

Research on Automatic Detection of Thesis Format Based on Rule Engine

Yangyang Xia^{1, a*}, Wei Zhang^{1, b}

¹School of Computer Science and Artificial Intelligence, WanJiang University of Technology, China;

^{a*}xiayangy@aliyun.com, ^b694660457@qq.com

Abstract. With the increasing development of academic research, thesis writing and publication have become important ways to measure scientific research achievements. However, non-standard thesis formats not only affect the reading experience, but may also lead to a decrease in the evaluation of academic achievements. Therefore, this article proposes a rule engine based automatic paper format detection system, aiming to improve the efficiency and accuracy of thesis format review and reduce manual review costs. This system is built on the basis of the Office Open XML document specification, which formulates a series of format checking rules for the basic format requirements of thesis. Then, the rule engine conducts in-depth checks on the thesis based on the preset format specification, automatically identifies format problems that do not comply with the specification, and outputs a detection report. It provides modification suggestions to assist users in quickly correcting format errors.

Keywords: thesis format specification, Open XML, rule engine, automatic detection

1. Introduction

In recent years, higher education in various countries has shown a trend of rapid expansion and development, and more and more people have the opportunity to receive higher education. Among them, in the undergraduate education stage, the graduation thesis serves as an important basis for measuring the quality of undergraduate education. In order to make undergraduate education more professional and improve the quality of undergraduate education, the education department will conduct spot checks on the graduation thesis, which means that universities need to invest a large amount of teacher resources every year to ensure the quality of the thesis.

The thesis format automatic detection system is designed to solve the problems of low efficiency and easy omission in manual inspection. The system base on Office Open XML, Also known as Open XML or OOXML, it is an XML based office document format that includes Word documents, Excel spreadsheets, PowerPoint presentations, as well as Chart, Diagram, Shape, and more. This specification was developed by Microsoft and adopted by ISO and IEC as ISO/IEC 29500, becoming an open international standard.

2. Related Research

2.1 Thesis Detection Related Technologies

At present, a large amount of research has been conducted on the analysis of discourse structure both domestically and internationally. In 2023, Jiang Feng et al. summarized the research on the analysis of discourse structure in English and Chinese, summarized the current trends and hotspots of research, and pointed out the problems and challenges in the analysis of Chinese discourse structure. After using support vector machine (SVM) as a classifier to extract features, Hernault et al. achieved a complete discourse structure analyzer for the first time. Li et al. used recurrent neural networks as local models to jointly model clauses, sentences, and the entire text, and constructed a discourse structure analyzer, making it a successful attempt to apply neural network models to discourse structure analysis. In terms of research on the standardization of graduation thesis format, Zhang Weiwei et al. used Word object model technology to achieve automated detection of thesis format, which can meet the requirements of their own school's graduation thesis detection. However,

the module recognition and localization method they used is to recognize and locate the structure of the graduation thesis in order, which is relatively rigid and cannot flexibly handle more complex structures. Moreover, the detection of some chapters (abstracts, tables of contents) is not comprehensive enough.

At present, there are differences in the format of undergraduate thesis among different schools and disciplines in China, and the final evaluation of thesis quality often heavily relies on the review of academic management personnel. This system uses rule engine related technologies to solve the problem of flexible configuration of thesis structure and format rules. This automated detection scheme provides theoretical and tool support for subsequent application validation.

2.2 Rule Engine Technology

A rule engine is a software system primarily used to execute a series of predefined rules to determine the way data is processed. These rules are typically based on business logic, policies, or regulatory requirements, allowing the system to automatically make decisions or recommend action plans without the need for manual intervention. Here are some core concepts related to rule engines:

Rules: Rules are the basic building blocks of a rule engine, usually expressed in the form of "if... then...", including the conditional part (facts and conditional judgments) and the action part (actions executed). For example, if the user is over 18 years old, access to adult content is allowed.

Action is one of the key components of a rule, which defines the actions or series of actions that the system needs to perform when the condition of the rule is met. Simply put, Action describes what to do if the condition is true. It is the part of the rule logic that produces actual effects, directly driving the system's response or changing the system's state.

Conditions: The conditions section of a rule used to evaluate whether a fact meets specific criteria or logic. The conditions can be simple comparisons or complex logical expressions.

Facts: Facts are data or information used by rule engines to evaluate rule conditions. These data can come from internal systems, external databases, user inputs, etc.

Rule repository: A place where all rules are stored, allowing users or administrators to add, modify, and delete rules. Maintaining centralized management and version control of rules is particularly important for maintaining large rule sets.

The main conceptual relationships of the rule engine are shown in Figure 1:

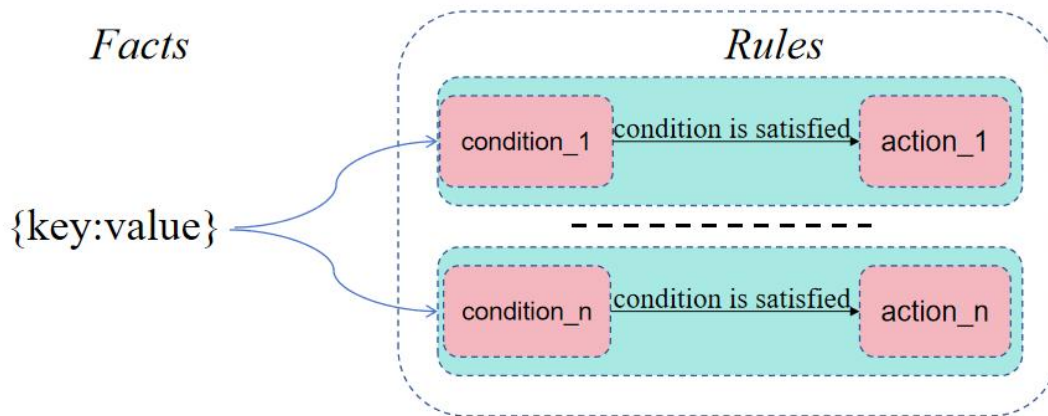


Fig. 1 Core Concept Relations

3. Design and Implementation of Detection System

3.1 Detection Rules

The core design of this system consists of three modules: the rule configuration module; Rule engine execution module, detection report module.

The rule configuration module is mainly used to abstractly define and configure rules for the structure of the paper. The structure of the paper may vary among different schools, different

degrees from the same school, and different majors. Due to the flexible configuration of rules used in this system, the main test subjects were selected for undergraduate theses, while the detection rule configuration and detection process for other theses are basically consistent with the above process.

The basic structure and corresponding format requirements of undergraduate thesis are shown in Table 1.

Table 1. Thesis Structure and detection point

Thesis Structure	Detection Point
Cover	Title, name, mentor, and font format of the content
Abstract	Word count requirements, title and paragraph formatting, font format, and keywords in both Chinese and English
Catalogue	Page number, title, and body paragraph formatting and font formatting
Content	Title, main text paragraphs, images, formulas, tables, title number
catalogue of figure, catalogue of table	Page number, title, and body paragraph formatting and font formatting
Summary	Title, main text paragraph
References	Number, paragraph format, font format, and structure of various types of literature
Appendix	Title numbering, paragraph formatting, font formatting, etc
Acknowledgments	Title, paragraph format, font format

Define the above detection points one by one as detection rules $R_1, R_2 \dots R_n$. In terms of detection logic, the relationship between rules is and, that is, detection rules defined in the same type of paper must meet the requirement that $R_1 \wedge R_2 \wedge R_n$ is true.

Considering the needs of detecting different types of thesis in system design, the detection rules for thesis of the same type are defined as a set of rules that contain the same type of thesis detection rules.

3.2 Rule Engine Execution Process

This system uses MVEL expression language as the execution tool for detection rules. It is a Java syntax based expression that supports operators for sets, arrays, and strings. After using MVEL expressions to execute rules, the specific execution process of the rule engine is shown in Figure 2.

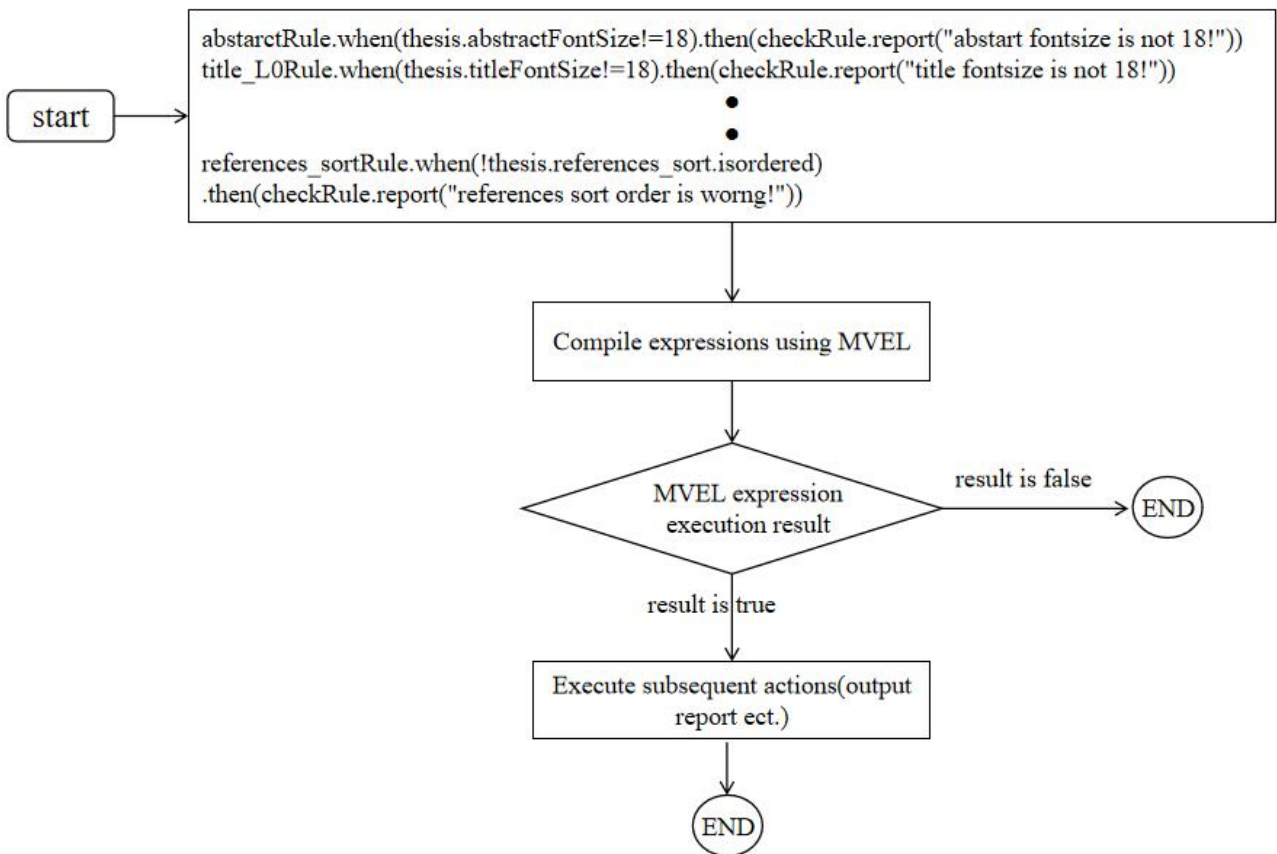


Fig. 2 MVEL Execution Process

3.3 System Implementation

The core of this system uses rule engine technology to abstractly define rules for paper format specifications. The core structure of the entire system is shown in Figure 3

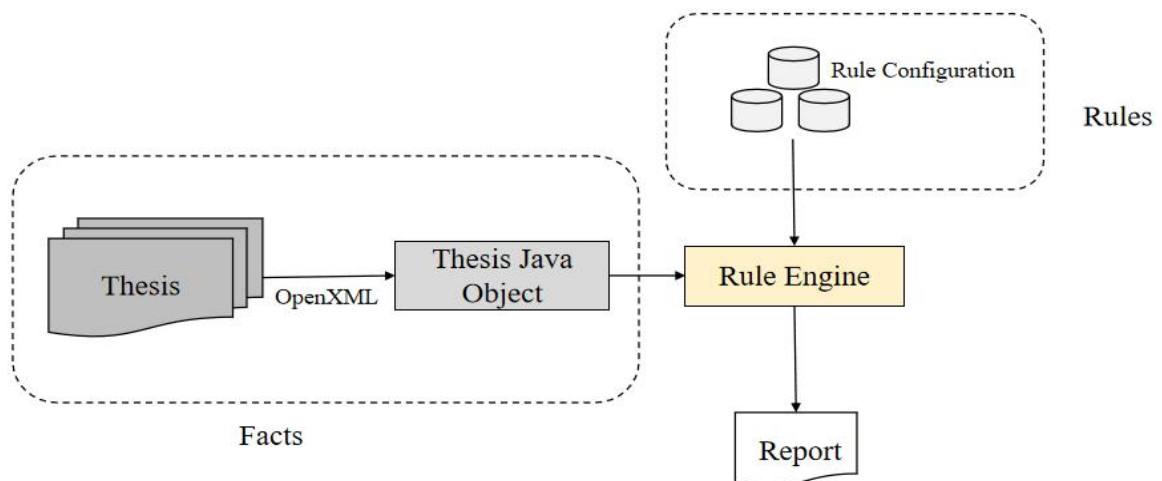


Fig. 3 System Modules

The main execution steps of the system include:

Step 1: According to the requirements of the education department, configure the rule group and rule items for thesis format detection in the system. The detection rule items include abstract, chapter titles, main text content, references, etc.

Step 2: Submit the test thesis in the system and select the corresponding detection rule group. The rule engine executes the detection rule items configured above, and the system automatically converts the rules into MVEL expressions and compiles them for execution. After completing the detection, the system outputs the corresponding detection report content.

Step 3: Based on the test report results, promptly correct any formatting issues in the thesis, including automatic correction and manual correction.

4. Result Analysis

This system uses MVEL as a rule execution tool and currently implements detection items such as title format, text font format, and reference format. During the testing phase, more than ten graduation theses were randomly selected for testing, and the first level title, font format of the main text, and reference format of the theses can be checked. Through this system detection, it can accurately display 3 formatting errors in the title of the thesis and 4 font size issues in the reference literature. Later, after manual review and comparison, it was found that there were still issues with undetectable or erroneous chart styles in the paper. For example, whether the annotations in the image meet the standards, whether all the content in the table needs to be centered, may require manual confirmation and adjustment according to the needs of the displayed content.

5. Summary

This article uses the Open XML protocol format to detect the format of the paper. Technically, it uses MVEL expression based simple rule engine technology to automatically detect the format of the abstract, chapter titles, main text paragraphs, references, and other structures in the paper and output detection reports. Finally, through manual comparison and verification, the structure of the abstract, chapter titles, and references is all effective. The charts in the main text need to rely on manual review.

The detection system can improve the efficiency of paper format checking. For deterministic format issues such as abstract format and chapter titles, it is possible to directly modify the format attributes of Open XML for correction. For the images in the paper, further research on text content format can be considered by combining OCR related technologies.

Acknowledgements

This work was supported by Wanjiang University of technology. Thanks to all the workers for their assistance.

References

- [1] Jiang F, Fan YX, Chu XM, Li PF, Zhu QM. Survey on English and Chinese Discourse Structure Analysis. *Journal of Software*,2023,34(09):4167-4194.
- [2] Hernault H,Prendinger H,du Verle DA,Ishizuka M. HILDA: A discourse parser using support vector machine classification. *Dialogue & Discourse*,2010,1(3):1-33.
- [3] Li JW,Li RM, Hovy E. Recursive deep models for discourse parsing. In:Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP).Doha:ACL,2014.2061–2069.
- [4] European Computer Manufactures Association. ECMA-376: Office Open XML File Format-Fundamentals And Markup Language Reference[EB/OL]. Brussels: ECMA, 2016[2022-2-18].
- [5] Zhang Weiwei.Design and Implementation of Thesis Format Specification Automatic Detection System[D].Dalian University of Technology,2019.
- [6] Jiménez P, Corchuelo R. On the design of an advanced business rule engine. *Software: Practice and Experience*, 2022, 52(10): 2097-2126.

- [7] Vu T M H, Thi T T P, Le Dinh T. Towards a Rule Modeling Framework for Context-aware Smart Service Systems[C]//ITM Web of Conferences. EDP Sciences, 2023, 51: 04005.
- [8] Cicolini L, Carloni F, Santambrogio M D, et al. One Automaton to Rule Them All: Beyond Multiple Regular Expressions Execution[C]//2024 IEEE/ACM International Symposium on Code Generation and Optimization (CGO). IEEE, 2024: 193-206.
- [9] Mirza A R, Sah M. Automated software system for checking the structure and format of ACM SIG documents[J]. New Review of Hypermedia and Multimedia, 2017, 23(2): 112-140.
- [10] Guo X G X. Research on logical structure annotation in English streaming document based on deep learning[J]. Journal of Computers, 2021, 32(4): 109-122.
- [11] N. Li, Q. Liang, Y.-M. Shi, The function of format information in document understanding, Journal of Beijing Information Science and Technology University 27(6)(2012) 1-7.