Automatic Metadata Extraction from Multilingual Enterprise Content

Melike Şah
Knowledge and Data Engineering Group,
School of Computer Science and Statistics,
Trinity College Dublin, Dublin, Ireland
Melike.Sah@scss.tcd.ie

Vincent Wade
Knowledge and Data Engineering Group,
School of Computer Science and Statistics,
Trinity College Dublin, Dublin, Ireland
Vincent.Wade@scss.tcd.ie

ABSTRACT
Enterprises provide professionally authored content about their products/services in different languages for use in web sites and customer care. For customer care, personalization/personalized information delivery is becoming important since it re-encourages users to return to the service provider. Personalization usually requires both contextual and descriptive metadata. But current metadata authored by content developers is usually quite simple. In this paper, we introduce an automatic metadata extraction framework, which can extract multilingual metadata from the enterprise content, for a personalized information retrieval system. We introduce two new ontologies for metadata creation and a novel semi-automatic topic vocabulary extraction algorithm. We demonstrate and evaluate our approach on the English and German Symantec Norton 360 technical content. Evaluations indicate that the proposed approach produces rich and high quality metadata for a personalized information retrieval system.

Categories and Subject Descriptors
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *semantic networks*; I.7.5 [**Document and Text Processing**]: Document Capture – *document analysis*.

General Terms
Algorithms, Design and Experimentation.

Keywords
Metadata Generation, Semantic Web, Ontologies, Personalization.

1. INTRODUCTION
Enterprises offer highly technical and professionally authored information about their products and services for use in technical manuals, web site documents, help files and customer care. With the advancements of the Web, today's users have different expectations from enterprises within a very competitive marketplace. Particularly users prefer advanced personalized customer support services in their preferred languages. Because of this, enterprises focus on customer support. There is also a growing interest to personalization within enterprises, because of the size of the available information to users, different kinds of user personas and personalization re-encourage users to return to the service provider. However, enterprises usually generate quite simple metadata during the authoring of content and this metadata is not sufficient for the personalization of the content.

An approach to support users is to provide better Information Retrieval (IR) systems. IR can be enhanced with descriptive metadata about content and semantic knowledge about document structures. Information can also be personalized to users with varying background and needs. In this context, personalized IR systems are becoming more popular since they try to improve traditional IR by decreasing information overhead of users and providing the best information according to their needs. For supporting personalized IR, rich and good quality metadata about content are necessary. However, it is difficult to automatically generate metadata for personalization since this requires deep understanding of concepts and context within the document. Our motivation is to support a personalized IR system by automatically generating rich machine processable metadata from multilingual enterprise content. In this paper, we propose an automatic metadata extraction framework, new ontologies for metadata creation and a new topic ontology extraction algorithm from DocBook indexterms. Our approach is evaluated on the English and German Symantec Norton 360 technical content.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0099-5/10/10...$10.00.

2. RELATED WORK
Automatic metadata generation techniques have been applied to many fields, such as learning object repositories [1] and digital libraries [3]. Generally, these techniques extract metadata from 1) *document content* by analyzing the object itself, 2) *document context* (i.e. from digital environment where the object resides or by analyzing semantic context), 3) *document usage* (i.e. log files, number of downloads, etc.), and 4) *document structure* by examining document storage structure. There is little research on automatic metadata generation from the enterprise content. In [2], researchers describe a process to automatically extract a domain specific vocabulary from unstructured enterprise domain using Linked Open Data. The process is based on extracting domain-specific terms by applying IR techniques. We approach to vocabulary generation differently. We introduce a new ontology for enterprise domain using existing structured content of the enterprise (DocBook). We also apply different techniques to extract metadata from the document content, structure and context for a personalized IR system for customer support.

3. ENTERPRISE CONTENT
Enterprises provide highly technical information about products and services of an organization. In general, this content is professionally authored and published in a number of formats in multiple

languages. In our case study, we use the English and German versions of the Symantec Norton 360 technical documentations. The content is formatted in XML and structured with DocBook DTD [5]. DocBook DTD is a common vocabulary for describing technical documentation using SGML or XML and widely used by many enterprise organizations, such as Sun Microsystems, Microsoft, Hewlett Packard, Red Hat and Symantec. It is very broad and complex since it covers numerous variations and options about the domain.

The objective is to generate a common vocabulary (ontology) for enterprise content and to extract rich multilingual metadata from DocBook documents in the form of RDF. The generated metadata then can be used for a personalized customer care. The next section explains new ontologies developed for this research.

4. ONTOLOGIES AND SCHEMAS

4.1 DocBook Ontology

In order to extract structural metadata about DocBook documents in RDF format, we created a new ontology from DocBook DTD, which we call DocBook Ontology. The DocBook Ontology is domain independent thus it does not have to be created every time an enterprise needs localization in different subject domains and is reusable across different domains of documents. The ontology is manually constructed in OWL Lite using the Protégé. The excerpt of the classes and properties of the DocBook Ontology are illustrated in Figure 1. The DocBook ontology can also serve as a common vocabulary for other enterprises since many enterprises use DocBook DTD for formatting their content.

DocBook documents mainly have *book*, *chapter* and *section* elements. The book is the most common top level element and has a number of sub-components such as *preface*, *chapter*, *appendix*, *bibliography*, etc. In our ontology, the book class have dc:hasPart relationship to its sub-components as shown in Figure 1. The components generally contain block elements. Therefore, we generated *blockelements* class and it is further divided into sub-classes, such as *para*, *procedure*, *tables*, etc. Each sub-class may have other sub-classes such as *para*, *informaltable*, *step*, etc. Components usually have dc:hasPart relationship to these BlockElements sub-classes. In addition, components may contain *sections* using dc:hasPart relationship. Sections can also contain block elements and are often recursive (e.g. nested); one section may use the content of another section. We have used *subsection* relationship to state this kind of relationships.

In DocBook documents, information is usually re-used such as different books may share same chapters. To enable sequential navigation of instances under a component, we introduce *Sequence* class. Every sequence instance has data about the parent instance (i.e. dc:hasPart relationships is used to determine parents) and the sequence number under this parent. In the DocBook ontology, we covered most of the elements and attributes defined by DocBook DTD version 5.0, which resulted in a complex ontology since DocBook DTD itself is very broad. In our ontology, all elements and attributes are optional as in the DocBook DTD. In addition, in order to facilitate re-usability and interoperability, we used RDF bindings of DC metadata elements within the DocBook ontology to describe some of the DocBook DTD elements. For this purpose, we created metadata mappings between DocBook DTD and DC elements, such as DocBook *author* attribute is mapped to *dc:creator*, *indexterm* element is mapped to *dc:subject*, etc. Besides, DC elements, *language*, *source*, *created* and *modified* are re-used.

Symantec uses a subset of the DocBook DTD to structure Norton 360 technical content. Thus, we generated a subset of the DocBook ontology, which we call the *Symantec Profile*. We extended the Symantec Profile with other schemas for creating metadata for personalization. For instance, every section instance must have metadata about the difficulty, interactivity level, interactivity type from IEEE LOM standard and metadata about processes from a Process Ontology and a metadata about subject from a Topic Ontology. Next sections explain these ontologies.

4.2 IEEE Learning Object Model (LOM)

DocBook documents need descriptive information about cognitive metadata such as difficulty that can be useful for personalization. However, such cognitive metadata is not supported by DocBook DTD. Thus, we reused IEEE LOM (http://ltsc.ieee.org/), which provides useful information about information objects. We analyzed the LOM elements and DocBook documents and choose three entities from LOM, such as difficulty, interactivity type and interactivity level for personalization.

4.3 Process Ontology

The semantics of DocBook DTD elements can reveal further information about the document. For example, *Procedure* element describes an interactive task using *Step* sub-elements. On the other hand, elements *Para*, *Note*, *Table*, etc. provide information about concepts. To model this, we developed a new ontology, called *Process Ontology*. The Process Ontology has a *ProcessType* class which is divided into two sub-classes: *Activity* and *Concept*. Activity class represents an action which is performed by a human agent and it has one instance, namely, *Task*. Concept class represents an information object that is covered by a document. It has six instances; *NarrativeText, Table, Image, Hyperlink, Example* and *Summary*. In DocBook ontology, sections have *processType* relation to ProcessType instances. In the Symantec Profile, every section must have at least one value for processType. This ontology can be used for personalization. Based on the user's information needs, best resources can be presented; if the user is looking for "Overview documents", then documents annotated with Summary can be ranked higher.

4.4 Semi-Automatically Extracting a Topic Vocabulary from Indexterms

DocBook documents may have *indexterm* element, which represent the subject of the document. In the DocBook DTD, originally indexterm is designed to present alphabetic index of keywords at the end of a book. Indexterms can also be facilitated to create a controlled vocabulary, since the primary term describes the main topic of the document and the secondary term is a sub-topic or sometimes an attribute of the primary term.

In our case study, English and German domains use *primary* and *secondary* terms to describe the subject of the document. We also observed that the secondary term may not be very informative by itself and indexterms for the same document may contain variations of same topic (pri: firewall, sec: rules and pri: rules, sec: firewall). Thus, we re-purposed indexterms to semi-automatically create a controlled vocabulary using Simple Knowledge Organization System (SKOS).

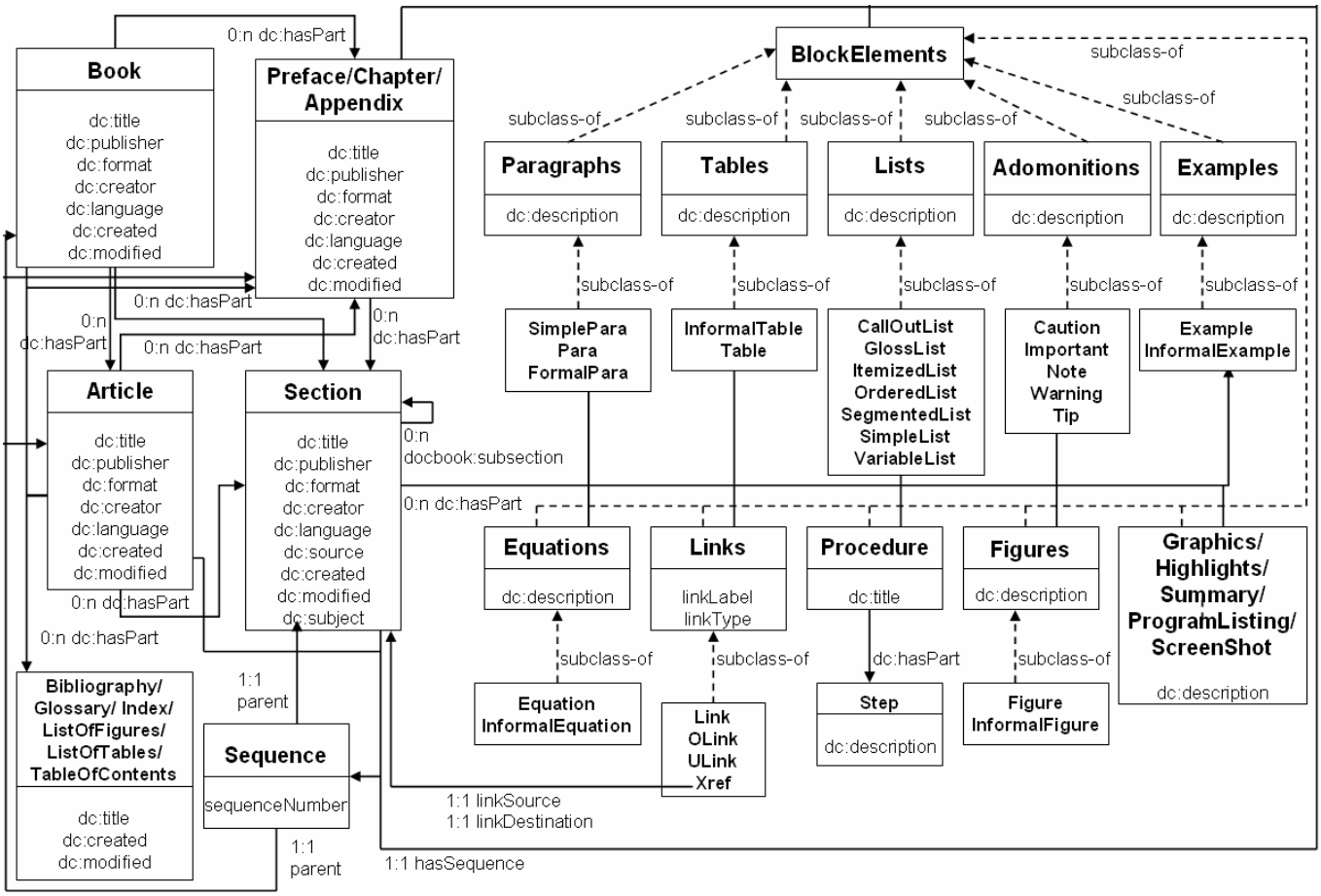

Figure 1. The overview of the DocBook ontology

The algorithm initially uses a script (Javascript) to extract primary and secondary terms from documents and combines them (primary+secondary). We combine two terms since the combined term is more informative and unambiguous. The algorithm then states that the primary and primary+secondary term are instances of *skos:Concept*. Subsequently, it is declared that the primary term is the *skos:broader* of the combined term and the combined term is the *skos:narrower* of the primary term. The document is annotated with the combined term using *dc:subject*. Indexterms contain variations of the same topics in the case study and the algorithm generates syntactically different but semantically duplicated terms. Manual cleaning is performed to remove duplicates using Protégé. Besides, terms do not have broader topics are analyzed and if possible manually replaced under a concept. We tested our algorithm on the English Norton 360 and generated a controlled vocabulary, which we call Topic Ontology. The Topic Ontology contains 1089 topics, 40 root topics and the longest depth in the hierarchy is four. This ontology can be used for user modelling; user's knowledge and interests can be linked to ontology concepts for personalization.

5. AUTOMATIC METADATA EXTRACTION FRAMEWORK

The automatic metadata generation framework works as follows: First DocBook documents are crawled and document names/IDs are extracted. Then, the document names are feed into the framework. Since, DocBook documents are valid XML, we parse these documents using Javascript and XML DOM. Following this, we use three Javascript functions to extract metadata as described below. Each script creates metadata in RDF turtle format. The created metadata by individual scripts are then unified and stored to a triple store using Jena or Protégé, which then can be accessed by applications using SPARQL. Our framework can also be applied to multilingual DocBook documents to extract multilingual metadata.

Extraction of Concept Instances and Structural Relations: By analyzing DocBook document structure, instances of DocBook ontology concepts are generated using unique DocBook element IDs (e.g. Book). Besides, relationships between instances are extracted by analyzing the document. For example, an instance may have dcterms:hasPart relations to sub-components or sections may have docbook:subsection relations to sub-sections.

Extraction of Instance Attributes and Properties: Title and subject metadata can be extracted by analyzing the DocBook document. In addition, links between documents are also extracted by extracting *link* elements and their attributes. Moreover, we extract creation date, modification date and source from system properties, and add publisher, creator, language and format metadata within this process.

Data Mining to Generate Metadata for Personalization: In order to support personalization, rich and useful metadata about resources are required. In our case study, we generate metadata for personalization using Process Ontology and LOM. By analyzing DocBook elements, metadata for processType can be created (Table 1). For example, if the document contains *Step* element, then processType is set to *process:Task*. In addition, for each processType, we estimate the covering percentage of each process type within the document by comparing size of the process type content to the size of the document (byte comparison). Moreover, we use fuzzy logic to create metadata values for difficulty, interactivity level and interactivity type. The details of this approach are not in the scope of this paper.

Table 1. ProcessType values based on DocBook DTD

DocBook DTD elements	ProcessType value
Step	process:Task
Summary	process:Summary
Example, InformalExample	process:Example
Paragraphs, Adomonitions, Lists, Tables	Process:NarrativeText
Figure, InformalFigure	process:Image
Link, OLink, ULink, Xref	process:Hyperlink

6. EVALUATIONS

To assess the metadata quality, we applied Ochoa and Duval's metadata quality metrics: completeness, accuracy, provenance, conformance to expectations, logical consistency/coherence, timeliness and accessibility [4]. The formulas of metrics are summarized in [4]. *Completeness* is the degree to which a metadata record represents all the information needed to have an ideal representation. Metadata standards and application profiles provide information about non-null metadata fields using mandatory element fields. In our case study, the ideal representation differs based on class types. For section instances, the ideal metadata record has the title, subject, dc:hasPart (at least a link to block components), difficulty, interactivity type, interactivity level, process type, format, source, created, modified, publisher, creator and language. *Weighted completeness* not only counts non-null metadata fields but also weights each field according to relative importance to the application. *Accuracy* measures the correctness of metadata values, usually by the manually entered values by experts. A user study is resumed but has not been completed. *Provenance* represents the origin of the metadata. We generate metadata about creation/modification date, creator, publisher and source that are useful for provenance. *Timeliness* represents the degree to which a metadata record remains current and useful over time. This could be calculated using the age of the record or frequency of usage. Accessibility metric measures the degree to which a metadata record is accessible both in terms of logical and physical accessibility. The logical accessibility measures readability and physical accessibility is how easy to find a metadata record in the repository (linkage). *Linkage* value is equal to number of other records that reference to it. *Consistency* measures the degree to which a metadata record matches a standard definition. For example, LOM suggests that if an object's interactivity type is active, then it should have high values of interactivity level. *Conformance to expectations* measures the degree to which the metadata record fulfills the requirements of a given community of use: vocabulary terms should be meaningful for users (we reuse DC and LOM, which are well established vocabularies), metadata values must be filled to perform a specific task (this can be measured by weighted completeness) and the amount of information should be enough to describe the information object for a specific task. Previously we discuss how metadata generated by process and topic ontologies can be used for personalization.

We applied the quality metrics to the metadata extracted from Symantec Norton 360 English and German contents (except linkage, all metrics are normalized from scale 0 to 10). Our framework extracted metadata about 639 and 671 documents from English and German domains respectively. Completeness in English domain is 9.94 and 9.91 in German domain. We observed that in both domains, there are number of documents which do not have a subject. Since subject is important, there is a small reduction in overall weighted completeness; 9.86 for English and 9.79 for German domains. In weighted completeness, the dc:title, dc:subject and dc:hasPart relations weighted 1, and other elements weighted 0.2 based on [4]. In addition, we applied LOM standard consistency suggestions to both domains and it was observed that metadata values are consistent with the standard (10.00). Since metadata is generated within the same week of evaluations, the metadata are up-to-date (timeliness is 10.00 (age) in both domains). Linkage between instances are also calculated, where in English domain, average linkage is 4.40 and in German domain average linkage is 3.58, which shows that documents can also be accessed through semantic relationships between them. Overall metadata quality tests showed that the extracted metadata in both domains have high completeness, up-to-date, consistent and contains an average of at least 3 links between instances.

7. CONCLUSIONS AND FUTURE WORK

We have presented a novel framework which extracts rich multilingual metadata for a personalized information retrieval system from enterprise content. Two new ontologies are also developed for metadata generation: DocBook ontology and Process Ontology, as well as a semi-automatic topic vocabulary extraction algorithm is introduced. We applied our framework to the English and German Symantec Norton 360 contents. Evaluations indicate that the proposed approach produces rich and high quality metadata. In future, we will perform a user study to measure the accuracy of the generated metadata.

Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University of Dublin, Trinity College.

8. REFERENCES

[1] Cardinaels, K., Meire, M, and Duval, E. 2005. Automating metadata generation: the simple indexing interface. *In Proc. of WWW*, 548-556.

[2] Dolby, J., Fokoue, A., Kalyanpur, A., Schonberg, E. and Srinivas, K. 2009. Extracting enterprise vocabularies using linked open data. In *Proceedings of ISWC*.

[3] Greenberg, J. 2003. Metadata extraction and harvesting: a comparison of two automatic metadata generation applications, *J. of Internet Cataloging*, 6, 4, 59-82.

[4] Ochoa, X. and Duval, E. 2006. Quality metrics for Learning Object Metadata. In *Proceedings of ED-MEDIA*.

[5] Walsh, N. and Muellner, L. 1999. The DocBook Definitive Guide, *O'Reilly Media*.