

# Automatic Mining of Cognitive Metadata using Fuzzy Inference

Melike Şah

Centre for Next Generation Localisation,  
Knowledge and Data Engineering Group,  
Trinity College Dublin, Ireland  
Melike.Sah@scss.tcd.ie

Vincent Wade

Centre for Next Generation Localisation,  
Knowledge and Data Engineering Group,  
Trinity College Dublin, Ireland  
Vincent.Wade@scss.tcd.ie

## ABSTRACT

Personalized search and browsing is increasingly vital especially for enterprises to be able to reach their customers. Key challenge in supporting personalization is the need for rich metadata such as cognitive metadata about documents. As we consider size of large knowledge bases, manual annotation is not scalable and feasible. On the other hand, automatic mining of cognitive metadata is challenging since it is very difficult to understand underlying intellectual knowledge about documents automatically. To alleviate this problem, we introduce a novel metadata extraction framework, which is based on fuzzy information granulation and fuzzy inference system for automatic cognitive metadata mining. The user evaluation study shows that our approach provides reasonable precision rates for difficulty, interactivity type, and interactivity level on the examined 100 documents. In addition, proposed fuzzy inference system achieves improved results compared to a rule-based reasoner for document difficulty metadata extraction (11% improvement).

## Categories and Subject Descriptors

I.2.3 [Artificial Intelligence]: Deduction and Theorem Proving – *uncertainty, fuzzy and probabilistic reasoning*.

## General Terms

Algorithms, Performance, Design.

## Keywords

Automatic metadata extraction, cognitive metadata, fuzzy inference, personalization, IEEE LOM.

## 1. INTRODUCTION

Enterprises provide 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 (i.e. title and subject), which is not sufficient to supply sophisticated user interfaces. Moreover, customers have high expectations from service providers. In particular, users prefer personalized customer support services in their preferred languages [1]. Hence, there is a growing interest to personalized customer care

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HT'11, June 6–9, 2011, Eindhoven, The Netherlands.

Copyright 2011 ACM 978-1-4503-0256-2/11/06...\$10.00.

because personalization helps users to stay more on the website and re-encourages them to return to the service provider.

Personalization needs rich metadata such as cognitive metadata. Although many automatic metadata extraction techniques are proposed, they mainly extract descriptive metadata (i.e. title, subject) [2-5], which is not enough to support personalization. There is much less research focused on automatic extraction of cognitive metadata since it is harder to understand the context and underlying intellectual knowledge about documents automatically. In addition, the feasibility of the approach such as pre-processing, training, extraction speed and precision are important for deployment.

This paper proposes an automatic metadata extraction framework based on fuzzy information granulation and fuzzy inference system to extract cognitive metadata from semi-structured documents. The framework extracts interactivity type, interactivity level and difficulty of documents, which is very useful to support personalization, such as personalization based on the user's competence (using document difficulty), based on the user's task (using document interactivity level) and based on the user's preference (using documents interactivity type). Our framework also uses different document parsing algorithms to generate descriptive, structural and administrative metadata as presented in [6].

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 explains enterprise content. Section 4 introduces the proposed fuzzy inference system. Section 5 presents evaluations prior to conclusions and future work.

## 2. RELATED WORK ON AUTOMATIC COGNITIVE/PEDAGOGICAL METADATA EXTRACTION

Even though there are many approaches that automatically extract descriptive metadata and general metadata fields [2-5], there are few approaches on the extraction of cognitive or pedagogical metadata automatically, where such metadata is very useful to support personalization. In the digital library content, these metadata values are usually manually provided by the content author, which is a labor intensive job and automated techniques are required for scalability.

Roy et al. [7] uses an automatic annotation tool for annotating learning objects with pedagogical metadata such as concepts/concept significance, type of concepts and learning resource type. Concept type (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 neural networks algorithms. Our framework also extracts document types as discussed in [6], but it is out of the scope of this paper.

Jovanovic et al. [8] present an ontology-based approach for automatic annotation of learning objects based on IEEE LOM. In their work, they generate metadata from presentation slides. 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.

[9] utilize parsing rules and NLP techniques to extract terms and phrases from sentences. In particular, system generates metadata about pedagogy-teaching method, pedagogy-grouping, pedagogy-assessment, pedagogy-process. However, the user study showed that pedagogic metadata fields had received very low metadata quality scores.

Metadata generation about document difficulty has not been automated by any metadata generation system before. To the best of our knowledge, our proposed fuzzy based method for the first time automates this process and achieves promising initial results as discussed in the evaluations section. On the other hand, text readability difficulty can be inferred automatically by using readability indexes [10] [11]. The Flesch Index measures text comprehension difficulty by analyzing word lengths and sentence lengths in a document [10]. 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 [11]. 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 documents are very structured and formatted by XML, which allow intellectual analysis of document semantics and reasoning using fuzzy inference. On the other hand, [12] [13] and [14] uses structured XML documents and fuzzy techniques for automatic ontology (taxonomy) generation. In our research, our focus is cognitive metadata extraction and not ontology generation. In conclusion, the proposed fuzzy granulation method and fuzzy inference system achieves very competitive results comparing to the state of the art with an average precision of 89.39% ranging from 82% to 96% depending on the metadata field.

### 3. 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 using XML [15]. DocBook was originally designed to enable the interchange of computer documentation between companies, such as Microsoft, Hewlett Packard, Sun Microsystems and Symantec use DocBook DTD.

DocBook DTD is not just limited to enterprise domain. It has been utilized for creating learning object resources in e-learning [16] [17].

DocBook documents are very structured and the content semantics are described in detail using DocBook tags. In our case study, we have used the English version of Symantec Norton 360 technical documentation. The content is formatted in XML with a subset of DocBook DTD as presented in Figure 1. Our objective is to automatically extract rich metadata from Symantec content for use in personalized customer support. [18] presents a personalized information retrieval system based on the automatically extracted metadata by our framework.

```
?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>
  <para id="v4567123">With the increase in ... </para>
  <procedure id="v7891234">
    <step id="v8765543">... </step>
  </procedure>...
</section>
```

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

### 4. MINING COGNITIVE METADATA USING FUZZY INFERENCE

DocBook documents need descriptive information about cognitive metadata such as difficulty that can be valuable for personalization. However such metadata is not supported by DocBook and we reused IEEE LOM [19] which provides useful metadata for personalization. After analyzing DocBook formatted enterprise content, we decided to use three entities: Difficulty, interactivity type and interactivity level. Difficulty describes how hard it is to work through the resource. Interactivity Type is the pre-dominant mode of learning supported by the resource (i.e. active, mixed and expositive). Interactivity Level is the degree of interactivity characterizing the resource. Information about these LOM elements could be very useful for personalization. For example, if we know the user has little experience with a topic, then a personalized IR can start showing the search results from easy to difficult documents. Assume the user explicitly provided the type of documents s/he wants during search, 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. Therefore, content can be re-organized based on the user’s needs. As well as, interactivity level of documents can be used to filter documents according to the task of the user. [18] presents possible use of these metadata elements for personalization.

LOM 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 large knowledge bases, 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 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 [20]. 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 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. First, we provide a brief summary of fuzzy sets and fuzzy logic concepts needed for the subsequent text.

#### 4.1 Basic Concepts of Fuzzy Logic

**Fuzzy Logic:** Fuzzy Logic (FL) is a multivalued logic to deal with reasoning using approximate values rather than crisp values of 0/1 and true/false. It is often referred as approximate reasoning. FL is used in the broad sense and almost synonymous with fuzzy set theory. Fuzzy set theory is a theory of classes with unsharp boundaries and it is considered as an extension of classical set theory. In the classical set theory, if an element  $x$  is a member of set  $A$ , membership function  $\mu_A(x) = 1$ , otherwise  $\mu_A(x) = 0$ . In many cases, however it is not quite clear whether  $x$  belongs to a set  $A$  or not since variables may not have sharp boundaries. Conversely, fuzzy sets provide graded membership between 0 and 1 as follows:

$$A = \int_X \mu_A(x) / x, \text{ if } A \text{ is continuous,}$$

$$A = \mu_1(x_1) / x_1 + \dots + \mu_n(x_n) / x_n \text{ or } A = \mu_1 x_1 + \dots + \mu_n x_n, \text{ if } A \text{ is finite} \quad (1)$$

where  $\mu_A(x): X \rightarrow [0,1]$  is Membership Function (MF),  $A$  is fuzzy set, which can be represented with a linguistic variable (i.e. young) and  $X$  is known as the universe of discourse. The higher the membership degree of  $x$  in fuzzy set  $A$ , the more true that  $x$  is in  $A$ .

**Fuzzy Set Operations:** As in classical logic, in FL, there are three basic operations on fuzzy sets: intersection, unification and complement. These operations can be represented as follows: Let  $A$  and  $B$  be fuzzy sets in a universe of discourse  $U$ , and  $\mu_A$  and  $\mu_B$  as the MFs of  $A$  and  $B$ , respectively, then

$$\begin{aligned} A \cap B &\Leftrightarrow \mu_{A \cap B} = \min(\mu_A, \mu_B) \\ A \cup B &\Leftrightarrow \mu_{A \cup B} = \max(\mu_A, \mu_B) \\ A' &\Leftrightarrow \mu_{A'} = 1 - \mu_A \end{aligned} \quad (2)$$

**Fuzzy Information Granulation (FIG):** In human cognition, there are three basic concepts: granulation, organization and causation. Granulation is partitioning whole into parts, organization involves combining parts into whole and causation involves causation with effects. According to Zadeh, this is how 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 [21]. But methodology and TFIG is mathematical. FIG can be characterised as follows: it is a mode of generalization which may be applied to any concept, method or theory. FIG involves granulation, fuzzification and reasoning. In granulation, a set is partitioned into granules. For example, human head can be partitioned into granules of forehead, eyes, nose, cheeks, etc. As in human reasoning and formation of concepts, in FIG, granules, their attributes and their values are fuzzy. For example, boundaries of granules are not sharply defined between cheek and nose. In addition, granules are associated with fuzzy attributes (i.e. length of hair) and fuzzy values (i.e. long, short, etc.). Hence fuzzification is the generalization of granules, their attributes and values. In this sense, a crisp granule is replaced by a fuzzy granule; a crisp set is replaced by a fuzzy set. In our opinion, TFIG can be applied to document content to support automatic and intelligent information analysis for metadata creation by using a fuzzy inference system.

**Mamdani Fuzzy Inference System (FIS):** A FIS is a way of mapping an input space to an output space by using FL. Two of the well known FIS systems are Mamdani and Takagi-Sugeno-Kang [14]. We utilized the most commonly used Mamdani FIS method because of its simple structure and implementation. Mamdani's method consists of four steps: fuzzification of input values, defining fuzzy if-then rules, fuzzy inference and defuzzification, which are summarized below.

**Fuzzification:** Fuzzification has different levels. In a broad sense, fuzzification can be seen as generalization; a crisp set is replaced by a fuzzy set. In the narrow sense, it is the process of finding membership degree of a non-fuzzy input value; converting real numbers to MFs,  $Fuz(\mathfrak{X}) \rightarrow F$ .

**If-Then Rules:** In simplest form, a fuzzy if-then rule has the pattern of:

$$\text{If } x \text{ is } A \text{ then } z \text{ is } C \quad (3)$$

where  $A$  and  $C$  are linguistic values defined by fuzzy sets in the universe of discourse  $X$  and  $Z$ ,  $x$  is the input variable and  $z$  is output variable. The input to the rule is a crisp (numeric) value given to the input variable,  $x$ . The output of the rule is a fuzzy set assigned to the output variable,  $z$ . Left side of the rule is referred as antecedent and the right side of the rule is referred as consequent. If the antecedent or consequent of a given rule has more than one input variable, the fuzzy operator is applied to obtain one number that represents the result of an antecedent or consequent for that rule. The fuzzy operator depends of how the parts of the antecedent are joined. For example if two parts of an antecedent are joint by conjunction (If  $x$  is  $A$  and  $y$  is  $B$  then  $z$  is  $C$ ), then fuzzy AND operation is applied ( $\min(A, B)$ ).

**Fuzzy Inference:** Assume the consequent of each rule is a fuzzy set. Then, according to the antecedent value of each rule, a fuzzy implication operator is applied to obtain a new fuzzy set for each rule's conclusion. The most commonly used implication methods are the minimum and the product. After obtaining each rule's conclusion, a fuzzy aggregation operator is applied to combine the outputs obtained by each rule into a single fuzzy set. For this purpose the maximum, the sum or the probabilistic sum operators can be used.

**Defuzzification:** At the end of the decision problem, we want to obtain a number and not a fuzzy set, therefore the system need to transform the final fuzzy set into a single numerical value, which is known as defuzzification. This can be performed by utilizing a

number of methods, such as Mean Of Maximum (MeOM), First Of Max (FOM), Last of Max (LOM), centroid, bisector, etc. Centroid is one of the most popular defuzzification method, which returns the center of the area under the output fuzzy set. In centroid method, outputs of all rules contribute to the result.

## 4.2 Proposed Fuzzy Information Granulation and Fuzzy Inference System

In algorithm 1, the step-by-step metadata generation process is provided. The proposed approach is represented more comprehensively as follows:

---

**Algorithm 1.** Fuzzy based metadata generation algorithm  
Input:  $D$  – Document formatted by DocBook DTD  
Output: A file that contains the generated metadata in RDF  
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 for reasoning. Our aim is to apply TFIG for reasoning on metadata values. 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, such as parts that are interactive (i.e. a web forum requests an input) and parts that are not (i.e. 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, DocBook DTD elements Paragraphs, Lists, Tables, Summary, Examples, Figures, Equation and Links within a DocBook document represent *Concepts* that are expositive and non-interactive content. DocBook DTD elements 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 documents into *Concepts* ( $C$ ) and *Activities* ( $A$ ) FIGs, which can be used for creating metadata.

$$Concept \text{ isfg } C, Activity \text{ isfg } A \quad (4)$$

**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 maximum values of  $C\_length$  (*Concept*) and  $A\_length$  (*Activity*) as follows:  $C\_length_{min}=1$ ,  $C\_length_{max}=37$ ,  $A\_length_{min}=0$ ,  $A\_length_{max}=35$ . As a result,  $U = [0, 37]$ .  $U$  can be partitioned into different intervals. For example, for four fuzzy set,  $U$  can be partitioned into four unequal intervals,  $u_i, i=1,4$ , 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. 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 among the combinations of fuzzy MFs, the best possible number of fuzzy sets and defuzzification methods, we also conducted an initial experiment to analyze the prediction error of the selected FIS 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 different FISs 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  $A\_length$ . For interactivity level, we only used four fuzzy sets since interactivity level is directly proportional with the membership degree of  $A\_length$  and it can take five output values. For example, in the case of there is no interactive content ( $A\_length=null$ ), the metadata value is very low interactivity and for other four input combinations, we use four fuzzy sets to represent the input space. The RMSE of each individual fuzzy model is evaluated using equation (5):

$$E_i = \sqrt{\frac{\sum_{j=1}^n (P_{ij} - T_j)^2}{n}} \quad (5)$$

where  $P_i$  is the numerical value predicted by the individual FIS  $I$  for sample case  $j$  (output of equation 6),  $T_j$  is the target metadata value for sample case  $j$  and  $n$  is the number of samples. For the calculation of RMSE, target metadata values that are provided by the expert are represented by numerical values, where we took the numerical value under the middle of maximum on output MFs. For example, for difficulty, very low = -100, low = -50, medium = 0, high = 50 and very high = 100 based on Figure 2. For a perfect match,  $P_{ij} = T_j$ ,  $E_i = 0$ . RSME of individual FISs for difficulty, interactivity level and interactivity type are presented in Tables 1, 2 and 3 respectively.

**Table 1. RMSE of LOM difficulty with respect to various MFs, number of fuzzy sets and defuzzification methods**

Input/Output Membership Function	Defuzzification Method	RMSE 3 fuzzy sets	RMSE 4 fuzzy sets	RMSE 5 fuzzy sets
Triangular	Centroid	23.79	19.93	28.28
	MeOM	25.0	20.91	23.71
	FOM	25.0	20.15	22.36
	LOM	25.0	22.36	27.38
Trapezoid	Centroid	14.79	19.35	26.48
	MeOM	15.24	23.26	27.95
	FOM	21.56	25.88	27.38
	LOM	26.17	27.74	29.58
Gaussian	Centroid	35.62	23.28	25.28
	MeOM	27.38	21.65	23.48
	FOM	27.38	22.36	22.38
	LOM	27.38	25.0	22.58
Generalized Bell	Centroid	31.21	20.32	20.25
	MeOM	25.61	20.15	20.91
	FOM	25.0	19.36	15.81
	LOM	27.38	22.36	27.38

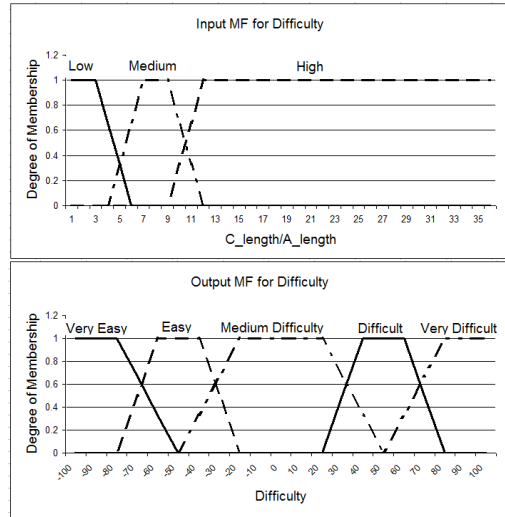
**Table 2. RMSE of LOM interactivity level with respect to various MFs, no of fuzzy sets and defuzzification methods**

Input/Output Membership Function	Defuzzification Method	RMSE 4 fuzzy sets
Triangular	Centroid	20.76
	MeOM	23.04
	FOM	25.00
	LOM	22.36
	Centroid	20.56
	MeOM	23.45

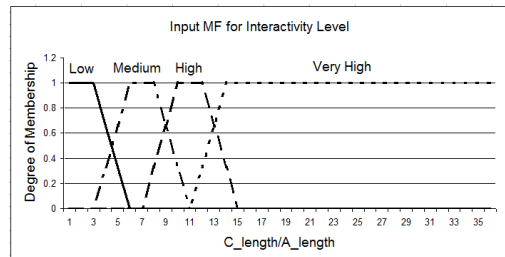
Trapezoid	FOM	27.56
	LOM	24.08
Gaussian	Centroid	25.30
	MeOM	23.04
	FOM	25.00
	LOM	22.36
Generalized Bell	Centroid	21.18
	MeOM	22.36
	FOM	22.36
	LOM	22.36

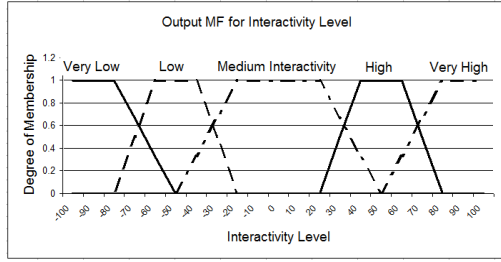
**Table 3. RMSE of LOM interactivity type with respect to various MFs, no of fuzzy sets and defuzzification methods**

Input/Output Membership Function	Defuzzification Method	RMSE 3 fuzzy sets	RMSE 4 fuzzy sets
Triangular	Centroid	0.0	0.0
	MeOM	0.0	0.0
	FOM	0.0	0.0
	LOM	0.0	0.0
Trapezoid	Centroid	0.0	3.70
	MeOM	22.36	2.23
	FOM	26.83	4.47
	LOM	17.88	0.0
Gaussian	Centroid	4.0	4.26
	MeOM	0.0	11.18
	FOM	0.0	22.36
	LOM	0.0	0.0
Generalized Bell	Centroid	3.15	0.0
	MeOM	0.0	0.0
	FOM	0.0	0.0
	LOM	0.0	0.0

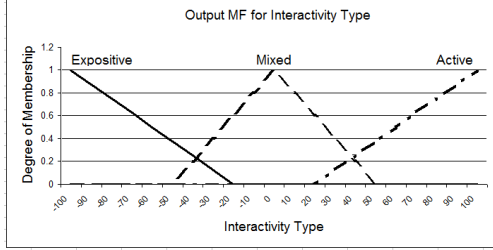
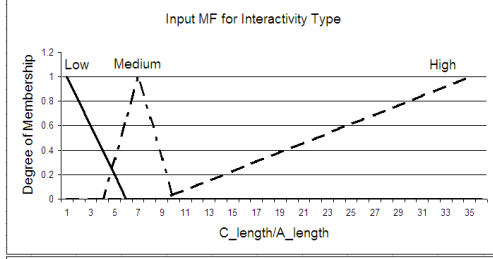


**(a) Input and output MFs for difficulty**





(b) Input and output MFs for interactivity level



(c) Input and output MFs for interactivity type

Figure 2. Input and output MFs for the proposed fuzzy inference system

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 2 (a)), which has the minimum RMSE. For interactivity level; we utilize trapezoid function as an input fuzzy set with *low*, *medium*, *high* and *very high* linguistic values and trapezoid MF as an output of the fuzzy model (Figure 2 (b)) that has the minimum RMSE. For interactivity type, various models gave RMSE of 0.0, thus we did not investigate the impact of five fuzzy sets since the complexity is increasing with increasing number of fuzzy sets. We choose 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 2 (c)). These FISs are employed for metadata generation.

**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. Assume, for difficulty metadata value generation, input is represented by four fuzzy sets with trapezoid MF as shown in Figure 2(a), and *C\_length*=4 and *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 *very high* linguistic value.

**Step 5:** We determined fuzzy rules using *C\_length* and *A\_length* attribute values of *Concept* and *Activity* FIG. In Tables 4, 5 and 6, the fuzzy if-then-rules of FISs for the difficulty, interactivity level and interactivity type 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 2 (a) (b) and (c).

Table 4. Fuzzy if-then-rules for LOM difficulty

Concept <i>C_length</i>	Activity ( <i>A_length</i> )			
	<i>Null</i>	<i>Low</i>	<i>Medium</i>	<i>High</i>
<i>Low</i>	V. Easy	Easy	Medium	V. Diff.
<i>Medium</i>	Easy	Medium	Difficult	V. Diff.
<i>High</i>	Easy	Medium	Difficult	V. Diff.

Table 5. Fuzzy if-then-rules for LOM interactivity level

Activity ( <i>A_length</i> )				
<i>Null</i>	<i>Low</i>	<i>Medium</i>	<i>High</i>	<i>Very high</i>
V. Low	Low	Medium	High	V. High

Table 6. Fuzzy if-then-rules for LOM interactivity type

Concept <i>C_length</i>	Activity ( <i>A_length</i> )			
	<i>Null</i>	<i>Low</i>	<i>Medium</i>	<i>High</i>
<i>Low</i>	Expositive	Active	Active	Active
<i>Medium</i>	Expositive	Mixed	Active	Active
<i>High</i>	Expositive	Mixed	Mixed	Active

**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. The process can be illustrated as follows. 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 and defuzzified to produce a single numeric output. In our experiments, centroid defuzzification method gave the minimum RMSE values as shown in Tables 1, 2 and 3. Therefore it is utilized in our FISs (Equation 6).

$$value = \frac{\sum_{i=1}^n \mu(rule_i) * output(i)}{\sum_{i=1}^n \mu(rule_i)} \quad (6)$$

where *value* is the numeric output of the FIS, *n* is the number of rules,  $\mu(rule_i)$  is the output weight of *rule<sub>i</sub>* and *output(i)* is the output value of *rule<sub>i</sub>*. The value of *output(i)* is determined by output variable value, where it is equal to the Middle of Maximum on the output fuzzy set of the linguistic value except very low/very high values of interactivity level and very easy/very difficult values of difficulty. For very low and very easy, *output(i)* equals to First of Max. For very high and very difficult, *output(i)* equals to Last of Max. For the above example, first rule has output of *medium*, *output*=0, since 0 is the Middle of Max on the output fuzzy set of *medium* as shown in Figure 2(a). Based on the example above, the outputs of rules can be defuzzified using the centroid method as follows:

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

The numeric output should be converted to a metadata value. This is performed based on the output fuzzy sets (Figure 2). 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” since it falls into that interval.

$$\begin{aligned}
& \text{if } -100 \leq \text{value} \leq -75, \text{ then difficulty} = V. \text{ Easy} / \text{Inter.level} = V. \text{ Low} \\
& \text{if } -75 < \text{value} < -25, \text{ then difficulty} = \text{Easy} / \text{Inter.level} = \text{Low} \\
& \text{if } -25 \leq \text{value} \leq 25, \text{ then difficulty} = \text{Medium Diff.} / \text{Inter.level} = \text{Medium} \\
& \text{if } 25 < \text{value} < 75, \text{ then difficulty} = \text{Difficult} / \text{Inter.level} = \text{High} \\
& \text{if } 75 \leq \text{value} \leq 100, \text{ then difficulty} = \text{Very Difficult} / \text{Inter.level} = V. \text{ High} \\
& \text{if } -100 \leq \text{value} \leq -50, \text{ then Interactivity Type} = \text{Expositive} \\
& \text{if } -50 < \text{value} < 50, \text{ then Interactivity Type} = \text{Mixed} \\
& \text{if } 50 \leq \text{value} \leq 100, \text{ then Interactivity Type} = \text{Active}
\end{aligned} \quad (7)$$

**Step 8:** Based on the metadata value and the interval that *value* falls into, our algorithm calculates a metadata confidence score. Let  $x_1 \leq x_p \leq x_2$ , 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 based on equations (7) and (8). If  $(x_1 \neq x_p \text{ and } x_2 \neq x_p)$ ,  $x_p$  is the Mean 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:

$$\text{if } (x_1 \neq x_p \text{ and } x_2 \neq x_p) \left\{ \begin{array}{l} \text{if } (\text{value} = x_p), \text{ then confidence} = 1 \\ \text{if } (\text{value} < x_p), \text{ then confidence} = 1 - \left( \frac{1}{x_p - x_1} \times \text{value} \right) \\ \text{if } (\text{value} > x_p), \text{ then confidence} = 1 - \left( \frac{1}{x_p - x_2} \times \text{value} \right) \end{array} \right\} \quad (9)$$

$$\text{if } (x_1 = x_p \text{ or } x_2 = x_p) \left\{ \text{confidence} = \left| \frac{\text{value}}{x_p} \right| \right\}$$

For output of “medium difficulty”,  $x_p = 0$ ,  $x_1 = -25$  and  $x_2 = 25$ . When  $\text{value} = 16.66$ , metadata confidence score for medium difficulty is  $\lfloor (1/(0-25)) \times 16.66 \rfloor - 1 = 0.33$  based on (9).

Finally, the extracted metadata is converted to RDF turtle format using IEEE LOM vocabulary and stored to a file.

### 4.3 Implementation

The FIS is implemented as a Web application using Javascript. Javascript has been chosen because of its XML DOM support for parsing and analyzing DocBook documents 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 the 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, number of fuzzy sets and defuzzification methods

have to be identified. However, the advantage of FL is that it allows expert knowledge and human reasoning to be intelligently interpreted by the inference system.

## 5. EVALUATIONS

The automatically extracted metadata has been evaluated in terms of *metadata quality* (Precision, Recall, F-measure), *Fuzzy Inference against Rule-based Reasoner*, *Prediction Error* and *User Perceived Quality*. This section explains the evaluation procedures and discusses the results.

### 5.1 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 PhD students from computer science) were asked to annotate randomly selected 100 documents from English version of Symantec Norton 360. Subjects had different levels of expertise in Symantec products (80% intermediate and 20% beginner). In the study, we asked metadata values of document difficulty, interactivity level and interactivity type. The subjects used an online Web annotation client for manual annotations. First, a briefing was given to explain the annotation task and meanings of metadata fields with a sample annotation task. Then, each user was asked to separately complete the manual annotations by using the online annotation client. Then, automatically extracted metadata was compared with the metadata produced by participants in terms of precision, recall and f-measure. Precision ( $P = \text{Correct}/A$ ) is the number of metadata fields correctly annotated (*Correct*) over the number of metadata fields automatically annotated (*A*) by the framework, recall ( $R = \text{Correct}/M$ ) is the number of metadata fields correctly annotated over the number of metadata fields manually annotated (*M*) and f-measure ( $F = 2 * ((P * R) / (P + R))$ ) 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 7, 8 and 9.

**Table 7. 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%
Mean	91.00%	88.76%	89.86%

**Table 8. 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%
Mean	79.46%	77.52%	78.47%

**Table 9. 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%
Mean	66.29%	64.52%	65.39%

**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. 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 scores), we took the median of annotations as a resultant value. Then, we compare the automatically annotated metadata against the agreed annotations in terms of precision, recall and f-measure as shown in Table 10. 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 10. Precision, recall and f-measure against the agreed annotations**

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

**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, 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 information granulation and inference approach created reasonably accurate metadata values compared to the agreed annotations of individual users with an overall average precision of 89.39%. Furthermore, our framework can be used to complement manual annotations. For example, the automatic metadata extraction framework can be used to generate metadata first, and then a metadata expert can validate the metadata values. In this way, the speed of annotation process can be improved significantly and metadata values will be more consistent throughout comparing to manual annotation.

## 5.2 Fuzzy Inference vs Rule-Based Reasoner

Our claim is that, cognitive metadata values are often fuzzy and fuzzy inference with approximate reasoning can provide better results. Therefore, the performance of the proposed fuzzy model can be compared with a rule-based reasoner, such as rules with sharp boundaries to assess the added value of fuzzy sets and fuzzy inference. For this purpose, we implemented a rule based reasoner, where it uses exactly the same rules as the fuzzy inference as discussed in Tables 4, 5 and 6, but the only difference is that the membership degrees of *A\_length* and *C\_length* are represented by crisp sets. For example membership degrees of *medium* for the linguistic variable difficulty is  $\{ \text{if } (4 < x < 12), \mu_{\text{medium}}(x) = 1, \text{ otherwise } \mu_{\text{medium}}(x) = 0 \}$ . We compared the generated metadata by the rule-based reasoner and the proposed FISs against the agreed annotations in terms of precision, recall and f-measure as shown in Table 11.

**Table 11. Fuzzy inference and rule-based inference comparison in terms of precision, recall and f-measure against the agreed annotations**

	Fuzzy Inf. Prec.	Rule-base Prec.	Fuzzy Inf. Rec.	Rule-base Rec.	Fuzzy Inf. F-Meas.	Rule-base F-Me.
I. Type	96.90%	96.90%	94%	94%	95.42%	95.42%
I. Level	88.65%	86.59%	86%	84%	87.30%	85.27%
Diff.	82.47%	71.13%	80%	69%	81.21%	70.04%
Avg	89.39%	84.87%	86.66%	82.33%	87.97%	83.57%

**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, 2) Interactivity level only depends on the length of the *Activity* granule and 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. Therefore, for difficulty, the fuzzy inference system performed improved performance than the rule-based reasoner with a precision of 82.47% and enhanced the performance of the rule-based reasoner ~11%. This shows that fuzzy inference suits better than the rule-based reasoner when input values do not have sharp boundaries such as metadata values of document difficulty.

## 5.3 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 equation (10):

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

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 presented in Figures 3, 4 and 5 respectively.

**Analysis:** As can be seen from Figure 3, 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. On agreed annotations, SSE of both fuzzy inference and rule-based reasoner is same, which is 6. Figure 4 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, user #4 and user #5 provided the closest annotations for interactivity level. For interactivity level, fuzzy inference slightly improved the rule-based reasoner such as on agreed annotations, SSE of fuzzy inference is 19 and SSE of rule-based is 24 for 100 documents. As illustrated in Figure 5, 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 is mainly because of different perceptions of users to very same documents. For difficulty metadata, fuzzy inference significantly improved the performance against the rule-based reasoner. On agreed annotations, SSE of fuzzy inference is 30 and SSE of rule-based is 45 for 100 documents.

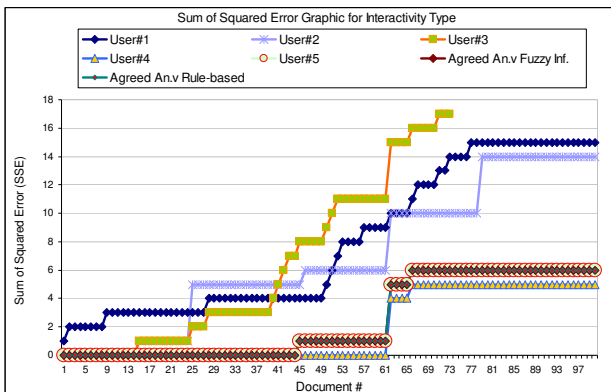


Figure 3. SSE graphic for LOM interactivity type

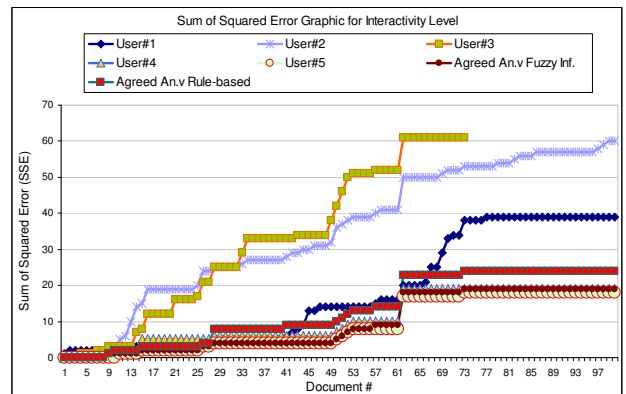


Figure 4. SSE graphic for LOM interactivity level

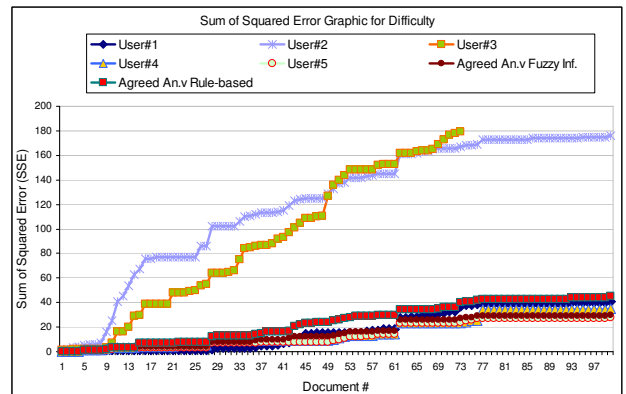


Figure 5. SSE graphic for LOM difficulty

## 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 cognitive 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. Average quality scores of 100 documents are calculated for each subject and presented in Figure 6.

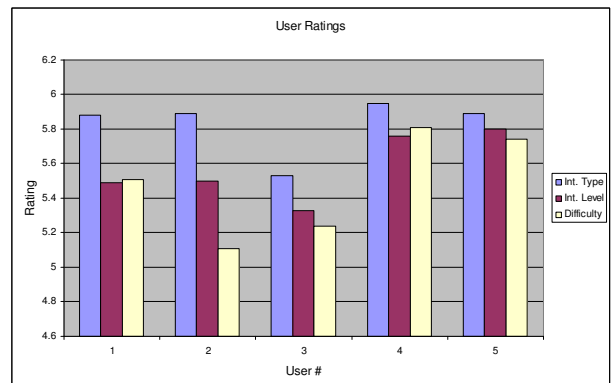


Figure 6. Average of subjects' ratings to Interactivity Type, Interactivity Level and Difficulty

**Analysis:** The results showed that metadata quality of the interactivity type found to be very high by all subjects. Metadata

quality of interactivity level and difficulty also rated 5.1 or over out of possible 6. The mean of user ratings were also calculated. 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 high user perceived quality scores out of the examined 100 documents.

## 6. CONCLUSIONS AND FUTURE WORK

We have presented a novel automatic metadata extraction framework based on fuzzy information granulation and fuzzy inference system for cognitive metadata mining in order to support personalization. The proposed method uses document semantics and approximate reasoning to predict difficulty, interactivity type and interactivity level of documents. The user evaluation study showed that our method achieves promising precision rates ranging from 82.47% to 96.90% depending on the metadata field. As well as, evaluation indicates that the user perceived metadata quality is high (5.57/6.00) for the examined 100 documents. The proposed fuzzy inference is also compared with a crisp rule-based reasoner; for the difficulty metadata mining, fuzzy inference achieved improved results and enhanced the performance ~11%. In future, our framework can be employed to complement manual annotations. For example, the automatic metadata extraction framework can be used to generate metadata first, and then metadata values that have a confidence score less than a threshold can be given to metadata experts for validation, which makes annotation process faster and makes metadata values more consistent in comparison to manual annotations.

## 7. ACKNOWLEDGMENTS

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. In addition, we would like to thank Stephen Curran for implementing the Web annotation client for evaluations.

## 8. REFERENCES

- [1] Jamison, N. 2010. Beyond the Customer Satisfaction Horizon Fostering Loyalty through Customer Service. <http://www.jamison-consulting.com/pdf/BeyondtheCustomerSatisfactionHorizon011210.pdf>
- [2] Khankasikam, K. (2010). A Hybrid Case-based and Rule-based for Metadata Extraction on Heterogeneous Thai Documents. In Proceedings of *IEEE International Conference on Computer and Automation Engineering*.
- [3] Flynn, P., Zhou, L., Maly, K., Zeil, S. and Zubair, M. (2007). Automated Template-Based Metadata Extraction Architecture. *ICADL. LNCS*, Vol. 4822, 327-336.
- [4] 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 Conference on Digital libraries*, 37-48.
- [5] Zhang, J., Niu, Y., Nie, H. (2009). Web Document Classification Based on Fuzzy k-NN Algorithm. In Proceedings of *International Conference on Computational Intelligence and Security*, 193-196
- [6] Sah, M. and Wade, V. 2010. Automatic Metadata Extraction from Multilingual Enterprise Content. In *Proceedings of International Conference on Information and Knowledge Management (CIKM)*, 1665-1668.
- [7] 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.
- [8] 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.
- [9] Yilmazel, O., Finneran, C. M., and Liddy, E. D. (2004). MetaExtract: An NLP System to Automatically Assign Metadata. In *Proceedings of ACM/IEEE Conference on Digital Libraries*, 241-242.
- [10] Flesch, R. (1948). A new readability yardstick, *Journal of Applied Psychology*, Vol. 32, 221-233.
- [11] 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.
- [12] Ceravolo, P., Nocerino, M.C., and Viviani, M. (2004). Knowledge Extraction from Semi-Structured Data Based on Fuzzy Techniques. In *International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, LNAI, Vol. 3215, 328-334.
- [13] 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.
- [14] 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 Proceedings of *International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, LNAI, Vol. 3681, 112-118.
- [15] Walsh, N. and Muellner, L. (1999). *The DocBook Definitive Guide*, O'Reilly Media.
- [16] Gueye, B., Rigaux, P., and Spyrtos, N. (2004). Taxonomy-Based Annotation of XML Documents: Application to eLearning Resources. In Proceedings of SETN, LNAI, Vol. 3025, 33-42.
- [17] 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.
- [18] Steichen, B., O'Connor, A., and Wade, V. (2011). Personalisation in the Wild – Providing Personalisation across Semantic, Social and Open-Web Resources. *ACM Conference on Hypertext and Hypermedia*.
- [19] IEEE Learning Object Model. (2002). [http://ltsc.ieee.org/wg12/files/LOM\\_1484\\_12\\_1\\_v1\\_Final\\_Draft.pdf](http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf)
- [20] Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, Vol. 8, 338-353.
- [21] 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.