DRTC Workshop on
Semantic Web
8^{th} – 10^{th} December, 2003
DRTC, Bangalore

Paper: E

Extracting Meaningful Metadata

Michael Shepherd
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada B3H 1W5
shepherd@cs.dal.ca

Abstract

The paper identifies the importance of context based metadata extraction for more meaningful web. It further discusses context thesaurus approach for metadata extraction.

1. Introduction

Information retrieval systems require the matching of a query representation against the representation of the documents in the data set. However, the three most popular models of information retrieval, the Boolean model, the vector space model and the probabilistic model, rely on a representation of the document content that is sometimes referred to simply as "a bag of words". This is due to the fact that the representations are based on vectors of keywords that occur in the documents, often with statistical weighting, but with no semantics involved in the representations.

In order to realize the Semantic Web, we must have semantic-based representations of the content of documents (or at least what a document is "about"). This means not just a platform for the exchange of representations, but representations themselves that are semantically rich.

This presentation discusses the Semantic Web platform and some current approaches to extraction of metadata, some attempts to extract more semantically-rich metadata and a research project to extract more meaningful metadata from genres of documents found on the Web.

2. The Semantic Web

Web-based documents are keyword indexed by search engines and, frankly, these search engines seem to work pretty well. However, keyword indexing is not sufficiently powerful to realize the Semantic Web and to permit reasoning across the Web. In order to realize the Semantic Web, we need semantically-rich representations of the documents. The approach taken on the Web is to use metadata, i.e., information about information.

Dublin Core [1] has become the *de facto* standard for metadata and the Resource Description Framework (RDF) [2] has become the standard platform for carrying this metadata. However, the Dublin Core has only fifteen elements, most of which are used to describe the "container" rather than the content of the resource. Only the subject and description elements of Dublin Core really provide for any semantic description of the content of the resource.

The Web is now estimated to have about two billion pages and the deep or hidden Web is estimated to be about 500 times larger than this [3]. If we wish to describe documents with meaningful metadata, the sheer size of the Web means it must be done automatically. This leaves us with two major questions:

- How do we identify important terms, automatically?
- How do we supply some semantics, automatically?

3. Approaches to Extracting Metadata

There are two different approaches being used (or researched) to automatically extract metadata. The first approach uses a controlled vocabulary or classification scheme to assign metadata about the contents of the resource while the second approach tries to extract metadata about the resource itself, such as, the author, title, email address, etc.

The first approach attempts to match terms from the resource against terms in the controlled vocabulary and only select those terms that appear in the controlled vocabulary. A more interesting method [4] uses machine learning techniques to classify the resource against the Dewey Decimal Classification (DDC) scheme to find the appropriate class for that resource. It then can use Qualified Dublin Core to express both the DDC term and its class mark as metadata in RDF.

The second approach uses rule-based parsing and machine learning techniques to parse the header of journal articles to extract elements such as the author, author affiliation, email, etc. [5] The header is defined as either everything up to the end of the introduction or the end of the first page, whichever comes first. The goal is to create elements that can be used in addition to the Dublin Core elements to better describe the resource.

However, in my opinion, neither of these approaches add significant semantics to the metadata describing the resource.

4. Extracting Meaningful Metadata

Although Dublin Core has become the *de facto* standard for metadata on the Web, it is not semantically rich (my opinion). Often other metadata standards supplement Dublin Core. One such is the Gateway to Educational Material (GEM) [6] which defines eight education-specific elements and a range of element qualifiers.

An interesting approach by Paik [7] combines natural language processing with machine learning techniques to extract GEM metadata values from lesson plans. Such an approach had a very high degree of success in assignment of values to GEM elements. This provides an attribute-value relationship that is much richer semantically than Dublin Core and allows more "intelligent" search, such as:

- Search by audience
- Search by grade level
- Search by type of pedagogy

5. Context Thesaurus

The approach taken in our current research is based on the concept of genre and stereotypic documents representing genres. A genre is defined by form and content. It is a socially recognized form of communication with a community. Examples of genres include newspapers, detective novels and cowboy movies. They also include the scientific paper.

Purcell [8] developed stereotypic forms of medical documents, such as the clinical research paper, and found that the contextual structure characterizes the text without influencing the presentation, i.e., most clinical research papers have the same contextual structure even if their presentations differ. She found that sentences associated with a particular context may occur in different sections of a document.

We followed up on that research using sixty Web-based documents from the *Journal of the American Medical Association*. We found that all of these documents have the same structure, and that we were able to identify the following eight subsections of the documents quite easily:

- Comment
- Conclusions
- Context
- Design
- Measures
- Methods
- Objective
- Results

The keywords were weighted and only terms with high weights were retained. We then used a context-thesaurus approach [9] to assign keywords associated with each of these attributes. The following table shows the coverage or number of documents (out of 60) in which significant terms were assigned to each attribute.

Attribute	***Coverage***
Comment	25
Conclusions	21
Context	39
Design	35
Measures	53
Methods	52
Objective	22
Results	60

Such an approach allows us to extract more meaningful metadata that can be used to ask more explicit questions, such as, "What methods were used to treat obesity?"

We are currently extending this work to map three different medical journals to a common stereotypic structure. These are the *Journal of the American Medical Association, British Medical Journal and the Annals of Internal Medicine*. We have found that they all can be mapped to the following structure:

- Title
- Authors
- Background
- Objective
- Methods
    - Statistical Methods
- Results
- Conclusions
- Limitations/Biases
- Acknowledgements/Collaborations

- References

This research includes a structural parse of each article to identify the sections of the common template. This is followed by a syntactic parse to identify noun phrases and verb phrases. These noun phrases are mapped into MeSH and the Mesh Main Header and Qualifier terms are assigned to the identified sections of the article.

This research has not yet been evaluated. The research question is, "Is context-based indexing better than just keyword extraction (bag of words)?" However, this is not a simple question as a "bag of words" lets you do document retrieval while the attribute-value pair lets you do question answering and reasoning across the Web as well as document retrieval.

6. Summary

While the context-thesaurus approach does appear to extract more meaningful metadata, there is a cautionary note to be sounded if this approach is used for passage or section-retrieval as opposed to document retrieval. Bishop [10] has found that the disaggregation of knowledge, i.e., splitting documents into sections and only showing those sections, may lead to the loss of context and completeness and may encourage a restricted view of the field.

For the Semantic Web to be realized we need meaningful metadata that is generated automatically. We must provide a context for and disambiguation of terms. This requires a combination of different approaches, including information retrieval, natural language processing, artificial intelligence, domain analysis and ontology development and use.

7. References

1. Dublin Core Metadata Initiative. http://dublincore.org/
2. Resource Description Framework . http://www.w3.org/TR/REC-rdf-syntax/
3. Bergman, M.K. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, 7(1), 2001.
4. Jenkins, Charlotte; et al., Automatic RDF Metadata Generation for Resource Discovery, http://www.scit.wlv.ac.uk/~ex1253/rdf_paper/
5. Han et al., Automatic Document Metadata Extraction Using Support Vector Machines. *JCDL 2003*, pp. 37-48.
6. Gateway to Educational Materials. http://www.thegateway.org/welcome.html
7. Paik, Woojin. Automatic Generation of Educational Metadata Elements to Enable Digital Libraries. Proceedings of the International Conference on Computers in Education (ICCE) 2001. Seoul, Korea. Nov. 12, 2001.
8. Purcell, G.P.; Rennels, G.D. and Shortliffe, E.H. Development and Evaluation of a Context-Based Document Representation for Searching the Medical Literature. *International Journal on Digital Libraries*. 1997, 1:288-296.
9. Shepherd, Michael; Watters, Carolyn; and Young, June. Context Thesaurus for the Extraction of Metadata from Medical Research Papers. Hawaii International Conference on System Sciences, 5-8 January 2004, Hawaii.
10. Bishop, A.P. Digital Libraries and Knowledge Disaggregation: The Use of Journal Article Components. *Digital Libraries 98*, Pitssburgh, PA, USA, 1998. ACM.