Metadata Extraction with Cue Model

Wan Malini Wan Isa[2], Jamaliah Abdul Hamid[1], Hamidah Ibrahim[2], Rusli Abdullah[2], Mohd. Hasan Selamat[2], Muhamad Taufik Abdullah[2] and Nurul Amelina Nasharuddin[2]

[1]Faculty of Educational Studies
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor
Tel: 03-89468177, Fax: 03-89468246
E-mail: aliah@putra.edu.my

[2]Faculty of Computer Science and Information Technology
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor
Tel: 03-89466555, Fax: 03-89466576
E-mail: wanmalini84@gmail.com, hamidah@fsktm.upm.edu.my, rusli@fsktm.upm.edu.my, hasan@fsktm.upm.edu.my, taufik@fsktm.upm.edu.my, nurulamelina@gmail.com

ABSTRACT

Metadata extraction from texts is important since it enables search for documents based on the metadata identified. It is impossible to retrieve journal articles without the metadata of documents, which includes information such as author, title, and journal publication details. This paper proposes a new technique to extract metadata of documents, called Metadata Extraction with Cue Model, which uses combinations of a few features to extract metadata automatically from documents. Automatic metadata extraction is increasingly important in today's influx of published material.

Keywords
Metadata extraction, Cue Model, machine learning, rule based, template matching

1.0 INTRODUCTION

Metadata is information used to describe a set of data. Example of metadata includes title, author, address, source and email. The extraction of metadata is important in electronic document management since it will enable the document to be stored and retrieved in a systematic manner. Hence metadata extraction is used in many document processing fields such as browsing, search, and filtering.

Most metadata extraction systems are still human intensive since they require expert decision to recognize relevant metadata but this is time consuming. The dependency on human expertise can some times result in the creation of inflexible databases based on pre-determined templates for metadata extraction, thus creating a limitation for system.

Several techniques have been used for automatic metadata extraction from documents. The main techniques fall into two categories; Machine Leaming based approach and Rule-Based approach.

Machine Learning (Hu, Li, Cao, Meyerzon, Zheng, 2005) involves having machines (or the system) to learn the information within a document in order for the system to propose rules for metadata extraction. The process however takes a very long time since the machine needs to learn a huge amount of documents before it can propose any rules. The flaw have been justified by another technique which is Rule based Extraction.

Rule based approach (Han, Manavoglu, Zha, Tsioutsiouliklis, Giles, Zhang, 2005) do not require any machine training processes due to the rules which have been pre-set earlier by the experts. This will allow the technique to be implemented straight away without the need to consider the exact nature of the information within the documents. However, the rules are limited and since the technique couldn't produce more rules dynamically on their own, the technique can only be applied to certain and specific types of documents.

Therefore an alternative metadata extraction model is needed to automatically extract metadata from text regardless of its genre and without using any pre-set template. Even though rules are still required, the rules are only to enable recognition of important cue metadata such as Part of Speech (POS) variation, cue words, line position, relative position and symbol.

This proposed new technique described in this paper extracts metadata automatically from document based on a combination of a few programming features that recognize parts of speech, cue words, line position, relative position and symbols to identify the metadata in a particular journal article. This technique is simpler and less time consuming than the existing conventional techniques. This new technique is called Metadata Extraction with Cue Model. The Cue Model will extract metadata directly from text solely based on selected text features.

2.0 RELATED WORK

Currently there are a few metadata extraction techniques available. The techniques of metadata extraction technique can be classified to two main categories which is Machine Learning and Rule based approached.

In 2003, machine learning method for automatic metadata extraction using Support Vector Machine has been proposed (Han, Giles, Manavoglu, Zha, Zhang and Fox, 2003). This Support Vector Machine classification-based method is to extract the metadata from the header of a research papers. The method first classifies each lines of the header into one or more of 15 classes. An iterative convergence procedure is then used to improve the line classification by using predicted class labels of its neighbor lines in the previous round. Further metadata extraction is done by seeking the best chunk boundaries of each line. The researchers found that discovery and use of the structural patterns of the data and domain based words clustering can improve the metadata extraction performance. The researchers then proposed an algorithm for metadata extraction based on two techniques: the Support Vector Machine Classification and Feature Extraction. Support Vector Machine attempts to find an optional separating hyper-plane to maximally separate two classes of training samples while Feature Extraction uses both word and line specific features to represent their data. The reseachers then designed a rule based, context-dependent word clustering method for specific feature generation, using the rules extracted from various domain database and text orthographic properties of words.

From the test done, they concluded that Support Vector Machine is known for good generalization performance and its ability in handling high dimensional data. They also discovered a few limitations for the method. Since the method generates labeled training data for the system, it requires a high amount of language computational resources in order to operate efficiently.

Two years later another new method was proposed to extract metadata, which is a rule based word clustering method (Han et al, 2005). The method suggest the clustering of similar words with the same syntactic or semantic categories through the use of cluster labels, which then act as features for text classification. Word clustering generalizes specific features by considering common characteristics and ignoring the specific characteristics among individual features of the words. By using cluster as features, words share more features representative of a target class, which can therefore be directly extracted as metadata.

Word clustering is based on domain database and word orthographic properties (Schone and Jurafsky, 2001), which contain a priori knowledge of a specific class. Domain corresponds to the class text classification tasks. A domain database can be a name word database for each metadata class. Specific words and characters are clustered based on their membership in the databases. For example, words "Mary", "Johnson" and "Tom", which appear in the name word database, are clustered and represented as ":name word:", the cluster label. This feature is called the *cluster feature representation*.

Word orthographic properties consider cases of the words, and digits or special characters the words contain. A word is a consecutive sequence of characters "@" character is an orthographic property of the email address, and is used to cluster the specific email addresses as ":email:".

To apply this method Han et al. first generate a domain database which defines two types of domain database that are External Domain Database and Constructed Domain Database. Then they designed the cluster based on the domain database and the words' orthographic properties. Finally they come up with the rules to match words from different domain database by checking the word orthographic properties and assigning them to an appropriate cluster. The rules consider multiple properties of the word to determine its cluster.

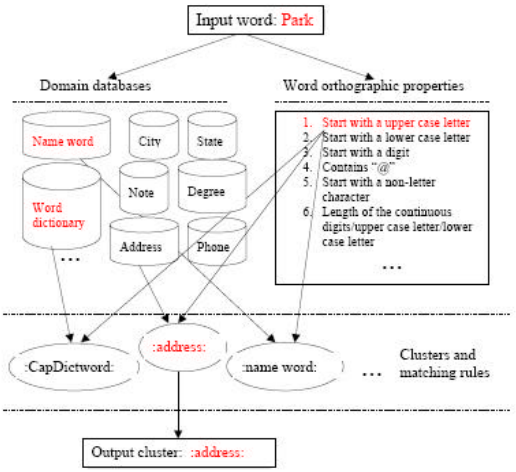

Figure 1: An example of cluster assignment for the word "Park" using rule-based word clustering method.

This method appears to have computational advantages since they only need to search the domain database and check word orthographic properties using simple rules. The method can be considered as an effective approach to reduce feature dimensionally while simultaneously also adhering to features sparseness, and thus resulting in improved text classification performance. But still since they're using a domain database, the size of the word depository will affect the range of documents that this method can be applied to.

Another new method proposed is a novel template matching for header metadata extraction from semi-structured documents stored in PDF (Huang, Jin, Yuan and Han, 2006). Their approach is to define a template and use it to guide Finite State Automation (FSA) to extract header metadata of an article. The metadata that they extracted are

limited to only title, author(s), affiliation(s), abstract, and keywords.

Huang et al proposed a layout information based approach to document desegregation. The method involves exploiting the layout of information while reading a document. The information includes the section properties, such as number of columns, section break type, and position of page number. Also the paragraph properties, such as flush left, flush right, flush centre, left indent, right indent, and first indent in the first line.

The layout for each metadata has a different characteristic. Title for example will always be on the upper portion of the first page using the largest bold font size, and positioned in the middle of the line and flush center using "enter" as the section break symbol. While the Author(s) is always located immediately under the title using the same font but smaller size compare to the title. Break symbols can be either a space, a comma or an "and". Author(s) and affiliation(s) may or may not always have superscripts.

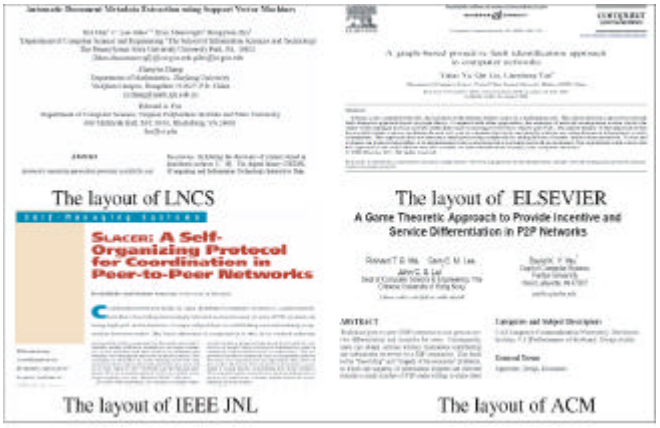

Figure 2: Examples of typical layout

Once the layout of document that has been defined, they formulated a template to match the layout information by using XML Schema. A variety of template for each metadata had to be built to cater to the differences in the layout information for different types of articles. At least 5 templates were built for each type of article.

```
<xsd:element name="template schema">
    <xsd:complexType>
    <xsd:sequence>
        <xsd:element name="title template"/>
        <xsd:element name="authors template"/>
        <xsd:element name="affiliations template"/>
        <xsd:element name="abstract template"/>
        <xsd:element name="keywords template"/>
    </xsd:sequence>
    </xsd:complexType >
</xsd:element>
```

Figure 3: The structure of template schema using XML

Template matching model is used to match the template with the text document to extract the metadata. The researchers also developed a template generation system to transform known semi-structured strings into sequences of data stream with layout information. These sequences are then manually saved as templates in template collection folder. Semi-structured documents are first translated into a sequence of data stream, which are then matched to the most similar template in the template collection folder. The metadata of the documents are parsed according to the template.

The proposed method by Huang et al can effectively extract metadata without any training cost. It can be incorporated in the architecture of active digital libraries and search engines base on metadata. However this method can be considered as inefficient and hard to be implemented due to the limited number of templates which can be created and then stored. When a new layout of journal is published, the developers need to design a new template to accommodate the new document structure. Although the system is capable of generating some temporary files to be user-included as new metadata throughout the metadata extraction process, this will however, require more storage space and resources which can too costly.

3.0 THE EXTRACTION WITH CUE MODEL

This new Metadata Extraction with Cue Model technique proposed in this paper can automatically extract metadata in any document and display it on an interface. Prior to the extraction, the input document is first converted into text format from the original format. The technique then performs the metadata extraction from the text. Then the extracted metadata such as title, author, address, source and email will be displayed on the screen. The process will not be using any relational database.

The Metadata Extraction with Cue Model technique will first extract the first 20 numbers of lines from the input document. We choose the first 20 lines because in most journal publications the metadata is found always at the top of the article. From the extracted lines, we analyze the features of each line to determine the set of features that they have. The features that will be analysed consist of part of speech variations, cue words, line position, relative position and symbol. After the process of analyzing the features of each line, the features or cues are then used as a basis to classify a line into certain line types, based on certain rules. The line types are title line, source line, author line, address line and email line. The extracted line will then be tested with the pre-defined rules to determine which metadata attribute it fulfilled. If it passes the test, the metadata will be extracted and be displayed on the output interface. The flow of the extraction is show in figure 4.

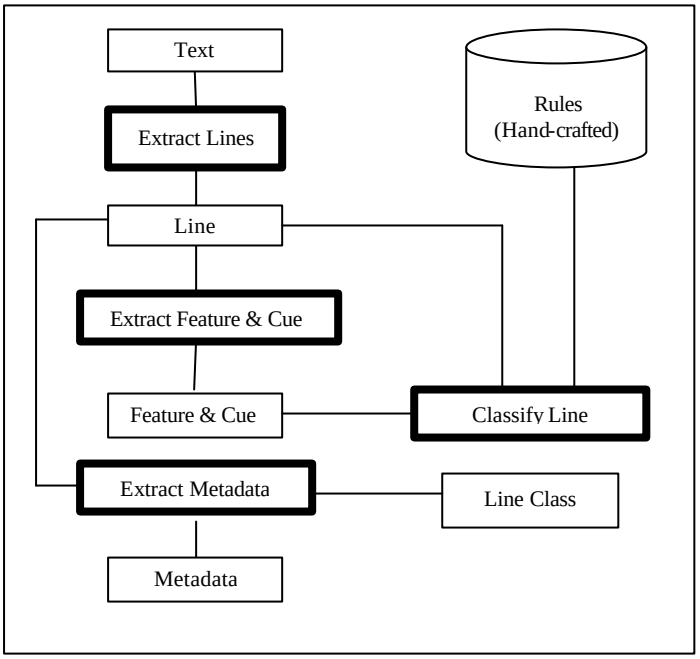

Figure 4: the flow of the extraction system

3.1 Set of Features

There are several features in Cue Model which are used to extract metadata from documents. The features are:

1. Part of Speech variation
   Each line contains terms with their own parts of speech. Parts of speech variation computes the total number of parts of speech in a line. Part of speech variations is useful in identifying the line that contains the authors' name. A line with low part of speech variation, usually no more than two, signals the author-line.

2. Cue words
   Cue words are certain word in the text that occurs consistently to signify a particular pattern. For example, the word "Journal" appears in multiple documents. In this case, "Journal" is the cue word of interest. It can be used to ascertain the source line. Certain source line does not contain the expected cue words. However, it often includes the year of publication. For instance "ACL '06" or "Language (4) 2004". This can also be used to signal the source line.

3. Line position
   The position of the line in text can also give the clue to the type of metadata involved. Line position is computed from the top to bottom. The value given is inversely proportionate with its position. Title is usually situated at the top portion of text, which means that it has high value of line position.

4. Relative position
   Contrary to line position which is computed directly from text, relative position is determined by comparing the position of a line type as compared to another. From the empirical observation, it is discovered that the position of a particular type of line could be relative to one another. For instance, the title line always comes before the author line, or written "title line< author line".

5. Symbols
   Symbol like "@" is used to recognized the email line.

A combination of features signifies a particular cue. Each cue distinguishes one type of line from another. Below are examples of some cue recognition.

Cue Name	Cue marker	Condition
Source	Cue words	Example: Journal, conference, proceeding.
	Line position	Always first position, if not then Title
	Check by POS	If contain Cardinal number example year, volume, page number
		If contain © or copyright
Title article	Check by POS	Accept combination of POS
		Never begin with selected preposition example by, up on.
	Line positioning	Always first unless proceeded by journal title.
Author	Proper noun	Must appear after title line
	Line positioning	Ignore separating comma
		If author is numbered, link to address.

Table 1: Example of some cue recognition

4.0 CONCLUSIONS AND FUTURE WORK

This Cue Model was developed to extract metadata from documents. Metadata extraction nowadays is important to aid in quick search and bibliography compilation. Users who rely on good metadata extraction might include students, lecturers and researchers. This is due to the fact that metadata is important and needed when it comes to writing references or searching for articles from depositories. Librarians can also use this technique to sort out existing documents in specific categories.

This new technique is intended to improve upon some existing techniques such as use machine learning technique,

relational databases and template matching. This Cue Model is simpler and less time consuming than conventional technique. At the present time, this technique will only be able to extract metadata from documents in a science domain and mainly from publications such as proceeding and journal.

For the next development, this system can be extended to other domains of publications such as agriculture, social science and economics. Besides that, the accuracy of the extraction of metadata is hoped to be incrementally improved through the development of new sets of features. This Cue Model technique is hoped to be useful in information retrieval and computer science field.

5.0 REFERENCES

Giuffrida, G., Shek, E. C., and Yang, J.(2000) Knowledge-based metadata extraction from Post-Script files. In *Proceedings of the 5th ACM Conference on Digital Libraries* 77-84.

Han, H., Manavoglu E., Zha H., Tsioutsiouliklis K., Giles C.L., Zhang X. (2005) Rule-based Word Clustering for Document Metadata Extraction. In *ACM Symposium on Applied Computing*

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., and Fox, E. A. (2003) Automatic Document Metadata Extraction using Support Vector Machines. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, 37-48*.

Hu, Y., Li H., Cao Y., Meyerzon D., and Zheng Q. (2005) Automatic extraction of titles from general documents using machine learning. In *Proceeding of the 5th ACM/IEEE-CS Joint Conference On Digital Libraries*

Huang, Z., Jin, H.,Yuan, P., Han, Z. (2006) Header Metadata Extraction From Semi-structured Documents using Template Matching. In *Proceedings of OTM 2006 Workshops, LNCS Issue 4278, 1776-1785*.

Ping Y., Ming Z., ZhiHong D. and DongQing Y. (2004) Metadata Extraction from Bibliographies Using Bigram HMM. *Lectures notes in Computer Science. Volume 3334/2004, pages 310-319*. Springer-Verlag Berlin Heidelberg

Yahya N.A and Buang R. (2006) Automated Metadata Extraction from Web Sources. In *Proceeding of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*.

Yunzhong L., Yaping L., Zhiping C .(2003) Using Hidden Markov Model for Information Extraction Based on Multiple Templates. In *Proceeding 2003 International Conference on Natural Language Processing and Knowledge Engineering.*