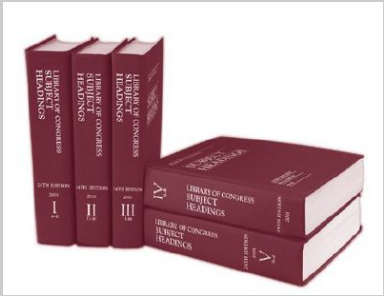
HIVE for LC Web Archives: Web Archives and Automatic Subject Indexing

IIPC, May 1, 2012

About the collaboration

- Metadata Research Center
  - University of North Carolina, School of Information and Library Science
- Library of Congress
  - Acquisitions and Bibliographic Access Directorate (includes LOC's Policy and Standards Division, which manages LCSH)
  - Technical assistance from LOC's Repository Development Group

Scope of the project



HIVE

- Matching a web set of web archives against a controlled vocabulary using text mining and natural language processing tools
- Public Policy Topics Web Archive - 500+ archived sites containing different viewpoints on a variety of American public policy topics
- Library of Congress Subject Headings - controlled vocabulary with hierarchical relationships designed to represent subject of materials in LOC's collections
- Helping Interdisciplinary Vocabulary Engineering (HIVE) is an IMLS-funded demonstration project exploring techniques for integrating multiple controlled vocabularies, including machine-aided indexing

Significance of the project

- Integrating machine-aided indexing into Library of Congress Web Archives cataloging workflow, which already includes manual LCSH assignment
- First HIVE experiment using LOC digital collections, LCSH, web archives
- First attempt at automated matching of LCSH to an LOC digital collection

Learning goals

- How well do general-purpose text mining techniques (such as those employed by HIVE) work with LCSH?
- How can LCSH be adapted for use by machine-aided indexing algorithms?
- How could these algorithms be adapted to work with LCSH?
- What is the best way to text mine LOC web archives specifically?

Web Archives

- Web sites are structurally different than other documents, presenting unique challenges to automatic indexing.
- Web pages may contain repeated information (headers/footers/menus)
  - *Can we correct this by only using the differences between pages?*
- Web sites do not have a clear beginning and end, so parameters must be defined
  - *How to define parameters (e.g., many pages deep (or hops) from the home page should we use to identify the topic of the site?)*

Evaluation process

- Compare LCSH headings extracted using the Maui algorithm to headings assigned manually
- Evaluate multiple parameter combinations and compared using F-measure (standard IR measure)

Parameters	**Values**
Differencing enabled	yes/no
Number of hops	0, 1, 2, 3
Stemmer	None, s-removal, Lovins, Porter
Minimum number of occurrences	1, 2
TF/IDF enabled	yes/no

Top results

Diff. Enabled	# Hops	Stemmer	Min Phrase Occur	TFxIDF enabled	Avg. # Correct Terms	Avg. F-measure
No	1	Porter	1	No	1.53	0.1050
Yes	3	Porter	1	No	1.51	0.1034
No	2	Porter	1	No	1.48	0.1011
Yes	1	Porter	1	No	1.44	0.0987
No	3	Porter	1	No	1.42	0.0987

Maui algorithm correctly identifies 1-2 LCSH headings per web archive.

Next steps

- LCSH is a large vocabulary, try using smaller subset.
  - Challenge: How best to build a subset?
- LCSH has unique features not present in other vocabularies
  - Pre-coordinated headings, subdivisions, qualifiers, meaningful patterns, punctuation
  - Challenge: Modify algorithm to better handle LCSH-specific features.

Q&A

Contact:

Rick Fitzgerald (LOC) (rfit@loc.gov)

Libby Dechman (LOC) (edec@loc.gov)

Craig Willis (UNC/MRC) (craig.willis@unc.edu)

Acknowledgements

- HIVE is supported by IMLS grant LG-07-08-0120-08
- The Maui algorithm was developed by Alyona Medelyan from the University of Waikato, New Zealand
- We would like to acknowledge Joan Boone, Jane Greenberg (UNC/MRC), Nicholas Taylor, Loche McLean, Pranay Pramod, Ed Summers (LOC) for their contributions to this project.

LIBRARY OF CONGRESS

Metadata Research Center <MRC>
UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

INSTITUTE of Museum and Library SERVICES

Maui/
Univ. Waikato

References

Library of Congress Web Archives: http://www.loc.gov/lcwa/

HIVE: http://ils.unc.edu/mrc/hive/

Maui algorithm: http://code.google.com/p/maui-indexer/