Automatic Bibliographic Metadata Extraction Approach

Cho Cho Khaing, May Aye Khine
University of Computer Studies, Yangon
chochokhaing.ygn@gmail.com,maya.khine@gmail.com

Abstract

Recently, many digital libraries have been constructed and published. Scientific papers often conclude with a section that lists referenced works in the form of a reference list or bibliography. This form of acknowledgment is crucial in helping readers and reviewers to relate the current work to its context within the research community's discourse. Such bibliographical references that appear in journal articles can provide valuable hints for subsequent information extraction. Therefore, automatic extraction of metadata and bibliographies is widely studies in recent years. Decision tree is now widely used machine learning approach in many areas. This paper applies the Decision tree for bibliographic metadata extraction and the experiment show that decision tree classifier achieves high accuracy result.

1. Introduction

Authors and publishers are beginning to make scientific publications available on the World Wide Web in increasing number. In order to search and exploit these disorganized digital documents, there is a rowing need to organize them efficiently. Organizing articles by their metadata is a good way and becomes more and more popular. Accordingly, automatic extraction of metadata from vast number of papers and papers' bibliographies has been widely studied in recent years. So many researchers are interested in improving the metadata extraction from papers ' bibliographies ,which is, segmenting a bibliography into individual fields such as author, title, publisher, date and so on. The field extraction from bibliographies is non-trivial because of the high variance in the structure of the current record-level search and the order of attributes is not fixed and not all attributes are present in all instances.

There have been much previous work in bibliographic metadata extraction; template matching, knowledge-based approach and machine-learning. In general machine learning methods are robust and adaptable and, theoretically, can be used on any document set. Generating the labeled training data is the rather expensive price that has to be paid for learning systems. Although regular expressions and rule-based systems do not require any training and are straightforward to implement, their dependence on the application domain and the need for an expert to set the rules or regular expressions causes these methods to have limited use.

The rest of the paper is organized as follows: section 2 presents the metadata extraction in bibliographic references; the background theory of decision tree is described in section3, section 4 describes the detail steps of the system. Section 5describes the experimental results and conclusion is presented in the following section.

2. Related Work

For a long time, librarians redacted bibliographical records or indexes to describe the available documents. Previous work on the topic of bibliographic meta-data extraction from research paper references can be subdivided into machine learning (ML) or rule-based [1, 2, 8, 3] approaches. Giuffrida et al. [9], for instance, developed a rule-based system for automatically extracting metadata from research papers in Postscript. They used rules like "titles are usually located on the upper portions of the first pages and they are usually in the largest font sizes". Liddy et al. [6] and Yilmazel el al. [12] performed metadata extraction from educational materials using rule-based natural language processing technologies. Mao et al. [7] also conducted automatic metadata extraction from research papers using rules on formatting information. The rule-based approach can achieve high performance. However, it also has disadvantages. It is less adaptive and robust when compared with the machine learning approach. Approaches based on ML try to derive the relationship between input and output strings according to a given set of samples and label future inputs using that knowledge. For the latter case a set

of adequate rules has to be derived manually by a domain expert via analyzing appropriate samples. Major benefits of systems based on ML are their high degree of adaptability and robustness with the drawback of required training whereas rule-based systems usually behave more rigidly and do not adapt very well. The wide assortment of applied machine learning techniques spans conditional random fields [6], hidden Markov models [7], support vector machines [5].

Han et al. [5], for instance, conducted metadata extraction with the machine learning approach. They viewed the problem as that of classifying the lines in a document into the categories of metadata and proposed using Support Vector Machines as the classifier. They mainly used linguistic information as features. They reported high extraction accuracy from research papers in terms of precision and recall.

This paper examined to use Decision trees classifier because Decision tree classifiers are used successfully in many diverse areas such as radar signal classification, character recognition, remote sensing, medical diagnosis, expert system, and speech recognition. The most important feature of decision tree is its capability to break down a complex decision making process into a collection of simpler decisions, thus providing a solution which is often easier to interpret.

3. Background Theory of Decision Tree

A tree approach is used for classification which is a model of data mining [11, 9]. To reach the classes through the shortest path the most suitable attribute is chosen to be the root and recursively the algorithm make calculations to determine the most suitable attribute in dataset to be the next node. Here 'the most suitable' adjective is to divide the rest of the database roughly into two or more equal parts. The attribute which is the closest to this is chosen as the most suitable attribute.

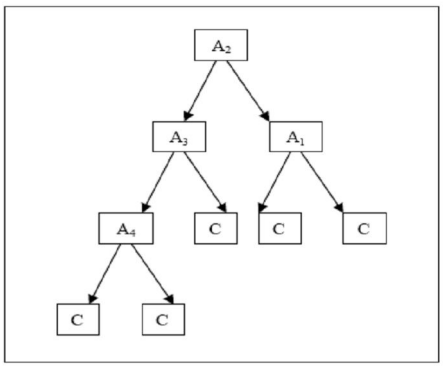

Figure1. A Simple Decision Tree

After choosing the root node, the algorithm adds arcs to the root for each predicate. Adding node to new branches is carried on as it is done for the first (root) node. During the processes if the pre-assigned criterion is reached, the arcing stops and the end of the final branch is labeled as one of the classes. Here, an algorithm differs from one another with the splitting criteria it exercises. In the literature, entropy and gini index are mostly employed by different algorithms as spitting criteria [10].

4. Bibliographic Metadata Extraction with Decision Tree Classifier

This system consists of three main steps: tokenization, feature extraction and classification.

1. Tokenization

Instead of the common practice of tokenizing a string into individual words, we use punctuation (except for hyphens and apostrophes) as token delimiters as shown in figure 2. This tokenization scheme often leads to phrases. There are a few advantages to this style of tokenization:
1) considering multiple words as a token allows more complex features to be used, thus giving a better chance of making a correct classification; and
2) reducing the number of tokens per reference string reduces the computational cost of this task.

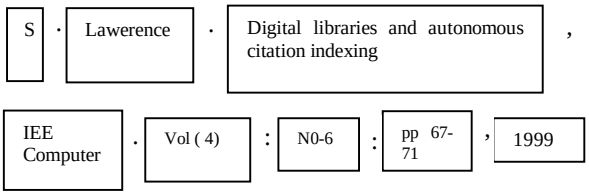

Figure 2.A tokenized reference string. Each string contains one token.

2. Feature extraction

During this step a simple set or heuristics rules were applied. Those rules are of the form 'If token consists of only four digit numbers it is a year'. For each of the tokens the following features are generated as shown in table 2.

Feature Name	Description
position: First, Mid, Last.	Indicates what the position of the token is within the citation.
IsAllUpper: y,n.	Indicates all characters are upper case alphabetical.
IsAllLower: y,n.	Indicates all characters are lower case alphabetical.
IsAllAlpha: y,n.	Indicates all characters are alphabetical.
IsCapitalized: y,n.	Indicates the word starts with an uppercase alphabetical character and continues with the remainder in lower case alphabetical characters.
IsAllNum: y,n.	Indicates all characters are digits.
FieldLength:	The number of characters the token has.
HasMonth: y,n.	Indicates whether the token contains any month words(eg.'January','Jan')
IsOneCap: y,n.	Indicates whether the token consists of only one capital letter eg. 'C'.
HasAbbrevation: y,n.	Indicates whether the token contains any words with more than one capital letter.eg. 'JCDL'
startPanctuation:	The punctuation that is preceded the current token. eg. Period, comma, hyphen, double quotes, opening brace, colon, others and none.
endPanctuation:	The punctuation that is immediately after the current token.
hasNumber: y,n.	Indicates whether the token contains any number.
IsName: y,n.	Indicates whether the token is name format. eg. Single capital letter follows by words that Start with capital letter.

Table1. List of Features.

3. Classification

This final stage is the heart of the system. Decision tree classifier is exploited because of its classification is computationally in expensive, consisting of a series of conditional statements. The classifier receives the sequence of features and then returns the class label associated to the tokens. The classifications are performed based on knowledge acquired in a learning process. The features extracted during the previous stage may not be adequate to identify sequential dependencies in the data fields. For instance a token should not be classified as title if previous token is classified as title. In order to taken into account sequential dependences between data fields, each token is not only used solely based on its own features, but also considering the features of adjacent tokens.

5. Experimental Results

5.1. Data Sets

The reference dataset was created by the Cora project (McCallum et al., 2000). It contains 500 references; during this system 350 for training and the rest 150 for testing are used. References contain 13 fields: author, title, editor, book title, date, journal, volume, tech, institution, pages, location, publisher, and date.

5.2. Evaluation Criteria

Researches in the IE field commonly report their results by using metrics such as precision, recall and the F value. In simple words, precision is the general correctness of the output and recall is the prediction of correct values.

Table 2 shows the results obtained from the experiments.

	Precision	Recall	F
author	98.3	99.8	.99
title	98.5	92.2	.96
book title	90.3	94.7	.97
date	98.6	99.8	.98
volume	90	98.5	.92
institution	97.1	65	.97
pages	96.5	99	.97
issue	90	96.7	.92
editor	92	82.3	.88
publisher	93	82	.85
journal	97.1	85.5	.93
location	96.5	88.4	.89
tech	94	72	.80

Table2. Extraction results of Cora data set

6. Conclusion

This paper described the classification based approach for automatic bibliographic metadata extraction. Experiments on the Cora data set demonstrated that decision tree based classification for bibliographic metadata extraction achieved high accuracy.

7. References

[1] D. Besagni and A. Belaid. Citation recognition for scientific publications in digital libraries. Document Image Analysis for Libraries, 00:244, 2004.

[2] D. Besagni, A. Belaid, and N. Benet. A segmentation method for bibliographic references by contextual tagging of fields. In ICDAR '03: Proceedings of the Seventh International Conference on Document Analysis and Recog-nition, page 384, Washington, DC, USA, 2003. IEEE Computer Society.

[3] F. Parmentier and A. Belaid. Logical structure recognition of scientific bibliographic references. In ICDAR '97: Pro-ceedings of the 4th International Conference on Document Analysis and Recognition, page 1072ff., Washington, DC, USA, 1997. IEEE Computer Society.

[4] Giuffrida, G., Shek, E. C., and Yang, J. Knowledge-based metadata extraction from PostScript files. In *Proceedings of the Fifth ACM Conference on Digital Librarie*s, 77-84, 2000.

[5] Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., and Fox, E. A. Automatic document metadata extraction using support vector machines. In *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Librarie*s, 37-48, 2003.

[6] Liddy, E. D., Sutton, S., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N. E., Diekema, A., McCracken, N., and Silverstein, J. Automatic Metadata generation & evaluation. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieva*l, 401-402, 2002.

[7] Mao, S., Kim, J. W., and Thoma, G. R. A dynamic feature generation system for automated metadata extraction in preservation of digital materials. In *Proceedings of the First International Workshop on Document Image Analysis for Librarie*s, 225-232, 2004.

[8] M.-Y. Day, R. T.-H. Tsai, C.-L. Sung, C.-C. Hsieh, C.-W. Lee, S.-H. Wu, K.-P. Wu, C.-S. Ong, and W.-L. Hsu. Ref-erence metadata extraction using a hierarchical knowledge representation framework, December 2006.

[9] Mitchell T.(1997). Machine Learning, McGraw-Hill International.

[10] Quinlan,J.Ross. (1987). Simplifying decision trees, International Journal of Man-Machine Studies,issue: 27(3), (pp. 221 – 234).

[11] Shafer J.C., Agrawal R., Mehta M.: "SPRINT: A Scalable Parallel Classifier for Data Mining", Proc. of the 22th International Conference on Very Large Databases, Mumbai (Bombay), India, Sept. 1996.

[12] Yilmazel, O., Finneran, C. M., and Liddy, E. D. MetaExtract: An NLP system to automatically assign metadata. In *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Librarie*s, 241-242, 2004.