Subject metadata support powered by Maui

Olena Medelyan
Computer Science Dept.
Auckland University of Technology
Auckland, New Zealand
omedelya@aut.ac.nz

Vye Perrone
University of Waikato Library
University of Waikato
Hamilton, New Zealand
vperrone@waikato.ac.nz

Ian H. Witten
Computer Science Dept.
University of Waikato
Hamilton, New Zealand
ihw@waikato.ac.nz

Categories and Subject Descriptors
H.3.2 [**Information storage and retrieval**]: Information Storage – *record classification*

General Terms
Algorithms, Design, Human Factors.

Keywords
Metadata extraction, subject heading extraction, keyword extraction, web interface.

1. INTRODUCTION
Selecting subject headings and keywords is a chore for all metadata editors, who often leave these fields blank or incomplete—even when there are no guidelines and any word or phrase can be chosen. For example, tags are absent from the vast majority of citations in the social scholarly reference repository CiteULike. Libraries employ professional cataloguers and indexers to ensure consistent subject metadata in their records. Because this task is time-consuming, professionals and volunteers alike would welcome high-quality automatically generated suggestions for the main topics of a document.

The multipurpose automatic indexing tool Maui has recently been released [1]. Its performance has been evaluated in several settings and shown to be as consistent with people in choosing topics as people are with each other. In other words, it is hard to differentiate a set of subject headings assigned by a person from one automatically generated by Maui from the document text. Thus Maui is likely to provide substantive support for those who enter subject metadata. We demonstrate a web interface, implemented in Java Script and powered by Maui's open source Java library, that performs this task.

2. THE MAUI WEB INTERFACE
The web interface is designed for those who specify metadata for one document at a time. The interface is opened in a separate browser window, next to the web form for metadata editing. Two cases are supported: when the abstract alone is available, or when the full document is available in text, PDF or Microsoft Word format. Internally, Maui operates on text files, but in this interface we handle other formats using the Apache PDFBox and POI libraries respectively.

After entering the required document, the user specifies the controlled vocabulary (or vocabularies) from which the subject headings should be derived. Maui incorporates several built-in controlled vocabularies:

- Agrovoc – the agricultural thesaurus developed by the UN Food and Agriculture Organization (FAO);
- Medical Subject Headings (MeSH) – the medical vocabulary used by the National Library of Medicine for indexing PubMed and other digital libraries.
- High Energy Physics thesaurus – used by digital libraries like the CERN Document Server.
- Library of Congress Subject Headings (LCSH).

Unlike the first three, LCSH is domain independent. Although Maui has not been evaluated on LCSH in [1], we find that— with rare exceptions involving incorrectly disambiguated terms—it chooses appropriate descriptors.

As well as assigning terms from the chosen vocabulary, Maui also extracts as keywords the most prominent words and phrases that appear in the document text. In this setting, no vocabulary is used.

The button *Run Maui* produces a page showing the first paragraph of the document (for reference), below which appear the list of automatically extracted subject headings and the list of keywords. The 5 top ranked topics are presented in each case, but the user can see more suggested topics by pressing *See next 5* until no further topics remain.

Maui only takes a few seconds to generate the results. The librarian (or a volunteer indexer) can simply copy and paste the appropriate topics for the document into the metadata field. For libraries, this has a clear advantage in that subject headings (along with descriptive information) can be added by less qualified people, for example when cataloguing theses or institutional repository items. Experienced cataloguers can ensure quality control by quickly reviewing entries, or by spot checking.

3. REFERENCES
[1] O. Medelyan. 2009. Human-competitive automatic topic indexing. PhD thesis. University of Waikato. http://maui-indexer.googlecode.com.

Conference'04, Month 1–2, 2004, City, State, Country.