Accepted Manuscript

Automatic Metadata Mining from Multilingual Enterprise Content

Melike Şah, Vincent Wade

PII: S1570-8268(11)00080-1
DOI: 10.1016/j.websem.2011.11.001
Reference: WEBSEM 253

To appear in: *Web Semantics: Science, Services and Agents on the World Wide Web*

Received Date: 11 April 2011
Revised Date: 4 November 2011
Accepted Date: 4 November 2011

Please cite this article as: M. Şah, V. Wade, Automatic Metadata Mining from Multilingual Enterprise Content, *Web Semantics: Science, Services and Agents on the World Wide Web* (2011), doi: 10.1016/j.websem.2011.11.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Automatic Metadata Mining from Multilingual Enterprise Content

Melike Şah and Vincent Wade
Knowledge and Data Engineering Group, Trinity College Dublin, Ireland
{Melike.Sah, Vincent.Wade}@scss.tcd.ie

Abstract. Personalization is increasingly vital especially for enterprises to be able to reach their customers. The key challenge in supporting personalization is the need for rich metadata, such as metadata about structural relationships, subject/concept relations between documents and cognitive metadata about documents (e.g. difficulty of a document). Manual annotation of large knowledge bases with such rich metadata is not scalable. As well as, automatic mining of cognitive metadata is challenging since it is very difficult to understand underlying intellectual knowledge about document automatically. On the other hand, the Web content is increasing becoming multilingual since growing amount of data generated on the Web is non-English. Current metadata extraction systems are generally based on English content and this requires to be revolutionized in order to adapt to the changing dynamics of the Web. To alleviate these problems, we introduce a novel automatic metadata extraction framework, which is based on a novel fuzzy based method for automatic cognitive metadata generation and uses different document parsing algorithms to extract rich metadata from multilingual enterprise content using the newly developed DocBook, Resource Type and Topic ontologies. Since the metadata generation process is based upon DocBook structured enterprise content, our framework is focused on enterprise documents and content which is loosely based on the DocBook type of formatting. DocBook is a common documentation formatting to formally produce corporate data and it is adopted by many enterprises. The proposed framework is illustrated and evaluated on English, German and French versions of the Symantec Norton 360 knowledge bases. The user study showed that the proposed fuzzy-based method generates reasonably accurate values with an average precision of 89.39% on the metadata values of document difficulty, document interactivity level and document interactivity type. The proposed fuzzy inference system achieves improved results compared to a rule-based reasoner for difficulty metadata extraction (~11% enhancement). In addition, user perceived metadata quality scores (mean of 5.57 out of 6) found to be high and automated metadata analysis showed that the extracted metadata is high quality and can be suitable for personalized information retrieval.

Keywords: Automatic metadata generation, ontologies, personalization, fuzzy information granulation and fuzzy inference.

1 Introduction

Personalization has proven to increase user motivation and user satisfaction with the information experience [47 - 51]. Within enterprises, there is also growing interest to personalized customer care since personalization helps users to stay more on the website and re-encourages them to return to the service provider. To support personalization, enterprises need rich metadata about the content. However, automatic generation of such metadata is challenging. We first discuss these challenges and our motivations. Then, we explain our rationale and approach for automatic metadata mining from multilingual enterprise content.

1.1 Problem Statement and Motivations

Enterprises generally produce highly technical and professionally authored content about their product and services for use in technical manuals, web sites, help files and customer care. However, they often generate simple/limited metadata about this content (e.g. title and subject of the document), which is not sufficient to support personalization. In contrast, the marketplace is very competitive and customer support is getting more important since users have high expectations from service providers. In particular, users prefer advanced personalized customer support services in their preferred languages [19] (e.g. personalized task assistance about a product in a preferred language). The ARCHING system [34] [41] is an example that illustrates personalized Information Retrieval (IR) and personalized presentation for real-life enterprise customer support scenarios. The system showed that adaptive selection and navigation techniques improved the user experience in terms of task assistance and user satisfaction, where users were more motivated to read and engage more with the system compared to a non-adaptive system. Many successful personalized systems are also discussed in [55]. To provide personalized services, enterprises face with poor metadata and lack of any cognitive metadata (i.e. metadata coming from perception, reasoning, or intuition such as difficulty of a document), which is very useful for personalization purposes.

Most enterprises use internal authoring mechanisms, such as content management systems to generate professionally authored content and often simple metadata about this content. In many cases metadata is structured but not descriptive enough that can be useful for personalization. Descriptive metadata about documents such as subject and difficulty can be created manually either at authoring stage or later by an expert. The problem is that within enterprises usually different parts of the content is authored and owned by different content authors. Thus it is difficult to provide immediate changes on metadata schema or alter metadata creation cycle dramatically. Alternatively, an expert can add rich metadata about the content following to authoring process. However, manual metadata generation is time-consuming, costly, requires trained staff and when metadata creation is not done by experts, the process may be error prone. Therefore, automatic metadata extraction techniques are required for scalability and many automated techniques are

ACCEPTED MANUSCRIPT

proposed to extract descriptive (e.g. title, author, topic), structural (e.g. instance relationships) and administrative metadata (e.g. modification date). Popular methods are rule based extraction [12] [18] [23] [24] [28], learning based systems [17] [25], natural language processing techniques [26] [37], Information Retrieval (IR) methods [40], ontology-based systems [1] [22] [7] and linguistic analysis (e.g. LDA) [30].

On the other hand, personalization needs rich cognitive metadata about documents, such as difficulty of the document, sentiment of the document, etc., which can be very useful for personalization. However, automatic cognitive metadata mining is difficult since it is harder to understand the context and underlying intellectual knowledge about documents automatically, as well as the feasibility of the approach such as pre-processing, training, extraction speed and precision are important measures for deployment. Moreover, cognitive metadata values themselves are often fuzzy and subjective to different users, which requires intellectual reasoning with unclear measures. As a result, there are few techniques for cognitive metadata extraction as we discuss in section 6.3. Additionally, increasing amount of data published on the Web is non-English and metadata extraction systems need to address this challenge by able to process and extract metadata from multilingual content in order to meet the changing dynamics of the Web content.

1.2 Rationale and Approach

In our research, we took an action research based approach; a specific business case with real content, real customer care processes and real end-users are taken. Our objective is to provide a personalized IR system to end-users using the enterprise content. But, first of all, rich metadata from the content have to be extracted to be employed by the personalized IR system. Then, a personalized customer care system can utilize the automatically extracted metadata by our framework as shown in [34] [41]. This paper focuses on the content analysis and metadata generation aspects of the research, where we propose an automatic metadata extraction framework which is based on a novel fuzzy based method for cognitive metadata extraction and different parsing algorithm to extract rich metadata from multilingual content (language neutral).

In our approach, we investigate fuzzy inference for cognitive metadata mining since fuzzy inference is tolerant to imprecision and may suit well for the extraction of cognitive metadata values that are often fuzzy and imprecise. We evaluate our fuzzy system against a rule-based reasoner to illustrate this advantage in the evaluations section. In particular, our framework uses fuzzy information granulation and Mamdani fuzzy inference system to automatically extract cognitive metadata about difficulty, interactivity level and interactivity type of semi-structured documents. These metadata elements are chosen since they are suited for personalization as they greatly enhance the ability of a personalization system to tailor the personalized presentations to the appropriate depth, degree of complexity and level of interactivity as well as they were suited for the envisioned personalized IR scenarios presented in [34] [41]". In addition, these metadata values can be mined automatically from DocBook formatted documents as we discuss in detail in section 2.3.3.2. On the other hand, our framework uses different document parsing algorithms to generate descriptive, structural and administrative metadata using the introduced DocBook, Resource Type and Topic ontologies for enterprise content, which can be applied to other domains.

The methods and approaches developed are evaluated with a case study on the English, German and French versions of Symantec Norton 360 technical knowledge base articles, where Symantec Norton 360 is just an example of a common kind of corporate documentation format that uses DocBook. Symantec Norton 360 provides structured data about product manuals and online help documentations of the security product Norton 360. The content is formatted in XML and structured with a subset of DocBook Document Type Definition (DTD) (see section 3 for more details). However, the paper points to the general applicability of the approach because of the use of standard metadata formats, content analysis and general algorithms. Evaluations indicate that the proposed framework extracts reasonably accurate metadata values with an average precision of 82% to 96% depending on the metadata field. In addition, the quality of the extracted metadata is found to be high and can be useful for supporting personalized IR on the English, German and French versions of the Symantec Norton 360 content.

A part of this work has been presented in [32] [42]. In this paper, we extend the explanations and evaluations of the framework. The rest of the paper is organized as follows: In section 2, we explain the proposed metadata extraction framework, the developed ontologies for metadata generation, and the metadata extraction techniques from enterprise content. Section 3 describes the case study on Symantec Norton 360 technical documentations. Section 4 introduces a novel fuzzy based metadata generation method. Section 5 presents the evaluations and experimental results. Section 6 discusses the related work. Section 7 is conclusions and future work.

2 Automatic Metadata Extraction from Enterprise Content

In this section, first we explain the enterprise content in section 2.1. Then, the architecture of the proposed framework is discussed in section 2.2. Finally, we explain the newly developed ontologies and metadata extraction process from DocBook documents in section 2.3.

2.1 Enterprise Content

One of the most commonly used eXtensible Markup Language (XML) schema for enterprise content is DocBook Document Type Definition (DTD). DocBook DTD is a unified vocabulary for describing documentations, which is defined by SGML or XML [35]. DocBook was originally designed to enable the interchange of computer documentation between companies. Sun Microsystems, Microsoft, Hewlett Packard, Red Hat and Symantec are among the organizations that use DocBook DTD to structure their content. DocBook DTD is very broad and complex, since it

ACCEPTED MANUSCRIPT

covers numerous variations and options about the domain. In addition, DocBook DTD is not just limited to enterprise domain. It has been used in other domains as well, such as for creating learning object resources in elearning [15] [29]. An example DocBook document is presented in Figure 1. The most common DocBook document elements are *book*, *chapter* and *section.* As can be seen from Figure 1, DocBook documents are very structured and the content is described in detail using DocBook tags.

////////////////////////////Figure 1 Here////////////////////////////////////

2.2 Automatic Metadata Extraction Framework

Our metadata extraction framework extracts metadata from documents formatted by DocBook DTD, where the architecture of the framework is presented in Figure 2. The system components are implemented as separate Web applications and framework works as follows: First DocBook documents are crawled and document names/IDs are extracted. The document names are then fed into the framework. Since DocBook documents are valid XML files, we parse these documents using XML Document Object Model, which is supported by Javascript. Following this, three different Web applications extract metadata from documents 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. The extracted metadata can be accessed by using SPARQL query language.

////////////////////////////Figure 2 Here ////////////////////////////

2.3 Metadata Extraction from DocBook Documents

To provide personalization, the domain has to be described in sufficient detail in a structured format. For this purpose, we developed three new ontologies namely, DocBook, Topic and Resource Type from enterprise domain. In addition, for interoperability, we re-used Dublin Core metadata [10] and IEEE Learning Object Model metadata [21] to describe descriptive and cognitive metadata about documents.

2.3.1 DocBook Ontology

In order to extract structural metadata from DocBook documents in RDF format, we created an ontology from DocBook DTD, which we call *DocBook Ontology* [52]. DocBook ontology describes how different documents are structured and related. In addition, it captures structured information about how document sub-parts are organized, which can be very useful for the re-organization and personalization of the content for needs of the user.

DocBook Ontology is developed from DocBook DTD version 5.0, where we took DocBook DTD in XML and developed a schema in OWL by utilizing existing vocabularies whereas possible and creating new classes, properties or relations where necessary. We constructed the ontology in OWL Lite using the Protégé ontology editor in our own namespace (http://cngl.ie/ontologies/2010/01/docbook). The DocBook ontology is domain independent; do not have to be created every time an enterprise needs localization in different subject domains. In addition, we can extract metadata from any enterprise content formatted by DocBook DTD using our metadata extraction framework and DocBook ontology. The excerpt of the classes and properties of the DocBook Ontology are illustrated in Figure 3.

////////////////////////// Figure 3 Here ////////////////////////////////

Generally, DocBook documents have *book*, *chapter* and *section* main elements. The book is the most common top level element and has a number of sub-components, for example *preface*, *chapter*, *appendix*, *bibliography*, etc. Components generally contain block elements (e.g. tables, figures, paragraphs, lists, etc.) and/or *sections*. Section elements contain block elements and are recursive (e.g. nested). We encoded these DocBook elements as OWL classes as shown in Figure 3. In addition, block elements, such as Tables, Lists, etc are sub-class of *BlockElements* class. In DocBook documents, information is usually re-used; sections may contain other sections or different chapters may share same sections. We have used *dc:hasPart* relation to represent part-of relationships and *docbook:subsection* for indicating relations between section instances. To enable sequential access to sub-instances of a concept, (e.g. sections of a chapter), we introduce *Sequence* class. Every sequence instance has data about the parent instance 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, instead of creating new entities/attributes/properties, we re-used Dublin Core (DC) metadata standard simplified version [10]. We created metadata mappings between DocBook DTD and DC. For instance, DocBook *title* attribute is mapped to *dc:title*, DocBook *author* attribute is mapped to *dc:creator*, DocBook *publisher* element is mapped to *dc:publisher*, DocBook *format* attribute is mapped to *dc:format* and DocBook *indexterm* element and its sub-elements *primary* and *secondary* is mapped to *dc:subject*.

Mining DocBook Instances and Instance Relationships: A script analyzes the DocBook document structure and instances of DocBook ontology concepts are generated such as instances of Book, Chapter, Section, Para, Procedure, etc. For this purpose, unique element IDs are utilized as instance names (e.g. ID:123456). Then, these unique instance names are used to generate global instance URIs using the base URI of the corpus domain (e.g. http://cngl.ie/ontologies/2010/01/symantec#123456). In addition, relationships between instances are extracted by parsing and analyzing the XML tree structure of DocBook documents. For example, an instance may have dcterms:hasPart relationship to sub-components (e.g. Chapters of a Book, Sections of a Chapter, etc.) or Section instances may have docbook:subsection relationship to sub-sections. This kind of metadata is known as structural

ACCEPTED MANUSCRIPT

metadata that represents the physical and logical structure of a complex object. Structural metadata can enable more effective utilization of resources within a user interface and can be useful to retrieve/browse relevant information objects according to their organization. A relevant work for automatic taxonomy-based annotation of DocBook documents are presented in [15], which is similar to our approach. However, in our approach, we do not only capture taxonomic relations, but we extract link type relations, instance attributes and properties, as well as analyze more detail semantics of the document to generate rich metadata for personalization.

Mining Instance Attributes and Properties: Instance attributes (e.g. title, subject, author, etc.) can be extracted by parsing and analyzing DocBook documents and from system properties. In particular, we analyze DocBook documents to extract document title, subject, creator and identifier. Creation date, modification date and source (e.g. file URL) are extracted from system properties and information about publisher, language and document format is added within this process. In addition, links between documents represented by *xref, link, olink* and *ulink* elements can be extracted by analyzing the content. Before creating a link between class instances, first the algorithm checks if the target document exists, in order to prevent broken links. Title, creator and subject provide descriptive metadata that is useful for identification and discovery of resources. Creation/modification date, file format, source and publisher provides administrative/provenance metadata that can help to manage a resource and track its origins. On the other hand, subjects of documents can be represented as semantic instances of an ontology rather than keywords (text). For IR, this can be very useful since relevant documents can be retrieved based on relationships between similar topics. To support IR, we semi-automatically developed a Topic Ontology as explained below.

2.3.2 Topic Ontology

In DocBook DTD subjects are represented as a list of keywords using *indexterm* element. *Indexterm* element has *primary* and *secondary* sub-elements to describe the subject of the document as shown in Figure 1. In the DocBook DTD, originally indexterm element is designed to present alphabetic index of *keywords* at the end of a book. However, indexterms can also be facilitated to generate 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. Therefore, we re-purposed indexterms to semi-automatically create a controlled vocabulary using Simple Knowledge Organization System (SKOS). SKOS is a simple knowledge representation standard to supply thesauri, classification schemes and taxonomies [33]. Representing subjects of documents with an ontology provide many advantages; a controlled vocabulary can be very useful for user modeling and providing personalized IR. For example, user's interests, expertise or knowledge of subject concepts can be obtained explicitly or implicitly and then relevant documents can be retrieved using the relationships between ontology concepts since documents are annotated with concepts of the ontology.

Algorithm 1. Generate_Topic_Taxonomy(*S*)
Input: *S* – number of Section documents formatted by DocBook DTD, i.e. $S = S_1, S_2, \ldots, S_n$ (n>1)
Output: *Topic_Ontology* – Topic taxonomy formatted by SKOS
Process:
1. **for** i=1 **to** n **do**
2. S_i*_DOM* = DOM tree of S_i (parsed by XML DOM using Javascript)
3. **if** (S_i*_DOM* **contains** *primary* **and** *secondary* tags) **then**
4. **for** j=1 **to** k **do**
5. $combined\_term_j = primary_j + secondary_j$
6. $Topic\_Ontology += primary_j\ secondary_j$, $combined\_term_j$ **isa skos:Concept**, $primary_j$ **isa skos:Concept**
7. $combined\_term_j$ **skos:broader** $primary_j$
8. $primary_j$ **skos:narrower** $combined\_term_j$
9. S_i **dcterms:subject** $combined\_term_j$
10. **end for**
11. **end if**
12. **end for**
13. **Store** Topic_Ontology to a file
14. **Delete duplicated** terms using Protégé
15. **Add skos:broader** and **skos:narrower** to uncategorized terms by analyzing the context using Protégé

In our semi-automatic topic taxonomy extraction algorithm which is presented in Algorithm 1, term extraction and generation of the initial taxonomy is automatic. Cleaning of duplicates and more categorization of the terms are performed manually. 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 secondary term itself may be vague but the combined term is more informative and unambiguous. The algorithm then states that the primary and the 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. In addition, the document is annotated with the combined term using dc:subject. We applied the algorithm to our case study in Norton 360 technical knowledge base articles. In this domain, indexterms contain variations of the same topics (e.g. primary: rules, secondary: firewall or primary: firewall, secondary: rules), our automatic extraction algorithm generates syntactically different but semantically duplicated terms (e.g. firewall rules and rules firewall). Therefore, manual cleaning is performed to remove duplicated terms using Protégé. It is also possible to automate cleaning in future. For example, string manipulation methods can be utilized in combination with a dictionary (e.g. WordNet) or techniques from multilingual and cross-lingual IR (e.g. Google search) for disambiguating terms in multiple languages. Besides, in cooperation who deliver products worldwide, their localization processes generate control vocabularies which contain dictionaries of such terms in multiple languages for technical terms. Where available these can also be used for disambiguation.

In our algorithm, uncategorized terms (i.e. terms do not have broader topics) are analyzed and if possible manually replaced under an existing concept by analyzing the context of the term. As a result, a controlled vocabulary is generated, which we call *Topic Ontology* that covers topics represented by the Symantec Norton 360 technical content.

ACCEPTED MANUSCRIPT

Since the enterprise content is formatted in multiple languages, we generate one Topic Ontology for each language by applying our algorithm to each language domain. For instance, for the English version of the Symantec Norton 360, the Topic Ontology contains 1089 topics, 40 root topics and the longest depth in the hierarchy is four.

To validate our approach, we compared the Topic Ontology that is generated from the English version Symantec Norton 360 technical content with a hand-crafted Norton 360 product ontology. The product ontology is independently developed by a researcher who has experience in Norton 360 and the Norton 360 technical documentation. The ontology is light-weight, contains 46 concepts and the longest depth in the hierarchy is three. Since our ontology has 1089 concepts and very complex, it is clear that the similarity of the two ontologies is very low, approaching to zero. Instead we compare the difference of the product ontology to the Topic Ontology based on the equation of [36]:

$$T - S = \{x \mid x \in T \wedge x \notin S\}, \quad D(T,S) = \frac{|T - S|}{T} \qquad (1)$$

where T represents senses set of all concepts in the target ontology and S represents senses set of all concepts in the source ontology. Senses sets (i.e. synonyms) can be extracted using WordNet. T-S represents how many distinct synonym words are in the target ontology and are not in the source ontology. $0 \le D(T,S) \le 1$, such that when no common elements, $D(T,S) = 1$, and if set T is a subset of set S ($T \subseteq S$), then $D(T,S) = 0$. It should be noted that this measurement is not symmetric.

According to equation 1, we calculated D(Product, Topic)=0.13. This result shows that the product ontology is a partial subset of the Topic Ontology and also the Topic Ontology covers the main concepts of the Norton 360 technical domain. The syntactic similarity [27] of ontology concepts based on Maedche and Staab's syntactical comparison, SM(Product, Topic)=0.8814, also confirms this result. In addition, we compared structural similarity of two ontologies according to Maedche and Staab's taxonomy comparison [27]. Simply, this measure calculates semantic correlation of different taxonomies by comparing super- and sub-concepts. Based on this measure, taxonomic similarity of the two ontologies is TSO(Product, Topic)=0.196. If we consider how complex the Topic Ontology is and the number of concepts within it, this result indicates that both ontologies at least share some taxonomic relationships among them.

The Topic Ontology is not only used to annotate technical knowledge based documents, but also utilized to annotate Norton 360 user forum threads. With the Enterprise 2.0, collaborative tools are becoming more important (e.g. forums, wikis, etc). To leverage this, the open-corpus harvesting system OCCS [45] has been used to collect the user generated content from Norton 360 user forum threads (http://community.norton.com/). Then, a crowd-sourcing collaborative annotation tool has been developed to capture user ratings/comments and to annotate these threads with concepts from the Topic Ontology. The rationale is investigating techniques to support personalized IR both on corporate data and user generated content since metadata extracted from the actual usage of documents (e.g. user ratings, comments, tags) can be very useful for personalization [34] [41]. However, this paper only focuses on automatic metadata extraction from multilingual corporate data.

On the other hand, the Topic Ontology can play an important role for the personalization. The topics can be used for user modelling; user profiles are linked to ontology. Subsequently, relevant Norton 360 corporate data and user generated Norton 360 forum content can be retrieved and personalized according to the needs of users. For instance, the ARCHING system [34] [41] presents an architecture for providing adaptive retrieval and composition of heterogeneous information sources of professionally developed corporate content, user generated content (forums) and pages in the wild (open Web). The approach enables adaptive selection and navigation according to multiple adaptation dimensions and across a variety of heterogeneous data sources. To achieve this, the ARCHING system utilizes the Topic Ontology as a high-level domain ontology for correlating corporate and user generated contents.

2.3.3 Metadata Extraction for Personalization

In order to support personalization, rich and useful metadata about resources are required. The problem is that DocBook documents provide limited descriptive metadata that can be used for personalization. However, their content is highly structured and provides very detailed semantic information about the context using a common vocabulary. Here, the context means the semantic meaning of the document. Therefore, the context of documents, concepts described within that documents can be understood automatically to a certain extent if DocBook documents are well markup-ed. To extract useful information for personalization from document content, we introduced a Resource Type Ontology and re-use IEEE Learning Object Model metadata as described below.

2.3.3.1 Resource Type Ontology

In a DocBook document, a document may describe a task using Procedure and Step elements or provide information about a concept using Para, Note, Table, Summary, etc. elements. We developed a new ontology, called *Resource Type Ontology* [53] to capture knowledge about resource types that are contained within the document (Figure 4). The resource Type Ontology has *ResourceType* class, which has sub-classes *Activity*, *Concept*, and object property *resourceType*. Activity class represents an action which is performed by a human agent. Activity class 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*.

////////////////////// Figure 4 Here ////////////////////////////////

Mining Process Type of Documents: Metadata about resourceType of documents are extracted by parsing and

ACCEPTED MANUSCRIPT

analyzing DocBook elements with XML DOM and Javascript; if the document contains *Step* element, then resourceType is set to *Task* or if the document contains *Summary* element, resourceType is *Summary* or if it contains *Para*, *Table*, *InformalTable*, *ItemizedList* or *Note*, then resourceType is *NarrativeText*. For each resourceType, we also estimate the covering percentage of this element by comparing it to the size of the document. The Resource Type Ontology metadata can be used to support personalized IR. One suggestion would be search intend based personalization. For instance, if the user's search intend is "to find overview documents", then, documents marked with resourcetype:Summary can be ranked higher.

2.3.3.2 Learning Object Model (LOM) Metadata

DocBook documents need descriptive information about cognitive metadata such as difficulty that can be useful for personalization. However such metadata is not supported by DocBook and we reused IEEE LOM which provides useful metadata for personalization. IEEE LOM standard defines a set of metadata elements for describing learning objects which are grouped under nine categories: General, Life Cycle, Meta-metadata, Educational, Technical, Rights, Relation, Annotation and Classification [21]. After analyzing DocBook formatted enterprise content, we decided to use three entities from LOM Educational category to describe metadata for personalization: Difficulty, interactivity type and interactivity level. We decided to use these three metadata elements since they suit for personalization as they greatly enhance the ability of a personalization system to tailor the personalized presentations to the appropriate depth, degree of complexity and level of interactivity as well as they were suited for the envisioned personalized IR scenarios presented in [34] [41]. Furthermore, their metadata values can be extracted automatically by analyzing DocBook formatted documents. For example, DocBook documents contain active content elements (e.g. procedures and steps) and expositive content elements (e.g. paragraph, etc.), where they can be used to predict LOM interactivity level and LOM interactivity type of documents. Interactivity type and interactivity level can be utilized for information preference based personalization as discussed in section 5.6. In addition, based on proportions of active/expositive content, LOM difficulty can be predicted. In particular, difficulty is very helpful for knowledge level based personalization.

IEEE LOM Difficulty describes how hard it is to work through the resource. For instance, a document describing a complex task is more difficult than a document containing simple text. IEEE LOM Interactivity Type is the pre-dominant mode of learning supported by the resource (i.e. active, mixed and expositive). IEEE LOM Interactivity Level is the degree of interactivity characterizing the resource, which is based on interactive (e.g. web forums) and non-interactive (e.g. text) content elements that are contained in the document. Interactivity level can take values from very low to very high. Information about these LOM elements could be very useful for personalization. The possible personalized IR services using IEEE LOM are discussed in section 5.6. Metadata values of these LOM entities are automatically created by a proposed fuzzy inference method which is explained in section 4.

3 A Case Study on Symantec Norton 360 Knowledge Base

To illustrate our metadata extraction framework, we have used the English, German and French versions of the security product Symantec Norton 360, which provide structured data about product manuals and online help documentations. The content is used by product users to learn/configure product features (e.g. how to configure Norton Safe Web, what feature is responsible for updating definitions, etc.) and to solve product related problems (e.g. how to find possible causes and/or solutions to this error message). The content is formatted in XML and structured with a subset of DocBook DTD. In these domains, every DocBook element (e.g. Section, Procedure, Book, etc.) has a unique ID. An example DocBook document is presented in Figure 1. Our objective is to automatically extract metadata from this multilingual Symantec content for use in a personalized IR system in multiple languages. Since the proposed framework can extract metadata from any valid DocBook document, multilingual metadata can be automatically generated.

Symantec uses a subset of DocBook DTD to structure Norton 360 technical content in different languages, thus we generated a simplified version of the DocBook ontology (subset), which we call *Symantec Profile*. The Symantec Profile contains constraints on a number of elements such as every section must have a title. Since in Symantec domain, DocBook documents do not have descriptive enough metadata for personalization, we extended Symantec Profile with the developed ontologies, the Resource Type Ontology and the Topic Ontology. In addition, we used the IEEE LOM metadata to describe cognitive information about documents such as difficulty, interactivity level and interactivity type. In Figure 5, the Symantec Profile and relationships between other metadata schemas are shown. Every section has metadata about the difficulty, interactivity level and interactivity type. Section instances are also annotated with metadata about resource types that are contained within the document using Resource Type Ontology. Moreover, each document is annotated with topics from the Topic Ontology.

////////////////////// Figure 5 Here ////////////////////////////////////

4 Mining Cognitive Metadata using Fuzzy Inference

LOM metadata elements, difficulty, interactivity level and interactivity type are usually manually provided by an expert in the field. As we consider number of information objects in a large knowledge base, this method is not scalable. However, automation of this process is difficult since these metadata values themselves are fuzzy and subjective to different users. This can be illustrated with an example. Let's take into account the *difficulty* of a

ACCEPTED MANUSCRIPT

document. Difficulty can take five values from very low to very high. However, these values do not have precise meanings since natural language describe perceptions that are intrinsically imprecise. The main source of imprecision is that unsharp class boundaries of metadata values, which is the result of fuzziness of perceptions. Fuzzy sets introduced by Lofti Zadeh, directly addresses the fuzziness of natural language, classes with unsharp boundaries with a scale of degrees [38]. Since, fuzzy sets are tolerant to imprecision and uncertainty, it allows reasoning with approximate values. Because of the nature of cognitive metadata and imprecise perceptions about their values, fuzzy reasoning may suit well for cognitive metadata extraction. We propose to use a fuzzy based method which is based on fuzzy information granulation; documents are partitioned into semantic parts (semantic granules), each semantic granule is associated with a fuzzy attributes and attribute values are represented by fuzzy sets. Finally, possible metadata values are inferred by using fuzzy if-then rules and Mamdani fuzzy inference system. A brief summary and explanations of fuzzy logic and fuzzy sets operations are summarized in [42].

On the other hand, although the cognitive metadata may be considered somewhat subjective, the intention is for this metadata to be used by an adaptive/personalized system. Therefore, in situations where the perception of the end user concerning a piece of content differs from the metadata value generated by our framework, the personalized system should be capable of adapting to adjust the end user model values accordingly (i.e. users perception of degree of difficulty, interactivity level and interactivity type). For example, if the end user was finding the content delivered by a personalized system too difficult (even though the cognitive metadata value had seemed value), it would be the personalized system's responsibility to adapt to the user preference/cognitive ability and select even easier cognitive value for content presentation. On the other hand, metadata generated from the actual usage of documents can also be utilized for personalization purposes. For instance, [46] presents metadata that is accumulated in time with the interactions of real users with the content (e.g. cognitive characteristics of users, context of use, interaction with the content, etc.).

4.1 Proposed Fuzzy Information Granulation and Fuzzy Inference System

In algorithm 2, the step-by-step metadata generation process is provided. The proposed approach is represented more comprehensively as follows:

Algorithm 2. Fuzzy based metadata generation algorithm
Input: *D* – Section document formatted by DocBook DTD
Output: A file that contains the generated metadata in RDF format
Process:
1. Granulation of documents into semantic fuzzy granules: Concept and Activity
2. Association of attributes (C_length and A_length) to Concept and Activity granules and determination of attribute value
D_DOM = DOM tree of *D* (parsed by XML DOM using Javascript)
For *i*=1 **to** *D_DOM*.childnodes.length
if (*D_DOM*.childnode[i].type==para ||
D_DOM.childnode[i].type==admonition) **then**
C_length++;
end if
if (*D_DOM*.childnode[i].type==table ||
D_DOM.childnode[i].type==list) **then**
for *j*=1 **to** *D_DOM*.childnode[i].childnodes.length
if (*D_DOM*.childnode[i].childnodes[j].type==row)
C_length++;
end if
end for
end if
If (*D_DOM*.childnode[i].type==procedure)
for *j*=1 **to** *D_DOM*.childnode[i].childnodes.length
if (*D_DOM*.childnode[i].childnodes[j].type==step)
A_length++;
end if
end for
end if
End for
return *C_length*, *A_length*
3. Determination of input and output fuzzy sets for the inference system
4. Fuzzification of C_length and A_length values
5. Determination of fuzzy if-then rules
6. Fuzzy inference and rule aggregation
7. Defuzzification and metadata generation
8. Metadata confidence score calculation and storing the generated metadata into a file

Step 1: The human mind informally decomposes the whole into parts (i.e. granules) for reasoning. In almost all cases human mind uses fuzzy measures for reasoning. Therefore, the theory of Fuzzy Information Granulation (TFIG) is inspired by the ability of the human mind to reason with granules [39]. But methodology and TFIG is mathematical. Fuzzy information granulation can be characterised as follows: it is a mode of generalization which may be applied to any concept, method or theory. It involves granulation, fuzzification and reasoning. Our aim is to apply the TFIG for reasoning on metadata values [39]. It can be explained as follows. Assume, you have been asked to provide information about the interactivity level of a document. First, the mind decomposes the document into semantic parts (i.e. granules), such as parts that are interactive (e.g. a web forum requests an input) and parts that are not (e.g. plain text). Then, the mind tries to find proportion of these parts and reasons over to find a solution using uncertain measures and approximation. This example illustrates how our approach aims to work. First, the document is decomposed to semantic parts. If we generalize, Paragraphs, Lists, Tables, Summary, Examples, Figures, Equation and Links within a DocBook document represent *Concepts* that are expositive and non-interactive content. Procedure and Step elements represent *Activities* (task) that are active and interactive content. Activities require more interaction and intellectual property. On the other hand, concepts are less interactive and can be absorbed by reading. For instance, an information object describing a complex task is more difficult and more interactive comparing to an information object that contains simple text. In addition, when there is no interactive content/object within the text, then interactivity level of the document is low. Furthermore, based on proportion of Concepts and Activities, the interactivity type of the document can be inferred, such as a document with no interactive content is expositive. Thus, we divide

ACCEPTED MANUSCRIPT

documents into *Concept* (*C*) and *Activity* (*A*) Fuzzy Information Granules (FIGs), which can be used for reasoning to create metadata.

$$Concept \text{ isfg } C,\ Activity \text{ isfg } A \quad (5)$$

Step 2: *C* and *A* FIGs are associated with length attributes, *C_length* and *A_length* respectively. To find the numeric value of length attributes, we parse and analyze the DocBook document. In particular, parsing algorithm counts un-nested XML tags to calculate attribute values as shown in Algorithm 2: *C_length* equals to the total number of paragraphs/admonitions plus total number of rows in a table/list and *A_length* equals to the number of Steps (i.e. tasks). It should be noted that our document parsing algorithm only uses number of concepts and tasks for reasoning, where these features are imprecise since number of words in a paragraph or task are unknown. However, this is how the human mind also works. We do not count how many words are exactly in a sentence when perceiving a document's difficulty. In addition, the advantage of approximation is that it can use uncertain data for reasoning and document processing is also very fast.

Step 3: To represent numeric input values of *C_length* and *A_length* with fuzzy sets, first the universe of discourse, *U*, should be identified. The universal set *U* is defined by calculating the minimum and the maximum number of *Activity* and *Concept* described within the test corpus (i.e. Symantec Norton 360). Then, *U* can be divided into intervals for defining fuzzy sets. For Symantec Norton 360 test corpus, minimum and maximum values of *C_length* and *A_length* are as follows: $C\_length_{\min} = 1$, $C\_length_{\max} = 37$, $A\_length_{\min} = 1$, $A\_length_{\max} = 35$. As a result, *U = [1, 37]. U* can be partitioned into different intervals. For example, for four fuzzy set, *U* can be partitioned into four unequal intervals, $u_i, i = \overline{1,4}$, such as represented by linguistic values of *low*, *medium*, *high* or *very high*. To find the best ranges for input fuzzy sets, we also analyzed the distribution of number of tasks and activities in the corpus of documents. We observed that ~70% documents contains only narrative text/hyperlinks, ~10% contains text+1-5 tasks, ~10% contains text+6-10 tasks and the remaining documents have text+10-35 tasks.

In order to select the best Fuzzy Inference System (FIS) among the combinations of fuzzy Membership Functions (MFs), the best possible number of fuzzy sets, defuzzification methods and different number of fuzzy sets, we also conducted an initial experiment to analyze the prediction error of the selected fuzzy system for difficulty, interactivity level and interactivity type. The best FIS was selected based on the analysis of Root Mean Square Error (RMSE) values. The experiment is conducted as follows: First, we asked an expert to annotate twenty documents from our case study with metadata values of difficulty, interactivity level and interactivity type. These metadata values are accepted as our sample case. Then, we run our FIS on the selected different membership functions, such as triangular, trapezoid, Gaussian and generalized bell and different defuzzification methods such as the centroid, Mean of Maxima (MeOM), First of Max (FOM) and Last of Max (LOM) with different number of fuzzy sets for representing *C_length* and *A_length*. The detail and results of this experiment is discussed in [42]. Based on this experiment, for difficulty, we use trapezoid function as an input fuzzy set (with low, medium and high linguistic values) and trapezoid MF as an output of the fuzzy inference model (Figure 6 (a)). For interactivity level, we utilize trapezoid function as an input (with low, medium, high and very high linguistic values) and trapezoid MF as an output of the fuzzy model (Figure 6 (b)). For interactivity type, triangular function is used as an input MF (with low, medium and high linguistic values) and triangular MF as an output of the fuzzy model (Figure 6 (c)). We used these best fuzzy models for metadata generation and in our evaluations.

////////////////////// Figure 6 Here ////////////////////////////

Step 4: Fuzzification of numeric input values of *C_length* and *A_length* variables. In this step, the membership grades of numeric values on input fuzzy sets are calculated. Let assume, for difficulty metadata generation, input is represented by three fuzzy sets with trapezoid MF as shown in Figure 6(a), and *C_length=4*, *A_length=10*. According to this example, *C_length* has a degree of membership of 0.66 for *low* linguistic value. *A_length* has a degree of membership of 0.66 to *medium* linguistic value and 0.33 degree of membership to *high* linguistic value as illustrated in Figure 7.

////////////////////// Figure 7 Here ////////////////////////////

Step 5: The fuzzy rules were determined experimentally by using *C_length* and *A_length* attribute values of *Concept* and *Activity* FIGs and analyzing RMSE of the FIS [42]. Rules that gave the minimum RMSE were chosen. In Tables 1, 2 and 3, the fuzzy if-then-rules for difficulty, interactivity level and interactivity type fuzzy inference are shown. The difficulty and interactivity level increase as the interactivity of the document increases. In addition, the consequent of each rule is represented by an output fuzzy set, where output fuzzy sets are shown in Figures 6 (a) (b) and (c).

////////////////////////////////////Table 1 here////////////////////////////////
////////////////////////////////////Table 2 here////////////////////////////////
////////////////////////////////////Table 3 here////////////////////////////////

Step 6: The rules which do not have empty antecedents are fired. Then, for each rule, the min implication operator is applied to obtain the output of the rule's consequent. According to the fuzzification example above, the following rules are fired for the prediction of the metadata value of LOM difficulty:

If C_length = Low and A_length = Medium, then Difficulty = Medium
$\rightarrow \min(0.66, 0.66) = 0.66$

If C_length = Low and A_length = High, then Difficulty = Difficult
$\rightarrow \min(0.66, 0.33) = 0.33$

Step 7: Outputs produced by each rule can be aggregated using the sum operator in order to produce a single output fuzzy set. Then, a deffuzification method can be used to obtain a single output. In our experiments, centroid defuzzification method gave the minimum RMSE values as discussed in [42]. Thus it is utilized in our FIS. For the

ACCEPTED MANUSCRIPT

above example, first rule has output of *medium*, *output=0,* since 0 is the MeOM on the output fuzzy set of *medium* as shown in Figure 6(a). Based on this example, the rule outputs can be deffuzified using the centroid method as:

$$value = \frac{0.66*0 + 0.33*50}{0.66 + 0.33} = 16.66$$

Finally, the numeric output should be converted to a metadata value. This is performed based on the output fuzzy sets (Figure 6). The algorithm checks which interval the output falls into and generates the metadata value based on equations (7) and (8). According to our example, the generated output is "MediumDifficulty".

$$\begin{aligned} &if\ -100 <= value <= -75, then\ \text{difficulty} = Very\ Easy\ /\ \text{int. level} = Very\ Low \\ &if\ -75 < value < -25, then\ \text{difficulty} = Easy\ /\ \text{int. level} = Low \\ &if\ -25 <= value <= 25, then\ \text{difficulty} = Medium\ Difficulty\ /\ \text{int. level} = Medium \\ &if\ 25 < value < 75, then\ \text{difficulty} = Difficult\ /\ \text{int. level} = High \\ &if\ 75 <= value <= 100, then\ \text{difficulty} = Very\ Difficult\ /\ \text{int. level} = Very\ High \end{aligned} \quad (7)$$

$$\begin{aligned} &if\ -100 <= value <= -50,\ then\ \text{Interactivity Type} = Expositive \\ &if\ -50 < value < 50,\ then\ \text{Interactivity Type} = Mixed \\ &if\ 50 <= value <= 100,\ then\ \text{Interactivity Type} = Active \end{aligned} \quad (8)$$

Step 8: Based on the metadata value and the interval that *m_value* falls into, our algorithm calculates a metadata confidence score. Let $x_1 \le x_p \le x_2$, where x_p is a point in the metadata interval $[x_1\ x_2]$ that has the highest membership degree on the output fuzzy set, x_1 is the left boundary of the metadata interval and x_2 is the right boundary of the metadata interval according to equations (7) and (8). If $(x_1 \neq x_p \text{ and } x_2 \neq x_p)$, x_p is the average of the maximum value (MeOM) on x-dimension of the output fuzzy set (e.g. for medium difficulty $x_p = 0$). If $(x_1 == x_p)$, then x_p is the smallest value for the maximum value on x-dimension (FOM) (e.g. for very easy $x_p = -100$). If $(x_2 == x_p)$, x_p is the largest value for the maximum value on x-dimension (LOM) (e.g. for very difficult $x_p = 100$). Then, confidence score is:

$$\begin{aligned} &if\ (x_1 \neq x_p \text{ and } x_2 \neq x_p) \begin{cases} if\ (value == x_p),\ then\ confidence = 1 \\ if\ (value < x_p), then\ confidence = \left(\left| \frac{1}{x_p - x_1} \times value \right| \right) - 1 \\ if\ (value > x_p),\ then\ confidence = \left(\left| \frac{1}{x_p - x_2} \times value \right| \right) - 1 \end{cases} \\ &if\ (x_1 == x_p \text{ or } x_2 == x_p), \left\{ confidence = \left| value \middle/ x_p \right| \right\} \end{aligned} \quad (9)$$

According to the equation (9), if the output of the FIS, *value*, is close to x_p, then a larger confidence score is generated (i.e. if $value == x_p$, *confidence*=1). If *value* is nearer to the boundaries of the output fuzzy set, then a lower confidence score is generated. For our example of "medium difficulty", $x_p = 0$, $x_1 = -25$, $x_2 = 25$ and *value= 16.66*. Then metadata confidence score for medium difficulty is $\left|(1/(0-25)) \times 16.66\right| - 1 = 0.33$, which is a low score since *value= 16.66* is closer to the boundary.

4.2 Implementation of Fuzzy Inference System (FIS)

The FIS is implemented as a Web application using Javascript. Using XML DOM support of Javascript, DocBook documents are parsed and analyzed for the calculation of *C_length* and *A_length* values. Fuzzy inference rules are deployed as if-then statements. Our approach is easy to implement comparing to machine learning based techniques which require pre-processing for feature extraction and training. In addition, since our technique roughly extracts features about the document (only number of concept and activities), the speed of the system is very fast. For instance, cognitive metadata about 700 documents are extracted within approximately 3-4 seconds. Although, we do not need to train the fuzzy system, rules have to be set by an expert and the best combination of fuzzy sets, membership functions, number of fuzzy sets and defuzzification methods for the task have to be identified previously. However, the advantage of fuzzy logic is that it allows knowledge to be intelligently interpreted by the inference system.

5 Evaluations

The automatically extracted metadata has been evaluated in terms of *Precision*, *Recall*, *F-measure*, *Prediction Error*, *User Perceived Quality* and *Fitness to Personalized IR*. This section explains the evaluation procedures and discusses the results.

5.1 Metadata Quality – Precision, Recall and F-Measure

Metadata quality can be assessed by comparing automatically extracted metadata values with the manually entered metadata. For this purpose, we conducted a preliminary user study. In the study, five subjects (3 post-docs and 2 final year PhD students from computer science) were asked to annotate randomly selected the same 100 documents from English version of Symantec Norton 360. Subjects had different levels of expertise. In *annotation*, user 1, 4, 5 are advanced and user 2, 3 are beginner. In *Symantec products,* user 1, 2, 4, 5 are intermediate and user 3 is beginner.

In the study, participants manually assigned metadata to document difficulty, interactivity level and interactivity type, which cannot be computed directly from content. By contrast, other metadata elements such as subject, title, creation date, modification date, source, URI and structural metadata (dc:hasPart, docbook:subsection, docbook:link) can be extracted with almost 100% precision by analyzing the document since they are manually provided by content authors. Therefore, we did not ask participants to manually annotate these metadata fields. Where content authors do not provide such descriptive metadata such as subject, title, etc., there are many examples of attempting to use NLP and statistical methods to mine such metadata [26] [37] [40]. Therefore, this paper has focused on metadata which have not been mined before yet which are particularly useful for personalization purposes.

ACCEPTED MANUSCRIPT

Before the experiment, we provided the required instructions (annotation guidelines) to the participants, which explain the annotation task and how to assess the meanings of difficulty, interactivity type and interactivity level of documents. We also provided example documents that were annotated by the expert in order to help them to assess the meanings of metadata values. In addition, we informed the participants about the minimum and maximum number of concepts/activities in the corpus, which assist them to understand min/max values of metadata fields (very low and very high). Variations of metadata values are inherently tackled to a certain degree by the fuzzy nature of estimation of metadata values; outputs of the system are linguistic values (e.g. low) rather than exact numerical numbers.

In the experiment, all participants annotated the same 100 documents, which also allow us to analyze individual perceptions and their annotation trends. The subjects used an online Web annotation client to separately complete the manual annotations. Then, automatically extracted metadata was compared with the metadata produced by participants in terms of precision, recall and f-measure. Precision is the number of metadata fields annotated correctly over the number of metadata fields automatically annotated by the framework, recall is the number of metadata fields annotated correctly over the number of metadata fields manually annotated and f-measure evaluates the overall performance by treating precision and recall equally. Out of 100 documents, 97 documents can be parsed and automatically annotated. During the analysis, we noticed that 3 documents had invalid XML syntax; therefore the framework could not extract metadata. Four of the participants annotated all of the documents. The user #3 annotated 73 documents. The results are shown in Tables 4, 5 and 6.

////////////////////////////////Table 4 here////////////////////////////////
////////////////////////////////Table 5 here////////////////////////////////
////////////////////////////////Table 6 here////////////////////////////////

Analysis: The results showed that interactivity type metadata quality scores are higher than interactivity level and difficulty scores with an average of 91% precision, 88.76% recall and 89.86% f-measure. On the other hand, there were sparse decisions on the metadata values of interactivity level and difficulty of documents, mainly because the perceptions of users on these values were different. Interactivity level prediction performance received an average of 79.46% precision, 77.52% recall and 78.47% f-measure. Whereas, difficulty metadata quality received the lowest scores among the three automatically generated LOM metadata fields with an average of 66.29% precision, 64.52% recall and 65.39% f-measure. If we looked at the individual precision measures, the results showed that different users annotated the same documents in a very different way. To understand the reason, we asked the participants about their annotation trends. We found out that, some documents contain embedded information about tasks within text sentences. Subject #2 and #3 treated these documents different than other three participants. In addition, some documents describe alternative ways of performing a task and again different users perceived interactivity level and difficulty of these documents in a different way. For example, some of the subjects rated difficulty low although the document describes complex activities since they perceived that they are alternative tasks and independent.

The study showed that cognitive metadata is subjective and there are different perceptions. Since different participants annotated documents separately and in their own way, individual measures provide different conclusions. To measure the overall metadata quality, we computed the agreed annotations of all five subjects. For each document, we counted the number of metadata entry values and took the metadata value that has the highest score as the agreed metadata value. In the case when there is no agreement (i.e. more than one highest score), we took the mean of annotations as a resultant value. In addition, Fleiss' kappa (κ) statistical measure [56] was applied for assessing the reliability of inter-agreement between five participants. According to [57], interpretation of κ values are as follows: $\kappa \leq 0$ represents poor agreement, $\kappa \in (0, 0.2]$ represents slight agreement, $\kappa \in (0.2, 0.4]$ represents fair agreement, $\kappa \in (0.4, 0.6]$ represents moderate agreement, $\kappa \in (0.6, 0.8]$ represents substantial agreement, $\kappa \in (0.8, 1)$ represents almost perfect agreement and $\kappa = 1$ represents perfect agreement. In our experiment, Fleiss' kappa measure for interactivity type was $\kappa = 0.83$ which showed almost perfect agreement. For interactivity level, Fleiss' kappa measurement was $\kappa = 0.62$ that means substantial agreement among participants. For difficulty, Fleiss' kappa measure was $\kappa = 0.61$ which also showed substantial inter-agreement. Inter-agreement statistics are also shown in Table 11. We compare the automatically annotated metadata against the agreed annotations in terms of precision, recall and f-measure as shown in Table 7. The quality scores of the agreed annotations increased considerably comparing to the average scores of individual annotations. Precision of interactivity type, interactivity level and difficulty was increased ~6% (96.90%), ~9% (88.65%) and ~16% (82.47%) respectively against the average of individuals' annotations.

////////////////////////////////Table 7 here////////////////////////////////

Discussions: We have observed that when manual annotation is not done carefully, it is prone to errors. In particular, some subjects annotated very similar documents with very different metadata values, which mean that manual annotations were not consistent. In addition, interpretation and perception of metadata values was subjective; same documents were annotated with different metadata values by different participant, which was our initial observation about cognitive metadata. From the user's feedback, it is pointed out that manual annotation is very time consuming. Despite drawbacks, our fuzzy based method generated reasonably accurate metadata values comparing to the agreed annotations of individual users with an overall average precision, recall and f-measure of 89.39%, 86.66% and 87.97% respectively.

Analysis of the Background Information: We also analyzed the implications of the background information on metadata values. Tables 8 and 9 illustrate the average precision rates of individual subjects that are grouped based on their expertise levels in annotation and Symantec

ACCEPTED MANUSCRIPT

products respectively. As expected, subjects with higher expertise provided better precision rates than subjects with less expertise. Since the background information might influence the results of the agreed annotations, we computed the agreed annotations of advanced and intermediate users for comparison to the agreed annotations of all subjects as shown in Table 10. Inter-agreement of advanced subjects in annotation was also computed based on Fleiss' kappa (κ) measure (see Table 11): κ =0.85 (almost perfect agreement), κ =0.64 (substantial agreement) and κ =0.65 (substantial agreement) for interactivity type, interactivity level and difficulty respectively. In addition, inter-agreement of intermediate participants in Symantec products was measured (see Table 11): κ = 0.84 (almost perfect agreement), κ = 0.62 (substantial agreement) and κ =0.61 (substantial agreement) for interactivity type, interactivity level and difficulty respectively. The results in Table 10 showed that background knowledge in Symantec products did not affect precision rates for interactivity level and difficulty (same as the agreed annotations of all subjects as shown in Table 7). Conversely, background information in annotation slightly improved the agreed annotations of all subjects (~1-2% improvement). Overall, the agreed annotations of participants who have some background knowledge are very similar to the agreed annotations of all subjects. This is because there are more subjects with background information and when the agreed annotations taken, the agreed annotations are in the favour of subject with background knowledge. Thus, the agreed annotations of all subjects (Table 7) are representative of the results.

//////////////////Table 8 Here //////////////////
//////////////////Table 9 Here //////////////////
//////////////////Table 10 Here //////////////////
//////////////////Table 11 Here //////////////////

5.2 Metadata Prediction Error

Precision, recall and f-measures does not give information about how the predicted metadata value is close to the manually provided metadata value. They only measure perfect matches. The Sum of Squared Error (SSE) can be employed to quantify the performance of each fuzzy inference model prediction comparing to manual annotations provided using the following equation:

$$SSE = \sum_{i=1}^{n} (P_i - M_i)^2 \qquad (10)$$

where *P* is the predicted metadata value, *M* is the manually provided metadata value and *n* is the total number of documents. For the calculation of SSE, metadata values are represented by numerical values. For example, difficulty and interactivity level values are represented from 1 to 5 scale (i.e. very low=1,…, very high=5) and interactivity type values are represented from 1 to 3 scale (i.e. expositive=1, mixed=2 and active=3). For a perfect match, $(P_i - M_i)^2 = 0$, and $(P_i - M_i)^2$ increases as the prediction value differs from the manually provided value. We calculated the SSE for interactivity type, interactivity level and difficulty as shown in Figures 8, 9 and 10 respectively.

Analysis: As can be seen from Figure 8, user #5 and user #4 annotated interactivity type of documents very similarly and user#5 provided exactly the same metadata values as the agreed annotations of all subjects (SSE: 6 for 100 documents). Figure 9 illustrates that user #3 and user #2 had a very similar pattern of annotations on interactivity level values, while user #4 and user #5 had different pattern but provided related annotations. Conversely, user #1 supplied values in the middle of two different patterns. When taken the agreed annotations of all users (SSE: 19 for 100 documents), user #4 and user #5 provided the closest annotations for interactivity type. As illustrated in Figure 10, for difficulty, users 1, 4, and 5 supplied very similar metadata values, while users 2 and 3 provided different but related annotations. If we analyze this graphic, there is a complete sparse decision about metadata values between the two groups, which could be mainly because of different perceptions of users to very same documents. For difficulty, SSE of agreed annotations on 100 documents is 30.

//////////////////Figure 8 Here //////////////////
//////////////////Figure 9 Here //////////////////
//////////////////Figure 10 Here //////////////////////////

5.3 Fuzzy Inference versus Rule-Based Reasoner

Rule-based systems are very popular for metadata generation. They are mainly computer programs that use programmed instructions (rules) to reach conclusions from a set of premises. In general, these premises are represented by crisp sets (true/false). Our claim is that, cognitive metadata values are often fuzzy, thus fuzzy sets and fuzzy inference can provide better metadata values. Therefore, the performance of the proposed fuzzy model can be compared with crisp rules with sharp boundaries to assess the added value of fuzzy inference. For this purpose, we implemented a rule based reasoner. The rule-based reasoner uses exactly the same rules as the fuzzy inference system but the rules discussed in Tables 1, 2 and 3 are represented by crisp sets as shown in Figure 11. We compared the generated metadata values by the rule-based reasoner and the proposed fuzzy inference model against the agreed annotations in terms of precision, recall and f-measure as illustrated in Table 12.

////////////////////// Figure 11 Here ////////////////////////////////////
//////////////////////////////Table 12 here//////////////////////////////

Analysis: Results showed that there was not a significant difference in the performance of rule-based reasoner for interactivity type and interactivity level. The reasons are: 1) Metadata values of interactivity type do not have sharp boundaries comparing to interactivity level and difficulty. If we look at Figure 6 (c), we see that input fuzzy sets have few overlaps 2) Interactivity level only depends on the length of the *Activity* granule, which means that on only one constraint, fuzzy sets improved the performance only ~2%. Conversely, metadata values of difficulty are often fuzzy and difficulty metadata values depend on both the length of *Activity* and *Concept* FIG. If we look at Figure 6(a), we see that input fuzzy sets overlap with a great

ACCEPTED MANUSCRIPT

degree as well. Therefore, for difficulty, the fuzzy inference system performed an improved performance than the rule-based reasoner with a precision of 82.47% and increased the performance of the rule-based reasoner ~11%. In conclusion, fuzzy inference improved the overall precision, recall and f-measure against the rule-based reasoner especially when the input data do not have sharp boundaries and depend on more than one constraint.

Furthermore, the SSE error graphics of the fuzzy inference system and the rule-based reasoner against the agreed annotations of users are compared as shown in Figure 12. For interactivity type, SSE graphics are exactly same. For interactivity level and difficulty, similar patterns are observed between both inference systems. However, the fuzzy inference predicted more accurate values than the rule-based reasoner.

////////////////////// Figure 12 Here //////////////////////////

5.4 User Perceived Quality

After manual annotation, a second experiment was conducted to assess user perceived metadata quality. In the experiment, the same five subjects were provided with automatically generated metadata about the same 100 documents which are generated by our framework. Then, they were asked for each document to indicate how well each metadata element represent the document from 1 to 6 scale, where 1 is very poor and 6 is very well. In particular, we asked metadata quality of difficulty, interactivity level, interactivity type and overall metadata quality (i.e. quality of all metadata fields that are extracted and discussed in this paper). Average quality scores of 100 documents are calculated for each subject and presented in Figure 13.

////////////////////// Figure 13 Here ////////////////////////////////

Analysis: The results showed that metadata quality of the interactivity type found to be very high by all subjects. There were sparse opinions about the quality of interactivity level and difficulty. However, metadata qualities about all of three fields were rated 5 or over out of possible 6. The mean of user ratings were also calculated as shown in Figure 14. The interactivity type was rated an average of 5.82 out of 6, interactivity level was rated an average of 5.57 out of 6 and difficulty was rated an average of 5.48 out of 6. In summary, all of the three metadata fields had a very high user perceived quality scores out of the examined 100 documents. Furthermore, the overall metadata quality ratings of users to each document were calculated, which is 5.57 out of 6. We took the average of overall metadata quality ratings of the five participants for each document. The results are shown in Figure 15. With the exception of one document, the overall metadata quality of the documents were rated 4 or over (from a range of 1-6). We found that the metadata parsing algorithm could not extract metadata about this particular document because of invalid XML syntax in that document. It is important to point out that the most of the documents have average rating between 5 and 6, which shows very high user perceived metadata quality among the examined documents.

////////////////////// Figure 14 Here ////////////////////////////
////////////////////// Figure 15 Here ////////////////////////

5.5 Automated Metadata Quality Assessment

Traditional metadata quality evaluation is based on comparing automatically generated metadata values with the manually provided ones, as we performed and discussed in sections 6.1, 6.2 and 6.3. This method works well on small-sized repositories but is not scalable for large repositories since humans “do not scale”. In order to deal with exponential growth of metadata records available and to retain some sort of metadata quality assurance, automated evaluator metrics are proposed. Ochoa and Duval [43] propose implementable and measurable quality metrics on LOM metadata, which is loosely based on a metadata quality framework proposed by Bruce and Hillmann [44]: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness and accessibility. These metrics can also be applied to other metadata element sets. The metrics assess the quality of metadata record, not the quality of metadata standard or information object itself. To evaluate the quality of the automatically extracted metadata from Symantec Norton 360 English, German and French knowledge bases by our framework, we applied Ochoa and Duval’s quality metrics.

////////////////////////////////Table 13 here////////////////////////////////

The formulas of metrics are summarized in Table 13. *Completeness* is the degree to which a metadata record represents all the information needed to have an ideal representation of an information object. Usually metadata standards and application profiles provide information about the ideal representation using mandatory element fields. In our case study on Symantec, the ideal representation differs based on class types as explained in section 4; for Docbook section instances, the ideal metadata record has the title, subject, dc:hasPart (at least one relation to block components), difficulty, interactivity type, interactivity level, resource type, format, source, created, modified, publisher, creator and language as shown in Figure 5. For other element types such as book, article, chapter, appendix and preface, the ideal metadata record has the title, at least one dc:hasPart relation to sub-components, format, source, created, modified, publisher, creator and language. In the simple completeness metric, all elements are equally relevant. However, some fields may have higher relative importance than other fields. *Weighted completeness* not only counts non-null metadata fields but also weight each field according to importance of the application. For instance important fields might have a weight of 1 (e.g. title, subject) and unimportant fields might have a weight of 0.2 according to [43].

Accuracy measures the correctness of metadata values. It usually measured by comparing with manually entered values as discussed in section 6.1. *Provenance* represents the origin of the metadata. In our approach, we support this by generating metadata about creation/modification date, creator, publisher and source URI. *Timeliness* represents the degree to which a metadata record remains current and how useful the metadata remains over time. This could be

ACCEPTED MANUSCRIPT

calculated in different ways, such as age of the record and frequency of usage. Since we do not have usage statistics, thus we calculated the age of the record. Accessibility metric measures the degree to which a metadata record is accessible both in terms of logical and physical accessibility. The logical accessibility measures how easy to understand information contained in the metadata record (human readability – how easy is to read the description of the metadata record) and physical accessibility is how easy to find a metadata record in the repository regardless of the accessing tool (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 and metadata values correlate positively among each other. For example, LOM standard suggests that if an information object has an interactivity type of 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 (where 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. The authors [43] use term frequency and inverse document frequency to measure the latter. However, we took a different approach and assess conformance to expectations in the next section by discussing requirements to personalized IR system and how the extracted metadata fulfills these requirements.

We applied the quality metrics to the metadata extracted from the Symantec Norton 360 English, German and French technical knowledge bases and results are shown in Table 14 (except linkage metric, all metrics are normalized from scale 0 to 10). In the evaluation, quality metrics are transformed to SPARQL queries and executed on the content. From the N360 English content, the metadata generation framework extracted 5182 DocBook Ontology instances and metadata about them. Out of 5182, only 639 of the metadata records represent actual documents (book, chapter, section, preface and appendix) and the rest is sub-component instances. In the same way, in the German domain, 6531 instances are extracted, 671 of which represent documents and in the French domain, 7054 instances are extracted, where 858 of them represent documents. We use metadata about actual documents for evaluations. In all domains, completeness is greater than 9.90; we observed that in all domains, there are number of documents which do not have a subject (i.e. if a document does not have a subject, this problem is handled by the personalized IR system [34] [41] by using statistical methods such as frequency/ inverse document frequency ($tf \times idf$) to create subject terms). Since subject is weighted higher in weighted completeness, there is a small reduction in overall weighted completeness. Please note that in weighted completeness, the dc:title, dc:subject and dc:hasPart relations weighted 1, and other elements weighted 0.2. In addition, we applied LOM standard consistency suggestions [21] to these contents and it was observed that metadata values generated by the proposed fuzzy based approach are consistent with the standard. Since metadata is generated within the same week of evaluations, the metadata records are up-to-date (timeliness). Linkage between instances are also calculated, where linkage in English domain is higher than other two domains such as average linkage is 4.40 in English domain, 3.58 in German domain and 2.79 in French domain. This shows that documents can also be accessed through semantic relationships between them in all domains. Overall metadata quality tests showed that the extracted metadata in all three domains have high completeness, up-to-date, consistent and contains an average of around 3 links between document instances (extracted from document structure and relationships between documents).

//////////////////////////////////Table 14 here//////////////////////////////

5.6 Metadata Fitness to Personalized Information Retrieval

According to the study of Guy et al. [16] on Eprints achieves, *"high quality metadata supports functional requirements of the system it is designed to support"* and metadata quality is about fitness to a task. In order to test metadata quality for personalized IR, first we analyzed functional requirements of metadata to personalized IR, which is listed Table 15. Table 15 shows what we are trying to achieve in the personalized IR system and what metadata is needed to support these functionalities.

//////////////////////////////////Table 15 here//////////////////////////////

In the following sub-sections, we discuss how the automatically extracted metadata can be used for personalized IR:

Metadata Fitness to Search: In our metadata generation framework, the title, subject and document types are extracted automatically. Therefore metadata supports the requirements for search. In addition, in our approach, information objects are annotated with ontology-based metadata, consisting of concepts, properties and values defined according to ontology. Thus, semantically relevant documents can be retrieved based on semantic relationships between information objects.

Metadata Fitness to Browsing: The ontology describes the semantic context of the document by defining concept classes and semantic relationships between them. In the case study, the DocBook Ontology provides structural information about how documents are organized and related. In addition, the Topic Ontology supplies a different point of view and classification of documents based on topics. Therefore, both ontologies can be utilized for supporting ontology-based browsing. Furthermore, semantically relevant documents can be browsed using the dc:hasPart, dc:subject, docbook:subsection and docbook:link relationships between documents.

Metadata Fitness to Personalization:

User Modeling: User modeling is an important part of the personalization. In traditional IR, user preferences are usually represented as raw keywords. In our approach, however, the user preferences can be represented as

ACCEPTED MANUSCRIPT

semantic entities with varying preference degrees. For this purpose, the DocBook Ontology or the Topic Ontology can be used, which have been used to annotate the corpus. This also provides a common representation ground for describing user preferences and content meaning, resulting in a fairly precise, meaningful, unified and interoperable representation model for personalized IR.

User Knowledge/Interest/Expertise-based IR: User's knowledge /interest/expertise can be represented as semantic entities to the DocBook ontology or the Topic Ontology as discussed above. In Adaptive Hypermedia, this kind of user profiling is known as overlay user model [4] and could be very useful for retrieving relevant documents based on a common interoperable vocabulary. User's knowledge, interest or expertise to ontology instances can be obtained explicitly (e.g. by relevance feedback) or implicitly (e.g. based on previous interactions with the IR system). Then, relevant documents can be presented based on the semantic structure of ontologies and user's ratings.

Document Difficulty-based IR: If we know the user has little experience with a topic or a document, then the system can show search results starting from easy to difficult documents. Particularly, in enterprise domain, documents may describe highly technical information. A beginner in a topic, may prefer to read easy documents first and an expert user may prefer moderate to difficult documents first in the search results. In the case study, all documents are automatically annotated with difficulty.

Interactivity Type-based IR: Assume the user explicitly provided the type of documents s/he wants such as "what" or "how" documents. What documents can provide expositive information, while "how" documents can provide active documents that explain how to perform a task. In the case study, all documents are annotated with interactivity type and users can be supported with personalized documents based on their information needs.

Document Interactivity Level-based IR: Interactivity level of documents can be used to filter results according to the preference of the user. In the case study documents are annotated with interactivity level automatically.

Document Resource Type-based IR: Metadata about concept types can be used to personalize search results based on search intents of the user. If search intend of the user is known (i.e. explicitly), then results can be personalized accordingly. For example, if the user's search intend is to "find an explanation", then resources describing narrative text can be ranked higher in search results. Or, if the user's search intend is to "find overview documents", then, documents marked with summary can be presented. In our framework, resource type metadata is extracted automatically.

Document Type-based IR: Preferred document types of the user can be used by search engine to customize search results. Google started to perform this kind of customization such as the user can select what kind of documents s/he wants to be displayed.

Document Re-Composition: We extract fine-grained metadata about sub-parts of documents such as Procedures, Paragraphs, Example, etc., which might be very useful for re-composition of the content for a particular use or an individual need.

Discussion: There are different stages of metadata creation that affects the overall quality. In our case study in Norton 360 domain for instance, automatic metadata extraction is the last chain of the previous steps. Metadata creation starts during the authoring process; the author provides the content and some metadata associated with it (e.g. title and subject topics). In addition to the metadata choices of authors (e.g. title), the engineering choices such as what information to include, how to supply metadata (controlled vocabularies) and in which format affect the overall metadata quality and consistency. For example, if information about target user personas of documents is available, then quality of the extracted metadata for personalization can be improved considerably. Therefore, metadata quality passes downstream from document creator to metadata extractor and to end-users.

In the given context, we applied the automated text extraction techniques to pull metadata (in RDF) from the text corpus using Semantic Web techniques. Particularly, we extract metadata from document content, DocBook document structure and document semantics. The evaluations showed that the extracted metadata is high quality and useful for personalized IR. It is also important to note that our approach extracts fine grained metadata about sub-parts of documents (e.g. Para, Steps, Summary, etc.), which can aid reuse of content-parts, also re-composition and personalization of content to individual needs. In addition, the extracted metadata can be reusable by other Semantic Web applications simply utilizing standard SPARQL. Furthermore, our approach and the proposed ontologies can be applied to other enterprise domains for automatic metadata extraction. For example, if different corporate sites publish their content in DocBook format (where DocBook is widely used by many enterprises as discussed in section 3), we can extract and aggregate rich metadata automatically from DocBook documents for the consumption of different applications such as cross-site personalized IR/browsing.

6 Related Work and Discussion

Generally, automated metadata extraction 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* (e.g. log files, number of downloads, etc.), and 4) *document structure* by examining document storage structure. Automated techniques can use different sources to create metadata and techniques depend on the application domain and the context. Most popular techniques are rule-based parsers, regular expressions (extraction of entities, properties, nouns and relations), natural language processing techniques, machine learning methods and ontology-based extraction. This section briefly summarizes related work.

6.1 Automatic Metadata Extraction Systems

Automatic Metadata Generation (AMG) [4] and A System for Automatic eXtraction of E-learning object Features (SAXEF) [2] are systems that automatically extract

ACCEPTED MANUSCRIPT

metadata from learning objects. AMG extracts metadata from the content and context of learning objects using object-based and context-based indexers. It generates IEEE LOM metadata, such as document type, package size, publication date, creation date, operating systems type, access right, main discipline, language, format, title and author's details. SAXEF is a Web application which extracts learning object features of any Web page found on internet. The application extracts main/secondary topic using word occurrence counting, synthetic/analytical level of document by measuring the ratio between the textual area and multimedia area, and media types/multimediality level by examining different media types found on the page, which is similar to resource type extraction of our framework. Our framework also extracts similar metadata but from enterprise content and in multiple languages. In addition, we apply fuzzy inference to generate cognitive metadata, which is different than these approaches.

6.2 Automatic Metadata Extraction Methods and Techniques

Existing automated techniques can be divided into five categories: Rule-based systems, learning based methods, natural language processing techniques, IR techniques, and ontology-based systems.

Rule-based/Template-based/Structure-based Parsers: Pre-defined knowledge based systems are very popular for metadata extraction. They generally use templates, layout/structure parsers and/or pre-defined patterns such as string matching techniques to identify metadata fields from documents. Rule-based techniques do not require training and straightforward to implement. However, the drawback is that when they are applied to heterogeneous input document types, it can result in complex rules that are difficult to create, test and maintain all possible combinations for high quality metadata extraction. Some of the rule-based systems are [24] [28] [12] [23] [18].

[24] uses layout and rule expressions to analyze and label the structure of a given document. The generated data is not in a structured format. The template based method achieves precision of 35% to 100% depending on the label. A similar rule-based labeling module is proposed by [28], which automatically generates descriptive metadata (title, author, affiliation and abstract) from scanned medical journals. The overall performance varies from 76.37% to 94.09%. [12] metadata approach is based on template based approach. They provided a template for each layout to extract metadata from PDF documents. The method performs an overall accuracy of 83% for documents with defined templates and 65% for documents without defined templates. [23] utilize case based and rule based reasoners to generate metadata from heterogeneous Thai documents. A rule based reasoner first analyzes document types and classify them into groups (e.g. thesis, research paper). Based on the document type, a rule based reasoner extract metadata (e.g. title, author). The system performs 62.31% - 90.78% depending on document type. [18] also uses a template matching method to extract metadata (title, author, affiliation abstract and keywords) from headers of PDF documents. The performance of the system evaluated on digital libraries of ACM, IEEE, Elsevier and LNCS, with an average precision of 70% to 91%. Our cognitive metadata generation method is also based on rules (fuzzy rules). Evaluations showed that our framework provides competitive results with an average precision of 82.47% to 96.90% depending on metadata field.

Learning-based Systems: [17] and [25] are some of the machine learning-based techniques that have been applied for metadata extraction. Han et al. uses Support Vector Machines (SVM) for metadata extraction from header parts of research papers [17]. The method automatically generates DC metadata elements with an overall precision of 90.38%. [25] generates DC Qualified metadata from pages using a combination of neural networks and TF / IDF statistical method to filter value of metadata fields. They achieved an overall precision of 84.29% on pages from BBC. Learning based systems perform well on relatively homogenous document sets that have similar structure and layout. However, they require extensive training before applied, which can be very costly.

Natural Language Processing (NLP) Techniques: NLP techniques are usually applied in conjunction with other techniques. For example, [26] uses NLP and machine learning techniques to generate metadata from educational resources. [37] utilize a rule-based system that uses shallow parsing rules and NLP techniques to extract terms and phrases from sentences.

Information Retrieval-based Methods: For subject/topic extraction, term frequency / inverse document frequency (TF/IDF) is a well known data mining technique, where the terms with highest rank is taken as the subject of the document. Fuzzy based techniques are also applied together with data mining techniques for improved extraction performance. For example, [40] uses TF/IDF for selecting document features, fuzzy membership grade to represent these features in the process of classification of Web documents with fuzzy k nearest neighbour algorithm.

Ontology-based Extraction: Many tools have been proposed for the automatic annotation of Web pages with named entities, properties and relationships based on ontologies. Artequakt [1], KIM [22], OpenCalais [54] and PANKOW [7] are examples of these tools. The Artequakt is a knowledge extraction tool that extracts knowledge about artists from the Web. In particular, the tool extracts entities and entity relationship using ontologies and lexical analysis. KIM is a semantic annotation platform for automatic semantic annotation. It is based on GATE framework [9], which is a mature software for named entity recognition based on ontology lexicon. OpenCalais is a web-service provided by Thomson Reuters for automatic metadata generation. It supplies a programatically accessable API for analyzing text and extracting semantic information from it in the form of entities and relationship instances. It also supports annotation of documents with linked data URIs. In our approach, we also extract DocBook Ontology concept instances, structural relationships between concept

ACCEPTED MANUSCRIPT

instances and relationship links by automatically analyzing document structure. Different than Artequakt, KIM and OpenCalais, we do not need to pre-extract concept lexicons or perform lexical analysis since DocBook documents (XML) are structured and already annotated with concept classes by content authors. Alternatively, Pattern-based ANnotation through Knowledge On the Web (PANKOW) is an unsupervised pattern-based approach to categorize Web content according to a given ontology. In this approach, a Web page is scanned for phrases that might be instances of an ontology. Then, series of linguistic patterns are applied and Google search API is used to disambiguate the annotation term according to the context of the page.

Discussions: Our framework not only extracts structural, descriptive, administrative and cognitive metadata from English content but from multilingual content. This is the novelty of our approach with respect to the state of the art, which is generally based on English content. It should be fair to acknowledge that in our framework, automatic multilingual metadata extraction comes from working with enterprise documents and content which is loosely based on the DocBook type of formatting. However, the most common format of documentation formally produced by enterprise adopts a DocBook type of formatting.

6.3 Cognitive/Pedagogical Metadata Extraction

There are many approaches that automatically extract DC metadata and general metadata fields of IEEE LOM as described above. However, there are few approaches on the extraction of cognitive or pedagogical metadata automatically, where such metadata can be very useful for the personalization. For example, according to pedagogic type or cognitive abilities of the user, personalized content can be presented. In the digital library content, these metadata values are usually manually provided by the content author, which is a labour intensive job and automated techniques are required for scalability. Our main focus and contribution is automatic cognitive metadata extraction. Thus, we discuss related work in automatic cognitive and pedagogical metadata extraction.

[31] uses an automatic annotation tool for annotating learning objects with pedagogical metadata such as concepts/concept significance, type of concepts, topic and learning resource type. Concept and topic of the document is extracted by using domain knowledge (concept ontology) and counting frequency of concepts and related concepts. Concept type identification (i.e. outcome, prerequisite, defined or used concept) is extracted by analyzing sentences using a shallow parsing approach and utilizing different inference rules. The performance of the algorithm mainly depends on developing all possible patterns for a concept type and it did not perform good enough (average precision of 60%) since it was incapable of handling all possible patterns. The system also extracts learning resource type metadata such as narrative text, questionnaire or experiment type by identifying document features, specific verbs, trigger words, phrases and special characters from text. These features are classified based on a neural network based method for metadata creation. The algorithm achieves a precision of 72.14% to 98.75% depending on the learning resource type and performance of different neural networks algorithms. In our framework, we also extract resource type metadata, which is very similar to learning resource type metadata of [31]. In our case, DocBook documents are already annotated with type of concepts (e.g. Paragraphs, etc.), thus we can extract the resource type of the document with 100% precision.

[20] presents an ontology-based approach for automatic annotation of presentation slides based on IEEE LOM. Specifically they extract pedagogic role (example, summary, references, etc.) Their approach is based on heuristic rules, where they observe the presence of specific terms along specific patterns. Their user study shows that they achieved a high precision (88%) on the experimental set for pedagogic role extraction.

[37] utilizes parsing rules and NLP techniques to extract terms and phrases from sentences. In particular, system generates metadata about creator, title, date, grade, duration, essential resources, pedagogy-teaching method, pedagogy-grouping, pedagogy-assessment, pedagogy-process, audience, standards, publisher, and relations. However, the user study showed that only two of the elements, title and keywords were provided significantly better results than manual annotation and pedagogic metadata fields had received very low quality scores.

Different than [31] [20] [37], we extract difficulty, interactivity level and interactivity type of the document automatically based on a fuzzy based method with an overall average high precision of 89.39%.

On the other hand, metadata about document difficulty has not been automated by any metadata generation system according to our knowledge. However, text readability difficulty can be inferred automatically by using readability indexes [11] [13]. The Flesch Index measures text comprehension difficulty by analyzing word lengths and sentence lengths in a document [11]. A fully automated semantic latent analysis (LSA) can also be utilized to assess text comprehension of the reader based on characteristics of sentences/words used and their coherence [13]. LSA can be costly in terms of processing needed to be performed. Although, text comprehension difficulty can be automatically extracted using these techniques, the difficulty of an interactive content cannot be extracted only using word/sentence analysis. The context of the document and semantic meaning of the content is required to be understood to a certain extent for difficulty analysis. In most enterprise domains, fortunately corporate data such as technical documentations are very structured and formatted by XML since it allows content delivery in multiple formats (e.g. online, CDs, hard copies, etc.) and languages. As a result, this allows automatic intellectual analysis of document semantics on corporate data. Recently, such formal documentation is being enhanced by additional (informal) content generated by corporate users through wikis and blogs (enterprise 2.0). Personalization systems can leverage such user generated content by correlating it with the formal documentation and metadata presented in this paper. Successful examples of this are [34] [41]. On the other hand, similar approaches that took advantage of structured XML documents and fuzzy techniques for automatic ontology (taxonomy)

ACCEPTED MANUSCRIPT

generation are presented in [5] [6] and [8]. In our research, our focus is on cognitive metadata extraction nd not ontology generation. In conclusion, our novel fuzzy granulation method and fuzzy reasoning method achieved very competitive results comparing to state of the art with an average precision of 89.39% ranging from 82.47% to 96.90% depending on the metadata field.

7 Conclusions and Future Work

We have presented an automatic metadata generation framework, which extracts structural, descriptive, administrative and cognitive metadata from enterprise content for use in a personalized IR system. In particular, we have developed a DocBook Ontology and Resource Type Ontology to extract structural and descriptive metadata from DocBook documents in RDF format. The DocBook ontology is domain independent and can be used by other DocBook applications. In addition, we propose an algorithm to semi-automatically extract a Topic Ontology from DocBook indexterms using Simple Knowledge Organization System (SKOS). Finally, a novel fuzzy based information granulation and fuzzy inference system is proposed for cognitive metadata generation. The proposed method uses document structure, document semantics, fuzzy information granulation, fuzzy if-then rules and Mamdani fuzzy inference system to predict difficulty, interactivity type and interactivity level of a document using approximate reasoning.

The framework has been applied to the English, German and French versions of the Symantec Norton 360 technical documentations and a user study is conducted on the English content using 100 documents. The user study shows that the proposed fuzzy inference system achieves promising precision rates ranging from 82.47% to 96.90% depending on the metadata field. The proposed fuzzy inference system also compared with a rule-based reasoner; for difficulty metadata extraction fuzzy inference improved the performance of the rule-based reasoner ~11%. User perceived quality of the extracted metadata found to be high with an overall metadata quality rating of 5.57 out of 6. Evaluations on the quality of metadata records on the English, German and French versions of the Symantec Norton 360 also showed that the extracted metadata is high quality. Finally, our analysis on metadata fitness to personalized IR demonstrated that the extracted metadata can be suitable for personalization, which has been already applied and presented in [34] [41]. In future, we will apply and test our framework on the Microsoft content for personalized IR support.

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.

References

1. Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H. and Shadbolt, N. R. (2003). Automatic Ontology-Based Knowledge Extraction from Web Documents. IEEE Intelligent Systems, Vol. 18, No. 1, 14-21.
2. Alfano, M., Lenzitti, B., and Visalli, N. (2007). SAXEF: A System for Automatic eXtraction of E-learning object Features. Journal of e-learning and Knowledge Society. Vol. 3, No.2, 83-92.
3. Brusilovsky, P. 2001. Adaptive Hypermedia. User Modeling and User Adapted Interaction, Vol. 11, 87-110.
4. Cardinaels, K., Meire, M, and Duval, E. (2005). Automating metadata generation: the simple indexing interface. In International World Wide Web Conference, 548-556.
5. Ceravolo, P., Nocerino, M.C., and Viviani, M. (2004). Knowledge Extraction from Semi-Stuctured Data Based on Fuzzy Techniques. In International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), LNAI, Vol. 3215, 328-334.
6. Ceravolo, P., Damiani, E., and Viviani, M. (2007). Bottom-Up Extraction and Trust-Based Refinement of Ontology Metadata. IEEE Transactions on Knowledge and Data Engineering. Vol. 19, No. 2, 149-163.
7. Cimiano, P., S. Handshuh, S. Staab (2004). Towards The Self-annotating Web. In International World Wide Web Conference.
8. Cui, Z. Damiani, E., Leida, M., and Viviani, M. (2005). OntoExtractor: A Fuzzy-Based Approach in Clustering Semi-structured Data Sources and Metadata Generation. In International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), LNAI, Vol. 3681, 112-118.
9. Cunningham, H., D. Maynard, K. Bontcheva and V. Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In 40th Anniversary Meeting of the Association for Computational Linguistics.
10. Dublin Core (2006). Dublin Core Metadata Element Set, Version 1.1. dublincore.org.
11. Flesch, R. (1948). A new readability yardstick, Journal of Applied Psychology, Vol. 32, 221–233.
12. Flynn, P., Zhou, L., Maly, K., Zeil, S. and Zubair, M. (2007). Automated Template-Based Metadata Extraction Architecture. ICADL. LNCS, Vol. 4822, 327-336.
13. Foltz, P.W., Kintsch W., and Landauer T.K. (1998). The measurement of textual coherence with latent semantic analysis. Discourse Processes, Vol. 25, No. 2, 285-307.
14. Fuzzy Inference Systems tutorial. http://www.mathworks.com/help/toolbox/fuzzy/fp351dup8.html
15. Gueye, B., Rigaux, P., and Spyratos, N. (2004). Taxonomy-Based Annotation of XML Documents: Application to elearning Resources. In SETN, LNAI, Vol. 3025, 33-42.
16. Guy, M., Powell, A. and Day, M. (2004). Improving the quality of metadata in Eprint archives. In Ariadne international Conference.
17. Han, H., and Lee Giles, C., Manavoglu, E., and Zha, H., Zhang, Z., and Fox, E.A. (2003). Automatic Document Metadata Extraction Using Support Vector Machines. In ACM/IEEE-CS joint conference on Digital libraries, 37-48.
18. Huang, Z., Jin, H., Yuan, P., and Han, Z. (2006). Header Metadata Extraction from Semi-Structured Documents Using Template Matching. In OTM Workshops, LNCS, 4278, 1776-1785.
19. Jamison, N. 2010. Beyond the Customer Satisfaction Horizon Fostering Loyalty through Customer Service. http://www.jamison-consulting.com/pdf/BeyondtheCustomer SatisfactionHorizon011210.pdf

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

20. Jovanovic, J. Gasevic, D., and Devedzic, V. (2006). Ontology Based Automatic Annotation of Learning Content. International Journal on Semantic Web and Information Systems, Vol. 2, No. 2, 91-119.
21. IEEE Learning Object Model. (2002). http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf
22. Kiryakov, A. B. Popov, D. Ognyanoff, D. Manov, A. Kirilov and M. Goranov (2004). Semantic annotation, indexing, and retrieval. Journal of Web Semantics. Vol. 2, 49-79.
23. Khankasikam, K. (2010). A Hybrid Case-based and Rule-based for Metadata Extraction on Heterogeneous Thai Documents. In IEEE International Conference on Computer and Automation Engineering.
24. Klink, S., Dengel, A., Kieninger, T. (2000). Document structure analysis based on layout and textual features. In IAPR International Workshop on Document Analysis Systems, 99–111.
25. Li, Y., Zhu, Q., and Cao, Y. (2004). Automatic metadata generation based on neural network. In International Conference on Information Security, 192-197.
26. Liddy, E. D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N. E., Diekema, A., McCracken, N., Silverstein, J., and Sutton, S. (2002). Automatic Metadata Generation and Evaluation. In ACM SIGIR Conference on Research and Development in Information Retrieval.
27. Maedche, A. and Staab, S. (2002). Measuring Similarity between Ontologies. In International Conference on Knowledge Engineering and Knowledge Management, Ontologies and the Semantic Web, LNCS, 2473, 251-263.
28. Mao, S., Kim, J.W., Thoma, G.R. (2004). A Dynamic Feature Generation System for Automated Metadata Extraction in Preservation of Digital Materials. In International Workshop on Document Image Analysis For Libraries, Vol. 225, IEEE Computer Society.
29. Martinez-Ortiz, I., Moreno-Ger, P, Sierra-Rodriguez, J.L. and Fernandez-Manjon, B. (2006). Using DocBook and XML Technologies to Create Adaptive Learning Content in Technical Domains. International Journal of Computer Science and Applications. Vol. 3, No. 2, 91-108.
30. Radovanovic, M., and Ivanovic, M. (2008). Text mining: Approaches and applications. Novi Sad Journal of Mathematics Vol. 38, No. 3, 227–234.
31. Roy, D., Sarkar, S. and Ghose, S. (2008). Automatic Extraction of Pedagogic Metadata from Learning Content. International Journal of Artificial Intelligence in Education, 97-118.
32. Sah, M., and Wade, V. (2010). Automatic Metadata Extraction From Multilingual Enterprise Content. In International Conference on Information and Knowledge Management (CIKM), 1665-1668.
33. SKOS Core Guide. (2005). http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/
34. Steichen, B. and Wade, V. (2010). Adaptive Retrieval and Composition of Socio-Semantic Content for Personalised Customer Care. International Workshop on Adaptation in Social and Semantic Web, 1-10.
35. Walsh, N. and Muellner, L. 1999. The DocBook Definitive Guide, O'Reilly Media.
36. Wang, J., Ali, F. and Srimani, P. K. (2010). An efficient method to measure the semantic similarity of ontologies. International Journal of Pervasive Computing and Communications, Vol. 6, No, 1, 88-103
37. Yilmazel, O., Finneran, C. M., and Liddy, E. D. (2004). MetaExtract: An NLP System to Automatically Assign Metadata. In ACM/IEEE Conference on Digital Libraries, 241-242.
38. Zadeh, L. A. (1965). Fuzzy sets. Information and Control. Vol. 8, 338-353.
39. Zadeh, L. A. (1997). Towards a Theory of Fuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic. Fuzzy Sets and Systems. Vol. 90, 11-127.
40. Zhang, J., Niu, Y., Nie, H. (2009). Web Document Classification Based on Fuzzy k-NN Algorithm. In International Conference on Computational Intelligence and Security, 193-196.
41. Steichen, B., O'Connor, A., and Wade, V. (2011). Personalisation in the Wild – Providing Personalisation across Semantic, Social and Open-Web Resources. In ACM Conference on Hypertext and Hypermedia.
42. Sah, M., and Wade, V. (2011). Automatic Mining of Cognitive Metadata using Fuzzy Inference. In ACM Conference on Hypertext and Hypermedia.
43. Ochoa, X. and Duval, E. (2006). Quality metrics for Learning Object Metadata. In World Conference on Educational Multimedia, Hypermedia, Telecommunications.
44. Bruce, T. and Hillmann, D. (2004). The continuum of metadata quality: defining, expressing, exploiting. In *Hillmann and Westbrooks eds, Metadata in Practice*.
45. Lawless, S., Hederman, L., and Wade, V. (2008). OCCS: Enabling the Dynamic Discovery, Harvesting and Delivery of Educational Content from Open Corpus Sources. In IEEE International Conference on Advanced Learning Technologies.
46. McCalla, G. (2004). The Ecological Approach to the Design of E-Learning Environments: Purpose-based Capture and Use of Information About Learners. Journal of Interactive Media in Education, Special Issue on the Educational Semantic Web.
47. Brusilovsky, P. (2007). Adaptive Navigation Support, The Adaptive Web: Methods and Strategies of Web Personalization. LNCS, 4321, 263-290.
48. Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch, S. (2007). Personalized Search on the World Wide Web, The Adaptive Web: Methods and Strategies of Web Personalization. LNCS, 4321, 195-230.
49. Bunt, A., Carenini, G., and Conati, C. (2007). Adaptive Content Presentation for the Web, The Adaptive Web: Methods and Strategies of Web Personalization. LNCS, 4321, 409-432.
50. Conlan, O., Wade, V. (2004). Evaluation of APeLS - An Adaptive eLearning Service Based on the Multi-model, Metadata-Driven Approach. Adaptive Hypermedia and Adaptive Web-based Systems, LNCS, 3137, 291-295.
51. O'Keeffe, I., Wade, V. (2009). Personalised Web Experiences: Seamless Adaptivity across Web Service Composition and Web Content. In User Modeling Adaptation and Personalization, 5535, 480-485.
52. DocBook Ontology, http://www.scss.tcd.ie/Melike.Sah/docbook.owl
53. Resource Type Ontology, http://www.scss.tcd.ie/Melike.Sah/resourcetype.owl
54. Open Calais, http://www.opencalais.com/
55. Brusilovsky, P., Kobsa, A. and Nejdl, W. (2007). The Adaptive Web: Methods and Strategies of Web Personalization. LNCS, 4321.
56. Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin, Vol. 76, No. 5, 378–382.
57. Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. In *Biometrics*. Vol. 33, 159–174.

ACCEPTED MANUSCRIPT

Table 1. Fuzzy if-then-rules for estimating LOM difficulty

Concept (*C_length*)	Activity (*A_length*)			
	Null	*Low*	*Medium*	*High*
Low	Very Easy	Easy	Medium	Very Difficult
Medium	Easy	Medium	Difficult	Very Difficult
High	Easy	Medium	Difficult	Very Difficult

Table 2. Fuzzy if-then-rules for estimating LOM interactivity level

Activity (*A_length*)				
Null	*Low*	*Medium*	*High*	*Very high*
Very Low	Low	Medium	High	Very High

Table 3. Fuzzy if-then-rules for estimating LOM interactivity type

Concept (*C_length*)	Activity (*A_length*)			
	Null	*Low*	*Medium*	*High*
Low	Expositive	Active	Active	Active
Medium	Expositive	Mixed	Active	Active
High	Expositive	Mixed	Mixed	Active

Table 4. Precision, recall and f-measure scores for LOM interactivity type

	Precision	**Recall**	**F-Measure**
User #1	84.53% (82/97)	82% (82/100)	83.24%
User #2	94.84% (92/97)	92% (92/100)	93.39%
User #3	80.82% (59/73)	80.82% (59/73)	80.82%
User #4	97.93% (95/97)	95% (95/100)	96.44%
User #5	96.90% (94/97)	94% (94/100)	95.42%
Average	91.00%	88.76%	89.86%

Table 5. Precision, recall and f-measure scores for LOM interactivity level

	Precision	**Recall**	**F-Measure**
User #1	78.35% (76/97)	76% (76/100)	77.15%
User #2	64.94% (63/97)	63% (63/100)	63.95%
User #3	72.60% (53/73)	72.60% (53/73)	72.60%
User #4	91.75% (89/97)	89% (89/100)	90.35%
User #5	89.69% (87/97)	87% (87/100)	88.32%
Average	79.46 %	77.52%	78.47%

Table 6. Precision, recall and f-measure scores for LOM difficulty

	Precision	**Recall**	**F-Measure**
User #1	74.22% (72/97)	72% (72/100)	73.09%
User #2	50.51% (49/97)	49% (49/100)	49.74%
User #3	35.61% (26/73)	35.61% (26/73)	35.61%
User #4	85.56% (83/97)	83% (83/100)	84.26%
User #5	85.56% (83/97)	83% (83/100)	84.26%
Average	66.29%	64.52%	65.39%

Table 7. Precision, recall and f-measure against the agreed annotations

	Precision	**Recall**	**F-Measure**
Int.Type	96.90% (94/97)	94% (94/100)	95.42%
Int.Level	88.65% (86/97)	86% (86/100)	87.30%
Difficulty	82.47% (80/97)	80% (80/100)	81.21%
Average	89.39%	86.66%	87.97%

Table 8. The average of precision rates depending on the background information of participants in annotation

	Interactivity Type	**Interactivity Level**	**Difficulty**
Beginner (user 2, 3)	87.83%	68.77%	43.06%
Advanced (user 1, 4, 5)	93.12%	86.59%	81.78%

ACCEPTED MANUSCRIPT

Table 9. The average of precision rates depending on the background information of participants in Symantec products

	Interactivity Type	Interactivity Level	Difficulty
Beginner (user 3)	80.82%	72.60%	35.61%
Intermediate (user 1, 2, 4, 5)	93.55%	81.18%	73.96%

Table 10. Precision rates of the agreed annotations depending on the background information of participants

	Interactivity Type	Interactivity Level	Difficulty
Advanced in annotation	97.93% (95/97)	90.72% (88/97)	84.53% (82/97)
Intermediate in Symantec products	97.93% (95/97)	88.65% (86/97)	82.47% (80/97)

Table 11. Inter-agreement statistics of participants based on Fleiss' kappa measure [56]

	Interactivity Type	Interactivity Level	Difficulty
All participants	κ =0.83	κ =0.62	κ =0.61
Participants advanced in annotation	κ =0.85	κ =0.64	κ =0.65
Participants intermediate in Symantec products	κ =0.84	κ =0.62	κ =0.61

Table 12. Fuzzy inference and rule-based inference comparison in terms of precision, recall and f-measure against the agreed annotations

	Fuzzy Inf. Precision	Rule-base Precision	Fuzzy Inf. Recall	Rule-base Recall	Fuzzy Inf. F-Measure	Rule-base F-Measure
Int. Type	96.90%	96.90%	94%	94%	95.42%	95.42%
Int. Level	88.65%	86.59%	86%	84%	87.30%	85.27%
Difficulty	82.47%	71.13%	80%	69%	81.21%	70.04%
Average	89.39%	84.87%	86.66%	82.33%	87.97%	83.57%

Table 13. Quality metrics formula based on [43]

Metric Name	Metric Formula
Simple Completeness	$\sum_{i=1}^{N} P(i) \Big/ N$, $P(i)=1$, if ith field $\neq$ null; $P(i)=0$, if ith field = null
Weighted Completeness	$\sum_{i=1}^{N} \alpha_i * P(i) \Big/ \sum_{i=1}^{N} \alpha_i$
Consistency	$\sum_{i=1}^{N} level\_of\_compliance(field_i, value_i) \Big/ N$
Timeliness (age)	$present\_age - publication\_year$
Accessibility (Linkage)	$Average\_linkage = \frac{no\_of\_referenced\_links}{total\_no\_of\_documents}$

Table 14. The results of quality metrics applied to Symantec Norton 360 English, German and French technical contents

Metrics #instances	N360 English (639 docs)	N360 German (671 docs)	N360 French (858 docs)
Completeness	9.94	9.91	9.90
Weighted Completness	9.86	9.79	9.73
Consistency	10.00	10.00	10.00
Timeliness (age)	10.00	10.00	10.00
Linkage(not normalized)	4.40	3.58	2.79

Table 15. Functional Requirements List

We would like users to be able to:		*We would like to be able to:*
Search records by: Title	Browse search results by: Document Tree (DocBook ontology)	Personalize based on: User Knowledge

ACCEPTED MANUSCRIPT

Keyword Document Type (i.e., chapter, section, image, etc.)	Topic Vocabulary (Topic Ontology) Document Type Relevant Documents	User Interest User Expertise Document Difficulty Docum. Interactivity Type Docum. Interactivity Level Process Type Document Type Document Re-composition

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

```
<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE section PUBLIC "-//OASIS//DTD DocBook V5.0//EN">
<section status="source" revision="12354" id="v123456">
  <title id="v234567">Spam filtering features</title>
    <indexterm>
       <primary>Norton 360</primary>
       <secondary>Allowed and Blocked lists</secondary>
    </indexterm>
     <para id="v4567123">With the increase in ... </para>
     <procedure id="v7891234">
          <step id="v8765543">...    </step>
    </procedure>
    ...
    </section>
```

Fig. 1. A fragment of a DocBook document from Symantec Norton 360

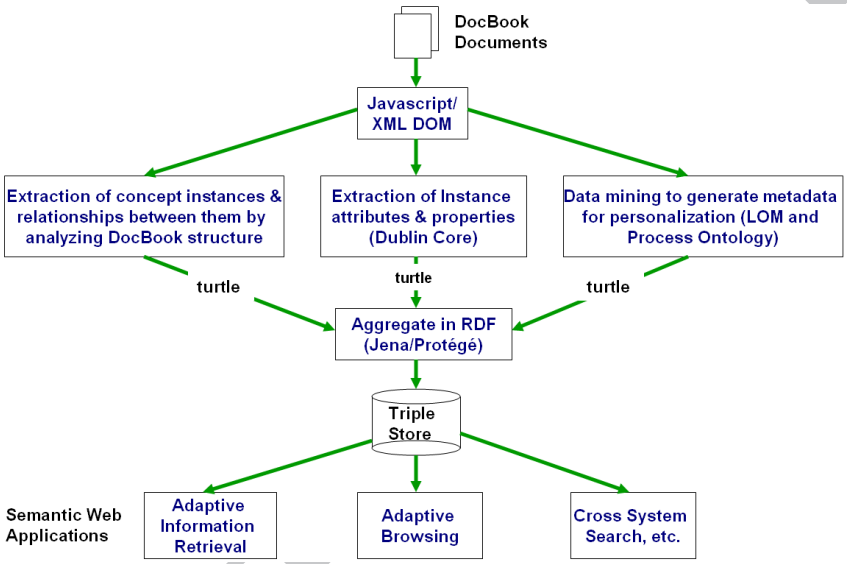


Fig. 2. A framework for automatic metadata extraction

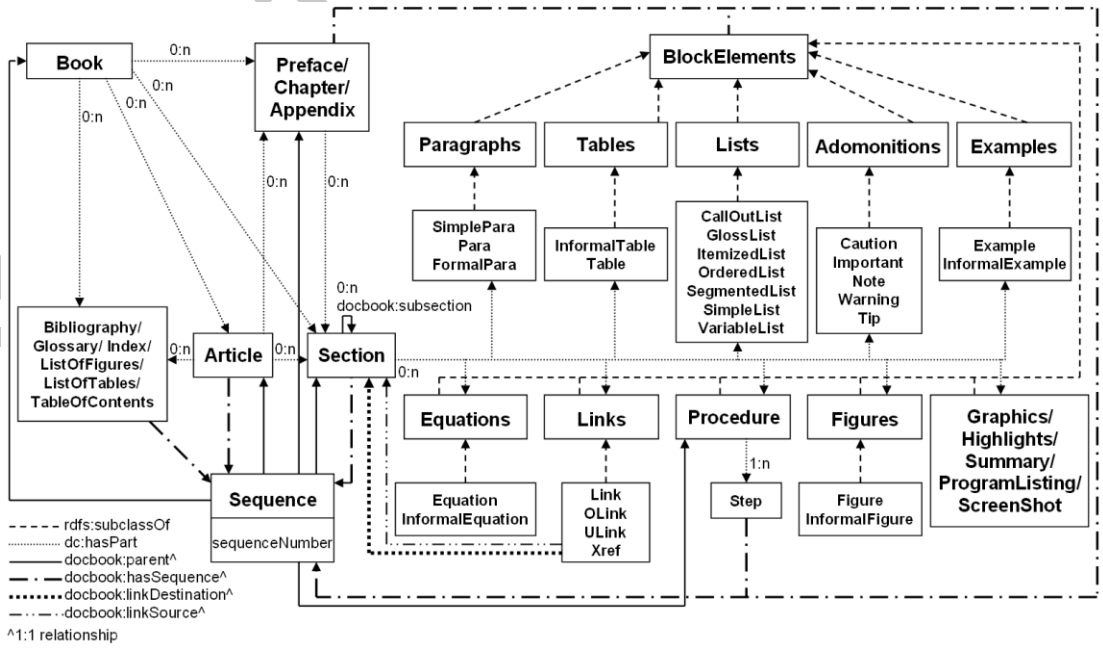


Fig. 3. The overview of the DocBook ontology

ACCEPTED MANUSCRIPT

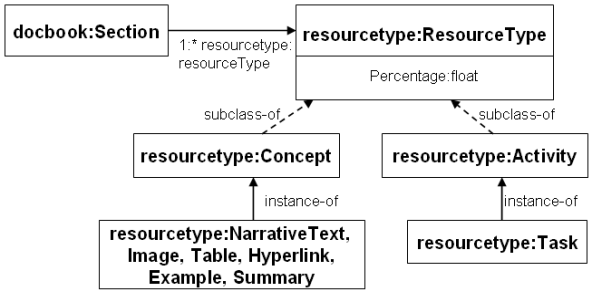


Fig. 4. The overview of the Resource Type Ontology

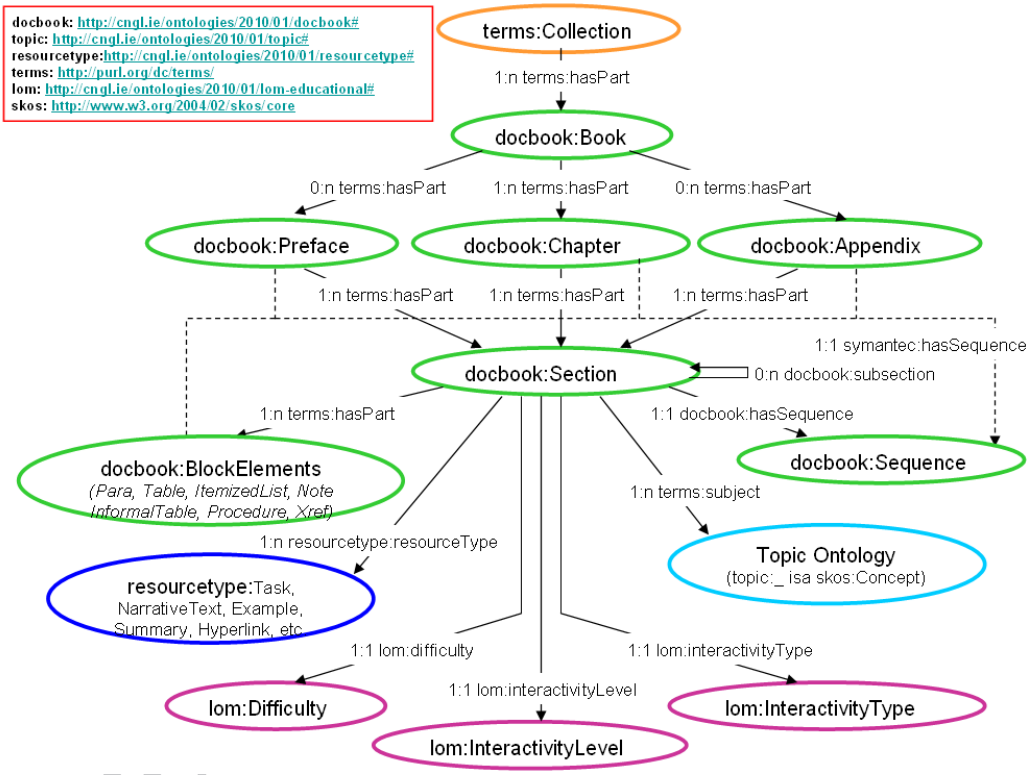
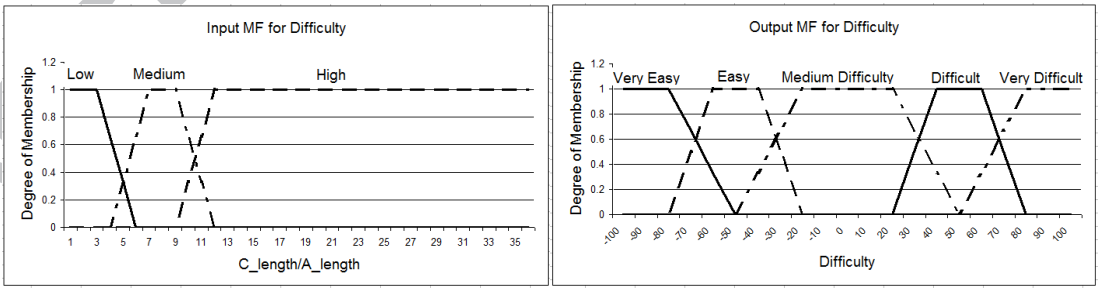
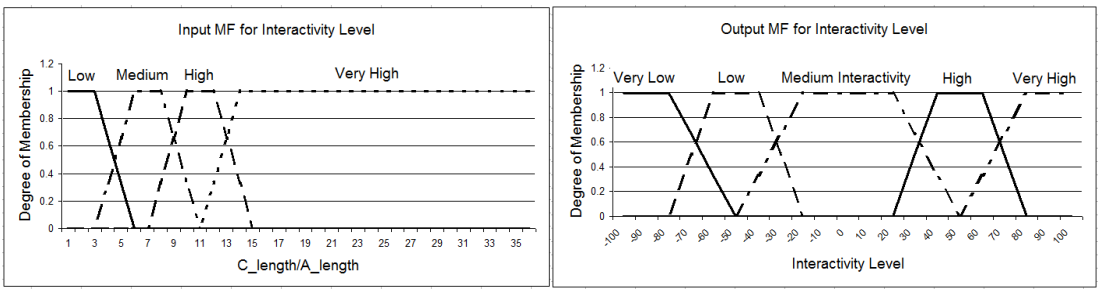


Fig. 5. Symantec Profile and relationships between other schemas

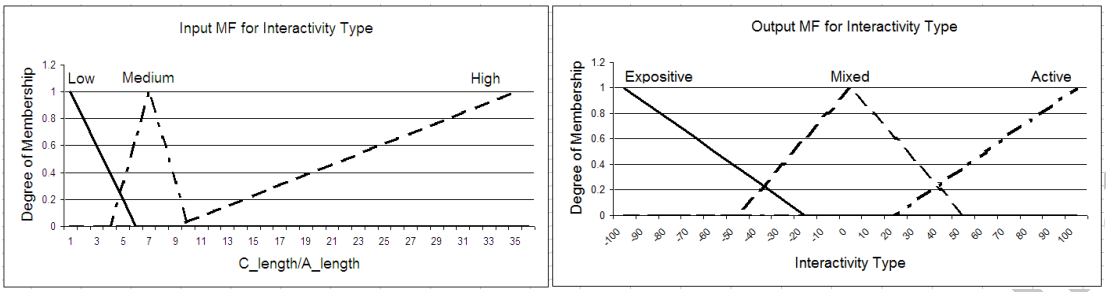


(a) Input (left) and output (right) Membership Functions for difficulty



ACCEPTED MANUSCRIPT

(b) Input (left) and output (right) Membership Functions for interactivity level



(c) Input (left) and output (right) Membership Functions for interactivity type

Fig. 6. Input and output Membership Functions for the proposed fuzzy inference system

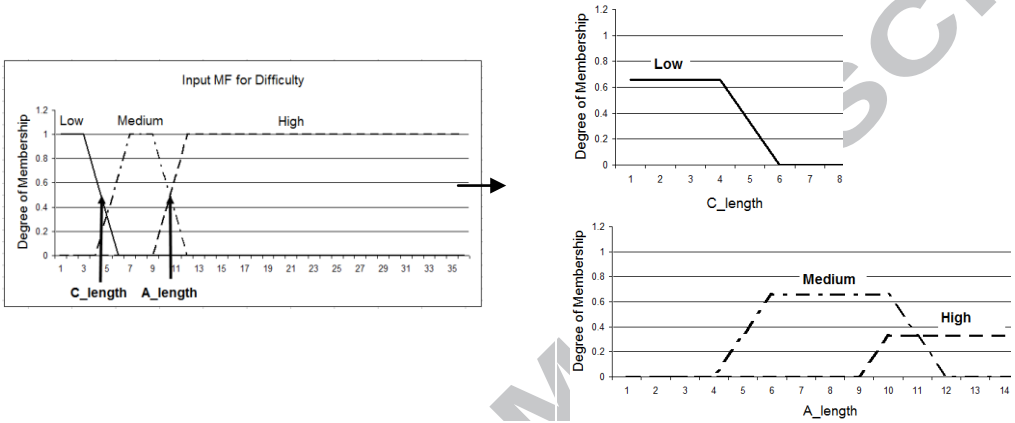


Fig. 7. Fuzzification of *C_length* and *A_length* input values

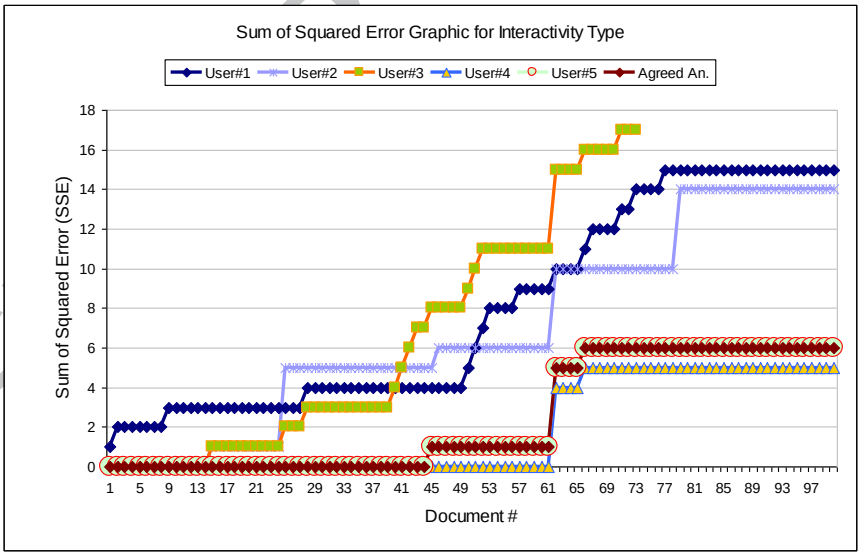


Fig. 8. Sum of Squared Error (SSE) graphic for LOM interactivity type

ACCEPTED MANUSCRIPT

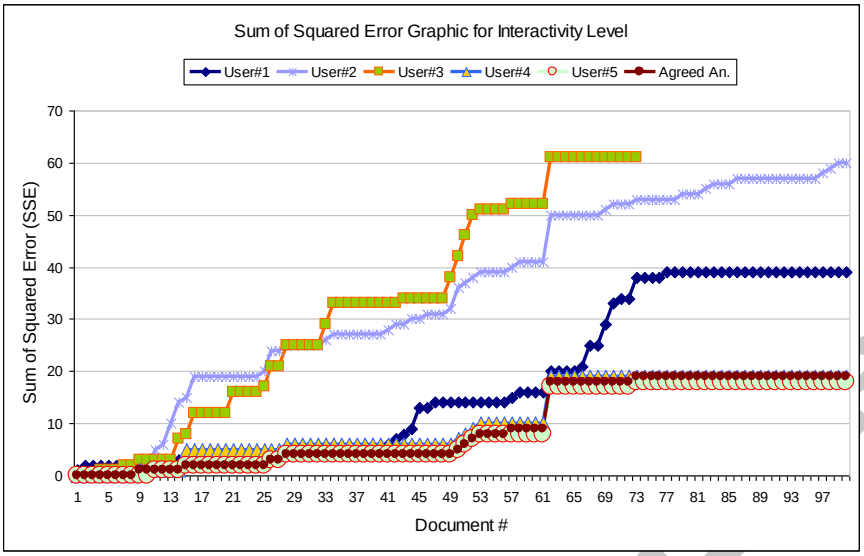


Fig. 9. Sum of Squared Error (SSE) graphic for LOM interactivity level

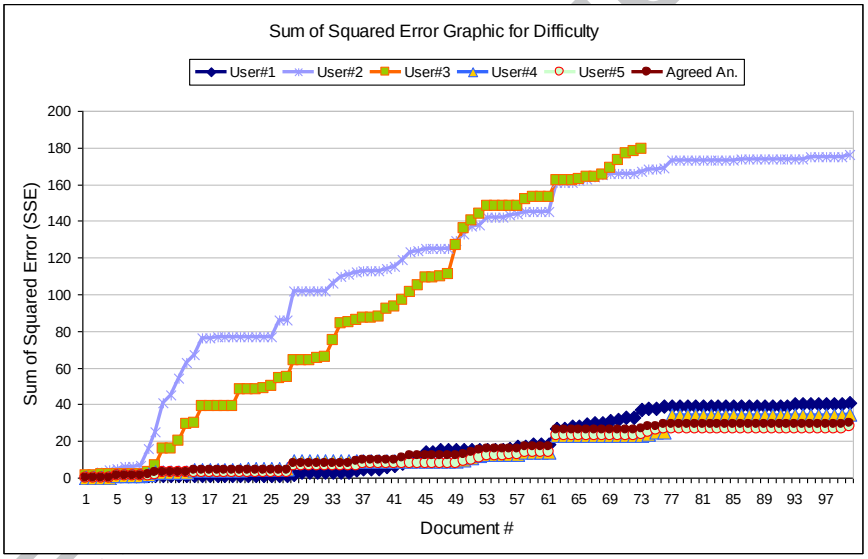


Fig. 10. Sum of Squared Error (SSE) graphic for LOM difficulty

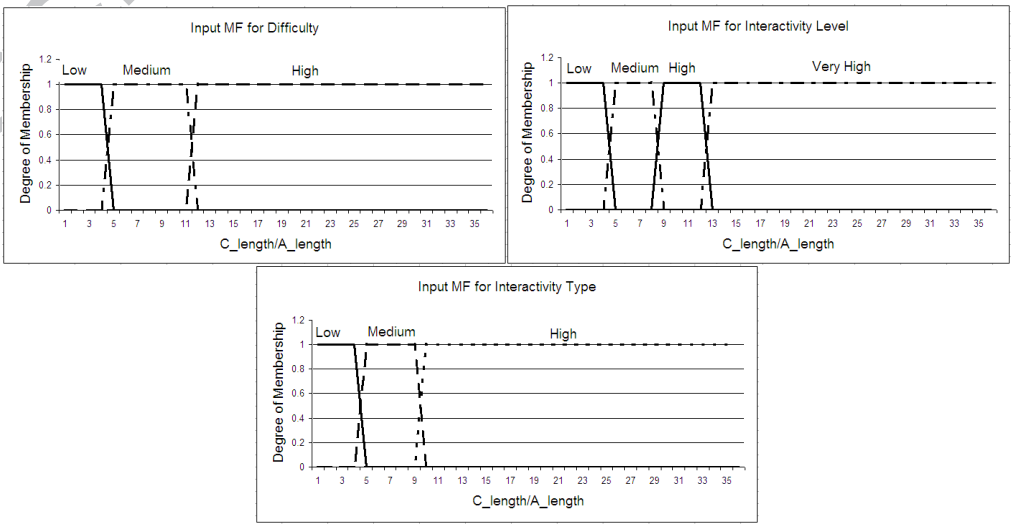
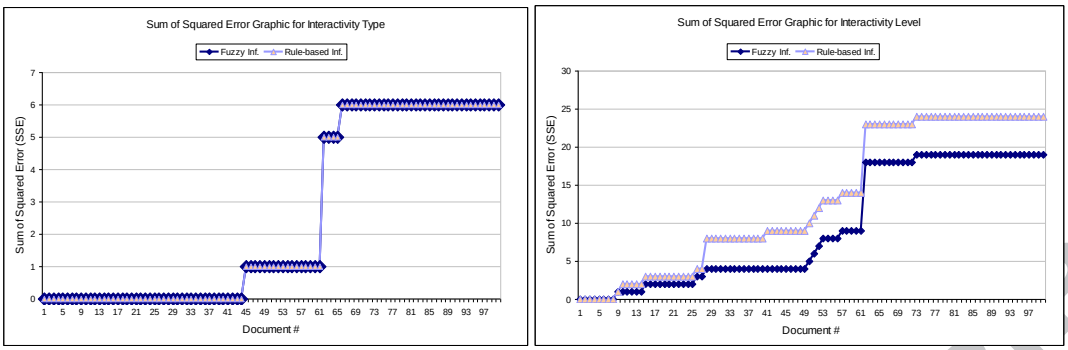
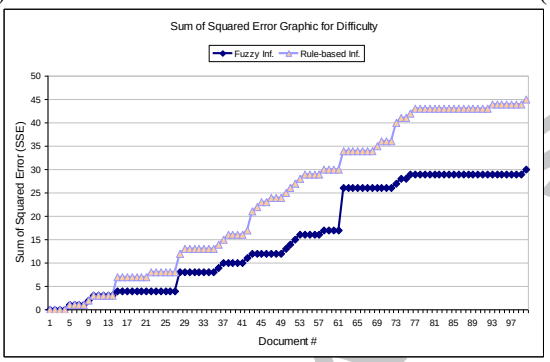


Fig. 11. Input sets (crisp sets) for the rule-based reasoner

ACCEPTED MANUSCRIPT



(a) (b)



(c)

Fig. 12. Sum of Squared Error (SSE) graphic of the fuzzy model and rule-based reasoner against the agreed annotations for LOM interactivity type (a), LOM interactivity level (b) and LOM difficulty (c)

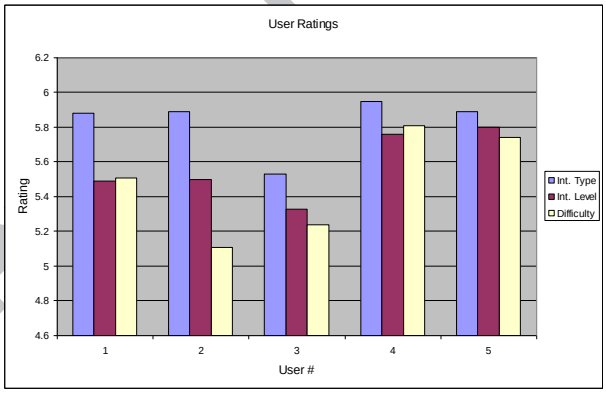


Fig. 13. Average of subjects' ratings to Interactivity Type, Interactivity Level and Difficulty

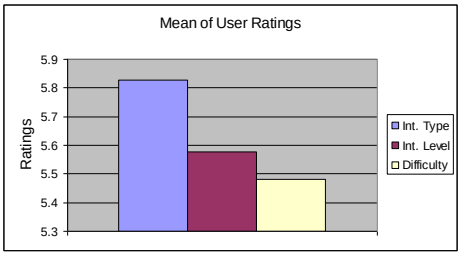


Fig. 14. Mean of users' ratings to interactivity type, interactivity level and difficulty

ACCEPTED MANUSCRIPT

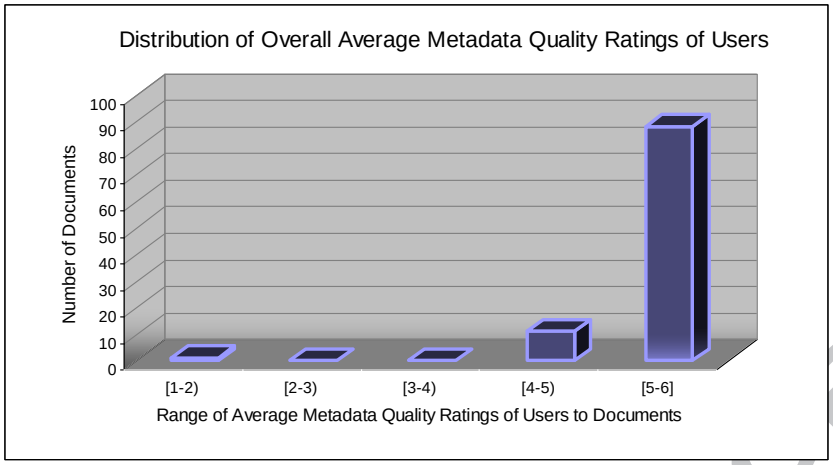


Fig. 15. Distribution of overall average metadata quality ratings of users to documents

ACCEPTED MANUSCRIPT