Automatic Subject Classification and Topic Specific Search Engines -- Research at KnowLib

Anders Ardö and Koraljka Golub
DELOS Workshop, Lund, 23 June 2004

KnowLib: Knowledge Discovery and Digital Library Research Group

Goals
- information systems
- digital library services
- knowledge discovery
- distributed knowledge organization technologies
  - usability of knowledge organization systems (thesauri, classifications, subject headings systems, ontologies...)
  - user interfaces

KnowLib Members
- Anders Ardö, Associate Professor
  Department of Information Technology, Lund University
- Koraljka Golub, PhD Student
  Department of Information Technology, Lund University
- Traugott Koch, Digital Library Scientist
  Knowledge Technologies Group, NetLab, Lund University Libraries
- Michael Ovnell, Chief Scientist
  Biblioteksjänst AB

KnowLib Projects: Log Analysis -- Renardus
- overall purpose: improve Renardus
- browsing and searching behaviour of users
- why log analysis?
  - catch unsupervised usage
  - evaluate the potential of thorough log analysis
    - own software developed

Renardus Home Page: www.renardus.org

Goals
- detailed usage patterns
- balance between browsing and searching and mixed activities
- hierarchical classification browsing behavior
  - usage degree of browsing support features
Main Navigation Features

• simple search
• advanced search
• subject browsing: DDC
  • intellectual mapping of classification systems used by the distributed subject gateways

Subject Browsing Support Features

• graphical fish-eye presentation of the classification hierarchy (Graph. Browse)
  – text version (Text Browse)
• search entry into the browsing structure (Search Browse)
• merging of results from individual subject gateways (Merge Browse)

First Step: Preparing the Log Files

• appx. 2 300 000 entries boiled down to 630 000 entries (appx. 165 000 sessions)

<table>
<thead>
<tr>
<th>Entries removed</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 107 378</td>
<td>images or style sheets</td>
</tr>
<tr>
<td>516 269</td>
<td>robots</td>
</tr>
<tr>
<td>17 586</td>
<td>HTTP code 301 (redirections)</td>
</tr>
<tr>
<td>12 647</td>
<td>malicious attacks</td>
</tr>
<tr>
<td>9 000</td>
<td>local IP-numbers</td>
</tr>
<tr>
<td>4 690</td>
<td>MS favicon.ico</td>
</tr>
<tr>
<td>408</td>
<td>HTTP code 408 and other errors</td>
</tr>
</tbody>
</table>

Dominance of Browsing Activities

• more than 80% of sessions are dominated by browsing
• among users starting at home page (21%), still 57% browse and only 12.5% search
• possible reasons:
  – indexing of browse pages by search engines
  • 71% start using Renardus at browsing pages
  • homepage design strongly “invites” for browsing

Major Conclusions

• clear dominance of browsing activities
• tendency to stay in the same group of activities
• good usage of the browsing support features, esp. graphical fish-eye browsing
• surprisingly low share of search activities needs to be further investigated
• log analysis can provide valuable insights

http://www.it.lth.se/knowlib/renardus-log/log-analysis.html
KnowLib Projects: KLIC-DDL…

- KLIC-DDL: KnowLib’s Intelligent Components of a Distributed Digital Library
  - architecture for a distributed digital library
  - implementation of information services using intelligent components
    - automated subject classification, text categorization
    - semi-intelligent information search agent with Web harvesting
    - subject specific search engines etc.

  [http://www.it.lth.se/knowlib/klic.htm]

KLIC-DDL: Automated Subject Classification…

- full-text Web-based documents
- established controlled vocabularies – browsing: DDC, FAST, Ei
- home-produced vocabularies: Materials Science, Carnivorous Plants
- machine learning: text categorization (TC)
- information retrieval: document clustering

[http://www.it.lth.se/knowlib/]

…KLIC-DDL: Automated Subject Classification

- explore heuristics
  - e.g. importance of metadata vs. title vs. anchor text
- compare results of ”All Engineering” with a TC algorithm
- compare browsing controlled vocabulary versus automatically clustered vocabulary
  - advantages and disadvantages of each approach
- explore SOMs as a browsing interface

[http://www.it.lth.se/knowlib/]

KLIC-DDL: Demonstrators

- Automatic subject classification of Web pages
- Multi-search demonstrator
  - the system analyses the query and dynamically generates indications based on which the user can modify his/her query
- Subject browsing of a harvest database
- Materials.dk [http://materials.dk/]

[http://www.it.lth.se/knowlib/demos.htm]