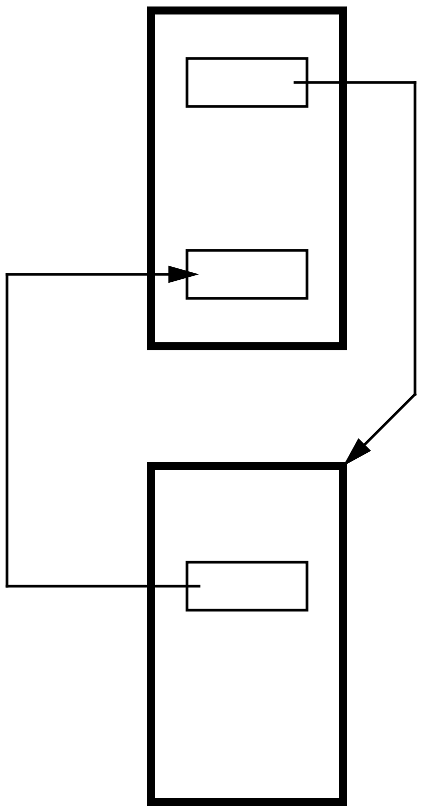
Mining the WWW

- **Main Themes**
  - Organization of Web Content
  - Applications
  - Techniques
- **Web Content Mining vs. Web Usage Mining**
- **Why is Web Mining Difficult?**
  - Heterogeneity of Data
  - Semi-structured (HTML, SGML etc.)
  - Dynamic and Rapidly Evolving

Disadvantages of Existing Hypertext Systems for Data Mining

- **Static Nature of Hyperlinks**
  - Direct Connections from one node to another (Extensional)
- **Stale Links**
  - 404 Not Found!!
- **Inexpressiveness**
  - Lack of Link Semantics
  - Link types set apriori
- **Expensive Construction**
  - Hand Construction
- **Inflexibility**
- **Duplication**
  - Rapid Proliferation of Links for Similar Purposes



What are the proposed strategies?

- ❑ Enhance the Hypertext Structure (Multilevel, Adaptive etc.)
- ❑ Utilize Semistructured DBMSs (Lore, Sch* etc.)
- ❑ Algorithmic Infrastructure: Latent Semantic Indexing

Example Applications

- ❑ AI + Internet Applications - Knowledge Repositories & Agents
- ❑ Adaptive Hypertext
- ❑ Automatic Information Filtering (PICS Labels)
- ❑ Advanced Search Engines (Altavista etc.)
- ❑ Information Clustering
- ❑ Information Discovery (Harvest System)

Multi-Level Hypertext

- **Use of Abstraction Barrier in Software Eng.,**
  - Declarative / Procedural
- **Declarative Models**
  - Focus on Data Structure
    (Semantic Nets, Frames)
- **Procedural Models**
  - Focus on Process
    (WAIS, cgi-bin)
- **Separate Link from Structure**
  - Hand Construction
- **Useful for**
  - Hypertext Linearization
  - Document Reuse
  - User Disorientation
  - Access Control, Corruption, Reorganization

Declarative Models

- ❑ Augments Text with a KR structure
- ❑ Semantic Nets
- ❑ Frames
- ❑ Petri Nets
- ❑ Bayesian Belief Networks

Semantic Nets

- Meaning of a concept comes from the ways in which it is connected to other concepts
- Vertices = Concepts; Edges = Relationships; Graph = Net
- Natural Analogy between Hypertext and Semantic Net
- Independent and Embedded Versions
- Coarse Grained Models
  - Associate Paragraphs with nodes
  - THOTH-II
  - Conceptual Graphs
- Fine Grained Models
  - SNITCH System
  - Use of connectionist schemes

Frames

- Slot and Filler Structures
- More Structure to Nodes as well as links
- Frame Systems advantageous over single frames
- Useful for Consistency Checking
- Frame Representation Languages (FRLs)
  - KRL
  - RLL
  - CYCL
- Inheritance Problems
  - leads to tangled hierarchies

Petri Nets

- **Process Model defined on a Directed Graph**
- **Vertices contain tokens; some are “transitions”**
- **Enabling/Firing Transitions**
- **Repeated Firing until “convergence”**
- **Useful for Access Control Mechanisms**
  - TRELLIS - Browsing Semantics
  - Security Purposes; similar to PICS labels
- **Timing Mechanisms**
  - TRELLIS - Controls Duration of Display
  - Modification of Hypertext with time/person

Bayesian Belief Networks

- **Use of DAGs to represent probabilistic relationships**
- **Can be applied to control conditional probabilities**
- **Used to express causality**
- **Strong use of Prior Probabilities**
- **Applications**
  - Associate Index terms with concepts for nodes
  - Microsoft Troubleshooting System on the Web
  - Hierarchical Organization of Terms in Handbooks
  - Adaptively Update Hypertext - useful in multi-user hypertext

Clustering Models

- ❑ **Augment set of nodes with clustering information**
- ❑ **No predefined formalized representation**
- ❑ **Clusters used as secondary structures**
  - ❑ Nested Context Model
  - ❑ Structure Nodes
  - ❑ Sequencing Nodes
  - ❑ Exploration Nodes
  - ❑ Guided Tours
  - ❑ Automatic Generation of Maps
- ❑ **Clustering Algorithms**
  - ❑ Semantic Compactness Clustering
  - ❑ Agglomerative Clustering
  - ❑ Aggregation Clustering with exceptions

Procedural Models

- **Invoke a process when attempted to traverse a link**
  - Query Based Browsing
  - Link Resolution Models
  - Generalized Link Models
  - Intensional Models
- **Based on research on Information Retrieval**
- **Two Varieties**
  - Compute connection dynamically
  - Compute connection + target dynamically
- **Computed Links**
- **Automated Link Construction**

Application Case Studies

- ❑ PICS (Platform for Internet Content Selection)
- ❑ Harvest Information Discovery System