Fusion of Libraries and the Web: Subject-based Information Retrieval in the Web 2.0 Era

Digital Library Research Division
Information Technology Center
University of Tokyo
Yoji Kiyota (Assistant Professor)
Agenda

1. What is happening on the information retrieval world in the Web 2.0 era?
2. Toward the Lib 2.0: what is needed to enhance the values of libraries?
3. A solution: fusion of folksonomy and taxonomy
4. Application for reference services: Littel Navigator
Agenda

1. **What is happening on the information retrieval world in *the Web 2.0 era***?
2. Toward *the Lib 2.0*: what is needed to enhance the values of libraries?
3. A solution: fusion of folksonomy and taxonomy
4. Application for reference services: Littel Navigator
History of information access methods

• Until the 20th century
  – To research in libraries
  – To go to bookstores
  – To ask someone
  – To ask consultant services
    (librarians, government offices, consultants, …)
  – To look up Web directories, BBS, databases, …

• From the 21st century: in addition,
  – To google (Web search engines)
  – To access QA sites, SNS, folksonomy, …

  *So-called the Web 2.0 paradigm*
What is Web 2.0?

Definition by Tim O’Reilly (2005)

• The Web as platform
• Harnessing collective intelligence
• Data is the next Intel Inside
• End of the software release cycle
• Lightweight programming models
• and more…

Just a buzzword? --- No, if we choose one of the definitions
What is Web 2.0?

Definition by Tim O’Reilly (2005)

• The Web as platform
• Harnessing collective intelligence
  – Improvements on web search engines
  – Folksonomy: organization by the wisdom of crowds
• Data is the next Intel Inside
• End of the software release cycle
• Lightweight programming models
Possibility of collective intelligence as an Information access method

• The broadest gate for information
  – Any keyword hits using Google
  – The best way to find cues?

• Diversity of information
  – Exceeds diversity of mass-media

• Self-organization of information
  – Improvements of web search engines
    • PageRank: democracy on the Web
  – Folksonomy: meta-tagging based on the wisdom of crowds
Folksonomy

= “folks” + “taxonomy”

Meta-tagging by ordinary people

• Web services diffusing from 2005
  
    Flickr, del.icio.us, YouTube, …

• Every participant assigns tags to contents, based on each viewpoint

• As a result, diverse tags are assigned to each content

• Example approaches of libraries
  
    – Ann Arbor District Library “Social OPAC”
Photos searched by “B747” Using Flicker

Related tags (co-occurrence)
Catalog Tag Cloud

This is a tag cloud of the 500 most popular tags in our catalog.


Top 10 Tags
1. fantasy (145)
2. manga (121)
3. anime (77)
4. shonen jump (60)
5. time travel (57)
6. shonen (55)
7. demons (36)
8. ghosts (36)
9. half-demons (36)
10. shape-changing (35)

10 Most Recent Tags
The following items are tagged with the term *shonen jump*:

**One piece. Vol. 3, Don't get fooled again**
- pirates, fantasy, manga, shonen jump, shonen

**One piece. Vol. 5, For whom the bell tolls**
- pirates, fantasy, manga, shonen jump, shonen

**One piece. Vol. 2, Buggy the Clown**
- pirates, fantasy, manga, shonen jump, shonen

**One piece. Vol. 1, Romance dawn**
- pirates, fantasy, manga, shonen jump, shonen

**Shonen Jump Naruto. 2, disc 2 [videorecording (DVD)]**
- anime, ninjas, fantasy, shonen jump, shonen

**One piece. Vol. 11, The meanest man in the East**
- pirates, fantasy, manga, shonen jump, shonen

**One piece. Vol. 4, The Black Cat Pirates**
- pirates, fantasy, manga, shonen jump, shonen
Wikipedia categories

- A multilingual encyclopedia edited by ordinary people
- Categories are assigned to each article
  - categories are regarded as folksonomy tags
  - multi-level folksonomy
  - multiple hyponyms

Price

Price in economics and business is the assigned numerical monetary value of a good, service or asset. The concept of price is central to microeconomics where it is one of the most important variables in resource allocation theory (also called price theory).....

Categories: [marketing][economics][market]
Two classification paradigms

**Folksonomy**
- Bottom-up approach
- Multiple hyponyms

Classification suitable for Web resources

**Taxonomy**
- Top-down approach
- Only one hyponym

Classification suitable for library materials
Limitation of the folksonomy

Browsing-oriented

- Floods of information without validation
- Prefer new information to old information
  - How many web pages survive 10 years?
    cf. web.archive.org
- Shallow organization
  - Meta noise of folksonomy
- Lack of methods for evaluation and validation

Can libraries complement the limitation of folksonomy?
Agenda

1. What is happening on the information retrieval world in the Web 2.0 era?

2. Toward the Lib 2.0: what is needed to enhance the values of libraries?

3. A solution: fusion of folksonomy and taxonomy

4. Application for reference services: Littel Navigator
Toward the Lib 2.0

Fusion of libraries and the Web
• To provide accesses to stocks in libraries
  – Induction from the Web to libraries
  – Enables validations of information on the Web
• To provide viewpoints of information
  – Various methods, including reference books, dictionaries, …

The role of library classification systems is very important!
Library classification systems (gaps between ideal and reality)

- Classification (UDC, LC, DDC, …)
  - ideal: browsing of organized bookshelves is useful for information retrieval
  - reality: lack of browsing methods online (cf. OPAC)

- Subject headings (LCSH, …)
  - ideal: concept-based search of catalogs
  - reality: lack of flexibility

a high hurdle for ordinary people?
A hint: pathfinders

• Firstly developed at MIT library (in 1970s)
• The most demanded information resources at the beginning of information retrieval?
• Manually created by librarians
  – low coverage

Can we create pathfinders automatically?
Possibilities of automation of pathfinder creation

• Templates are available
• Subject headings and library classifications play the major role for pathfinder generation

↓

Possible if subject headings are estimated for query keywords
• databases of reference books, elementary books, …

A Problem

too few subject headings!
Agenda

1. What is happening on the information retrieval world in *the Web 2.0 era*?

2. Toward *the Lib 2.0*: what is needed to enhance the values of libraries?

3. **A solution**: fusion of **folksonomy** and **taxonomy**

4. Application for reference services: Littel Navigator
A solution: fusion of Wikipedia categories and library classifications

Expansion of classifications using Wikipedia
• Wikipedia: folksonomy-based organization
  – The most organized information resource on the Web
  – Compatibility of other Web resources
• Library classifications: taxonomy-based organization
  – Reflect the knowledge structure of human history
  – Powerful tools for demanded information resources

Automated Subject Induction from Query Keywords through Wikipedia Categories and Subject Headings
-> recommendation of useful information
Start point of retrieval

information resources on the Web
folksonomy

Wikipedia

library classification and subject headings

information resources in libraries
• elementary books
• reference books
• journals
• past literatures

integration for deeper retrieval

reliability on information retrieval
1. What is happening on the information retrieval world in *the Web 2.0 era*?
2. Toward *the Lib 2.0*: what is needed to enhance the values of libraries?
3. A solution: fusion of folksonomy and taxonomy
4. **Application for reference services:** *Littel Navigator*
What is Littel Navigator?

A search engine for hints of information retrieval
= A pathfinder generator

• Fusion of various information sources
  – Subject headings, Library classifications, reference books
  – Web sites, Wikipedia...

• Automated theme induction from any keyword
  – Traversal of the Wikipedia network

• Navigation from vague keywords to specific materials
  – Stored history of user inputs
Input keywords related to what you want to know.
Induced themes related to the keywords

Induction paths of the themes

Information resources related to earthquake
Reports for 10th anniversary of Hanshin-Awaji earthquake, Kobe University
A journal for seismology (volumes in a library, OPAC)
Meta search of various databases
A search result of Google Scholar (articles for Hanshin-Awaji Earthquake)
A search result of JapanKnowledge (encyclopedia)
地震

「阪神・淡路大震災」に関する情報

経済

経済学

災害
History of Information Retrieval
Conclusion

• Fusion of library classification systems and the Web 2.0 paradigm enables “A search engine of hints for information retrieval
  – Various viewpoints for information retrieval
  – Derivation from Web resources to reliable resources

• Complement each other
  – Web: any keywords can be uses as cues
  – Library: reliability and organization