Web Science and Web Technology
„User Intentions and Intentional Structures on the Web“

Markus Strohmaier

Univ. Ass. / Assistant Professor
Knowledge Management Institute
Graz University of Technology, Austria

e-mail: markus.strohmaier@tugraz.at
web: http://www.kmi.tugraz.at/staff/markus
Overview

Agenda

- Different degrees of explicitness in intentional artifacts - Studying user goals in a large search query log.pdf
Selected Research Questions

1. How can we identify goals in search query logs?
2. How can we represent search goals in a semi-formal goal graph?
3. How can we represent the search process as a traverse through the goal graph?
4. How can goals direct search behaviour?
5. Is there a difference between explicit and implicit intentional search queries, and how can it be identified?
6. How can we construct large scale goal graphs?

Result of joint work with students
P. Prettenhofer, A. UsSaed, C. Körner, M. Kröll
“The aggregate results of every search ever entered, every result list ever tendered, and every path taken as a result. [...] This information represents [...] a place holder for the intentions of humankind - a massive database of desires, needs, wants, and likes that can be discovered, subpoenaed, archived, tracked, and exploited to all sorts of ends.


BUT: The degree to which the goals and intentions of users can be reconstructed varies considerably

„Car“ ➔ Lower degree of intentional explicitness

„get loan to buy a used car in Miami“ ➔ Higher degree of intentional explicitness

An intentional artifact is an electronic artifact produced by users or user behaviour that contains recognizable “traces of intent”, i.e. traces of users’ goals and intentions.
Research Overview

Research Questions:

- How can we distinguish and identify different degrees of intentional explicitness in search queries?

Following an exploratory, qualitative research style, we aimed to:

- Inspect a large search query log
- Analyze the way goals are expressed in search queries
- Identify possible distinction criteria
- Study differences between implicit and explicit intentional queries

Expected results:

- Increased empirically-coupled knowledge about user goals in search queries
An Exploratory Study

Data Sources:

The **AOL Search Dataset** [Pass 2006]

- ~20 mio search queries
- 657,426 unique user IDs
- \{UserID, query, timestamp, (ItemRank, URL)\}*
- Collected between March 1, 2006 and May 31 2006 by AOL
Explicit vs. Implicit Intentional Queries

Explicit intentional queries:
• Related to a specific goal in a recognizable, unambiguous way. Recognizable refers to what [Kirsh 1990] defines as “trivial to identify” by a subject within a given attention span.
• On a more practical level, this idealized definition is related to “better queries”, or queries that have “more precise goals” (R. Baeza-Yates at the “Future of Web Search” workshop 2006, Barcelona).

Implicit intentional queries:
• A query, where it is difficult or extremely hard to elicit some specific goal from the intentional artifact.
• Examples include blank queries, or queries such as “car” or “travel”

Why is this distinction important?
• Disambiguation
• switching between explicit and implicit intentional queries
• Narrowing the cognitive gap
Different degrees of explicitness in search queries

- Search queries exhibit considerable variety with respect to degree of explicitness

  car, car Miami, car Miami dealer, buy a car in Miami, buy a used car in Miami, get loan to buy a used car in Miami

- Explicit vs. Implicit intentional queries

  car, car Miami, car Miami dealer, buy a car in Miami, buy a used car in Miami, get loan to buy a used car in Miami

- Arbitrary selection criteria:
  A query is an explicit intentional query if
  - The query contains at least one verb
  - The query corresponds to our definition of goals
    
    Definition Goal: “a condition or state of affairs in the world that some agent would like to achieve or avoid”
Results of Human Subject Study(1)

4 independent raters labeled 3000 queries

- “bug killing devices”
- “mothers working from home”
- “how to lose weight”

<table>
<thead>
<tr>
<th>Pair</th>
<th>( \kappa )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A - B</td>
<td>0.86</td>
</tr>
<tr>
<td>A - C</td>
<td>0.87</td>
</tr>
<tr>
<td>A - D</td>
<td>0.88</td>
</tr>
<tr>
<td>B - C</td>
<td>0.83</td>
</tr>
<tr>
<td>B - D</td>
<td>0.84</td>
</tr>
<tr>
<td>C - D</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Queries Not Containing Goals → Queries Containing Goals

Not containing goals
Controversial queries
Containing goals
Different degrees of intentional explicitness – Studying user goals in a large search query log

**Approach: Binary classification of search queries**

**Input:**
- Search Queries
  - Car, car Miami, car Miami dealer, buy a car in Miami,

**Classifier**

**Output:**
- Explicit Intentional Queries
  - buy a car in Miami
- Implicit Intentional Queries
  - Car, car Miami, car Miami dealer
An Experimental Classification Approach 1/2

Eliminating noise in the query log

<table>
<thead>
<tr>
<th>Pre-processing step</th>
<th># of queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of queries</td>
<td>21,011,038</td>
</tr>
<tr>
<td>Empty queries</td>
<td>20,527,902</td>
</tr>
<tr>
<td>Short queries</td>
<td>7,242,610</td>
</tr>
<tr>
<td>URL queries</td>
<td>6,631,084</td>
</tr>
<tr>
<td>Syntax check</td>
<td>5,880,900</td>
</tr>
<tr>
<td>Queries containing lyrics or movie titles</td>
<td>5,754,994</td>
</tr>
<tr>
<td>Corrected misspellings</td>
<td>5,405,547</td>
</tr>
<tr>
<td>Verb filter</td>
<td>1,002,861</td>
</tr>
</tbody>
</table>

“Buying a car” -> “buying/VBG a/DT car/NN”

Part-of-Speech tagging queries:

- Only queries with query length > 2 (removes ~ 60% of queries)
- Stochastic approach
- Maximum Entropy tagger
- Tag set: Penn Treebank with 45 word classes

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>Noun, sing. or mass</td>
<td>car</td>
</tr>
<tr>
<td>VB</td>
<td>Verb, base form</td>
<td>eat</td>
</tr>
<tr>
<td>VBG</td>
<td>Verb, gerund</td>
<td>eating</td>
</tr>
<tr>
<td>VBP</td>
<td>Verb, 3sg pres</td>
<td>eats</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
<td>yellow</td>
</tr>
<tr>
<td>WRB</td>
<td>Wh-adverb</td>
<td>how, where</td>
</tr>
<tr>
<td>TO</td>
<td>“to”</td>
<td>to</td>
</tr>
</tbody>
</table>

Table 1. A sample of Penn Treebank tags (from [14])
Supervised Learning of Goal Features:
Finding $f(x) = y$: in this case, a function from $x = \text{queries}$ to $y = \text{classes}$ [explicit/implicit intentional queries]

- Part-of-speech trigrams (instead of word n-grams) to transform queries into a feature vector space
  "buying/VBG a/DT car/NN"

- Set of stemmed words -stopwords
  [buy, car]
Classification Approach(2)

Linear Support Vector Machine [Dumais98]
- Robust and effective in the area of text classification
- Weka Machine Learning Toolkit [Witten05]
- No feature selection

Performance:
- 10 trails – 3 fold Cross Validation
- Values averaged
- Precision, Recall and F1-Measure for the class: “queries containing goals”

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1 – Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.77</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>Nr.</td>
<td>POS</td>
<td>SOW</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>5</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>7</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>9</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>19</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>
Result Set

Applying the learnt classifier results in:
- Result set containing 118,420 entries
- 97,454 (82.3%) of them are unique

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Goal Instance</th>
<th>#Users</th>
<th>Nr.</th>
<th>Goal Instance</th>
<th>#Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>add screen name</td>
<td>194</td>
<td>11</td>
<td>cancel aol account</td>
<td>46</td>
</tr>
<tr>
<td>2</td>
<td>create screen name</td>
<td>98</td>
<td>12</td>
<td>check my computer</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>rent to own</td>
<td>85</td>
<td>13</td>
<td>skating with celebrities</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>listen to music</td>
<td>78</td>
<td>14</td>
<td>discover credit card</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>pimp my ride</td>
<td>64</td>
<td>15</td>
<td>pimp my myspace</td>
<td>34</td>
</tr>
<tr>
<td>6</td>
<td>pimp my space</td>
<td>61</td>
<td>16</td>
<td>change my password</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>assist to sell</td>
<td>57</td>
<td>17</td>
<td>how to gain weight</td>
<td>32</td>
</tr>
<tr>
<td>8</td>
<td>wedding cake toppers</td>
<td>53</td>
<td>18</td>
<td>enterprize car rental</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>cancel aol service</td>
<td>50</td>
<td>19</td>
<td>manage my account</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>“deleted”</td>
<td>47</td>
<td>20</td>
<td>trick my truck</td>
<td>30</td>
</tr>
</tbody>
</table>
Result Set

Applying the learnt classifier results in:

- Result set containing 118,420 entries
- 97,454 (82.3%) of them are unique

Co-Occurences of most frequent Verbs & Nouns:

<table>
<thead>
<tr>
<th></th>
<th>make (8763)</th>
<th>buy (8557)</th>
<th>find (8545)</th>
<th>get (6562)</th>
<th>do (6391)</th>
<th>listen (2485)</th>
<th>learn (2014)</th>
<th>sell (1962)</th>
<th>use (1688)</th>
<th>play (1598)</th>
</tr>
</thead>
<tbody>
<tr>
<td>home</td>
<td>210</td>
<td>237</td>
<td>169</td>
<td>65</td>
<td>70</td>
<td>18</td>
<td>12</td>
<td>141</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>card</td>
<td>208</td>
<td>117</td>
<td>25</td>
<td>103</td>
<td>62</td>
<td>0</td>
<td>16</td>
<td>38</td>
<td>22</td>
<td>63</td>
</tr>
<tr>
<td>name</td>
<td>96</td>
<td>12</td>
<td>192</td>
<td>41</td>
<td>72</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>screen</td>
<td>96</td>
<td>10</td>
<td>30</td>
<td>26</td>
<td>69</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>credit</td>
<td>5</td>
<td>66</td>
<td>20</td>
<td>130</td>
<td>40</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>music</td>
<td>58</td>
<td>58</td>
<td>57</td>
<td>33</td>
<td>51</td>
<td>477</td>
<td>34</td>
<td>8</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>money</td>
<td>631</td>
<td>43</td>
<td>60</td>
<td>68</td>
<td>52</td>
<td>0</td>
<td>10</td>
<td>15</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>weight</td>
<td>19</td>
<td>6</td>
<td>17</td>
<td>13</td>
<td>52</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>school</td>
<td>19</td>
<td>17</td>
<td>104</td>
<td>55</td>
<td>44</td>
<td>27</td>
<td>28</td>
<td>1</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>car</td>
<td>32</td>
<td>224</td>
<td>94</td>
<td>54</td>
<td>25</td>
<td>2</td>
<td>5</td>
<td>90</td>
<td>15</td>
<td>4</td>
</tr>
</tbody>
</table>
Result Set

Rank / Frequency of Users / Goal
Results II

What kind of queries did our experimental approach classify correctly / incorrectly?

Queries in the Condensed Dataset

True positives

Correctly Classified Intentional Queries

“buying groceries online”

“how to get revenge on neighbor within limits of law”

“helping children handle death of a loved one”

“cleaning the ak-47”

“coughing up blood”

“dealing with the guilt of cheating”

False positives

Incorrectly Classified Intentional Queries

“saving privat ryan”

“driving school Illinois”

“stem cell transplant”

“founding fathers temple”

“recovering the satellites lyrics”

Table 5. Examples of incorrectly classified queries

Table 4. Examples of correctly classified queries
Results III

Do more explicit intentional queries yield different click-through result sets?

Top 16 websites in the entire dataset
low proportion of explicit intentional queries
(1.69-3.01%)

Examples: 43things.com (from #388 up to rank #15), ehow.com (from #64 up to #2), hgtv.com (from #97 up to #7), and medhelp.org (from #104 up to #16).

Top 16 websites in the condensed dataset
high proportion of explicit intentional queries
(49.6-58.4%)

Experimental Classification
Observations:

- AOL goals seem to deal with health related issues:
- 43Things goals appear to exhibit a more positive sentiment:
Observations:

- Users seem to have different time frames in mind - AOL Users often appear to seek immediate answers:

- 43Things Users do not seem to underlie these time constraints:
### Verbs in AOL vs. 43Things

Top N most frequent verbs of both goal corpora

**Observations:**
- AOL verbs seem to deal with technical issues
- 43Things contains more verbs reflecting social activity

<table>
<thead>
<tr>
<th>#Verbs</th>
<th>AOL</th>
<th>Overlap</th>
<th>43Things</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>buy, listen, sell, use, play</td>
<td>make, find, get, do, learn</td>
<td>be, go, have, read, see</td>
</tr>
<tr>
<td>50</td>
<td>listen, change, look, move, add, remove, clean, install, apply, draw, put, are, set, convert, rent, tell, fix, pimp, wed, check, cook, deleted</td>
<td>get, be, learn, go, make, have, do, read, see, find, buy, take, write, start, stop, eat, want, keep, create, build, play, use, lose, is, grow, know, sell</td>
<td>s, become, meet, finish, live, watch, run, give, spend, try, own, improve, love, organize, save, speak, join, visit, attend, ride, let, work, am</td>
</tr>
</tbody>
</table>
Next Steps

In the medium term, …

• We intend to develop methods and techniques that support
  – Intentional Classification of Artifacts and
  – Extraction of Goals from Intentional Artifacts and Social Corpora

• to construct large-scale
  – Goal Association Graphs (X is associated with Y) and
  – Intentional Networks (different intentional relations and nodes, see image to the right)

• that enable novel applications, such as
  – Intentional Query Expansion
  – Intentional Metadata
  – Intentional Analysis

<table>
<thead>
<tr>
<th></th>
<th>Entire Dataset</th>
<th>Condensed Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Queries</td>
<td>20,464,002</td>
<td>279,260</td>
</tr>
<tr>
<td>Explicit Intentional</td>
<td>346,349</td>
<td>123,612</td>
</tr>
<tr>
<td>Queries</td>
<td>616,869</td>
<td>163,689</td>
</tr>
<tr>
<td>Implicit Intentional</td>
<td>19,877,133</td>
<td>116,172</td>
</tr>
<tr>
<td>Queries</td>
<td>20,475,552</td>
<td>140,747</td>
</tr>
<tr>
<td>Explicit Intentional</td>
<td>1.66% ± 1.01%</td>
<td>49.6% ± 58.4%</td>
</tr>
<tr>
<td>Queries, 95%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>confidence interval</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Users</td>
<td>657,426</td>
<td>94,487</td>
</tr>
</tbody>
</table>

Table 3. Statistical overview of the condensed dataset
### Idea I

- We have two unequally distributed types of queries
  - Explicit and implicit intentional queries
    - Distinction: Explicit intentional queries contain verbs and thereby allow to more easily grasp the intentionality behind a query than implicit intentional queries
    - Example: „car shop“ vs. „buy car“, „sell car“, „repair car“

<table>
<thead>
<tr>
<th>Implicit Intentional Queries</th>
<th>Explicit Intentional Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>implicit intentional queries</td>
<td>Explicit intentional queries</td>
</tr>
<tr>
<td>how to get rich breast milk [1]</td>
<td>how to have good breast milk [1]</td>
</tr>
<tr>
<td>breast milk [1]</td>
<td>how to have good breast milk [1]</td>
</tr>
<tr>
<td>target [1]</td>
<td>breast feeding and going back to work [1]</td>
</tr>
<tr>
<td>walmart [1]</td>
<td>breast feeding and going back to work [1]</td>
</tr>
<tr>
<td>yellow breast milk [1]</td>
<td>breast feeding and going back to work [1]</td>
</tr>
<tr>
<td>yellow breast milk [1]</td>
<td>breast feeding and going back to work [1]</td>
</tr>
<tr>
<td>yellow breast milk [1]</td>
<td>breast feeding and going back to work [1]</td>
</tr>
<tr>
<td>breast feeding and going back to work [1]</td>
<td>breast feeding and going back to work [1]</td>
</tr>
<tr>
<td>breast feeding and going back to work [1]</td>
<td>breast feeding and going back to work [1]</td>
</tr>
</tbody>
</table>

How can we automatically construct a network model based on this information?
Idea II

• Observation:
  – Users rarely express their goals via explicit intentional queries (1-4%)
  – Implicit intentional queries make the dominant part of queries (96-99%)

• Assumption:
  – Whenever a user issues an explicit intentional query, it is “surrounded“ by implicit intentional queries (i.e. queries issued before and after a query of interest within a search session), reflecting the users‘ process of iterative search and query refinement
  – This assumption only holds for informational queries
  – We believe explicit intentional queries represent informational queries

• Intuition:
  – Let's treat the implicit intentional queries in an explicit intentional query’s environment as tags
  – An explicit intentional query’s environment is represented by the queries a user has issued within a certain time frame or set of queries before and after issuing an explicit intentional query
Idea III

Approach: Treat the set of all queries \( \{q_{-n} \ldots q_i \ldots q_n\} (n=0) \) within the \( n \)th environment of the explicit intentional query \( q_i \) as tags for \( q_i \)

With \( n = 6 \), this approach results in tagging “how to have good breast milk” with the following tags (excerpt):

[Breast milk], [Yellow breast milk], [Breast feeding and going back to work], [Nestle formula], [Free nestle formula], [Good start], [What fenugreek]

http://www.verybestbaby.com
http://www.breastfeeding.com
Modes of Tagging

General Alternatives
1. Only using queries that occur after the explicit intentional query ($q_n$ with $n > 0$)
2. Only using queries that occur before the explicit intentional query ($q_n$ with $n < 0$)
3. Using queries that occur before and after the explicit intentional query ($q_n$ with $n > 0$ and $n < 0$)

These different modes of tagging have different implications for intentional query expansion:
If Mode 2 or 3 is meaningful, then suggesting adequate explicit intentional queries based on a set previously entered implicit intentional queries might be feasible.
Crystalization of Explicit Intentional Queries - The Iceberg Metaphor

1-3%
Explicit intentional queries

97-99%
Implicit intentional queries

- Breast milk
- Free nestle formula
- Nestle formula
- What fenugreek
- Breast feeding and going back to work
- Yellow breast milk
- Good start
- How to have good breast milk
Network Analysis

• Analyzing the tripartite graph of Search
  – Consisting of users, explicit intentional queries and tags

Based on this conceptualization, the following two-mode networks can be folded into one mode networks:

- Intentional Queries – Tags
- Users – Intentional Queries
- Users - Tags
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 of goals only.
### Terminology / 0

<table>
<thead>
<tr>
<th>id</th>
<th>query</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>fluffy roofs house designs</td>
<td>2006-05-27 13:37:19</td>
</tr>
<tr>
<td>3</td>
<td>english cottage house plans</td>
<td>2006-05-27 13:45:14</td>
</tr>
<tr>
<td>4</td>
<td>old world english cottage house plans</td>
<td>2006-05-27 14:02:02</td>
</tr>
<tr>
<td>5</td>
<td>build an english cottage</td>
<td>2006-05-27 14:09:58</td>
</tr>
<tr>
<td>6</td>
<td>english cottages</td>
<td>2006-05-27 14:15:23</td>
</tr>
<tr>
<td>7</td>
<td>domain furniture</td>
<td>2006-05-27 20:56:23</td>
</tr>
<tr>
<td>8</td>
<td>floral design clock and ethan allen</td>
<td>2006-05-27 21:08:38</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>2006-05-28 12:33:51</td>
</tr>
</tbody>
</table>

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

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 that is very sparse
  no relations between goals of different users

- $d = 3$ in this example
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
Approach C

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

• $T = R_m$

• $T_{\text{build an english cottage}} = \{\text{house, plans, old, world}\}$
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 \(\rightarrow T = R_m \subseteq W(g)\)

• \(T_{\text{build an english cottage}} =\)
  \{build, cottage, english, house, plans, old, world\}

• Good approach for now
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?
• Sub graph of result of approach A
Observations / 2

- Sub graph of result of approach E
Intentional Metadata: Research Opportunities

- Intentional XHTML
  - `<a href="url">label</a>` link relationships with `rel="..."` 'rel' is extensible and multivalued `rel="relationship1 relationship2"

- Intentional Microformats

- Intentional (Search?)Browser-Plugins

- Intentional Tags / Geotags

- Intentional Webservice Descriptions
  - e.g. WSMO including
    - `wg` mediators (web service – goal) + `gg` mediators (goal -goal)

- Automatic Intentional Metadata Creation (Corpus-based)

- Intentional Weblog Tags

- SPECIFIC EXAMPLES
Administrative Issues

Final Exam in two weeks

– little repetition of previous home assignments, + New Questions
– Understanding of basic network constructs (e.g. centrality, weak ties, formal models of small world networks, etc)
– Scope: All lecture slides and mandatory readings
– You need to register for the exam with TUG online
– Date: 23.6. 12:00, HS i12
– Results: within 4 weeks
– „Einsichtnahme“: Fr, 25.7. 2008, 10:00 – 11:00

Next week: Guest lecture

Web Technologies II, Peter Scheir, KMI

The semantic web represents a current research effort to increase the capability of machines to make sense of content on the web. In this class, Peter Scheir will give a guest lecture on the basic principles underlying the semantic web vision, including RDF, OWL and other standards.
Any questions?

See you next week!