

Construction of Goal Association Graphs from Search Query Logs

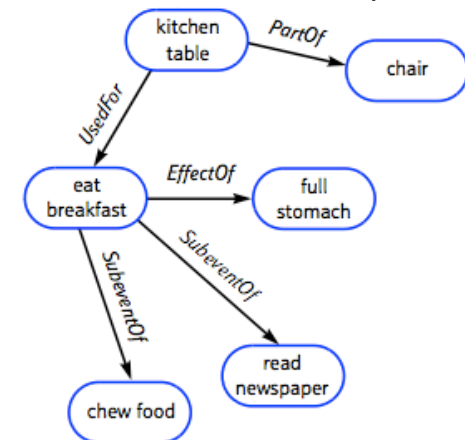
Christian Körner

MSc student

Graz University of Technology

Motivation / 1

- Assuming the availability of automated techniques to separate goals from other queries, it would be interesting to study if and how relations between goals can be inferred.
- Related work:
 - [Baeza-Yates2007] generates graphs from search query logs. Does not infer semantic relations (e.g. links between documents)
 - [Liu2004]: ConceptNet – semantic network for commonsense knowledge

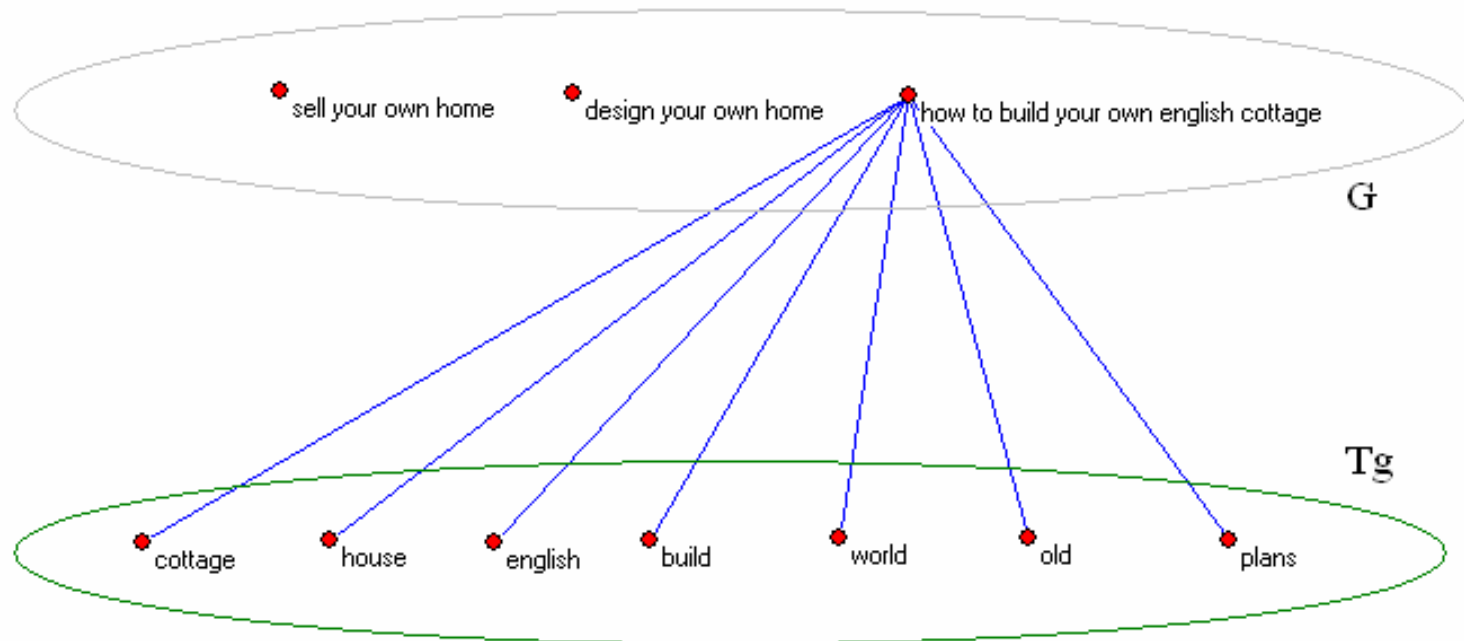


Motivation / 2

- Identifying intentional relations may play a role in query recommendation or in the formation of social search communities sharing similar goals
 - E.g. Web communities which deal with „How to build an english cottage“

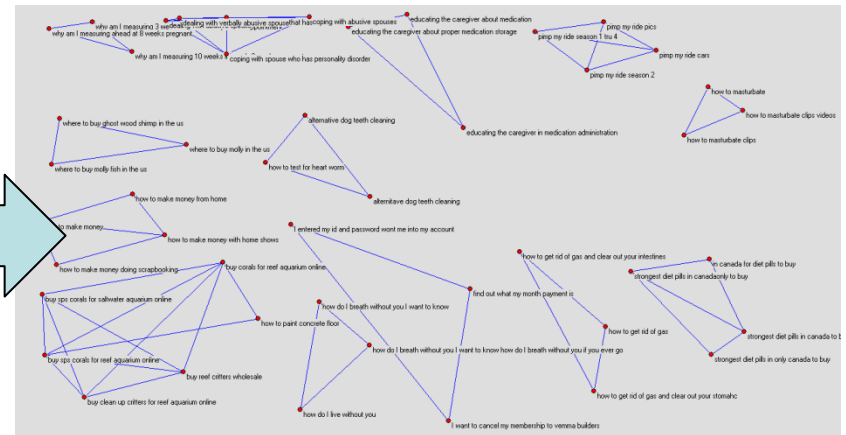
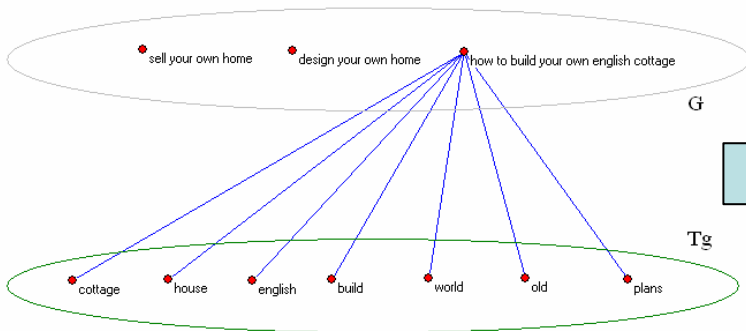
The Graph Construction Process / 1

- Idea: use tags to build a 2-mode graph
 - First mode: goals
 - Second mode: tags



The Graph Construction Process / 2

- We fold the 2-mode network into a 1-mode network consisting only goals



Terminology / 0

id	query	date
1	fluffy roofs house designs	2006-05-27 13:37:19
2	cute house plans	2006-05-27 13:39:15
3	english cottage house plans	2006-05-27 13:45:14
4	old world english cottage house plans	2006-05-27 14:02:02
5	build an english cottage	2006-05-27 14:09:58
6	english cottages	2006-05-27 14:15:23
7	domain furniture	2006-05-27 20:56:23
8	floral design clock and ethan allen	2006-05-27 21:08:38
9	[REDACTED]	2006-05-27 21:10:55
10	[REDACTED]	2006-05-28 12:33:51

Excerpt of the AOL search query log sorted by time of occurrence. User id was omitted and sensitive queries were blacked out.

Terminology / 1

- $q \in Q$ denotes a query, Q_n the set of n queries in a query log
- Q consists of 2 disjoint sets G and I with $g \in G$ and $i \in I$
 - G is the set of queries containing explicit user goals (“build my own english cottage”)
 - I is the set of queries not containing explicit goals (“english cottage house plans”)

Terminology / 2

- Tag set T_g where each t_g shares an intentional relation to a query g
- $N_{g,d}$ is the set of queries which are within a certain distance d of a query g

Terminology illustrated

id	query	date
1	fluffy roofs house designs	2006-05-27 13:37:19
2	cute house plans	2006-05-27 13:39:15
3	english cottage house plans	2006-05-27 13:45:14
4	old world english cottage house plans	2006-05-27 14:02:02
5	build an english cottage	2006-05-27 14:09:58
6	english cottages	2006-05-27 14:15:23
7	domain furniture	2006-05-27 20:56:23
8	floral design clock and ethan allen	2006-05-27 21:08:38
9		2006-05-27 21:10:55
10		2006-05-28 12:33:51

$g \in G$ (blue arrow pointing to row 5)
 $d = 3$ (red arrows indicating distance from row 5 to rows 2 and 8)
 Q (green arrow pointing to row 1)
 $N_{g,d}$ (red arrow pointing to the set of rows {2, 3, 4, 6, 7, 8})

Excerpt of the AOL search query log. User Ids were omitted. Queries are sorted by time of occurrence. Sensitive queries were blackened out.

Approaches

- The constructed 2 - mode networks depend heavily on the tags.
- Tag generation is the most important step!
- So far five different approaches labeled A – E
- Each approach generates another set of tags T_g for a given goal g

Approach A

- Simply uses the queries in the neighborhood $N_{g,d}$ as tags
- $T_{\text{build an english cottage}} = \{\text{cute house plans, english cottage house plans, ...}\}$
- Problem: resulting 2-mode graph is very sparse
no relations between goals of different users

- $d = 3$ in this example

id	query	date
1	fluffy roofs house designs	2006-05-27 13:37:19
2	cute house plans	2006-05-27 13:39:15
3	english cottage house plans	2006-05-27 13:45:14
4	old world english cottage house plans	2006-05-27 14:02:02
5	build an english cottage	2006-05-27 14:09:58
6	english cottages	2006-05-27 14:15:23
7	domain furniture	2006-05-27 20:56:23
8	floral design clock and ethan allen	2006-05-27 21:08:38
9		2006-05-27 21:10:55
10		2006-05-28 12:33:51

Approach B

- Uses tokens as tags e.g. single words of the neighboring queries
- $W(q \in Q)$ denotes set of distinct words $w \in W$ of query q
- $T_{\text{build an english cottage}} = \{\text{and, cottage, cute, english, house, plans, old, world, ...}\}$
- Problem: noise

id	query	date
1	fluffy roofs house designs	2006-05-27 13:37:19
2	cute house plans	2006-05-27 13:39:15
3	english cottage house plans	2006-05-27 13:45:14
4	old world english cottage house plans	2006-05-27 14:02:02
5	build an english cottage	2006-05-27 14:09:58
6	english cottages	2006-05-27 14:15:23
7	domain furniture	2006-05-27 20:56:23
8	floral design clock and ethan allen	2006-05-27 21:08:38
9		2006-05-27 21:10:55
10		2006-05-28 12:33:51

Approach C

- Tokens are single words
- A set of stop words S removes noise e.g. the words „the“, „a“, „and“ etc.
- $T = W(N_{g,r}) \setminus S$
- $T_{\text{build an english cottage}} = \{\text{cottage, cute, english, house, plans, old, world, ...}\}$
- Only “and” removed in this example

id	query	date
1	fluffy roofs house designs	2006-05-27 13:37:19
2	cute house plans	2006-05-27 13:39:15
3	english cottage house plans	2006-05-27 13:45:14
4	old world english cottage house plans	2006-05-27 14:02:02
5	build an english cottage	2006-05-27 14:09:58
6	english cottages	2006-05-27 14:15:23
7	domain furniture	2006-05-27 20:56:23
8	floral design clock and ethan allen	2006-05-27 21:08:38
9		2006-05-27 21:10:55
10		2006-05-28 12:33:51

Approach D

- Observation: Not all neighboring queries share an intentional relationship with the goal g
- Introduce set R_m that satisfies $|W(g) \cap W(N_{g,d})| \geq m$ where m specifies the minimum intersection size (raw similarity according to [Rijsbergen1997])
- $T = R_m$
- $T_{\text{build an english cottage}} = \{\text{house, plans, old, world}\}$

id	query	date
1	fluffy roofs house designs	2006-05-27 13:37:19
2	cute house plans	2006-05-27 13:39:15
3	english cottage house plans	2006-05-27 13:45:14
4	old world english cottage house plans	2006-05-27 14:02:02
5	build an english cottage	2006-05-27 14:09:58
6	english cottages	2006-05-27 14:15:23
7	domain furniture	2006-05-27 20:56:23
8	floral design clock and ethan allen	2006-05-27 21:08:38
9		2006-05-27 21:10:55
10		2006-05-28 12:33:51

Approach E

- Again $|W(g) \cap W(N_{g,d})| \geq m$
- Words from the query g are added to the tag set T as well $\rightarrow T = R_m \in W(g)$
- $T_{\text{build an english cottage}} = \{\text{build, cottage, english, house, plans, old, world}\}$
- Good approach for now

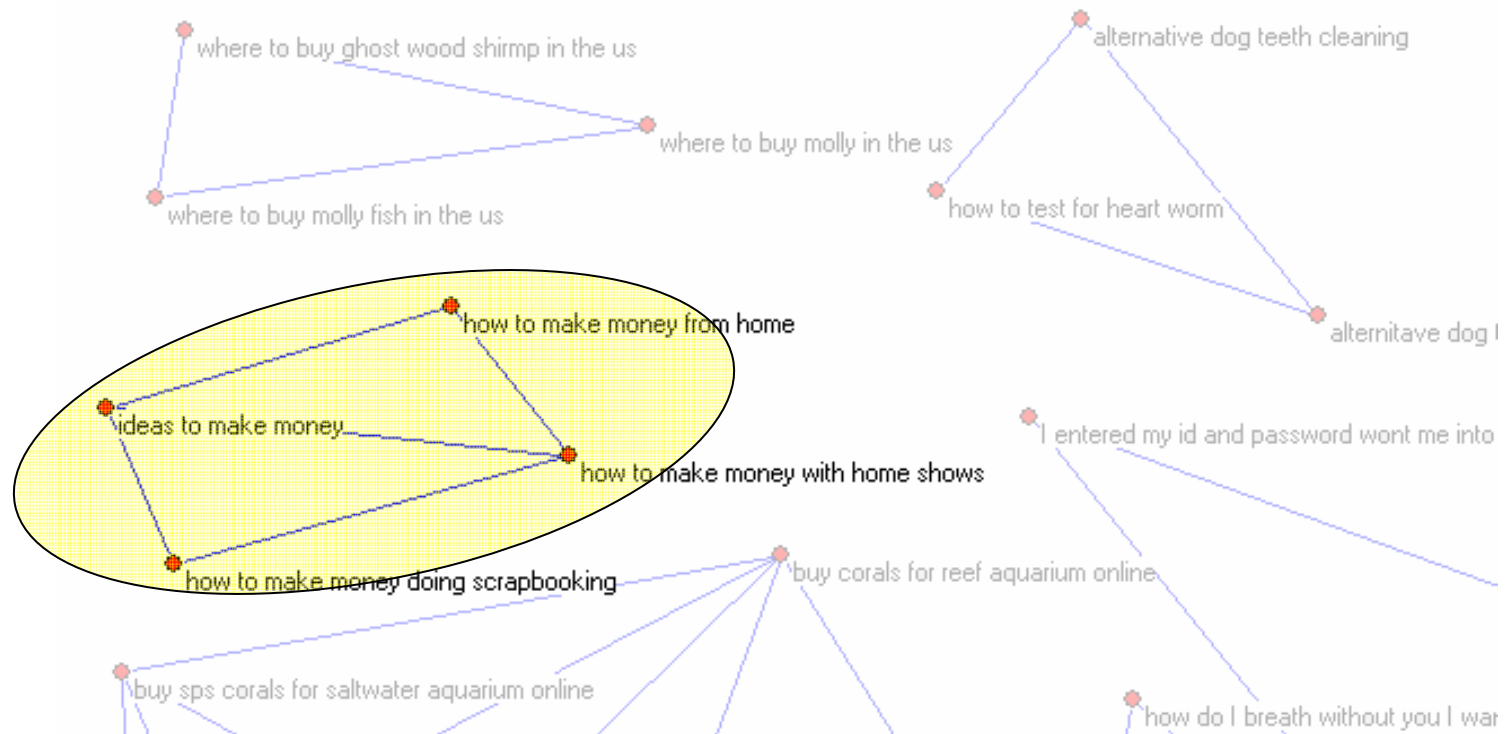
id	query	date
1	fluffy roofs house designs	2006-05-27 13:37:19
2	cute house plans	2006-05-27 13:39:15
3	english cottage house plans	2006-05-27 13:45:14
4	old world english cottage house plans	2006-05-27 14:02:02
5	build an english cottage	2006-05-27 14:09:58
6	english cottages	2006-05-27 14:15:23
7	domain furniture	2006-05-27 20:56:23
8	floral design clock and ethan allen	2006-05-27 21:08:38
9		2006-05-27 21:10:55
10		2006-05-28 12:33:51

Interesting research questions

- What are good tags and how do we generate them automatically?
- How do the parameters influence the quality of the tag generation?
- How can the resulting graph be evaluated?

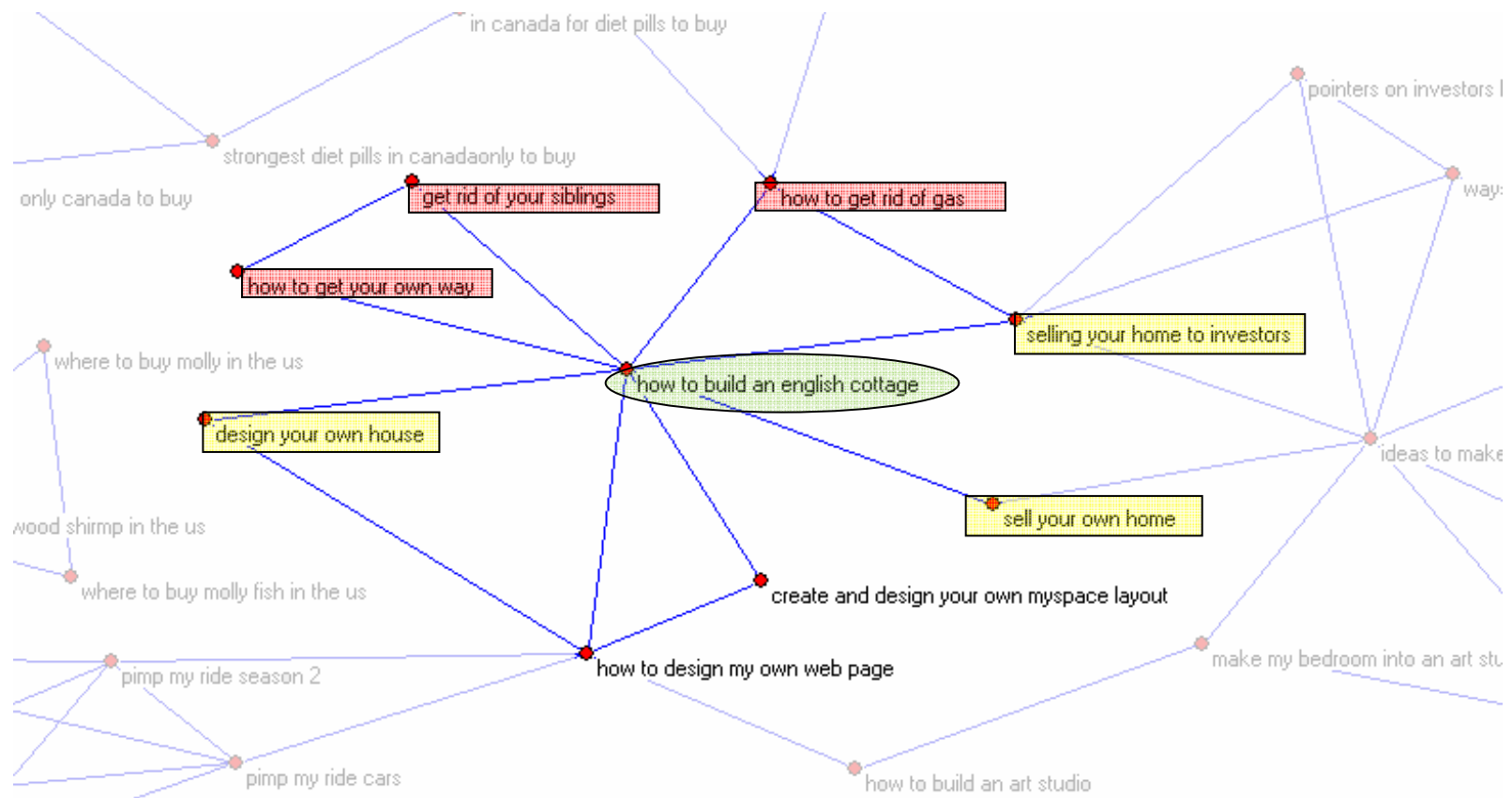
Observations / 1

- Sub graph of result of approach A



Observations / 2

- Sub graph of result of approach E



Outlook

- Advance the formalization
- Evaluate the graphs using facilities such as diameter, KNC-plot [Kumar2008] etc.
- Experiment with different approaches and multiple parameters and evaluate the results

Thank you for your attention!

References

- [Baeza-Yates2007] Baeza-Yates, R., Tiberi, A.: Extracting Semantic Relations From Query Logs, KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007
- [Kumar2008] Kumar, R., Tomkins, A., Vee, E., Connectivity structure of bipartite graphs via the KNC-plot, WSDM '08: Proceedings of the international conference on Web search and web data mining, 2008
- [Liu2004] Liu, H., Singh, P.: ConceptNet — A Practical Commonsense Reasoning Tool-Kit, BT Technology Journal, 2004
- [Rijsbergen1997] Van Rijsbergen, C.: Information Retrieval, 2nd edition, Dept. of Computer Science, University of Glasgow, 1997