Multi-Resolution Topic Maps for Information Navigation

Azadeh Shakery
Multi-Resolution Topic Map

High Level Concepts
Community Finding Algorithm: Previous Work

- Clique Percolation Method (CPM)

[k-clique communities of unweighted undirected graphs]

- Not too restrictive (compared to cliques)
- Based on the density of links
- Does not yield cut-nodes or cut-links (whose removal would disjoin the community)
- Allows overlaps
Problems with the CPM

- A couple of very big clusters, a lot of singletons
- Clusters of lower resolutions have more words (somehow counter intuitive)
Community Finding Algorithm: New Idea

- Define the resolution based on the number of documents covered by each topic
- Topics at the same resolution cover approximately the same number of documents
- Topics at higher resolutions cover more documents
Topic Finding

- Extract representative words from each document
- Use frequent item-set mining algorithms to find frequent word sets
- Define the topics based on these frequent word sets
Topic Finding: Specific Method

- Use words in the titles which are neither very frequent nor very rare.
- Use RPMine implemented in *illimine* to find frequent item-sets

[Dong Xin, Jiawei Han, Xifeng Yan, Hong Cheng, Mining Compressed Frequent-Pattern Set, Proceedings of 2005 Int. Conf. on Very Large Data Bases (VLDB 2005)]
Topic Summarization

- Use the words for summarization
- Sort the words based on the placement of the words in the titles

\[ w(w_1, w_2) = \sum_{\text{titles containing } w_1, w_2} \frac{1}{\text{dist}(w_1, w_2)} \]

- Sort the words based on these weights
Topic Summarization - Example

- Information language retrieval cross
  - (e) inform language: -16.6 -----> 3
  - (f) inform retrieval: 110.2 --------> 1
  - (g) inform cross: -7.33333 -------> 5
  - (h) language retrieval: 12.9595 ---> 4
  - (i) language cross: -20 ------------> 2
  - (j) retrieval cross: -6.25 ------------> 6
Demo