# Constructing Large Scale Goal Graphs from Search Query Logs

Christian Körner

MSc student

Graz University of Technology

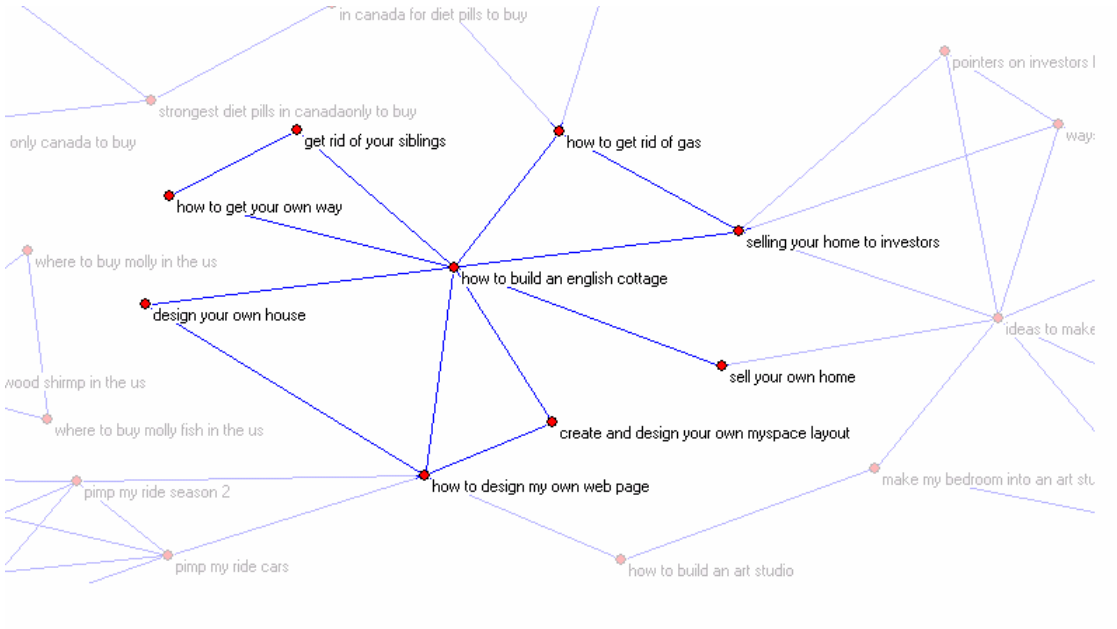
# Input: Search Query Logs

| id | query | date |
|---|---|---|
| 1 | fluffy roofs house designs | 2006-05-27 13:37:14 |
| 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 13:56:10 |
| 5 | how to build an english cottage | 2006-05-27 14:11:12 |
| 6 | english cottages | 2006-05-27 14:15:23 |
| 7 | domain furniture | 2006-05-27 20:56:23 |
| [redacted] | | |

<u>Explicit intentional query</u>:
how to build an english cottage

<u>Implicit intentional queries</u>:
fluffy roofs house designs
cute house plans
english cottage house plans

# Output: Large Scale Goal Graph




- Applications:
  - Goal Recommendation
  - Goal Prediction

# The Problem

- Constructing large,„high quality“ goal graphs from a search query log

- Questions:
  - How can we generate „high quality“ goal graphs?
  - What characterizes a „high quality“ goal graph?
    - Qualitative / quantitative evaluation

# Motivation

- Based on an automated technique to separate explicit from implicit user goals it would be interesting to study if and how relations between explicit user goals can be inferred.

# Related Work

- [Baeza-Yates2007] generates graphs from search query logs. Does not infer semantic or intentional relations (e.g. links between documents)
- [Liu2004]: ConceptNet – semantic network for commonsense knowledge
- [Dai2006]: detects online commercial intention
- [Jansen2006]: defining search sessions on web search engines

# The Idea

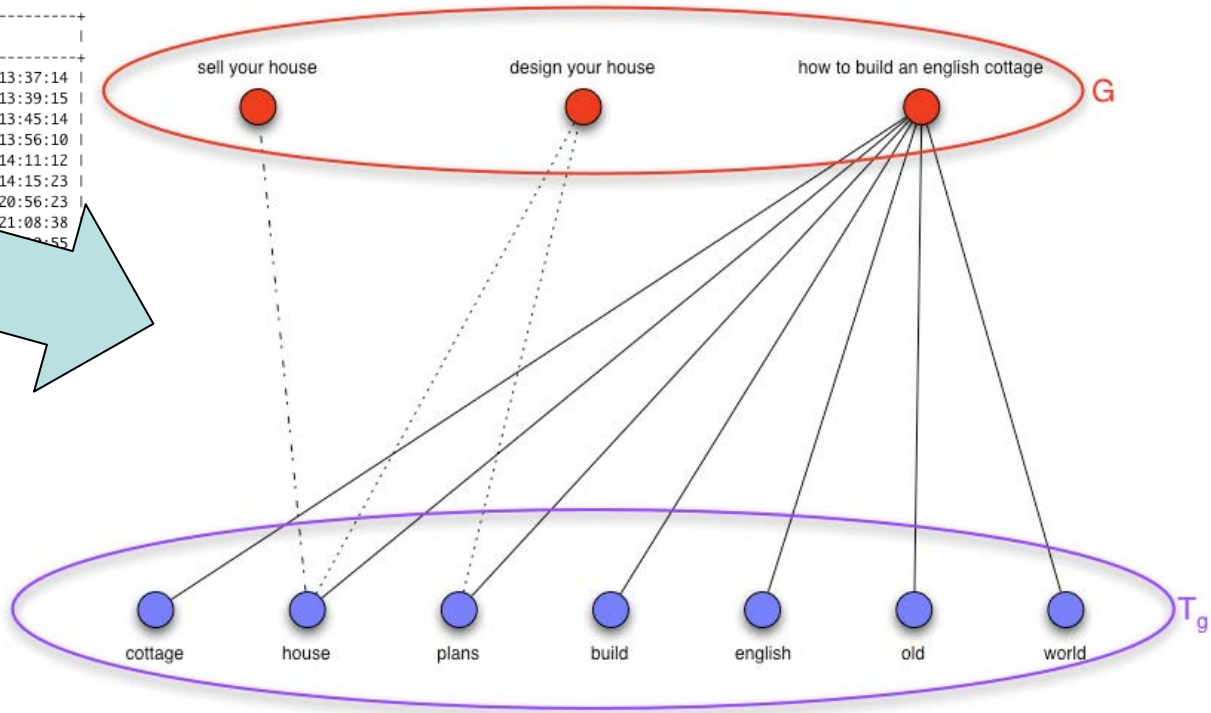
- Given:
  - The AOL Search Query Log [Pass2006]
    - 21.011.340 web queries
    - 657.426 users
  - An improved set of explicit intentional queries [Strohmaier2008]
    - 118.420 queries
    - Precision: 0.77
    - Recall: 0.63
- Assumption:
  - Related goals share a similar „word space“ (Tags) which can be accessed by examining “neighboring” queries of a goal

→ Construct a bipartite graph

# The Graph Construction Process / 1

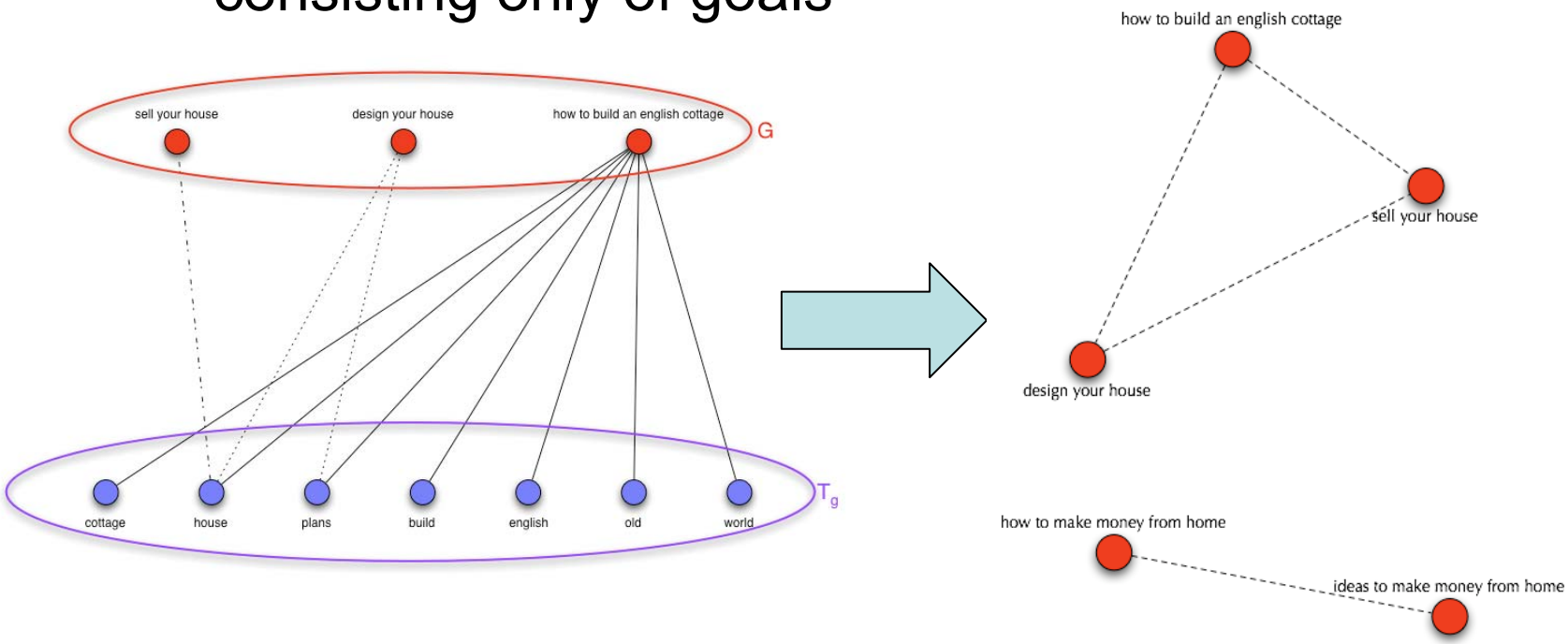
- Idea: use tags to build a bipartite graph
  - First mode: goals
  - Second mode: tags

| id | query | date |
|---|---|---|
| 1 | fluffy roofs house designs | 2006-05-27 13:37:14 |
| 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 13:56:10 |
| 5 | how to build an english cottage | 2006-05-27 14:11:12 |
| 6 | english cottages | 2006-05-27 14:15:23 |
| 7 | domain furniture | 2006-05-27 20:56:23 |
| 8 | [redacted] | 27 21:08:38 |
| 9 | [redacted] | :55 |



# The Graph Construction Process / 2

- We fold the bipartite graph into a unipartite graph consisting only of goals




Folding: multiplying the affiliation matrix A with its transpose ($=A*A^T$)

# 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 occurence. User id was omitted and sensitive queries were blacked out.

# Terminology illustrated

Q

d= 3

g □ G

$N_{g,d}$

| 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. User Ids were omitted. Queries are sorted by time of occurence. Sensitive queries were blackened out.

# Parameters for the tag generation

- d(istance): specifies the neighborhood which should be searched for new tags
- i(ntersection): if a query shares *i* words with the original goal query it is used for the tag generation (raw similarity - [Rijsbergen1997] )
- Include goal words: if this parameter is set to true the words of the goal are added as tags as well

# Parameters for the tag generation / 3

| 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 |

- assignTags(3, 2, true)

# Evaluation Strategy – Work in Progress

Two Approaches:

- Quantitative Evaluation
- Qualitative Evaluation

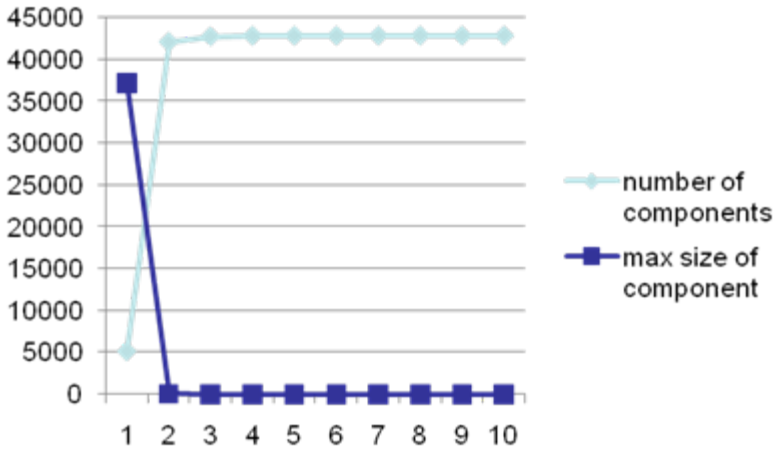
# Quantitative Evaluation / 1

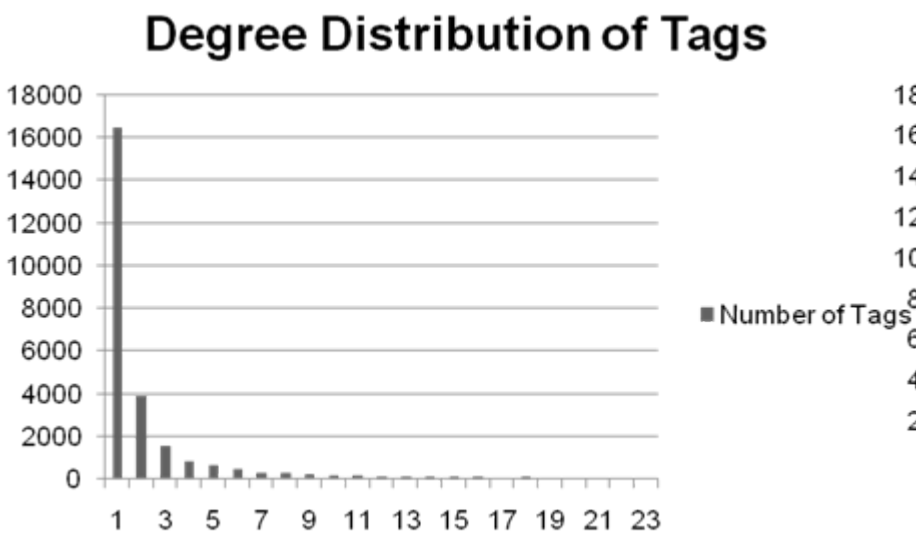
- Evaluation of graph characteristics such as
  - Degree Distribution
  - KNC Plot [Kumar2008]
  - Clustering Coefficient [Wasserman1994]

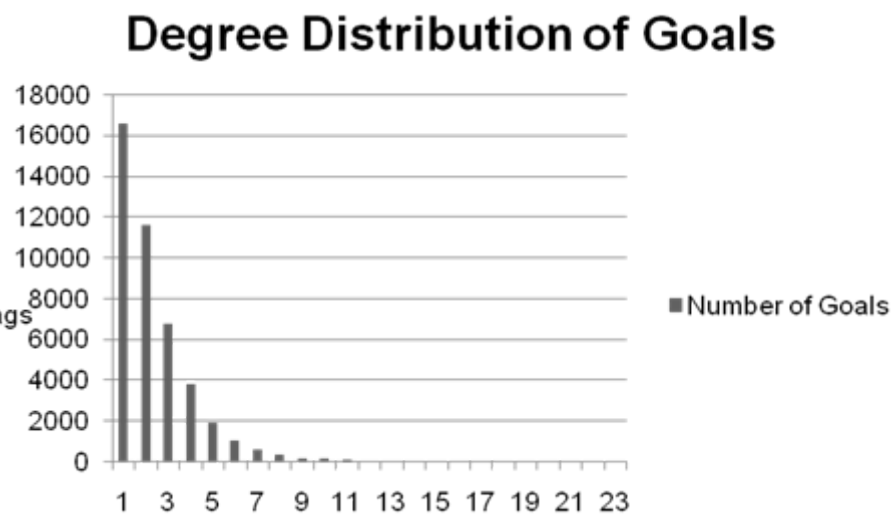
# Quantitative Evaluation / 2

- assignTags(1,1,false)

- # goals in the graph: 42849
- # tags in the graph: 26339
- # edges in the graph: 101890



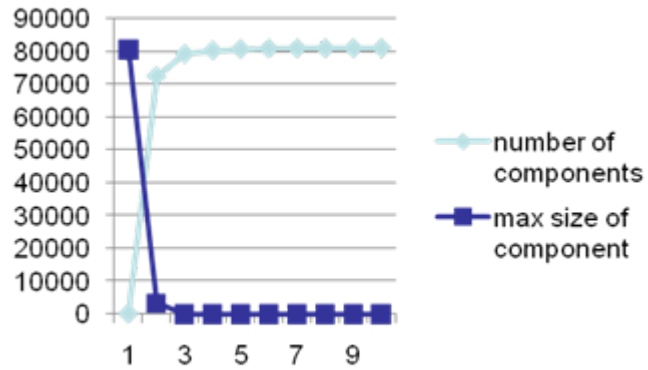



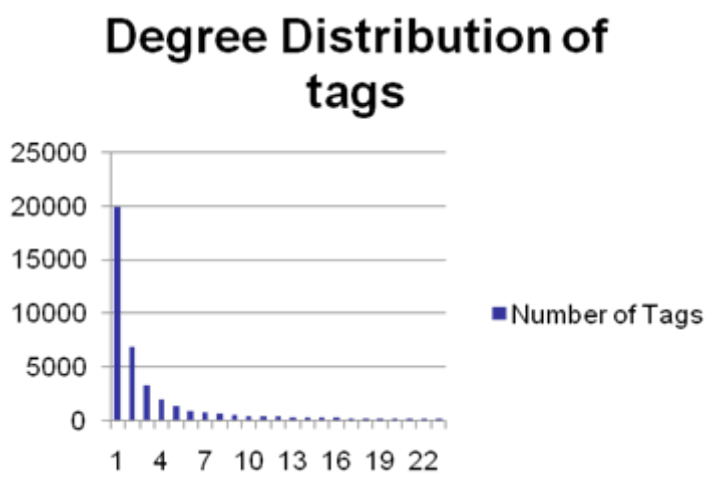



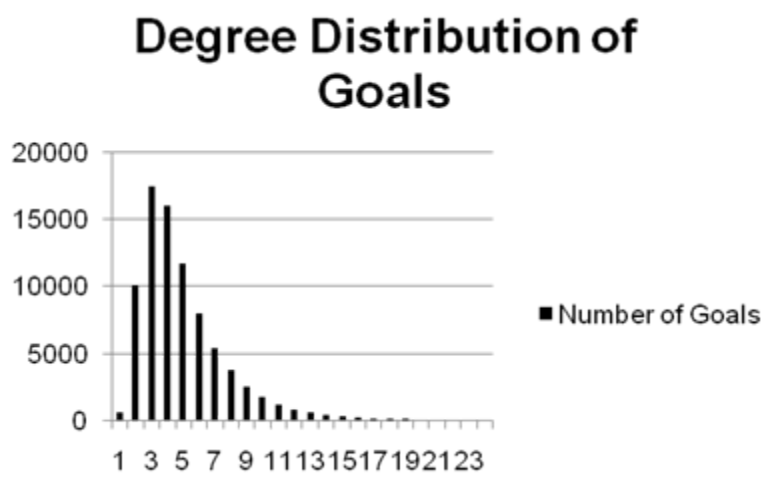
# Quantitative Evaluation / 3

- assignTags(3,2,true)

- # goals in the graph: 80993
- # tags in the graph: 41138
- # edges in the graph: 403705



# Qualitative Evaluation

- Goal Prediction Interface
  - Given a query, get the most probable goals which are associated to it.

- Implementation of a Web Service
  - Services like:
    - Get goals which are associated to a given tag
      - http://localhost/goals/connectedTo/exampleTag?parameter=value
    - Get goals which are associated to a given query
      - http://localhost/goals/associatedTo/exampleQuery
    - Etc.

# Demo of the Goal Prediction Application

Demo of the Goal Prediction Application

# Outlook

- Desiderata for “high quality” goal graphs
- Web service for accessing the generated graphs
- Empirical study to evaluate the quality of the generated graphs

Thank you for your attention!

# References / 1


- [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
- [Dai2006] Detecting online commercial intention (OCI), H.K. Dai et. al. Proceedings of the 15th international conference on World Wide Web, 2006
- [Pass2006] G.Pass et. al.: A Picture of Search - Proc. 1st Intl. Conf. on Scalable Information Systems, 2006
- [Jansen2006] Jansen et al, How to Define Searching Sessions on Web Search, WEBKDD’06, August 20, 2006
- [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
- [Strohmaier2008] M. Strohmaier, P. Prettenhofer, M. Lux.: Different Degrees of Explicitness in Intentional Artifacts - Studying User Goals in a Large Search Query Log-- Proceedings of the CSKGOI'08 Workshop on Commonsense Knowledge and Goal Oriented Interfaces - 2008

# References / 2


- [Rijsbergen1997] Van Rijsbergen, C.: Information Retrieval, 2nd edition, Dept. of Computer Science, University of Glasgow, 1997
- [Wasserman1994] Wasserman S., Faust K., Social Network Analysis, Cambridge University Press, 1994

# Implementation of a Goal Prediction Application / 1

Single words *s* of the user search query *S* are mapped on $t \in T_g$.

Each *t* is connected to a number of goals $g \in G$ with a weight of *w.*

3 weights which influence the results:

- Intersection Weight
  - A goal *g* is more likely to be predicted to the more tags *t* it is connected .

$iw_g = |\text{connected tags}| / |\text{tags}|$ $iw_g = |T_s| / |S|$

# Implementation of a Goal Prediction Application / 2

- Graph Weight
  - A goal g is more likely to be predicted if the weight of the edge is higher than other goals.

$gw_g = \sum$ all weights of the edges to the tags / |tags| * MAX_WEIGHT

$gw_g = \sum w_{ts} / (|S| * \max(w))$

- Length Weight:
  - A goal is more likely to be predicted the more words of the original user query it contains.

$iw_g$ = |number of tags in goal| / |number of words in goal|

$iw_g = |T_s| / |S|$