Far from the Madding Subjects: A Comparison of Controlled and Open Authority for Metadata Remediation

Ying Wang
Xiangrui Meng
Remediation Overview

- Why remediating the metadata?
  - The goal is to improve searching and browsing, and support faceted search.

- What to remediate?
  - In this summer we focused on 2 aspects: consistency of metadata and enhancement of metadata.
Why a full knowledge base?

We can extract proper names (person, location, event) from text.

However, to really use them we need to match these names to a knowledge base.

By matching the names to the knowledge base, we uniquely determine the entity being referenced, and we obtain additional information.
A Simple Work Flow

LoC Collections

Proper Names

Knowledge Base (e.g. Wikipedia)

Proper Names

Match
Comparison of Three Candidate Knowledge Bases

<table>
<thead>
<tr>
<th>Knowledge base</th>
<th>Reliability of information</th>
<th>Size</th>
<th>Internal links</th>
<th>Classified</th>
<th>Descriptive information</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC Authority Files</td>
<td>Very high</td>
<td>Huge</td>
<td>No</td>
<td>Yes</td>
<td>Low</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>High</td>
<td>Large</td>
<td>Yes</td>
<td>No</td>
<td>High</td>
</tr>
<tr>
<td>Freebase</td>
<td>Reasonably high</td>
<td>Huge</td>
<td>No</td>
<td>Yes</td>
<td>Low</td>
</tr>
</tbody>
</table>
Pros and Cons – LC Authority Files

Pros:
- LC name authority files contains a huge amount of items, with high reliability.
- The items are well classified.

Cons:
- Each item has limited description, making it hard to disambiguate.
- LC authority files was not designed for detailed description of headings, but rather to uniquely identify them.
Pros and Cons – Wikipedia

Pros:
- Wikipedia redirects and internal links are extremely useful for our purpose.
- Each item in Wikipedia comes with a detailed description.

Cons:
- Wikipedia data is relatively hard to access.
- Wikipedia pages are not fully classified.
- Wikipedia has a smaller number of items.
Pros and Cons – Freebase

Pros:
- Freebase collects its data from different sources, with Wikipedia being the major one.
- 8,367,566 items are completely classified into 2,221 bases.
- The most related information are extracted for each base.

Cons:
- Limited description per item.
- Data is collected from different sources, making redundancy and consistency a problem.
So… What Works Best for us?

- We definitely want the Wikipedia links and redirects.
- We want the topics to be classified.
- We want the most related information extracted.
- The data should be easy to access.

- Is there even such a thing???
A Simple Solution!

Freebase Wikipedia Extraction (WEX) classifies Wikipedia pages.

WEX also contains other useful Wikipedia data in a very accessible format.

So, Wikipedia plus WEX is the solution!
Testing with Abraham Lincoln Collection

- Here we extracted all person names from the Abraham Lincoln collection.

- On the right shows the number of matched names.
Test Result

Using Wiki + WEX, we matched a total of 1500 unique person out of 23,407 recognized names.

We sampled and validated 300 matched names. The accuracy is about 90%.

Most incorrect matches can be eliminated by simple criterions.
From Data to Application

- What do we do with the enhanced metadata?
  - Les Fletcher suggested building a mobile phone application, so that when people visit a place they can use the app to search for books (or other artifacts) that is related to that particular location (or person).
  - David Gleich suggested hyperlinking the text in LoC to Wikipedia, so while browsing LoC collections the user can find out related information from Wikipedia.
Conclusion

- We compared 3 knowledge bases for metadata remediation and obtained preliminary results.
- The study shows that Wikipedia + Freebase WEX works brilliant for our purpose.
- The experiment on Lincoln collection yields promising results.
- The enhanced metadata has great potential in future applications.
Next Step

- Extend current work (person, events) to include more categories (geographical features, musical works, etc.).

- Study the problem of disambiguation: when a name matches several items in the knowledge base, which one shall we pick?

*This issue will be discussed in my next talk, so don’t go away!* 😊