【iiWAS2014】

Extracting Facets from Textual Contents for Faceted Search over XML Data

Takahiro Komamizu, Toshiyuki Amagasa, Hiroyuki Kitagawa
University of Tsukuba, Japan

Background

- XML (Extensible Markup Language)
  - a de facto standard data format of (semi-)structured data
  - represented as a tree
  - e.g., RDF/XML, KEGG, Swiss-Prot
- Search over XML data
  - find XML subtrees that meet users' demands

Search demands over XML data

- Ad-hoc search demands
  - A user has concrete search demands, which can be specified by various means.
- Exploratory search demands
  - A user has vague search demands, so he can specify ambiguous requirements.

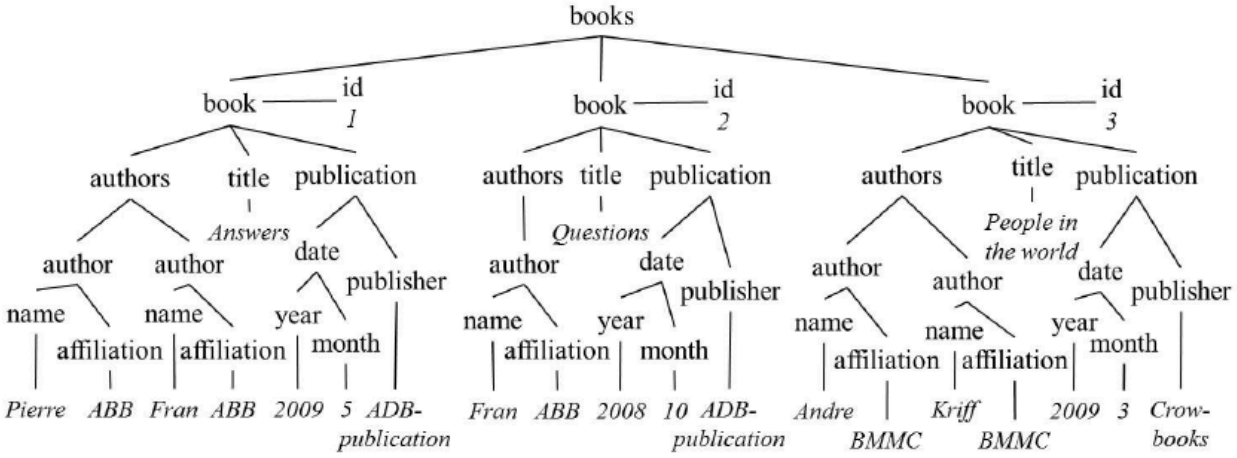
Search over XML data

- For ad-hoc search demands
  - existing methods work
    - path-based search (e.g., XPath and XQuery)
    - keyword-based search (e.g., LCA families)
- For exploratory search demands
  - A user is required to perform several searches by modifying her queries.
  - System support is necessary.

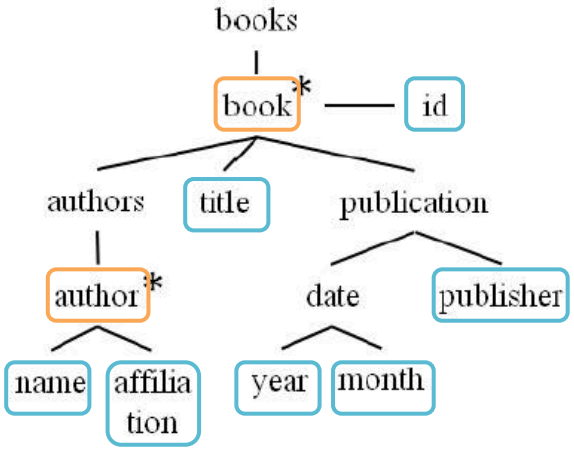
    ➔ We applied faceted search.

Our previous work [iiWAS'11]

- Applying faceted search over XML data
  - extracting target XML subtrees and XML elements as facets, and operation families



XML data



Structural information
(e.g., DataGuide)

Motivation of this paper

- The previous work does not use long texts such as titles of papers.
  - because such facets work as identifiers, so not useful in faceted search context
- However, such long texts still also contain useful facet-values.
  - ➢ This paper attempts to utilize the textual contents in order to improve search performance.

What we want to do

previous work

title facet

- XML search (1)
- faceted XML search (1)
- XML keyword search (1)
- XML query suggestion (1)
- XML search log analysis (1)
- indexing for XML keyword search (1)
- RDF search (1)
- faceted RDF search (1)
- RDF keyword search (1)

utilize

in this work

title facet

- search (8)
- XML (6)
- RDF (3)
- faceted (2)
- keyword (3)
- query (1)
- suggestion (1)
- log (1)
- analysis (1)

Our task: choose terms should be shown

Contributions of this paper

- Propose a facet-value extraction scheme from textual contents of XML data
- Propose an evaluation scheme for exploratory search in terms of specificity
- Evaluate faceted search using extracted facet-value comparing with the previous work as well as keyword search

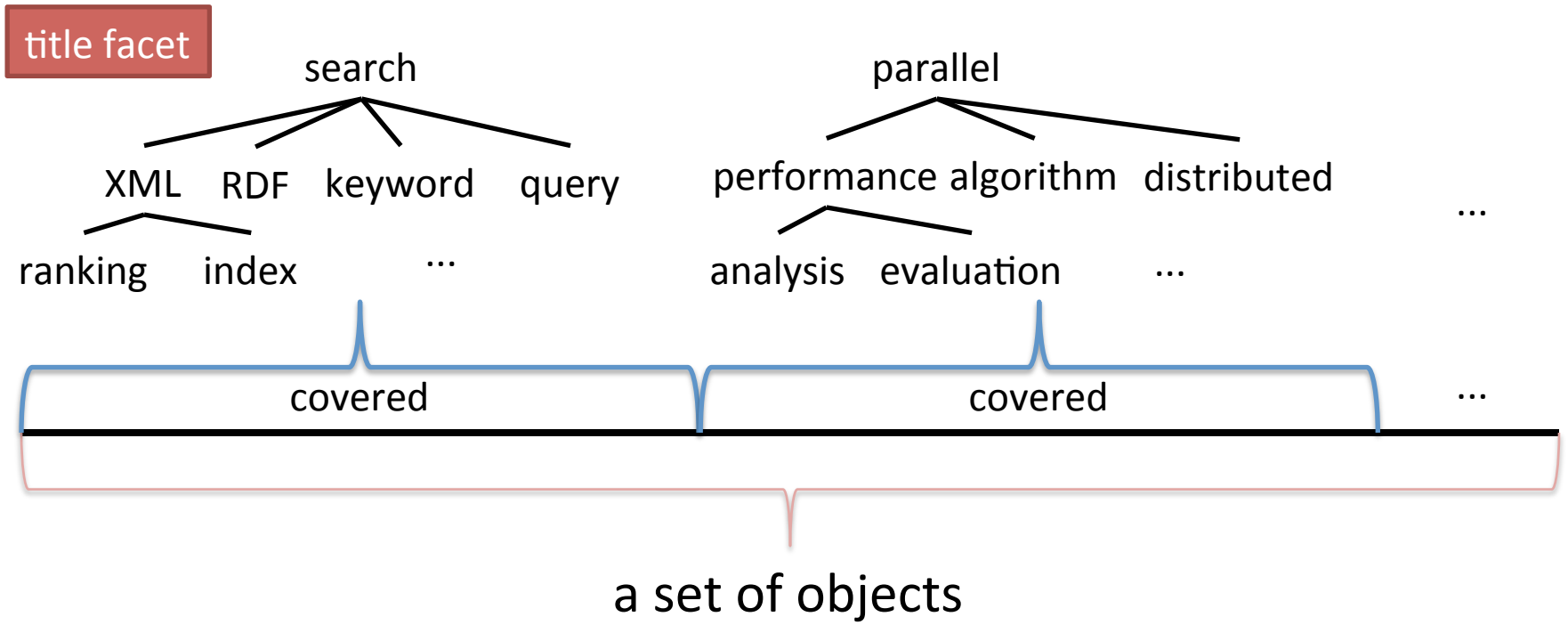
Subsumption

[Sanderson & Croft, SIGIR'99]

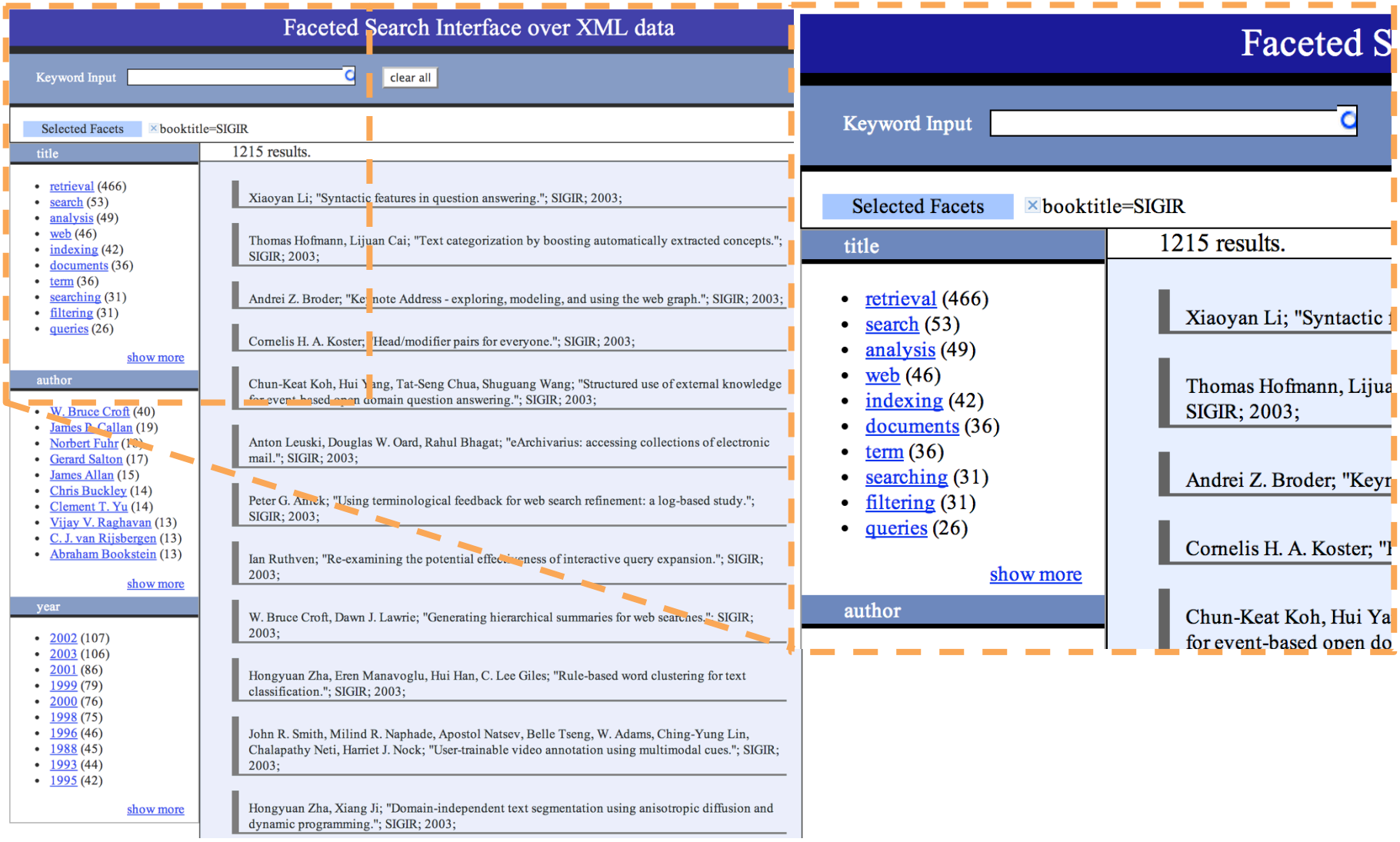
- Concept hierarchy construction scheme from a number of textual documents
- Probabilistic approach using co-occurrences of terms
  - *x* subsumes *y* if
    - $p(x \mid y) > \tau_s$, and $p(x \mid y) > \tau_d \cdot p(y \mid x)$
    - according to their experiments, $\tau_s = 0.8$, $\tau_d = 1.2$

How we use Subsumption

- Covering as many objects as possible
  - Continuously apply Subsumption for uncovered objects



Snapshot of our system



Evaluation: user study

- We observe how our system improves search performance based on proposed scheme.
- Still, designing tasks for user study in terms of exploratory search demands is difficult.
  - Tasks should be designed to be explorative, vague, or searchable in a trial-and-error manner.

Task design principle

- Template-based designing
  - based on [Kules et al., JCDL'09]

> Imagine that you are taking a class called ____. For this class, you need to write a paper on the topic ____. Use the catalog to find two possible topics for your paper. Find three books for each topic.

- Examiner fills the blanks
- Examinees have to explore related (or sub) topics and papers for the given topic.

Specification level of a task

- Definition: overall selectivity of terms contained in the task

$$sl(T) = \frac{\left|\bigcap_{t \in T} \sigma_{keyword=t}(D)\right|}{|D|}$$

  - *sl*(*T*) is a specification level of a task *T* consisting of several terms {t_1,t_2,...}.
  - σ returns a set of objects having keyword *t*.
  - *D* is the total set of objects.

Terms and specification level

(from titles of DBLP dataset)

term	specification level
analysis	0.03534
design	0.03198
database	0.01713
graph	0.00755
large	0.00746
security	0.00549
neural, networks	0.00484
case, study	0.00482
logic, programming	0.00366
user, interface	0.00173
knowledge, representation	0.00148
relational, database	0.00115
world, wide, web	0.00110
support, vector, machines	0.00052
inductive, logic, programming	0.00035
analysis, case, study	0.00026

Task example

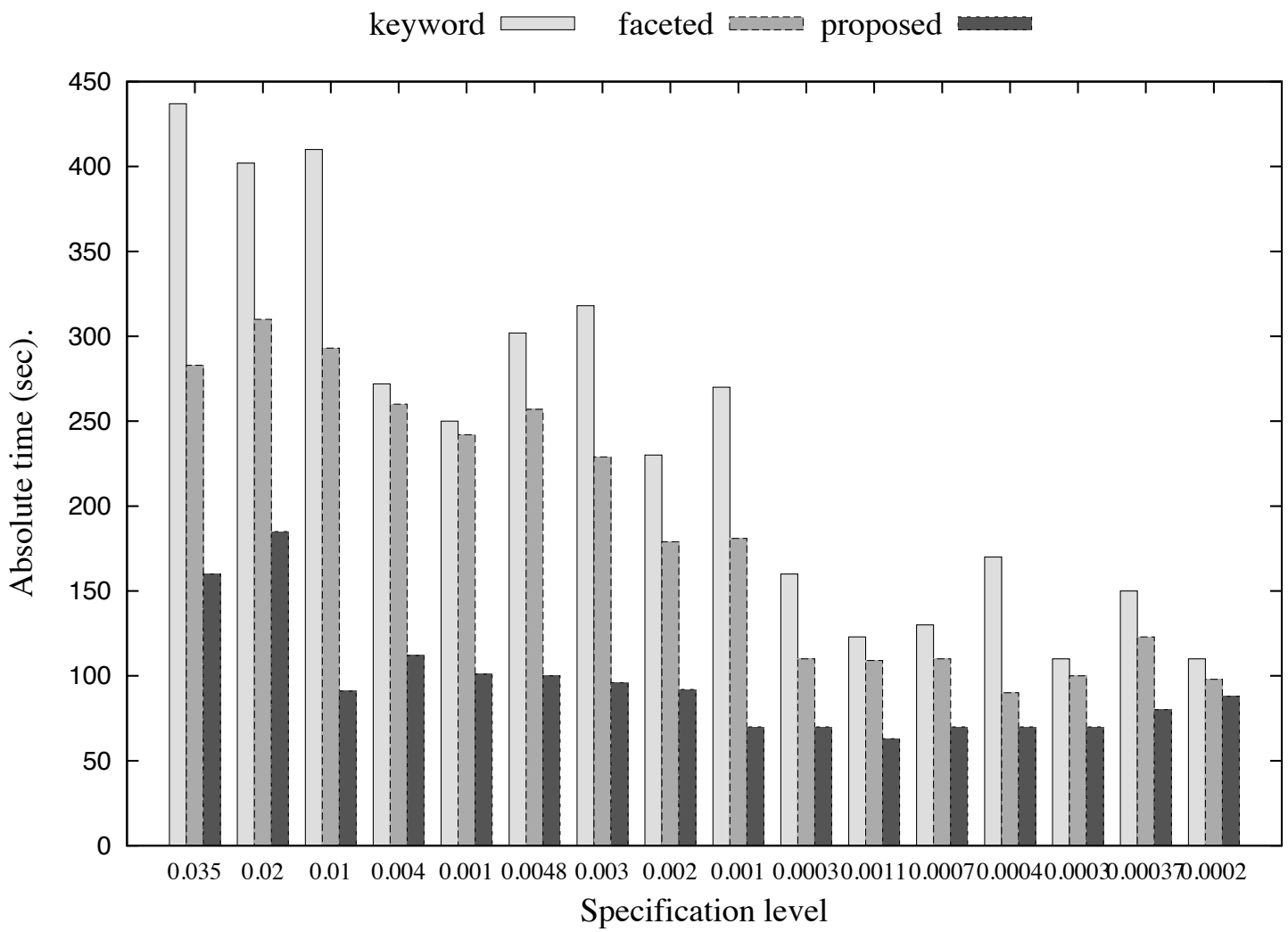
- Given specification level requirement, the second black is filled the closest term.
  - e.g., 0.0005 → support vector machines
- Examiner fills the first blank
  - e.g., Introduction to Machine Learning

> Imagine that you are taking a class called Introduction to Machine Learning. For this class, you need to write a paper on the topic support vector machines. Use the database to find two possible topics for your paper. Find three books for each topic.

Evaluation methodology

- Measure time until achieving given tasks
  - choosing tasks having various specification levels
- Competitors
  - conventional faceted search (our previous work, keyword search enabled)
  - keyword search
- #examinees: 5

Evaluation results



Conclusion

- Propose a facet-value extraction scheme from textual contents of XML data.
  - Subsumption-based approach
- Propose an evaluation scheme for exploratory search systems
  - specification level concept
- Experimentally show our proposed scheme outperforms conventional systems

THANK YOU
FOR YOUR ATTENTIONS