Database Issues for the Web
- Dasfaa 2003 Panel -

Dik Lun Lee
Department of Computer Science
Hong Kong University of Science and Technology

Characteristics of the Web

- Publication of information by:
  - Businesses, large and small
  - Individuals

- No global rules or structures: structures and rules are spontaneously created by web authors
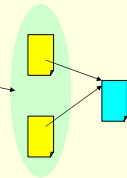
- Result: we can only discover, not control, the web

Why is Google so Successful

- Exploit collective opinions of web authors
  - Use link analysis to identify web pages of high authority and/or quality
  - Use anchor texts as an indicator for web page theme
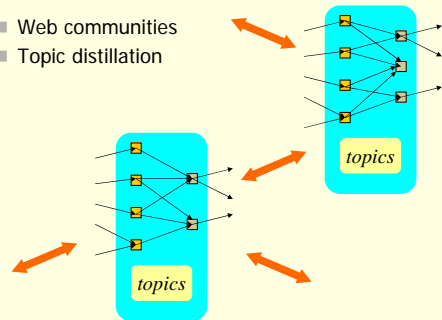- Both are created by human authors
- Large scale ensures robustness

Human-created index/knowledge

Sub-consciously created Works only under large scale

Beyond Google

- Web communities
- Topic distillation

topics

topics

Research Issues: mining meaning

- Derivation of concise and precise web communities
  - Link and content analysis
  - Integration of multiple sources of "intelligence": Click-stream, search engine log, human-created directories: Open Directory and ontological databases
- Creation of multiple hierarchies between communities
- Topic drifting: restructuring of communities and their corresponding topics; migration of "drifted" pages to another community
- Implementation: centralized versus peer-to-peer

Research Issues: Mobile Web

- Modeling of physical space to support, e.g., semantic nearest neighbors (time and space)
  - I am at the Kansei airport now, lead me to the platform for the Kyoto-bound train
  - Fine-grain, step by step instructions
  - Continuous monitoring to ensure I am at the right place at the right time
- Processing nearest-neighbor, k-nn, window and continuous queries
  - Indexing of locations of data objects
  - Specialized index for NN and k-NN queries
  - Caching: validation of answers as user moves along
  - Connection: point-to-point, broadcast, continuous

Research Issues: Wireless Web

- Wireless mesh networks: nodes are connected wirelessly; peer-to-peer network with limited reach
  - Applications: sensor networks, low-cost substitution of 3G networks
  - Dissemination of large volume of data
    - How to announce availability of new data or updates?
    - Data broadcast; message hopping across the network
    - Filter on-the-air
  - Finding information on the network
    - Limited bandwidth and batter power: continuous active querying is discouraged

Summary

Personalization, query tracking

Ubiquitous information access

Clustering, term suggestion, query tracking

Mobile/wireless web

google, link analysis

Ubiquitous information publication