Data Mining and the Web

Susan Dumais
Microsoft Research

KDD’97 Panel - Aug 17, 1997

The Web as a Text Database

- BIG and doubling every year
  - 70 million observations (urls)
  - 50 million variables (words)
  - very sparse
- BAD and UGLY
  - uncontrolled quality, widely distributed, rapidly changing, heterogeneous/complex data types, no consistent semantics or structure within or across objects, etc.

“Data Mining” the Web

- Today:
  - Search and meta-search engines
  - Hand-crafted hierarchies
  - Special-purpose information discovery and extraction algorithms (e.g., home pages, authority pages, interesting pages, fun cities)

Data Mining the Web

- To Come:
  - Inter-document associations uncovered by:
    - Automatic classification
    - Generating fixed or ad hoc structures (e.g., clustering)
    - Exploring similarity neighborhoods (e.g., visualization)
  - Highly interactive interfaces
    - Analysis of interrelations among objects
    - Interest specification/Query formulation problem

What we Need to Get There

- Better Text Mining Tools (for the Web)
  - Robust, scalable methods for *feature selection* - word statistics, learned indexing features, tags
  - Integration w/ databases
  - Web *mining services* (rich API to Web indices)
  - Model/Pattern specification and summarization
    - Content/topical interests
    - Patterns of interest - new, different, central

What we Need to Get There

- Going Beyond Text
  - Metadata
    - Date, size, author, site, time etc.
  - Structure - reflects prior human knowledge
    - Link structure (in-links, out-links)
  - People - individually and collectively
    - Ratings/preferences
    - User models, usage patterns
  - Integration of the above