IDEAS 2014

A Scheme of Automated Object and Facet Extraction for Faceted Search over XML Data

Takahiro Komamizu, Toshiyuki Amagasa, Hiroyuki Kitagawa

University of Tsukuba
Background

- XML has become a de facto standard data format for semi-structured data.
- Searching over XML data requires knowledge of either or both of:
  - structure (or schema) of the XML data
  - query languages of XML data (e.g., XPath or XQuery)
  - processors of XML data to search (e.g., SAX or DOM)
- Furthermore, users are expected to have concrete search demands.
  - Otherwise, the users may not explore the XML data to find desired subtrees, since the users are not capable to express their vague demands.

Research Objective

We develop a system, which assists users to navigate the XML data.
Applying **faceted search** for XML data

- Faceted search is one of the successful exploratory search methods.
- By using faceted search, a user can search objects by clicking interesting attributes (called **facets**) shown on the interface.

**Proposed approach**

- We develop a **framework** for faceted search over XML data
  - extracts object candidates, and facet candidates
  - provides an interface for the system manager to select which candidates to be objects and facets.
  - generates the faceted search interface for selected objects and facets
Faceted Search: Basic Data Structure

- Ordinal faceted search expects **record structure**
- Each record corresponds to an **object** and some attributes are regarded as **facets** (e.g., author, year, and publisher)

<table>
<thead>
<tr>
<th>title</th>
<th>author</th>
<th>year</th>
<th>publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML Search</td>
<td>John A. Smith</td>
<td>2012</td>
<td>AAC publisher</td>
</tr>
<tr>
<td>XML: An introduction</td>
<td>John A. Smith</td>
<td>2010</td>
<td>CCD publisher</td>
</tr>
<tr>
<td>XML Data Management</td>
<td>Anna F. Doe</td>
<td>2012</td>
<td>CCD publisher</td>
</tr>
<tr>
<td>RDF Search</td>
<td>Anna F. Doe</td>
<td>2014</td>
<td>AAC publisher</td>
</tr>
</tbody>
</table>
Faceted Search: Search Paradigm

- A user selects a facet and its value. (e.g., year and “2012”)

<table>
<thead>
<tr>
<th>title</th>
<th>author</th>
<th>year</th>
<th>publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML Search</td>
<td>John A. Smith</td>
<td>2012</td>
<td>AAC publisher</td>
</tr>
<tr>
<td>XML: An introduction</td>
<td>John A. Smith</td>
<td>2010</td>
<td>CCD publisher</td>
</tr>
<tr>
<td>XML Data Management</td>
<td>Anna F. Doe</td>
<td>2012</td>
<td>CCD publisher</td>
</tr>
<tr>
<td>RDF Search</td>
<td>Anna F. Doe</td>
<td>2014</td>
<td>AAC publisher</td>
</tr>
</tbody>
</table>

\[ \downarrow \sigma_{\text{year} = “2012”}(D) \]

<table>
<thead>
<tr>
<th>title</th>
<th>author</th>
<th>year</th>
<th>publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML Search</td>
<td>John A. Smith</td>
<td>2012</td>
<td>AAC publisher</td>
</tr>
<tr>
<td>XML Data Management</td>
<td>Anna F. Doe</td>
<td>2012</td>
<td>CCD publisher</td>
</tr>
</tbody>
</table>
Problem Statement

Problems on applying faceted search for XML data

- For faceted search, we need to define **object** and **facet** beforehand.
  - Unlike record-structured data, XML data do not explicitly have neither objects nor facets of objects in the structure.
- As well, we need operations to search over XML data via faceted search interface.
  ⇒ We proposed a framework[5]

Problems this paper deals with

- The framework is *semi-automatic* (detail will be in a next few slides) and a system manager is still required a burden to decide which XML elements to be objects, facets or none.
  ⇒ We, in this paper, want to reduce this burden by automating object extraction and facet extraction.
Faceted Search for XML Data

Task

- determine which XML subtrees used for results objects
- determine which XML subtrees of objects for facets

Framework overview

1. extract **structural information** from XML data
2. determine **objects** and **facets** on structural information
   - Candidate objects: XML elements which occur multiple times under single parental elements.
   - Candidate facets: XML elements which occur under object elements.
3. (a system manager) determines objects and facets from the candidates
4. (users) search the objects through defined operations on the interface (see [5] for detail)
Structural Information

- Structural information of XML data is a structural summary (or schema) of the XML data, which describes how the XML data is organized (e.g., DTD, XML Schema, and DataGuide).

(a) XML data: paper list

(b) Structural information

- The number labels beside vertices in structural information denote the average frequency of the XML elements under their parental elements in the XML data.
A Scheme of Automated Object and Facet Extraction for Faceted Search over XML Data

Takahiro Komamizu, Toshiyuki Amagasa, Hiroyuki Kitagawa

An example

(a) XML data: paper list

(b) Structural information

<table>
<thead>
<tr>
<th>object candidates</th>
<th>facet candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>paper</td>
<td>authors, author, title, venue, year</td>
</tr>
<tr>
<td>author</td>
<td></td>
</tr>
</tbody>
</table>

\[\downarrow\text{selected by a system manager}\]

<table>
<thead>
<tr>
<th>object</th>
<th>facets</th>
</tr>
</thead>
<tbody>
<tr>
<td>paper</td>
<td>author, venue, year</td>
</tr>
</tbody>
</table>
Motivation

Conventionally, objects and facets on faceted search over XML data have been determined manually.

Our framework enables to reduce the effort, by picking up possible objects and facets from XML data, but the process is still semi-automatic.

So, determining objects and facets still requires large effort.

Automation of extracting objects and facets is desirable.

Objective

Automate object and facet extraction process.
Basic Ideas

- Our automation scheme is based on the following observations:
  - XML elements consistently occurring frequently under their parental elements tend to be result objects.
  - XML elements with id-like textual contents tend not to be facets.
    - id-like: unique value on each object
  - XML elements whose names are not meaningful (e.g., ee or sub) should be avoided to be facets.

- Ideas
  - filtering out candidates by frequency threshold.
  - filtering out unrecognizable candidates by external resources.
    - e.g., Wikipedia and WordNet.
Proposed Approach - Frequency-based Object Extraction -

- Extract an XML element as an object, when the average occurrence of the XML element under its parental element is greater than the given threshold.
- Example: threshold = 1.7

- paper is extracted as object.
Given an object, extract an descendant XML element as a facet, when the average number of occurrence of textual values (afv for short) is greater than the given threshold.

Example: threshold = 1.2, object = paper
- afv(author) = 1.67, afv(title) = 1, afv(year) = 1.5, ...

- author, venue and year are extracted as facets of paper object.
Given an object, extract an descendant XML element as a facet, when the maximum semantic similarity between the name of the element and any term in semantic information (e.g., Wikipedia entries) is greater than the given threshold.

- Example of semantic similarity: inverse of distance in WordNet graph
- Example: threshold = 0.8, I is semantic information, object = paper
  - $\text{sem\_sim}(\text{author, I}) = 1$, $\text{sem\_sim}(\text{pid, I}) = 0.2$, ...

- author, authors, author, title, venue and year are extracted as facets of paper object.
Experimental Settings

- **Purpose**: check how accurately extracting objects and facets
  - observe the effect of thresholds
  - comparison of proposed approaches
- **Dataset**: XML data available on UW XML Repository, and QCDml
  - in UW XML Repository: Protein Sequence Database, SwissProt, Yahoo! Auction data, DBLP, University Courses (including reed, uwm, wsu), and SIGMOD Record
- **Measurement**: precision, recall, and f-score
  - ground-truth data are manually provided
- **Methodology**
  - extract objects and facets using the proposed approaches
  - calculate accuracy of the extracted objects and facets
Frequency-based Object Extraction

- Extracting objects by changing frequency threshold.

(a) Precision  
(b) Recall  
(c) F-Score

- All of results on recall marks 1, thus necessary objects are extracted.
- When increase the threshold, the precision increases and thus the f-score increases as well.
Facet Extraction

- Compares the proposed approaches, frequency-based, semantic-based, and hybrid of them.
- The frequency threshold: 1.2, and the semantic similarity: 0.8.

- Frequency-based extracts nicely in most cases.
- Semantic-based extracts too many facets.
- Semantic-based increases the accuracy of Frequency-based.

![Graphs showing precision, recall, and F-score](chart.png)
Conclusion

- Proposed
  - An automated object and facet extraction scheme on the framework [5] of faceted search for XML data

- Future work
  - Improve the automatic extraction.
  - Extraction of textual facets from textual contents in facets.
    - Identify facets for textual facets.