Extending faceted navigation for RDF data

Mohammed HALALO
Contents

- Problem
- Faceted browsing
- A faceted interface for RDF data
- Automatic facet ranking
- Evaluation
- Conclusion
- Review
Problem

- Semantic Web data:
  - expressed in RDF
  - very large
  - highly interconnected
  - heterogeneous
    (no one fixed schema)
Navigation technique should be
- Scalable
- support graph-based navigation
- Generic:
  - not depend on a fixed schema
  - allow exploration of the dataset without a-priori knowledge of its structure
Current solutions:

1. Keyword search, e.g. Swoogle
   (-) not for higher search activities such as learning and investigating
2. Explicit queries, e.g. Sesame
   (-) difficult and requires schema knowledge
3. Graph visualisation, e.g. IsaViz
   (-) does not scale to large datasets
4. Faceted browsing
   (-) Manually constructed
   (-) domain-dependent
   (-) and do not fully support graph-based navigation
Solution

The paper contributes into:

1. Improving the current faceted browsing techniques for RDF data
2. Developing a technique for automatic facet ranking
Faceted browsing

- the information space is partitioned using orthogonal conceptual dimensions of the data.
- These dimensions are called **facets** and represent important characteristics of the information elements.
- Each facet has multiple **restriction values** and the user selects a restriction value to constrain relevant items in the information space.
Search

About 2,040,000 results (0.33 seconds)

Clip art

Related searches: semantic web

Page 2
Ali Atwa, Ammar Mansour Bouslim, Hassan Rostom Salim (select)

Aliases
- Ammar Mansour Bouslim
- Hassan Rostom Salim

Build
- Build type medium

Caution
- Ali Atwa was indicted for his role and participation in the June 14, 1985, hijacking of a commercial airliner which resulted in the assault on various passengers and crew members, and the murder of one U.S. citizen.

Citizenship
- Lebanese

Date of birth used
- Approximately 1960

Eyes
- Brown

Hair
- Unknown

Height
- 5'8"

Language
- Arabic

Place of birth
- Lebanon

Reward
- The Rewards For Justice Program, United States Department of State, is offering a reward of up to $5 million for information leading directly to the apprehension and/or conviction of Ali Atwa.

Remarks
Functionality

- Selection operators:

(a) Basic selection

(b) Existential selection

(c) Join selection
Functionality

- Selection operators (cont’d)

(a) Inverse basic selection

(b) Inverse existential selection

(c) Inverse join selection
Functionality cont’d

“all unmarried thirty-year-olds who know somebody –working in Ireland– who knows Stefan”
Definition:

(RDF Graph). An RDF graph $G$ is defined as

$G = (V,E,L, l)$

where $V$ is the set of vertices (subjects and objects),

$E$ is the set of edges (predicates),

$L$ is the set of labels,

$l : E \rightarrow \text{label}$ is the labeling function for predicates and with $V$ and $E$ disjoint.

The projections, $source : E \rightarrow V$ and $target : E \rightarrow V$, return the source and target nodes of edges.
### Operator definitions

<table>
<thead>
<tr>
<th>operator</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic selection</td>
<td>$\text{select}(l, v') = { v \in V \mid \forall e \in E : \text{label}(e) = l, \text{source}(e) = v \text{, target}(e) = v' } $</td>
</tr>
<tr>
<td>inv. basic selection</td>
<td>$\text{select}^-(l, v') = { v \in V \mid \forall e \in E : \text{label}(e) = l, \text{source}(e) = v' \text{, target}(e) = v } $</td>
</tr>
<tr>
<td>existential</td>
<td>$\text{exists}(l) = { v \in V \mid \forall e \in E : \text{label}(e) = l, \text{source}(e) = v } $</td>
</tr>
<tr>
<td>inv. existential</td>
<td>$\text{exists}^-(l) = { v \in V \mid \forall e \in E : \text{label}(e) = l, \text{target}(e) = v } $</td>
</tr>
<tr>
<td>not-existential</td>
<td>$\text{not}(l) = V - \text{exists}(l) $</td>
</tr>
<tr>
<td>inv. not-existential</td>
<td>$\text{not}^-(l) = V - \text{exists}^-(l) $</td>
</tr>
<tr>
<td>join</td>
<td>$\text{join}(l, V') = { v \in V \mid \forall e \in E : \text{label}(e) = l, \text{source}(e) = v \text{, target}(e) \in V' } $</td>
</tr>
<tr>
<td>inv. join</td>
<td>$\text{join}^-(l, V') = { v \in V \mid \forall e \in E : \text{label}(e) = l, \text{source}(e) \in V' \text{, target}(e) = v } $</td>
</tr>
<tr>
<td>intersection</td>
<td>$\text{intersect}(V', V'') = V' \cap V'' $</td>
</tr>
</tbody>
</table>

As an example, all thirty-year-olds without a spouse would be selected by: $\text{intersect}(\text{select}(\text{age}, 30), \text{not}(\text{spouse}))$.  

---

Table 1: Operator definitions
Dataset is large

A lot of facets

Users will not be able to navigate easily

we need a technique to suggest good facets

which facets are more useful and more important than others?!
We need to find among all predicates,
- best represent the dataset (the best **descriptors**),
- most efficiently navigate the dataset (the best **navigators**).
What are suitable descriptors of a data set?

- page number of an article?

Intuitive facets describe a property that is

- temporal (e.g. year of publication, date of birth)
- Spatial (conference location, place of birth)
- personal (author, friend)
- material (topic, color)
- energetic (activity, action).
Navigators

- A suitable facet allows efficient navigation through the dataset
- Faceted browsing is like constructing and traversing a decision tree

Fig. 6: Faceted browsing as decision tree traversal
Navigators

- navigation quality based on:
  - Predicate balance
  - Object cardinality
  - Predicate frequency

- All metrics range from [0..1]
- Combined into a final score through (weighted) multiplication
- recomputed at each step of the decision tree
- BUT: facets should be ranked on their both navigational and descriptive values (not done!)
Predicate balance

Tree navigation is most efficient when the tree is well balanced

\[ balance(p) = 1 - \frac{\text{var} [ n_s(p, o_1), \ldots , n_s(p, o_n)]}{1 + \text{var} [ n_s(p, o_1), \ldots , n_s(p, o_n)]} \]

- \( p \) : predicate
- \( n_s \) : number of subjects
- \( o \) : object
A suitable predicate has a limited (higher than one) amount of object values to choose from

\[ card(p) = \begin{cases} 0 & \text{if } n_o(p) \leq 1 \\ \exp \left( \frac{(n_o(p) - \mu)^2}{2\sigma^2} \right) & \text{otherwise} \end{cases} \]

<table>
<thead>
<tr>
<th>predicate</th>
<th>objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>title</td>
<td>4215</td>
</tr>
<tr>
<td>url</td>
<td>4211</td>
</tr>
<tr>
<td>author</td>
<td>4037</td>
</tr>
<tr>
<td>pages</td>
<td>2168</td>
</tr>
<tr>
<td>text</td>
<td>1069</td>
</tr>
<tr>
<td>booktitle</td>
<td>1010</td>
</tr>
<tr>
<td>number</td>
<td>349</td>
</tr>
<tr>
<td>address</td>
<td>341</td>
</tr>
<tr>
<td>journal</td>
<td>312</td>
</tr>
<tr>
<td>editor</td>
<td>284</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>type</td>
<td>13</td>
</tr>
</tbody>
</table>
A suitable predicate occurs frequently inside the collection:

- the more distinct resources covered by the predicate, the more useful it is in dividing the information space

\[ freq(p) = \frac{n_s(p)}{n_s}. \]
## Evaluation

- Formal Evaluation

<table>
<thead>
<tr>
<th>operator</th>
<th>BrowseRDF</th>
<th>Flamenco</th>
<th>mSpace</th>
<th>Ontogator</th>
<th>Spectacle</th>
<th>Seamark</th>
</tr>
</thead>
<tbody>
<tr>
<td>selection</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>inv. selection</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>existential</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>inv. exist.</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>not-exist.</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>inv. not-exist.</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>join</td>
<td>+</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
</tr>
<tr>
<td>inv. join</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>intersection</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Table 3: Expressiveness of faceted browsing interfaces
Evaluation

Experimental evaluation

- 15 test subjects: beginner (8), good (3) and expert (4) (in RDF)
- None were familiar with the dataset (FBI fugitives)
- Three interfaces: keyword search, manual (N3) query construction, and our faceted browser.
- They were asked to perform a set of small tasks
- In each interface the tasks were similar but not exactly the same
- All subjects were filmed
- two minute time-limit per task.

“find the number of people with brown eyes”, or “find the people with Kenyan nationality”
Experimental evaluation results

Chart Title

- easiest to use
- most flexible
- most dead-ends
- most helpful
- preference

- keyword
- query
- faceted

- solved
Faceted browsing is a data exploration technique for large datasets.

We have extended the expressiveness of existing faceted browsing techniques.

Developed metrics for automatic facet ranking.

Experimental evaluation shows better usability than current interfaces.
Review

Like

- Good expressive RDF graph definition
- Good selection formal expressions
- Auto facet ranking

Dislike

- New example every section
- Concept introduced but not used (descriptors)
- Unavailable tool
- Do we really need RDF browsing?!
Faceted browsing is a data exploration technique for large datasets. We have extended the expressiveness of existing faceted browsing techniques by developing metrics for automatic facet ranking. Experimental evaluation shows better usability than current interfaces.