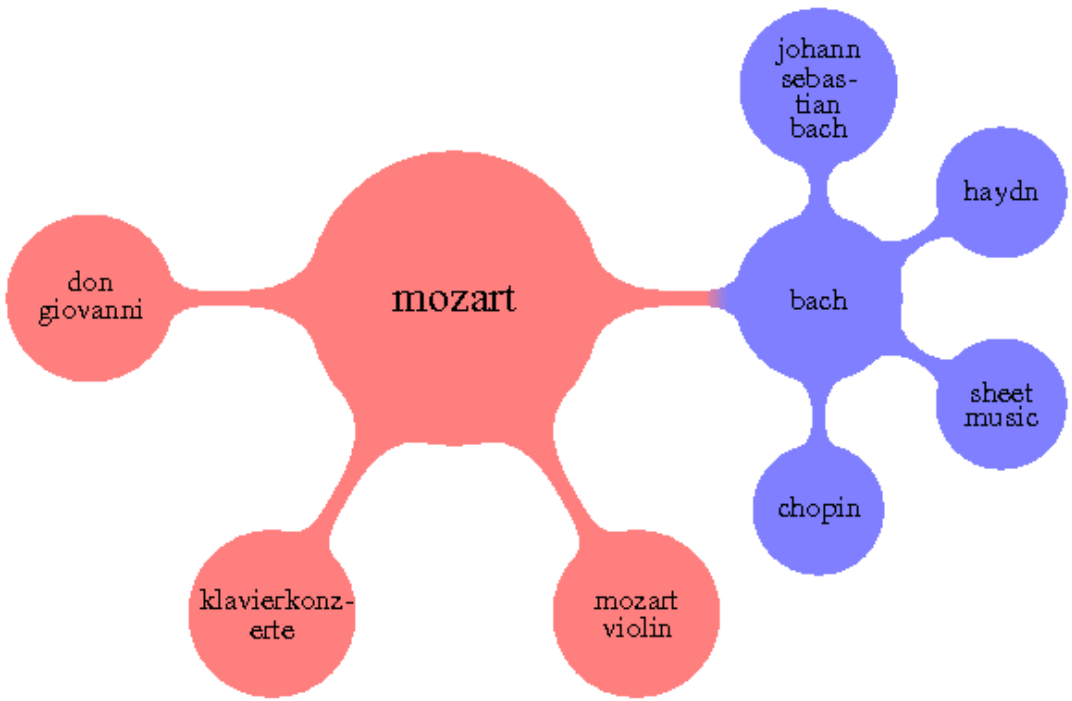
Partial Domain Knowledge (Web Site)



A Different Model



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;83et8b7j010eh4vlht3ucj8dl1;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)

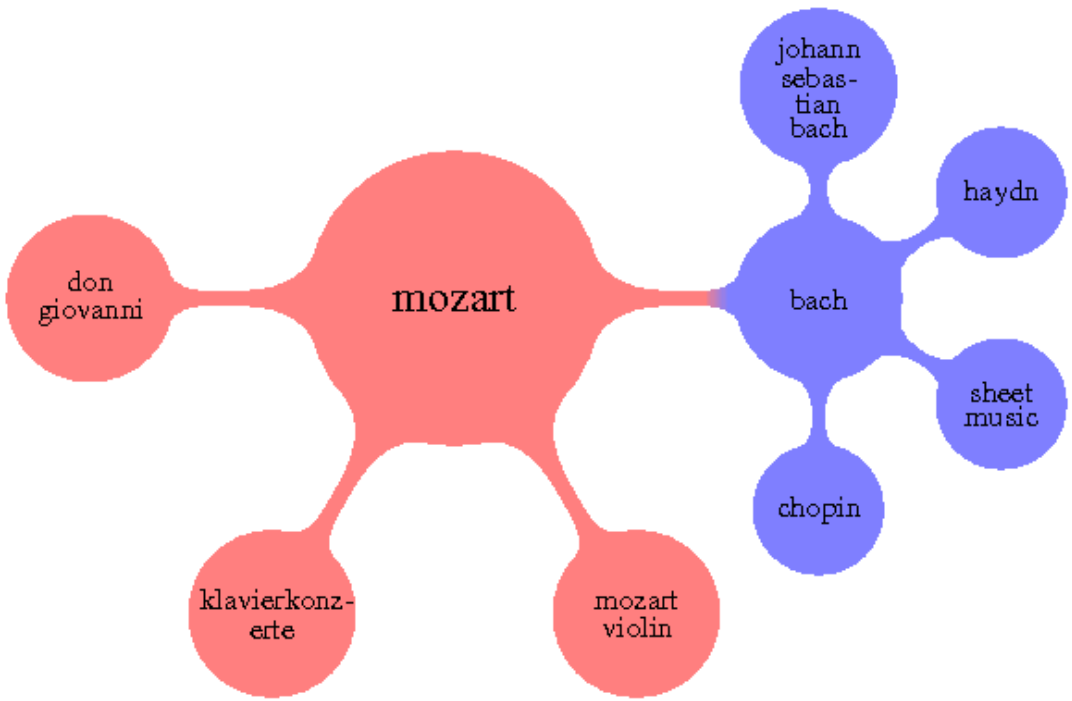
Using Log Data to Acquire a Domain Model - Example

```
...
903779;guest;83.33.xxx.xxx;83et8b7j010eh4vlht3ucj8dl1;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

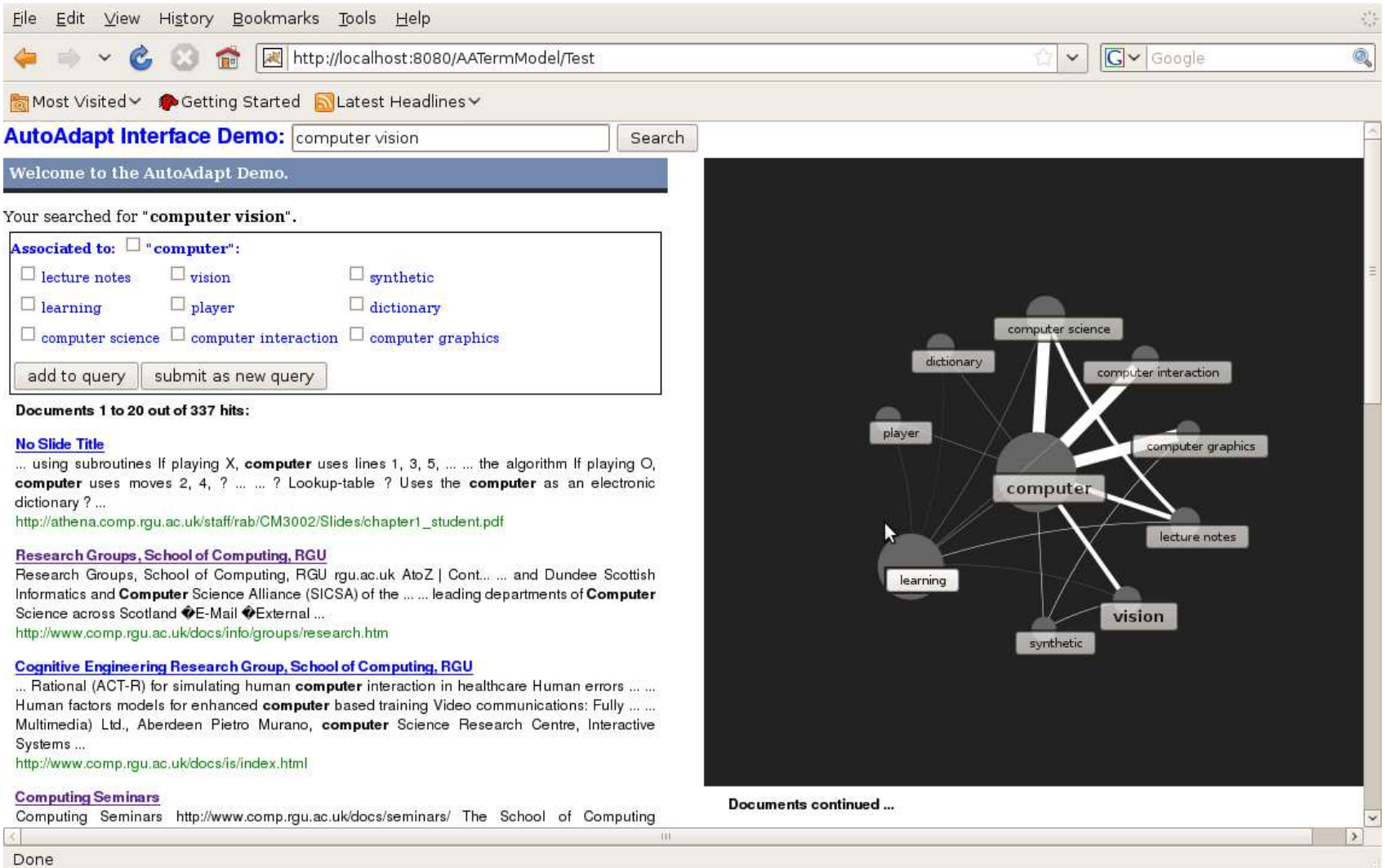
- We use query logs collected on different collections, e.g.
  - *University of Essex* intranet search engine:
    more than 2 million queries (since Nov 2007)
  - *The European Library*:
    1.8 million interactions (Jan 2007 - Jun 2008)
- Query log analysis (not discussed here)
- Bootstrap (adaptive) 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"* $\rightarrow$ *"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:

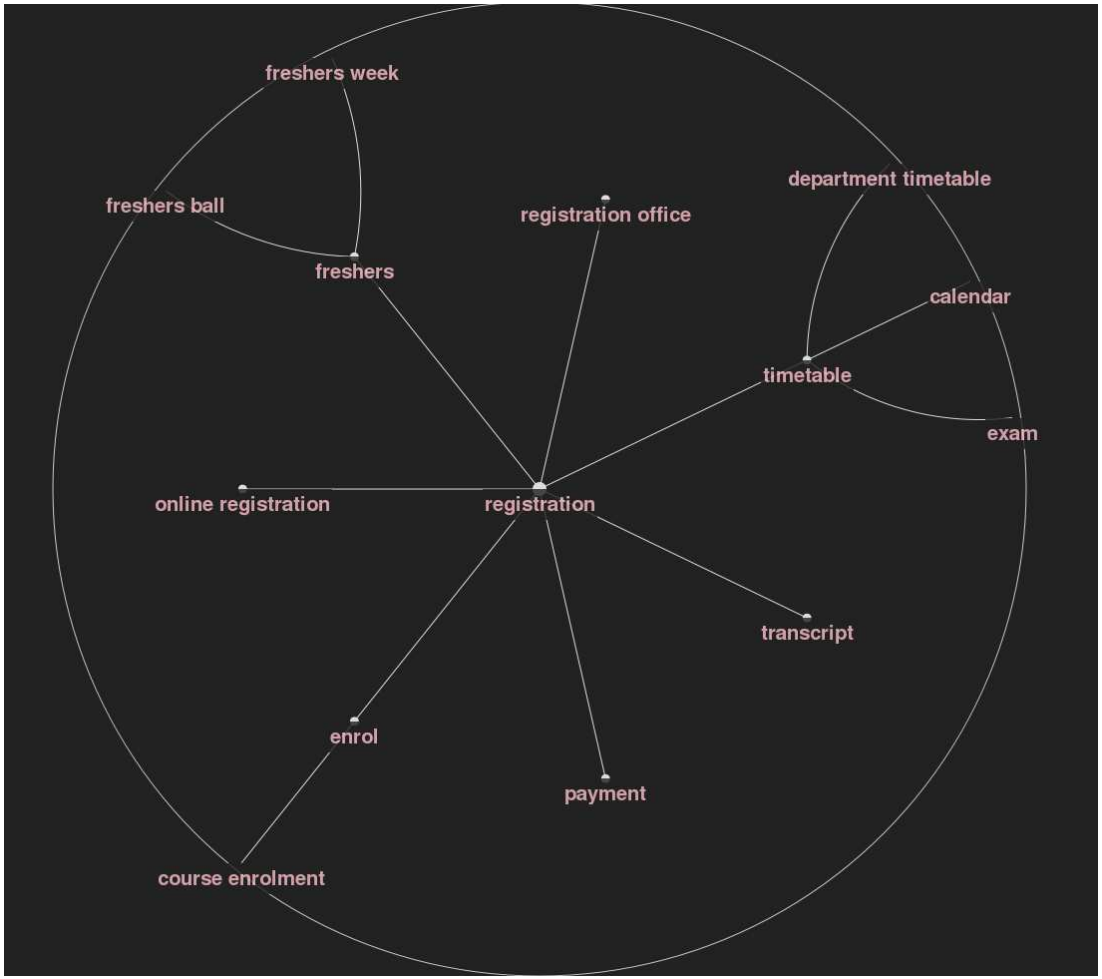
- Exploiting Maximum Likelihood Estimates (MLE)
- Formal Concept Analysis (FCA)
- Ant Colony Optimization analogy (ACO)
- Enhanced Query Flow Graph (QFG)
- Hybrid Approach: Documents + Query Logs
- Adaptive Intranet Navigation

... no time to look at any of these approaches ...

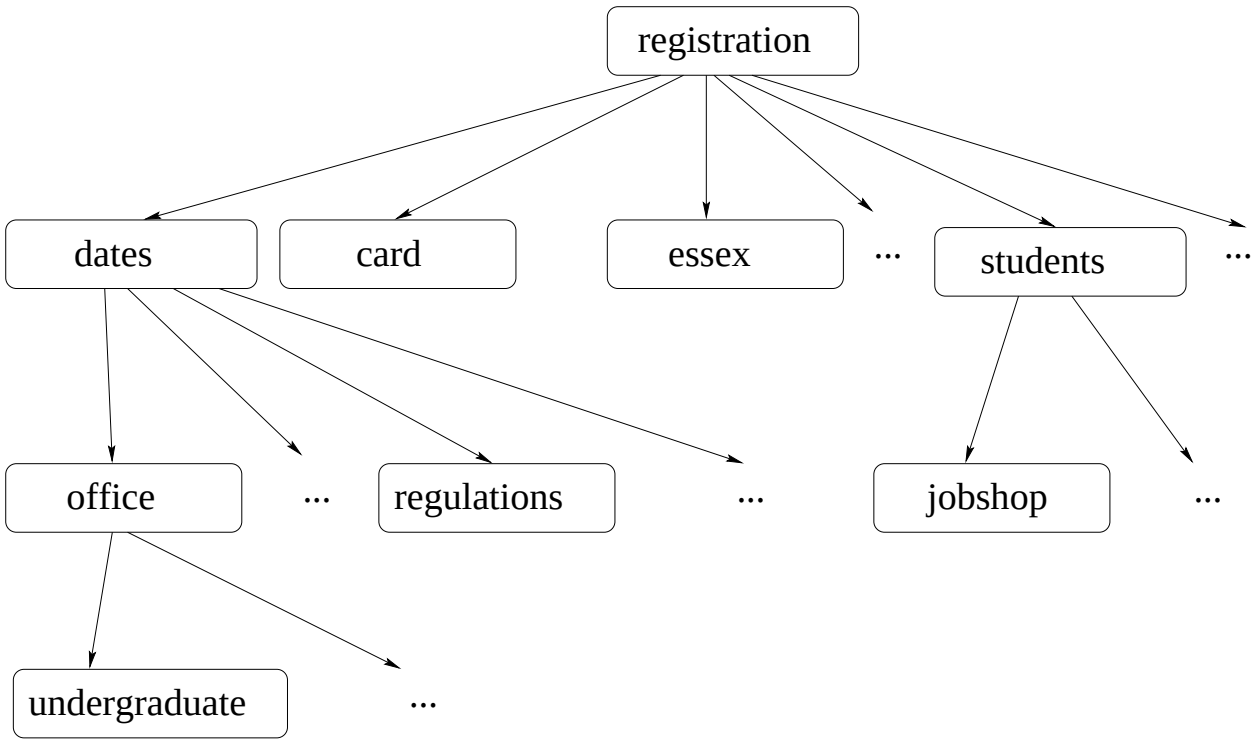
MLE: Domain Model derived from Query Logs

q1	**q2**	**MLE**
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

MLE: Domain Model derived from Query Logs II



MLE: Reminder - Original Domain Knowledge



User Study[1]

- Studied 3-year log file of Essex search engine (about 1.6 million queries)
- Sampled frequent and less frequent queries
- User study to assess quality of derived term suggestions using a variety of log-based and document-based methods (e.g. extracting suggestions from snippets of best matches, association rules approach, query flow graphs, maximum likelihood estimation...)
- Maximum Likelihood approach very accurate (but sparse)
- Query Flow Graphs better coverage and consistent
- Session-based approach seems ok, but more fine-grained session identification is better

[1] (Kruschwitz et al., 2013) in JASIST

FCA Approach to Adaptation[2]

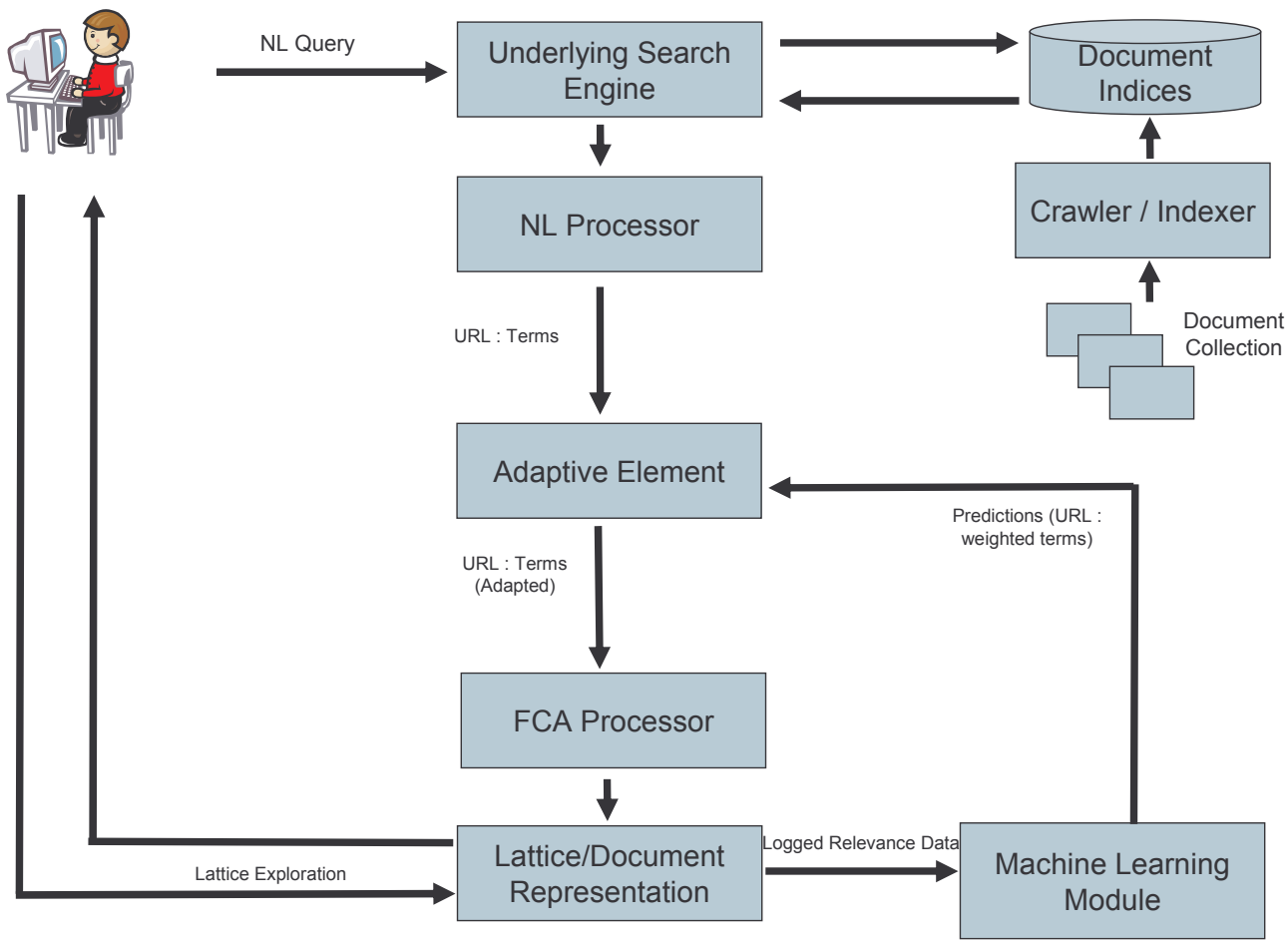
- Lattice structure representing terms and corresponding documents
- Concept in lattice defined by objects (URLs) and attributes (terms)

[2](Lungley, 2012) PhD thesis, University of Essex

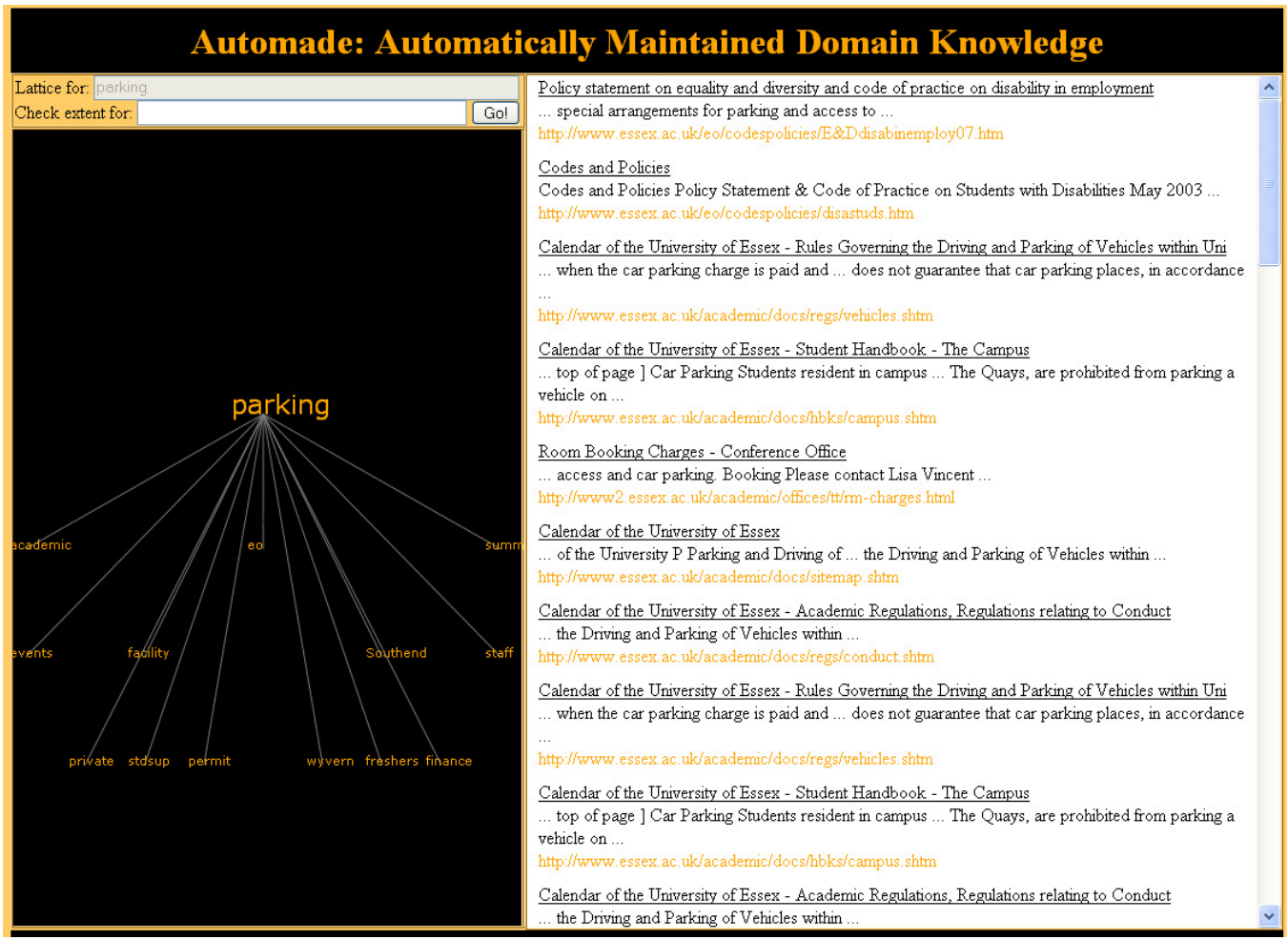
FCA Approach to Adaptation II

- 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



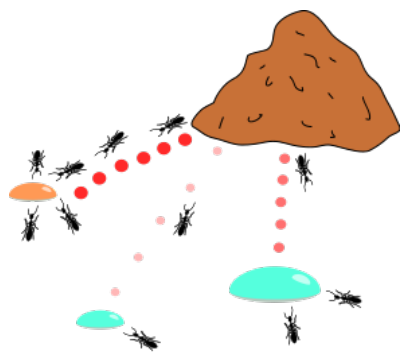
FCA: Results[3]

- User study similar to MLE evaluation (using MSN logs)
- 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
- Drawback: complexity
- Task-based evaluation (using Essex Web site): mixed results

[3] (Lungley et al., 2012) at ECIR 2012

Ant Colony Optimisation[4]

- Biologically inspired model. Ants wander randomly, and upon finding food return to their colony while laying down **pheromone trails**. Trails are then followed by ants and **reinforced** if they find food eventually. Trails also **evaporate** over time.
- Idea: learn associations as they become popular, allow for forgetting relations as well



[4](Albakour et al., 2011) at ICTIR 2011

ACO: Results[5]

- 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)

[5](Kruschwitz et al., 2011)

ACO: Longitudinal Study on Field Force Engineers[6]

- Applying ACO in a realistic call centre setting
- BT engineers both in the field and the call centre
- Search engine indexing different information silos
- A/B testing applied to MLE, AR and ACO
- Finding: low uptake , but ...
- Higher uptake of ACO suggestions than the alternatives
- Statistically significant increase after training phase

... more details in the book chapter

[6](Albakour et al., 2013)

Enriched Query Flow Graphs[7]

- Build *query flow graphs* (QFG) from the query logs
- Update the weights of the edges based on the number of clicks
- Experimented with different co-efficient factors of query click bands

$$w(q, q') = \frac{C_0.\varphi_0(q, q') + C_1.\varphi_1(q, q') + C_k.\varphi_k(q, q')}{\Sigma_{r \in R_q} \Sigma_i C_i.\varphi_i(q, r)}$$

- Evaluation framework: *AutoEval*
- Results: overall improvement over standard QFG; boosting queries that are followed by a *single click* has a positive impact; eliminating queries with *no click* has a negative impact

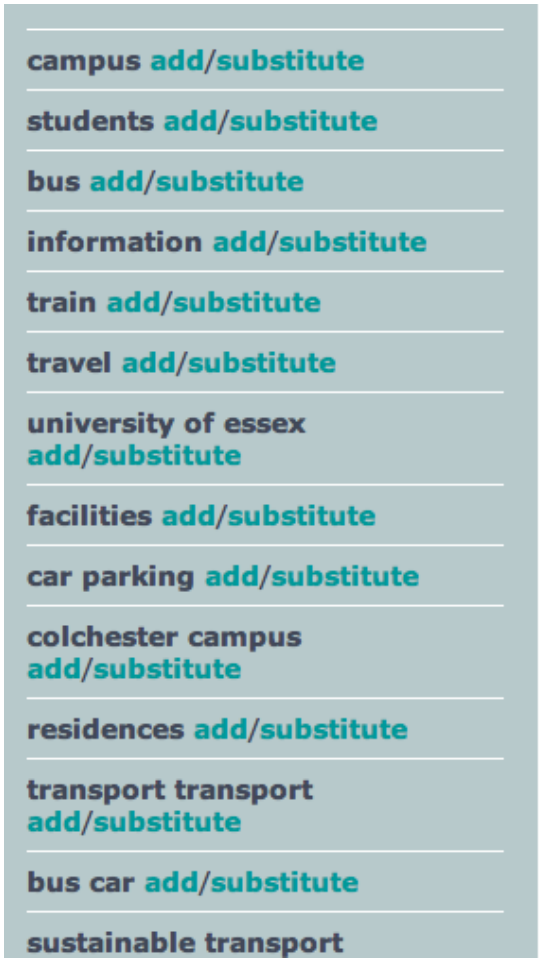
[7] (Albakour et al., 2011) at AIRS 2011

Automatic Domain Model Adaptation

Variety of adaptation models, including:

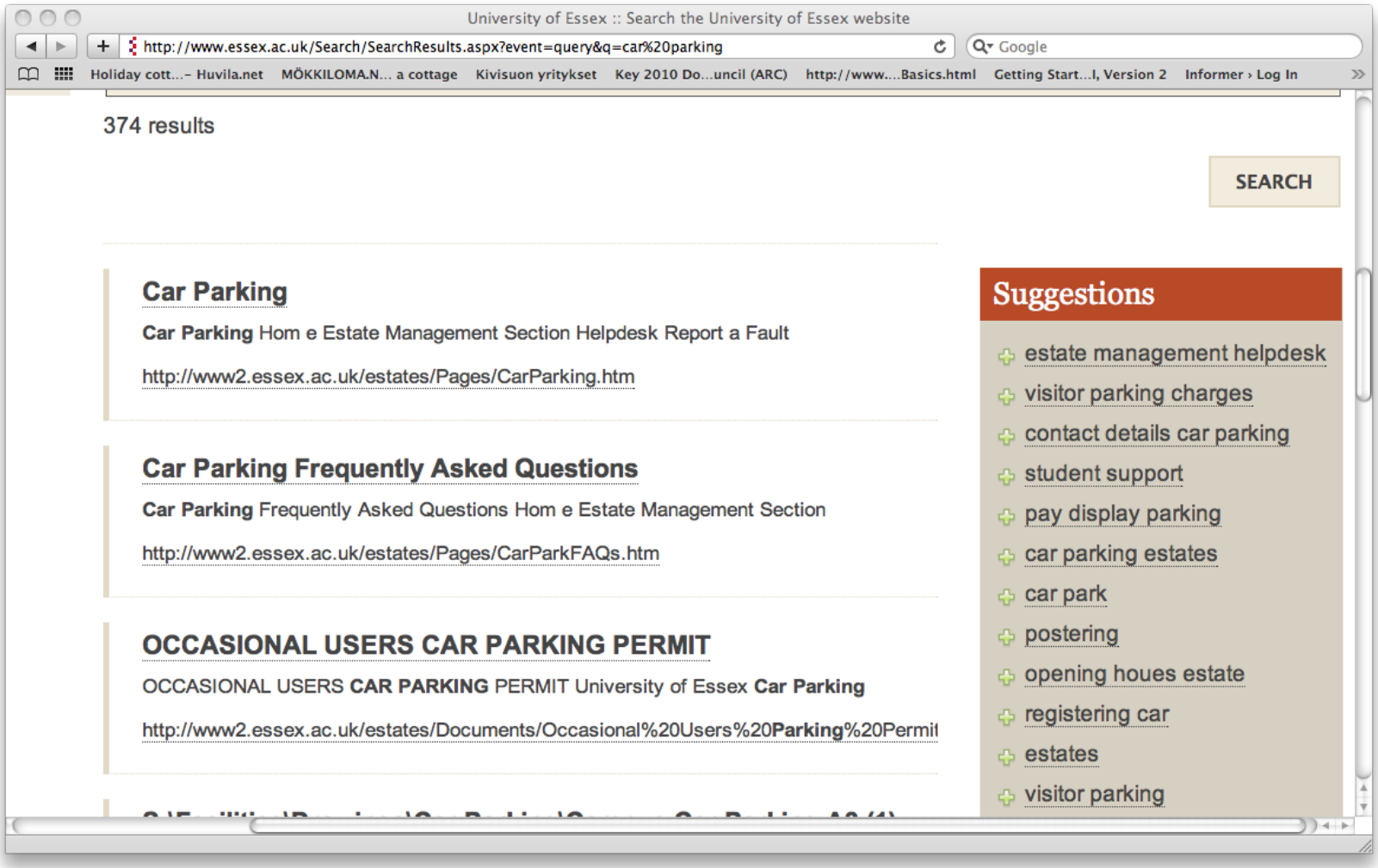
- Exploiting Maximum Likelihood Estimates (MLE)
- Formal Concept Analysis (FCA)
- Ant Colony Optimization analogy (ACO)
- Enhanced Query Flow Graph (QFG)
- Hybrid Approach: Documents + Query Logs
- Adaptive Intranet Navigation

Hybrid Approach: Motivation[8]



[8](Adeyanju et al., 2012) at SIGIR 2012

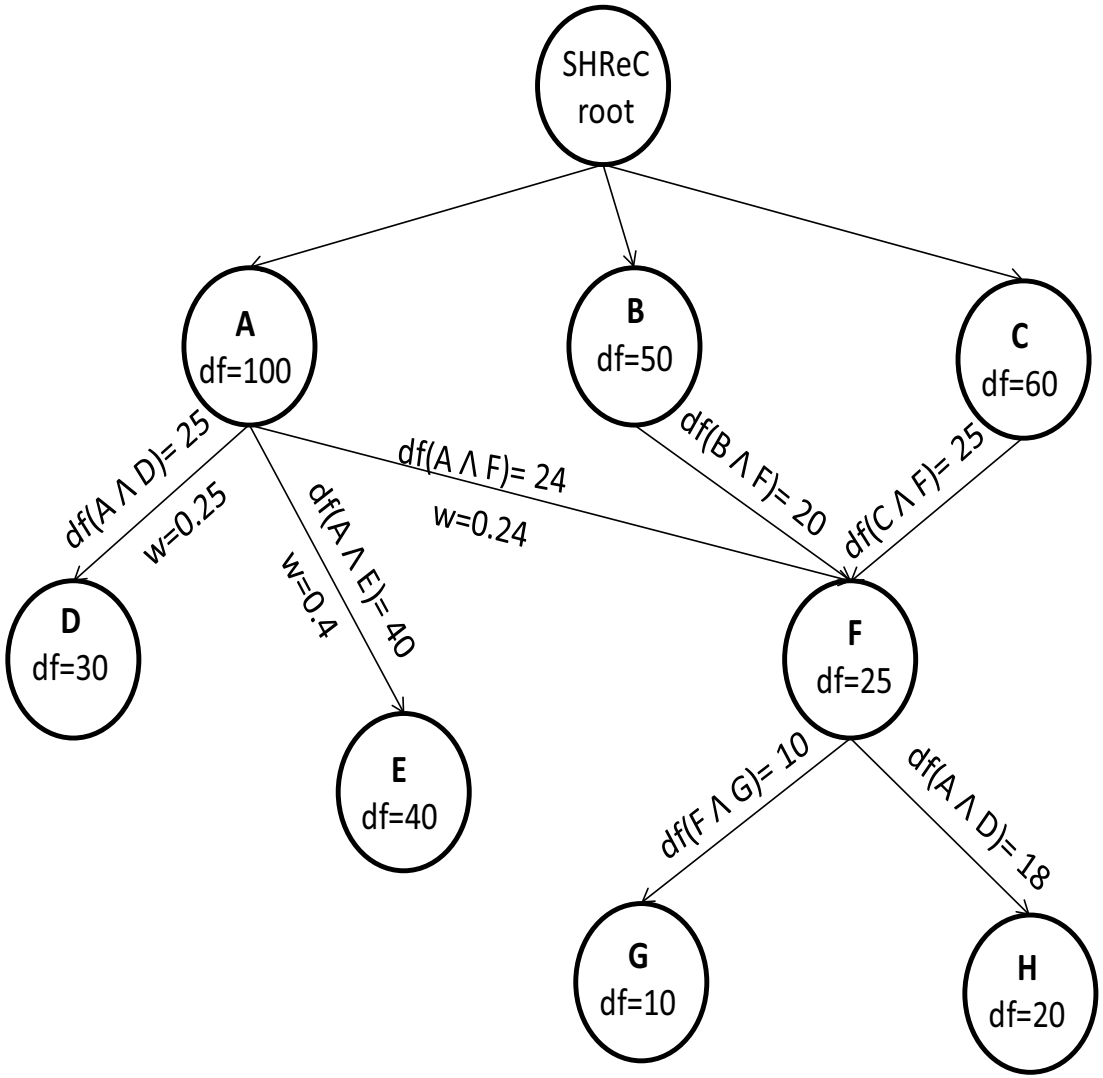
Hybrid Approach: Motivation II



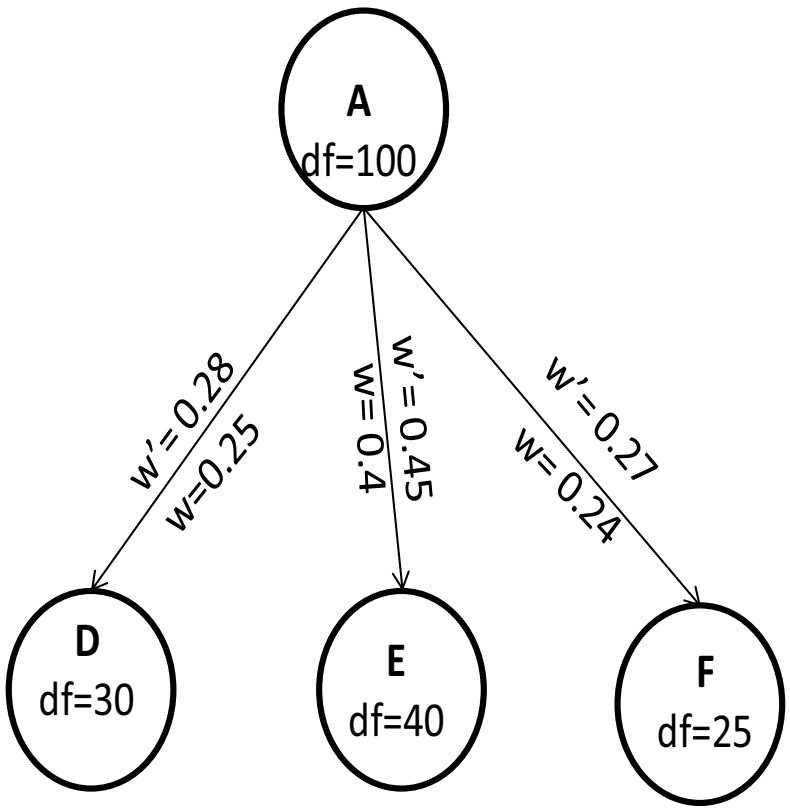
Hybrid Approach: Document-Based + Log-Based

- Observation: models appear to be complementary
- Various pros and cons for each approach
- Hence: bootstrap an initial model (using subsumption hierarchies)
- Evolve model over time using log data
- Result is a hybrid model

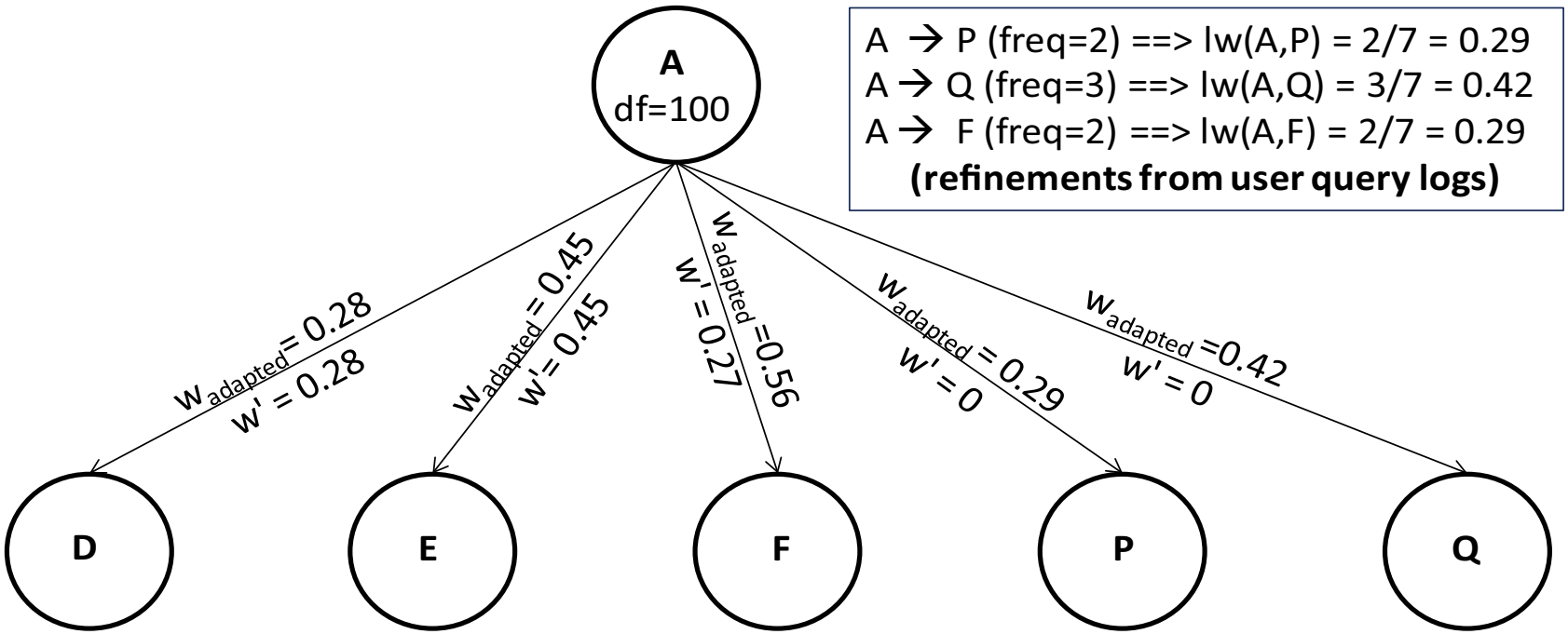
Example: Initial Subsumption Hierarchy Model



Example: Normalised Subsumption Hierarchy Model



Example: Automatically Adapted Model



... all the details in our SIGIR'12 paper

Adaptive Intranet Navigation: Motivation[9]

- Not search but browsing/navigation
- Domain model has exactly the same structure, but nodes are URLs and not queries
- Again: domain model learned from query logs (capture knowledge of *intranet* users to help others)
- Instead of proposing query suggestions: propose links
- Therefore: build a clickgraph using same methods (MLE, ACO, QFG ...)

[9] (Saad & Kruschwitz) at IRFC 2011 and ECIR 2013

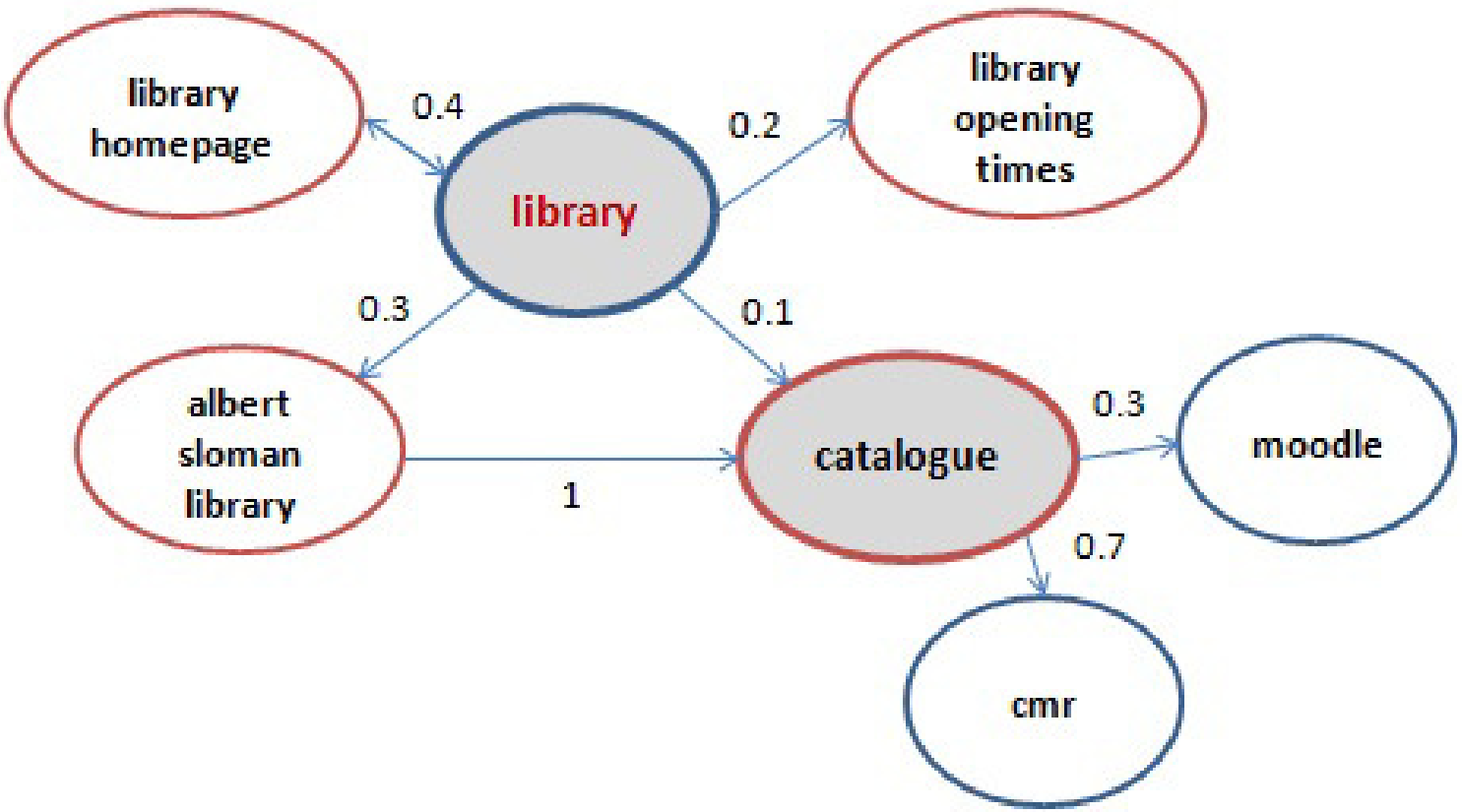
Adaptive Intranet Navigation: Example

Adaptive Intranet Navigation: Results

- Task-based evaluations within a local Web site
- Main measure: navigation trails (steps and time taken)
- Finding of experiment one: adding suggestions (using MLE) outperforms standard Web site without suggestions
- Finding of experiment two: ACO outperforms MLE approach

Adaptive Intranet Navigation: Summarisation

Adaptive Intranet Navigation: Reminder



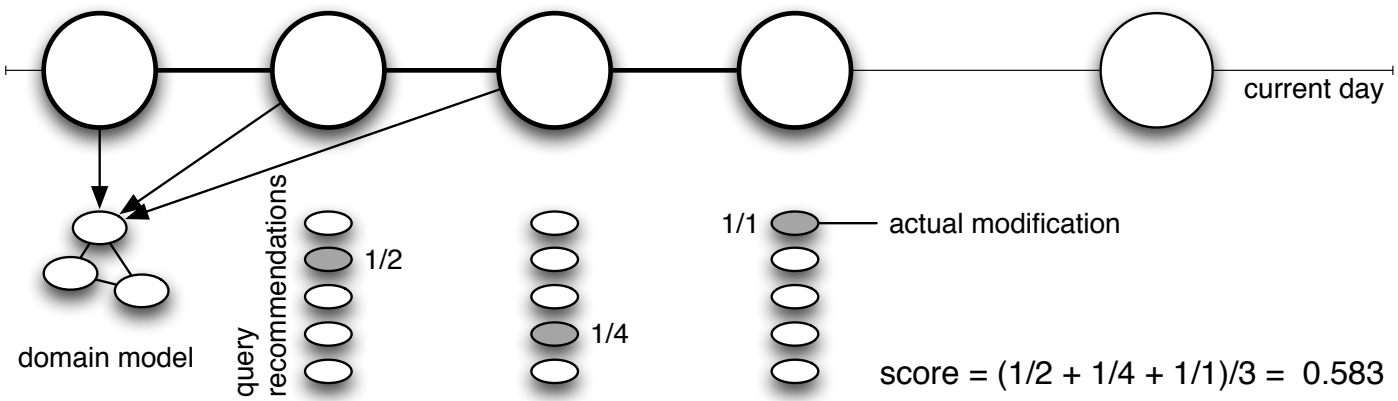
Adaptive Intranet Navigation: Summarisation[10]

- Again: domain model of *terms*
- Applying ACO to summarise documents as you are browsing
- Single document *vs.* multi-document summaries
- Finding: potential for *navigation*
- Profile-based summaries lead to significantly shorter interactions

[10](Alhindi *et al.* at ECIR 2013) and (Alhindi *et al.* under review)

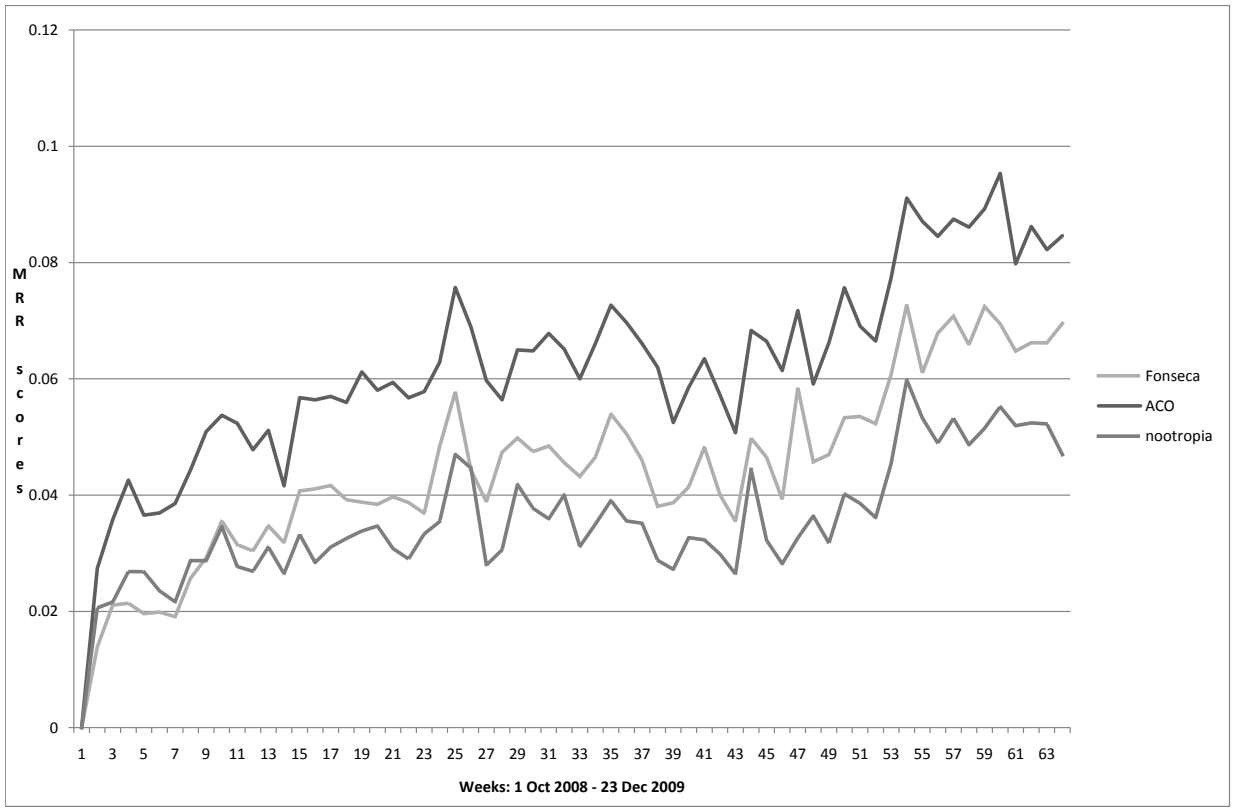
AutoEval: Evaluate Adaptive Search[11]

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

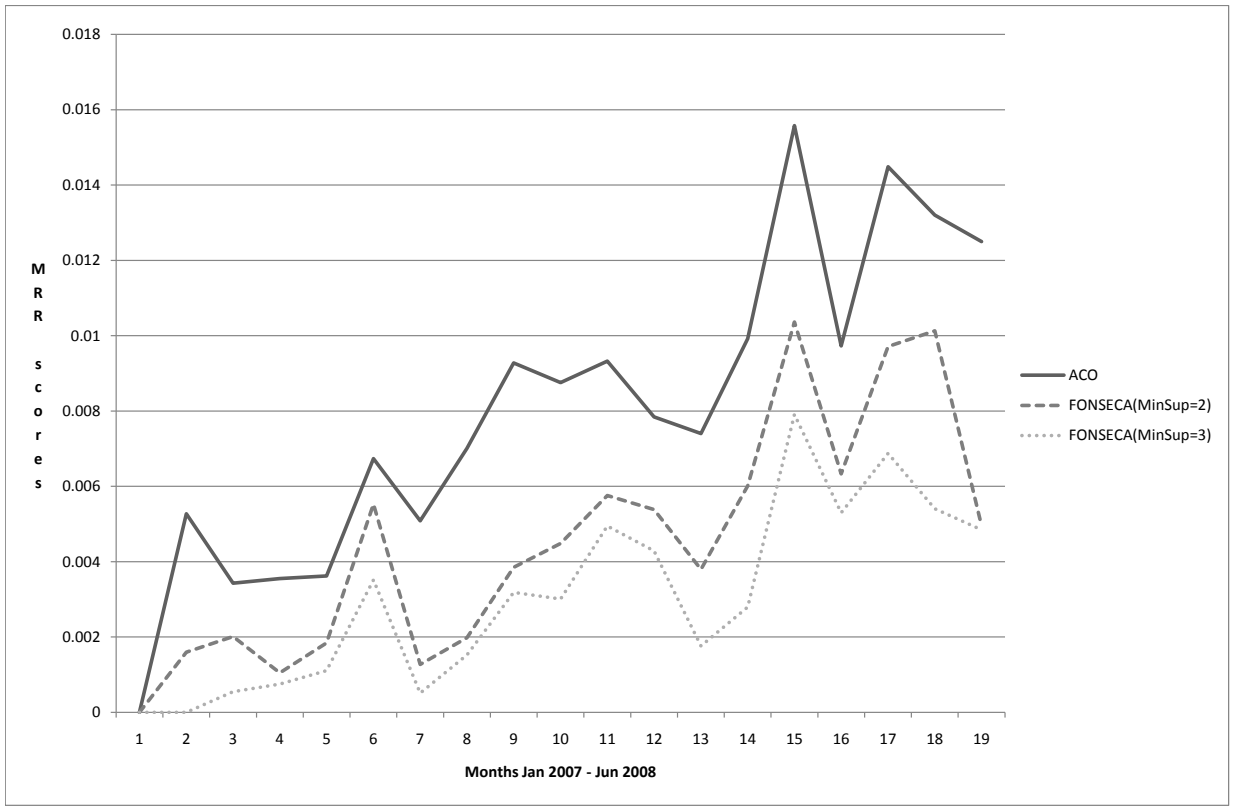


[11](Albakour et al., 2011) at ECIR 2011

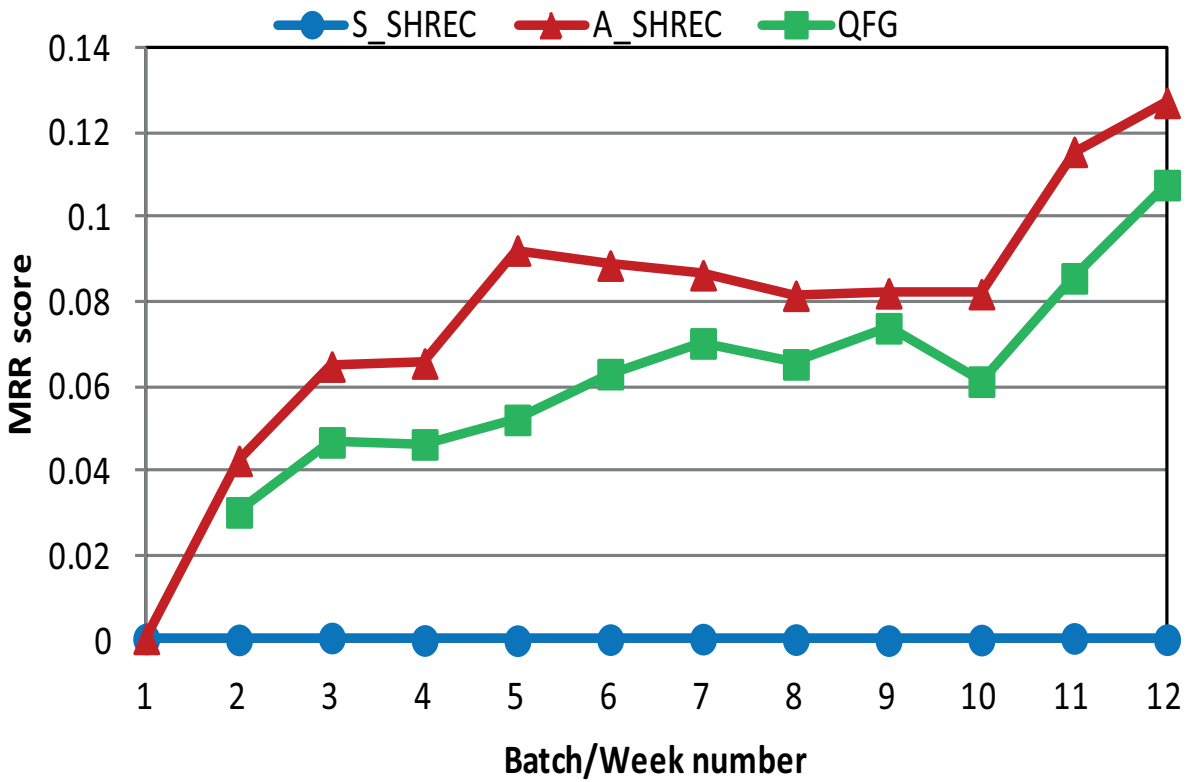
AutoEval Results (Web Site)



AutoEval Results (Digital Library)



AutoEval (Hybrid Approach)



Current Work (Selection)

- University of Essex site search
  (Dyaa Albakour, Deirdre Lungley)
- Prototype in operation for BT's mobile workforce, since summer 2012 see (Albakour *et al.*, 2013)
- Profile-based summarisation (Azhar Alhindi)
- KTP project with MBS Group / Signal combining adaptive profiling with search, summarisation and filtering
- New KTP project just started with the Minority Rights Group on civilian-led monitoring of human rights violations
- EU FP7: SENSEI project on summarisation of conversations in social media
- Centre for Doctoral Training: Intelligent Games and Game Intelligence (IGGI), just started with first cohort of 12 PhD students

Conclusions

- Adaptive search by exploiting query logs
- Focus on site search, digital libraries etc.
- 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

Acknowledgements

- Dyaa Albakour, Nikolaos Nanas, Ibrahim Adeyanju, Dawei Song, Anne De Roeck, Maria Fasli, Jinzhong Niu, Stephen Dignum (EPSRC-funded AutoAdapt research project)
- Deirdre Lungley, Sharhida Saad, Azhar Alhindi (Language and Computation Group at Essex)
- Johannes Leveling (DCU)

References (ACO, QFG)

- U. Kruschwitz, M-D. Albakour, J. Niu, J. Leveling, N. Nanas, Y. Kim, D. Song, M. Fasli, and A. De Roeck. *Moving Towards Adaptive Search in Digital Libraries*. In Advanced Language Technologies for Digital Libraries, volume 6699 of LNCS, pages 41-60. Springer. 2011.
- M-D. Albakour, U. Kruschwitz, N. Nanas, D. Song, M. Fasli and A. De Roeck. *Exploring Ant Colony Optimisation for Adaptive Interactive Search*. In Proceedings of the 3rd International Conference on the Theory of Information Retrieval (ICTIR), volume 6931 of LNCS, Springer, 2011.
- M-D. Albakour, U. Kruschwitz, et al. *Enriching Query Flow Graphs with Click Information*. In Proceedings of the 7th Asian Information Retrieval Societies Conference (AIRS), Springer, 2011.

References (ACO in BT Prototype)

- M-D. Albakour, G. Ducatel, and U. Kruschwitz. *The Role of Search for Field Force Knowledge Management*. In Transforming Field and Service Operations, pages 117–132. Springer. 2013.

References (FCA)

- D. Lungley. Adaptive Information Retrieval Employing Formal Concept Analysis. PhD thesis. University of Essex, 2012.
- D. Lungley & U. Kruschwitz. Automatically Maintained Domain Knowledge: Initial Findings. ECIR 2009, volume 5478 of LNCS, pages 739-743, Springer, 2009.
- D. Lungley, U. Kruschwitz & D. Song. Learning Adaptive Domain Models from Click Data to Bootstrap Interactive Web Search. ECIR 2012, volume 7224 of LNCS, pages 527-530, Springer, 2012.

References (Hybrid Domain Model + Navigation)

- I. Adeyanju, D. Song, M-D. Albakour, U. Kruschwitz, A. De Roeck and M. Fasli. *Adaptation of the Concept Hierarchy Model with Search Logs for Query Recommendation on Intranets*. In Proceedings of SIGIR 2012, Portland, 2012.
- S. Z. Saad & U. Kruschwitz. *Applying Web Usage Mining for Adaptive Intranet Navigation*. In Proceedings of the 2nd Information Retrieval Facility Conference, volume 6653 of LNCS, Springer, 2011.
- S. Z. Saad & U. Kruschwitz. *Exploiting Click Logs for Adaptive Intranet Navigation*. In Proceedings of ECIR 2013, Moscow.
- A. Alhindi, U. Kruschwitz & C.Fox. *A Pilot Study on Using Profile-Based Summarisation for Interactive Search Assistance*. In Proceedings of ECIR 2013, Moscow.

References (Evaluation Methodology, Survey Paper)

- M-D. Albakour, U. Kruschwitz, N. Nanas, Y. Kim, D. Song, M. Fasli, and A. De Roeck. *AutoEval: An Evaluation Methodology for Evaluating Query Suggestions Using Query Logs*. In Proceedings of ECIR 2011, volume 6611 of LNCS, pages 605-610. Springer, 2011.
- M. Clark, Y. Kim, U. Kruschwitz, D. Song, M-D. Albakour, S. Dignum, U. Cervino Beresi, M. Fasli, and A. De Roeck. *Automatically Structuring Domain Knowledge from Text: an Overview of Current Research*. Information Processing and Management (IP&M). 48(3): 552-568, May 2012.

References (Essex at TREC)

- M-D. Albakour, U. Kruschwitz, J. Niu, M. Fasli. *University of Essex at the TREC 2010 Session Track*. In Proceedings of the Nineteenth Text Retrieval Conference (TREC 2010), NIST, 2011.
- M-D. Albakour, U. Kruschwitz, N. Nanas, B. Neville, D. Lungley, M. Fasli. *University of Essex at the TREC 2011 Session Track*. In Proceedings of the Twenteeth Text Retrieval Conference (TREC 2011), NIST, 2012.
- I. Adeyanju, F. M. Nardini, M-D. Albakour, D. Song, U. Kruschwitz. *RGU-ISTI-Essex the TREC 2011 Session Track*. In Proceedings of the Twenteeth Text Retrieval Conference (TREC 2011), NIST, 2012.
- M-D. Albakour and U. Kruschwitz. *University of Essex at the TREC 2012 Session Track*. In Proceedings of the Twenty-First Text Retrieval Conference (TREC 2012), NIST, 2013.

References (Query Suggestion Evaluation)

- U. Kruschwitz, D. Lungley, M-D. Albakour and D. Song *Deriving Query Suggestions for Site Search*. Journal of the American Society for Information Science and Technology (JASIST), 64(10):1975–1994, October 2013.

See you again next month?