

# Extraction of Topic Maps from Web pages

---

Motohiro Mase  
Tokyo Institute of Technology  
Seiji Yamada  
National Institute of Informatics  
Katsumi Nitta  
Tokyo Institute of Technology

## Introduction(1/2)

---

- Information gathering
  - a huge amount of Web pages
- Web users' problems[GVU98]
  - finding out the necessary information
    - Search engine: Google, Yahoo!
    - Web personalization[Speretta04, Widyantoro99]
    - Web navigation[Armstrong95, Lieberman95]
    - etc



– organizing the gathered information



## Introduction (2/2)

- Organizing the gathered information

- Managing histories by tasks
  - Browsing Icons[Matthias01],
  - Google Search history, Google Web history
- Visualization of histories
  - Domain Tree Browser[Gandhi00], VISVIP[Cugini99]
- Search the information that user have seen
  - Stuff I've Seen[Dumais03]
  
- Indexing by tasks , time-line, terms



- Topic Maps

- Indexing by topics and the relations of the topics

2

## Topic Maps(1/2)

- ISO/IEC 13250

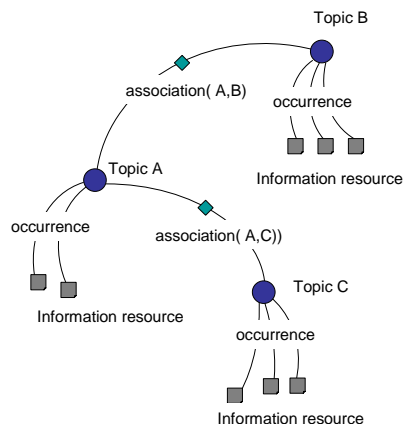
- A solution for organizing, accessing and retrieving information

- Composed of three elements

- topic
  - any concepts and subjects
- association
  - relation between topics
- occurrence
  - connection between topics and information resources related to them

- High flexibility

- user can define topics and associations freely

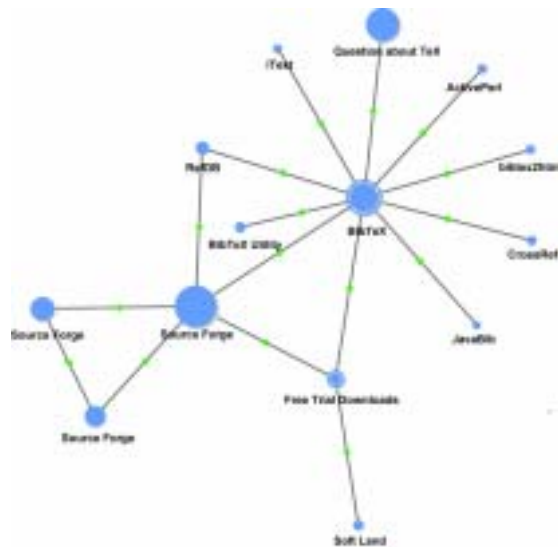


3

## Topic Maps (2/3)

---

### Searching BibTeX tools



4

## Topic Maps(3/3)

---

- How to create Topic Maps from Web pages
  - manually
    - cost
  - Automatically
    - cover the existing ontology[Reynolds02]
    - XML, RDF : structured data
- Web pages
  - HTML : semi-structured data

5

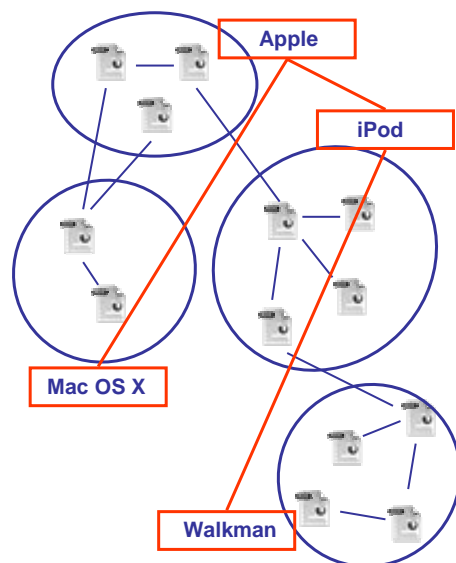
## Purpose

- **Extracting the topic maps from Web pages**
  - assist user to organize and manage the gathered information from Web
- **prototypes of topic maps**
  - Users modifies and completes the prototypes
  - Exhaustively extracted topics and associations
    - Adding lack of topics and associations
    - Removing unnecessary items

6

## Approach(1/2)

- **Topic Maps**
  - Topics
  - Associations
- **To extract topic maps**
  - divide by topics
  - relations between the topics
- **Clustering method**
  - Contents-based
    - similarities
  - Structure-based
    - relations



7

Structure-based clustering

**Web link structure**

+

similarity between contents of pages

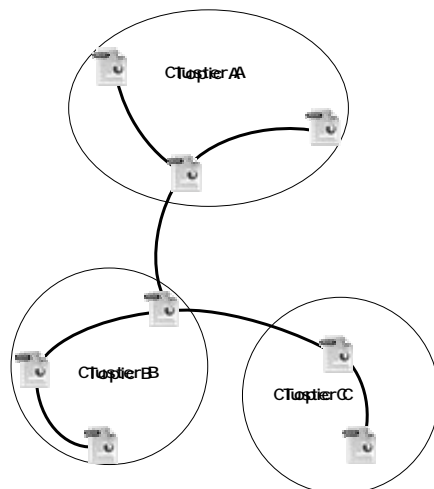
similarity of document vectors

**weight by types of Web links**

### Web link structure

---

- The primitive relationships between the topics of the linked pages
- To extract the relationship of the topics
  - Merging the linked clusters



## Types of Web links (1/2)

---

- Relevance between the topics of linked pages
  - types of links
  - distance between directories in which the pages are located
- The authors of Web sites probably make directories on Web sites with following intention
  - Web pages are classified into directories along the topics of pages
  - Topics of pages in child directories are specialization of the topics in their parent directory
  - Topics of pages in closer directories are similar than in distant directories
- form dense clusters by prioritizing links of pages in closer directories

---

10

## Types of Web links(2/2)

---

- three types of Web links
  - upward/downward
    - link between a page and in ancestor or descendant directory in same Web sites
  - crosswise
    - link between pages in same Web sites, except upward/downward links
  - outward
    - link between pages in two different Web sites
- distance of directories
  - The number of directories on shortest path between any two pages in tree structure which consists of pages and directories as nodes
- To generate dense clusters
  - Prioritize crosswise over upward/downward and
  - Prioritize links which have lower values for distance of directories

---

11

## Proposed Method

---

### Structured-based clustering method

- Newman's method

Structure of links

### Applying Weights to links

- Common similarity of document vectors
- Weight by types of links

Similarity of contents

---

12

## Newman's method

---

- Newman's method[Newman2004]
  - merging clusters to maximize modularity Q
  - Modularity Q: function to check the validity of network division

$$Q = \sum_i (e_{ii} - a_i^2) \qquad a_i = \sum_j e_{ij}$$

$e_{ij}$ : the fraction of the edges between the cluster  $i, j$  to all edges in network

- High value of Q: vertices are closely bound together in each clusters
- Applying the weights to edges

---

13

## Weight of edge

- weight of edge:

$$\text{weight}(p, q) = \alpha \times \text{similarity}(p, q) + (1 - \alpha) \times \text{link weight}(p, q) \quad (\alpha = 0.5)$$

- content similarity:

$$\text{similarity}(p, q) = \frac{d_p \bullet d_q}{\|d_p\| \times \|d_q\|} \quad d : \text{document vector of page}$$

- weight by types of links

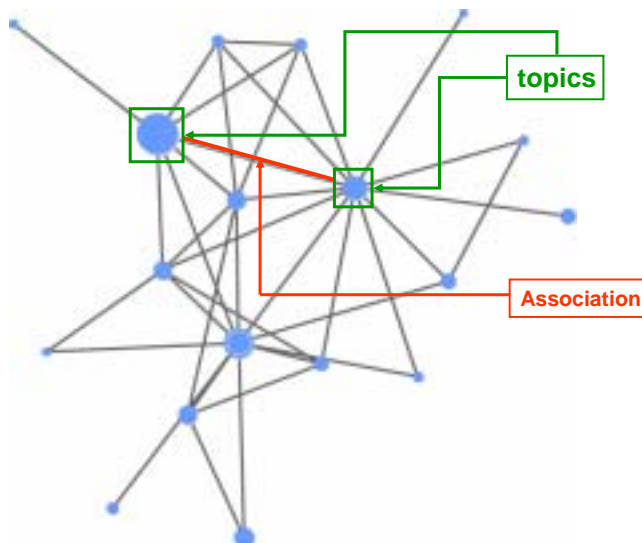
$$\text{link weight}(p, q) = \begin{cases} \frac{0.25}{(D+1)} \times C_{ld} & (\text{upward / downward}) \\ \frac{0.5}{(D+1)} \times C_{ld} & (\text{crosswise}) \\ \frac{0.4}{(D+1)} \times C_{ld} & (\text{outward : } \tau_{sim} < \text{similarity}(p, q)) \\ 0 & (\text{outward : } \text{similarity}(p, q) < \tau_{sim}) \end{cases}$$

$d$  : distance of directories in which the pages are located

$C_{ld}$  : type of link direction (2 : two way, 1 : one way)

- Defined by our policy that is prioritizing links of pages in closer directories

## Build the prototypes of topic maps





## Experiment(1/2)

---

- Evaluate the framework of topic maps
  - Proposed method
  - Newman's methods
- Measure
  - Number of valid topics and associations
    - Good frameworks cover topics and associations in detail
    - Suitable for user's modification
- Task
  - Naming the topics and the associations appropriately
  - Valid topics and associations are appropriately named items
  
  - Topic
    - The list of the Web pages (URL, title, content of the pages)
  - Association
    - The information of the two topics( anchor texts of links)

---

16

## Experiment(2/2)

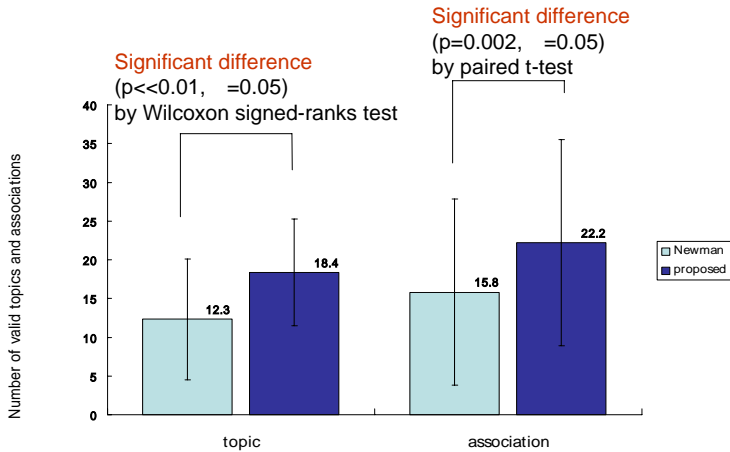
---

- Experiment Settings
  - Participants: 16
  - Browsing history data: 2 / participant
  - Web pages of history data: 5 / data
  - Web page set:
    - Expanding the outbound and inbound links from the Web history pages by 4 steps
    - Selecting 3 links to expand each step
  - Topic Maps: 4(2data x 2methods)
  - Evaluation order: random

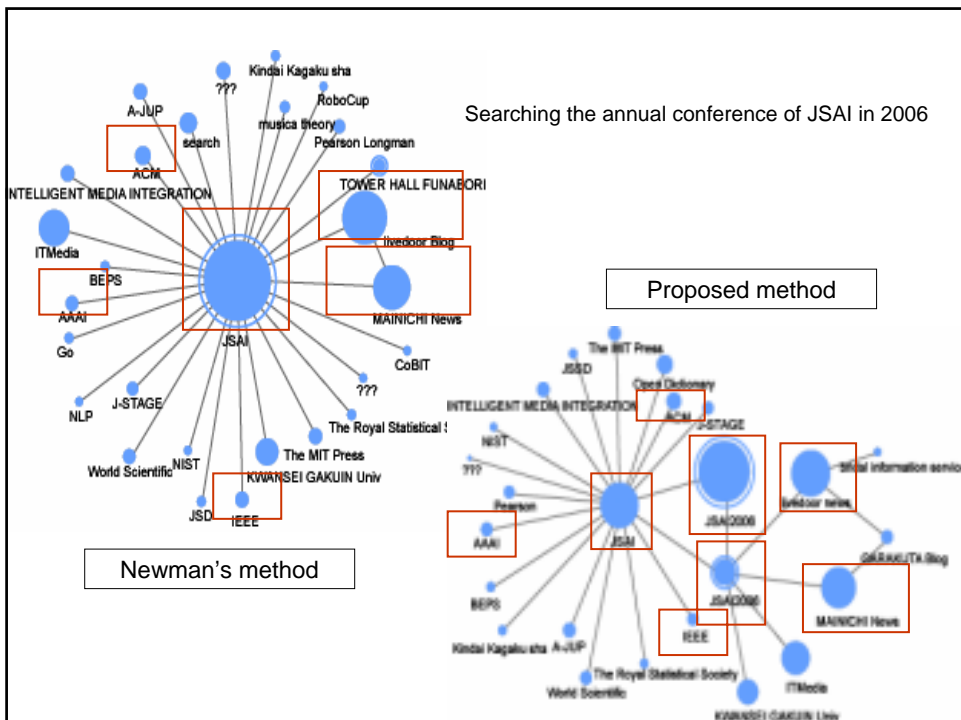
---

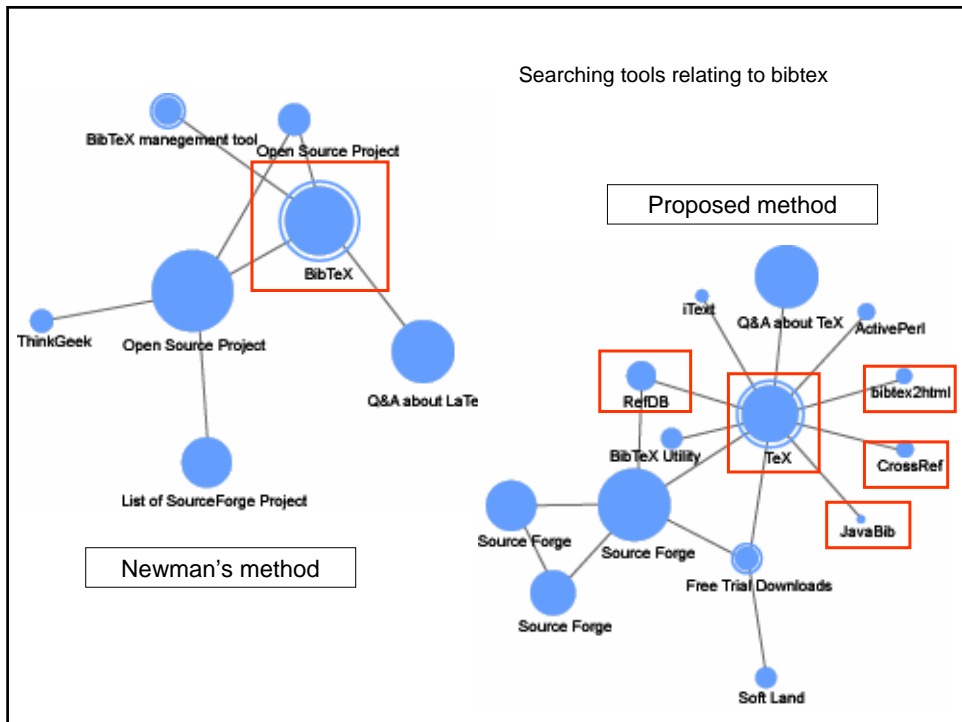
17

# Result: Number of valid topics, associations



The proposed method can extract the prototypes of topic maps, which have more valid topics and associations than by Newman's method





## Discussion

- The proposed method can extract more topics and associations than Newman's method
- The proposed method is good for representing useful topics and associations

## Conclusion

---

- We proposed the method to extract the prototypes of topic maps
- We conducted the experiment to evaluate the proposed method by comparison with Newman's method
- The result of experiment showed that the proposed method can extract the prototypes of the topic maps, which have enough valid topics and associations

## Current Work

---

- Utilizing tags or categories of Web pages
  - Types of links and distance of directories
- News site
  - Directories according to date of article
    - <http://plusd.itmedia.co.jp/pcuser/0710/31/news024.html>
  - Use tags or categories as directories
    - ex. notepc
      - <http://plusd.itmedia.co.jp/pcuser/notepc/news024.html>

---

ITmedia / PCUSER

Defaults Grouping

- Desktop PC
- Note PC
- Mac
- Printer
- Peripheral Device
- PC Parts
- Software
- Akihabara
- Event
- Mother board

