Building Topic Hierarchy of e-Documents using Text Mining Technology

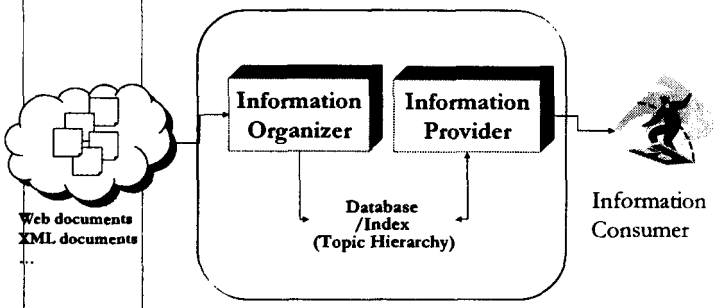
Han-joon Kim

Department of Electrical and Computer Engineering
University of Seoul

khj@uos.ac.kr

1

Information System



2

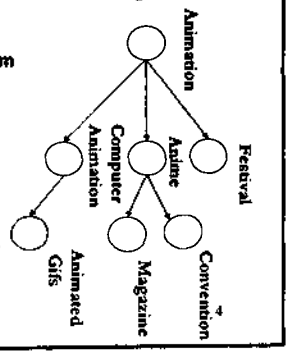
Information Organization

- **Indexing**
  - **Key word-based**
    - **Polysemy/Synonymy problem**
  - **Topic-based**
    - **Topic hierarchy: Yahoo-style**

3

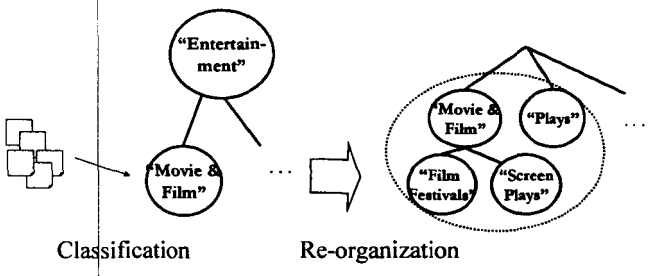
Introduction

- **Topic hierarchy paradigm**
  - **Summary view of a large amount of information**
- **Management of a large document collection**
  - **Topic hierarchy based Information System**
    - **Web search engine: Yahoo!, Lycos, Open Directory**
    - **Digital libraries: Library of Congress Catalog**
  - **Topic-oriented indexing**
  - **Browsing directories**
- **Maintenance of topic hierarchy**
  - **limitation of manual maintenance**



4

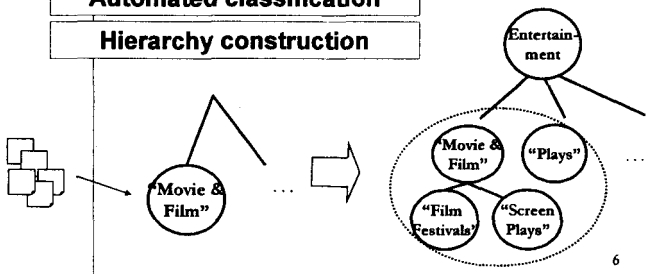
Information Organization: topic-based



5

Information Organization

- **Issues : Text-Mining approach**
  - **Topic discovery (clustering)**
  - **Automated classification**
  - **Hierarchy construction**



6

Clustering (Topic Discovery)

- **Document clustering**
  - **Grouping a set of documents into classes of similar documents**

Document Collection

Clustering

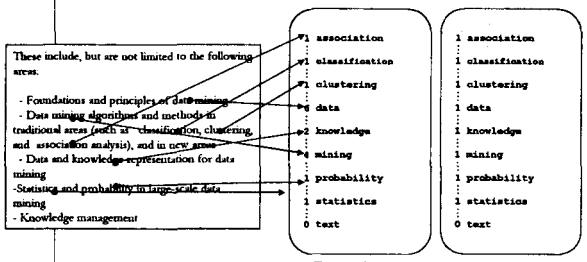
A *B* *C* *D* *E*

Clusters

7

Preliminaries

- **Document representation**



Unstructured document

Bag-of-words representation

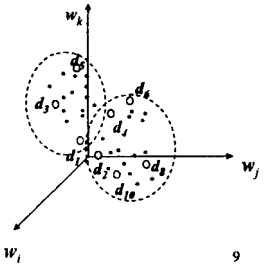
Boolean representation

Vector representation

8

Clustering (Topic Discovery)

- **Conventional clustering methods**
  - **Hierarchical clustering methods**
  - **Depends on the standard Euclidean distance metric**
  - **Not suitable for topic discovery**
- **Goal**
  - **Generate user-specific clusters !**



Supervised Clustering (Topic Discovery)

Document Collection

Clustering

Human Knowledge

A' B' C' D' E'

Clusters ⇨ **Topics (categories)**

Distance Functions

- **Euclidean distance function**

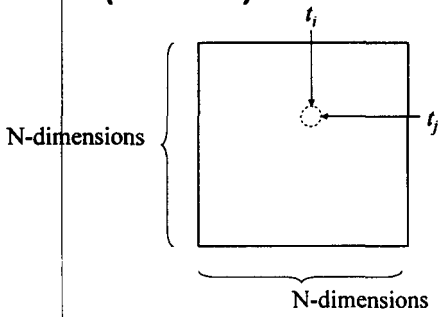
$$dist_{\mathbf{W}}(\vec{d}_x, \vec{d}_y) = \sqrt{(\vec{d}_x - \vec{d}_y)^T \cdot \square \cdot (\vec{d}_x - \vec{d}_y)}$$

- **Quadratic distance function**

$$dist_{\mathbf{W}}(\vec{d}_x, \vec{d}_y) = \sqrt{(\vec{d}_x - \vec{d}_y)^T \, \mathbf{W} \, (\vec{d}_x - \vec{d}_y)}$$

Weight Matrix

- **Representing the inter-correlation of terms (features)**



Quadratic Distance Function

- **Examples**
  - Vocabulay = {'*digital*', '*image*', '*picture*'}
  - "*e-book with digital images*"
    - $d_1 = (1, 1, 0)$
  - "*picture book*"
    - $d_2 = (0, 0, 1)$
  - **Weight Matrix**
  - **Distance values**

$$W_{novelist} = \begin{array}{c} \begin{matrix} d & i & p \end{matrix} \\ \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.8 \\ 0.0 & 0.8 & 1.0 \end{pmatrix} \end{array} \begin{matrix} d \\ i \\ p \end{matrix}$$

$$W_{scientist} = \begin{pmatrix} 1.0 & 0.5 & 0.0 \\ 0.5 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{pmatrix} \begin{matrix} d \\ i \\ p \end{matrix}$$

$$dist_{standard}(d_1, d_2) = \sqrt{3}$$

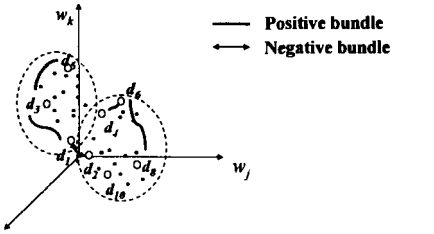
$$dist_{novelist}(d_1, d_2) = \sqrt{1.2}$$

$$dist_{scientist}(d_1, d_2) = \sqrt{3.5}$$

13

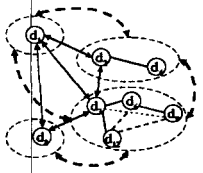
Human Knowledge

- **Document bundles (or bundles)**
  - **Positive/Negative bundles**

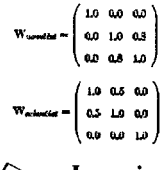


14

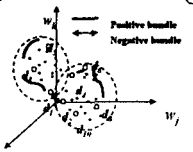
Supervised Clustering: overview



Bundle Knowledge



Learning weight matrix

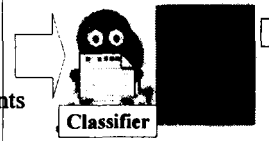


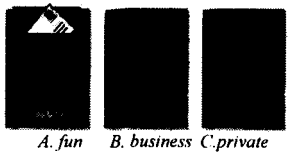
Clustering

15

Automated Classification

Training e-Documents





Unknown e-Documents

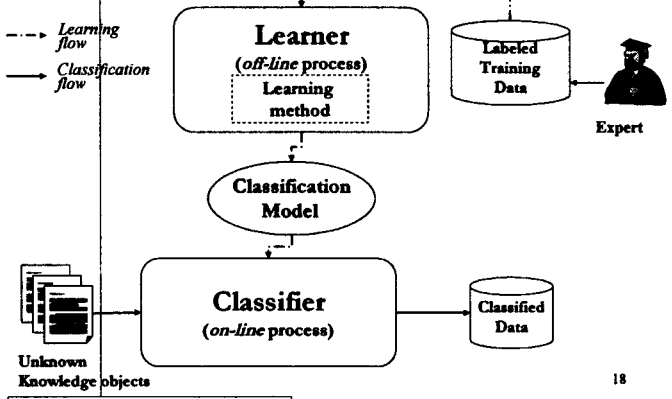
16

Automated Classification

Factors to Consider	Manual	Automated
Cost (per document)	High	Low
Speed	Slow	Fast
Consistency	Variable	High
Quality	Variable	Variable

17

Machine Learning based approach (Basic architecture)



18

Automated Classification

- **Related algorithms**
  - Decision trees
  - Neural networks
  - Support vector machines
  - Bayesian statistics
  - etc

19

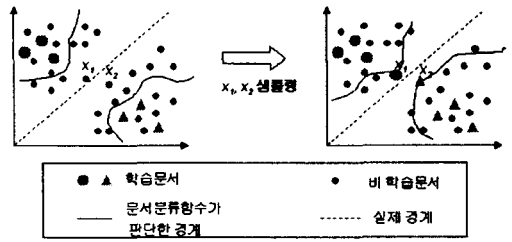
Automated Classification

- **Critical Issue**
  - Difficulty in obtaining good quality training data
  - Related researches
    - Active learning
      - Actively selecting the best training data out of unclassified ones
    - EM algorithms
      - Augment of labeled (training) documents with classifying unlabeled documents

20

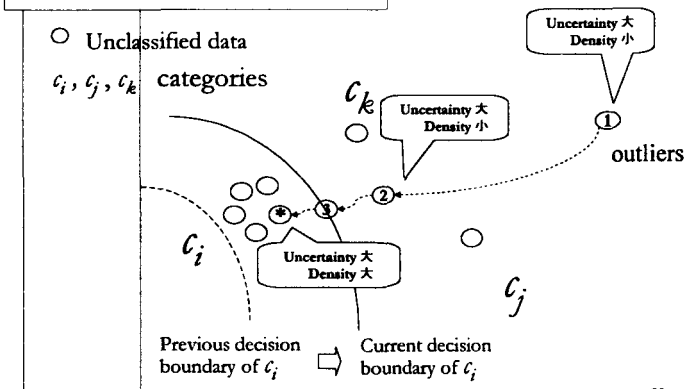
Active Learning

- **Uncertainty-based sampling**



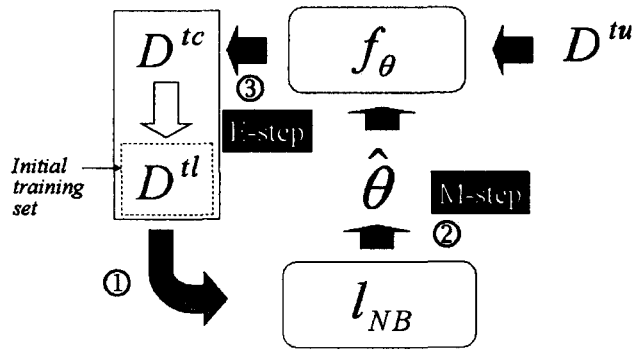
21

Uncertainty/Density-based Selective Sampling



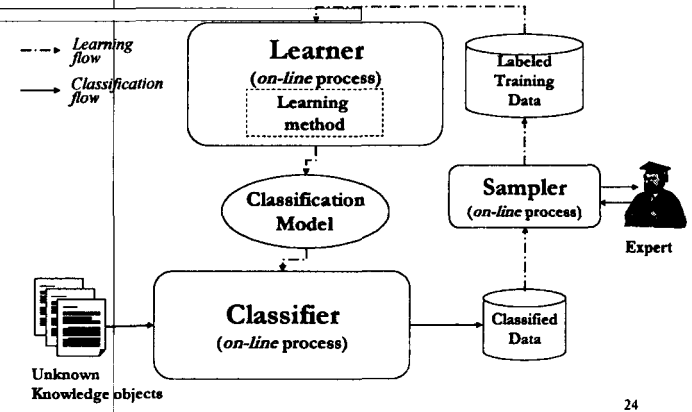
22

EM (Expectation Maximization) algorithm



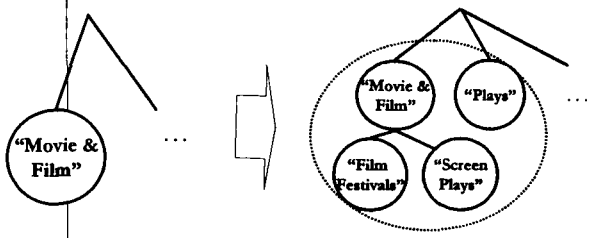
23

Machine Learning based approach (The proposed architecture)



24

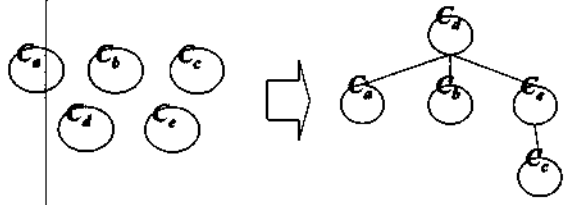
Topic Hierarchy Construction



- **Goal : find category subsumption relations**
  - **Solution: generate fuzzy partial ordering**

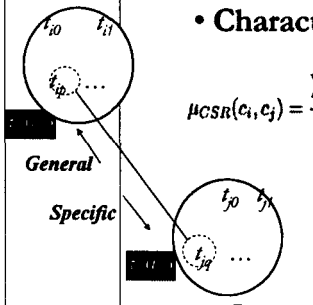
Topic Hierarchy Construction

- **Goal : find category subsumption relation**



Category Subsumption Relation

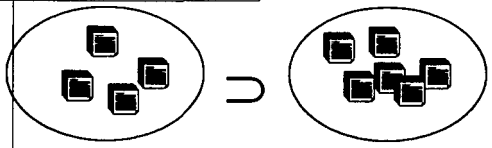
Category A



- **Characteristic function**

$$\mu_{CSR}(c_i, c_j) = \frac{\sum_{t_i \in V_{c_i}, t_j \in V_{c_j}} \tau_{c_i}(t_i) \cdot \tau_{c_j}(t_j) \cdot Pr(D_{t_j} \subset D_{t_i})}{\sum_{t_i \in V_{c_i}, t_j \in V_{c_j}} \tau_{c_i}(t_i) \cdot \tau_{c_j}(t_j)}$$

Term Subsumption Relation

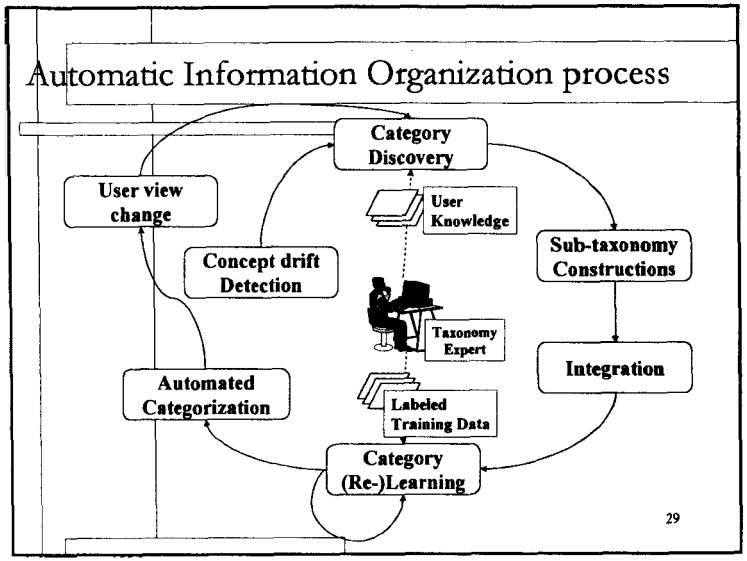


D_{t_i} D_{t_j}

t_i : more general

/

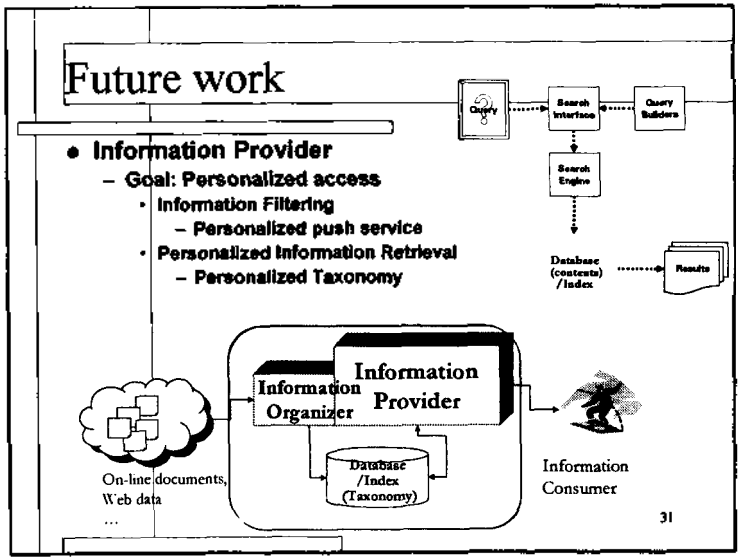
t_j



Summary

- **Text-mining approach to e-documents organization based on topic hierarchy**
  - **Machine-Learning & information Theory-based**
    - ***'Category (topic) discovery' problem***
      - → document bundle-based user-constraint document clustering
    - ***'Automatic categorization' problem***
      - → Accelerated EM with CU-based active learning
    - ***'Hierarchy Construction' problem***
      - → Unsupervised learning of category subsumption relation

30



Building Topic Hierarchy of e-Documents using Text Mining Technology

Thank you

Han-joon Kim
Email: khj@uos.ac.kr
Tel: 02-2210-5632

32