ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.0846
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

Research on Construction Quality Problem Prevention

Shaohua Jiang[1]*, Jingqi Zhang[2]

[1] *Department of Construction Management, Faculty of Infrastructure Engineering, Dalian University of Technology, Dalian 116024, China, E-mail address: shjiang@dlut.edu.cn*
[2] *Department of Construction Management, Faculty of Infrastructure Engineering, Dalian University of Technology, Dalian 116024, China, E-mail address: JingqiZhang@mail.dlut.edu.cn*

Abstract: A project's success is directly guaranteed by the prevention of construction-related problems. Nevertheless, the prevention of quality issues frequently overlooks how issues are coupled with one another, which might result in a domino effect of quality issues. In order to solve the above problems, this work first preprocesses unstructured text data with quality problem coupling. Then the pre-processing data is used to build a knowledge base for the prevention of construction quality problems. Then the text similarity algorithm is used to mine the coupling relationship between the qualities and enrich the information in the database. Finally, some text is used as test object to verify the validity of the method. This study enriches the research around the prevention of building quality problems.
Key words: Construction quality problems; Text mining; Knowledge base

1. INTRODUCTION

Quality issues in the Architecture, Engineering, and Construction (AEC) industry, including structural flaws and functional impairments, pose safety risks and can significantly reduce a project's lifespan[1]. Traditional quality management methods, relying on manual inspection and experience, are inadequate for analyzing vast data and identifying interconnected problems. This can lead to a domino effect of quality defects[2]. Current methods fail to address the complex and dynamic nature of AEC projects, impacting efficiency and escalating costs. With the advent of the digital era, it's crucial for the AEC industry to embrace digital tools in construction quality management to effectively prevent and control quality issues[3].

Data can be categorized into structured and unstructured types. Structured data are easily processed by computers due to their clear, numerical structure. In contrast, unstructured data, prevalent in construction quality management, often contains vital but underutilized information in texts like laws, regulations, and construction logs[4]. Traditionally, extracting value from these texts required manual effort by experienced engineers, a process becoming inefficient with the growing complexity of construction and the surge in quality reports. However, mining and analyzing unstructured data meticulously can reveal intricate connections in quality issues, enhancing our understanding and evaluation of potential risks.

Drawing on the successful experience of manufacturing industry and other fields, this paper tries to introduce the commonly used text mining technology in information extraction into the prevention of building quality problems. The initial unstructured text data is structured, the required knowledge is

automatically extracted and stored for reuse, and the knowledge base is built. It is helpful for text similarity algorithm to mine the coupling relation of construction engineering quality problems and enrich the built knowledge base. Thus, the automation level of construction engineering problem prevention can be improved, and efficient and accurate engineering quality management can be realized.

This research holds paramount significance in the field of AEC, as it addresses a critical gap in current quality management practices. Despite the advances in digital technology, the AEC industry has been slow in adopting data-driven approaches for quality problem prevention, particularly in harnessing the wealth of information embedded in unstructured data sources. Our study is one of the first to systematically apply text mining techniques to identify and analyze the coupling relationships of quality issues in construction projects. The outcome of this research not only contributes to the theoretical understanding of construction quality problem prevention but also provides practical insights for industry professionals. By integrating a novel approach to process and analyze unstructured data, this research delineates a clear pathway towards the implementation of more efficient and effective quality management strategies in the AEC industry. In doing so, it fills a vital research gap and sets the stage for future advancements in this domain. The urgency and relevance of this research are underscored by the increasing complexity of construction projects and the escalating costs associated with quality defects, making our findings not just timely but essential for the evolution of construction quality management.

The rest of this study is organized as follows. The second part describes the methodology for constructing a knowledge base for the prevention of building quality problems. This is followed by validating its performance through illustrative examples in the third part. Finally, the conclusions of this paper and future research directions are presented.

2. Methodology

2.1. Text Collection and Domain Dictionary Construction Based on OCR Technology

OCR technology, a critical component in digitizing textual data from images, has advanced significantly with the emergence of sophisticated software solutions[5]. The OCR recognition process is shown in Figure 1. Abbyy Fine Reader, developed by the Russian company Abbyy, exemplifies this technological progress. This professional OCR software offers an array of features for professionals, including accurate digitization of document images, efficient retrieval, editing capabilities, secure protection, and seamless sharing and collaboration functions[6].

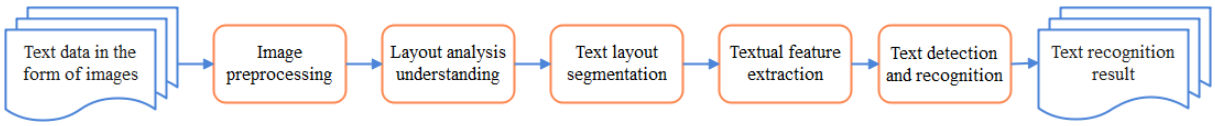

Figure 1. OCR recognition process diagram

In practical engineering contexts, especially within the construction industry, data often exists in the form of images or scanned documents, which are not inherently processable by computers. Recognizing this challenge, Abbyy Fine Reader has been selected for its exceptional text recognition capabilities, making it an ideal tool for processing quality-related text images[7].

The construction industry, with its abundance of proper nouns and unique processes, presents a specific challenge for OCR technology. Relying on general dictionaries for OCR recognition often results in a cumbersome manual review process due to the industry's specialized vocabulary.

Consequently, developing a construction field-specific dictionary becomes an imperative[8]. There are principally two methods to construct such a lexicon: the corpus-based approach and the knowledge-based approach.

The corpus-based method leverages syntactic patterns and seed lists of opinion words within a large corpus to identify domain-specific words, offering the advantage of discovering industry-specific terminology. Meanwhile, the knowledge-based approach relies on existing lexical resources like Word Net or How Net, starting with a manually collected seed set of sentiment words and expanding it by identifying synonyms and antonyms in the knowledge base. However, the latter approach faces limitations, especially in the context of Chinese lexical resources[9]. To address these limitations, this study adopts the corpus-based approach for the semi-automated construction of a dictionary tailored to the construction field.

This research utilizes the user dictionary feature of ABBYY FineReader software, supplemented by the widely-used Chinese word segmentation tool, jieba, for text parsing. It involves annotating the segmentation results to create a specialized construction field dictionary. This approach significantly enhances the accuracy of recognizing technical terms compared to direct OCR.

2.2. Regular Expression Based Text Processing

Regular expressions, a powerful tool for text processing, were first proposed in 1956 by Stephen Kleene, an eminent American mathematical scientist[10]. They are specifically designed to automatically match text content that adheres to a predetermined set of rules. Essentially, a regular expression is a string composed of ordinary characters and special characters with distinct meanings, together forming a logical structure that characterizes the target text[11]. This enables precise and efficient extraction of knowledge from text data with known features.

Although not an independent computer language, regular expressions serve as a versatile text processing tool. They are adept at screening, modifying, replacing, and segmenting the target text to extract specific content, ultimately yielding a logical expression structure that aligns with the target string[12]. Regular expressions usually include two types of characters: ordinary characters, which represent themselves without any special functionality, and special characters or metacharacters, each imbued with a unique meaning to perform specific functions.

The application of regular expressions in knowledge extraction involves parsing text that typically contains information like names, quality issues, phenomena, causes, etc[13]. The initial step is to discern the logical structure of the target text, thus facilitating the creation of effective regular expressions. For instance, in the context of quality issues, relevant information can be grouped into a 'quality problem information group.'

During the development of regular expressions, tools like RegexBuddy 4 prove invaluable. This regular expression editing tool, with its visual test window, vividly highlights the logical structures matching the regular expression. It not only aids in the design and editing of regular expressions but also supports their export to numerous programming languages[14]. Additionally, RegexBuddy 4 offers a variety of programming functions related to regular expressions, which can be selected directly for use.

When identifying regular expressions, the clarity provided by RegexBuddy 4's visual test window is instrumental. It not only highlights the logical structure matching the regular expression but also assists in identifying specific elements like the name of a quality problem[15]. The process of writing regular expressions involves a step-by-step approach to determine the logic of writing, thereby visualizing the meaning of the regular expression.

Moreover, RegexBuddy 4's support for regular expression export across many programming languages enhances its utility. It offers an array of programming functions related to regular

expressions, as illustrated in Figure 2. An example of its application can be seen in the Python 2.7 language environment, where a partial string interception is achieved through regular expression matching. This expansion of the original text enriches its content and provides a more comprehensive understanding of the significance and application of regular expressions in text processing and knowledge extraction.

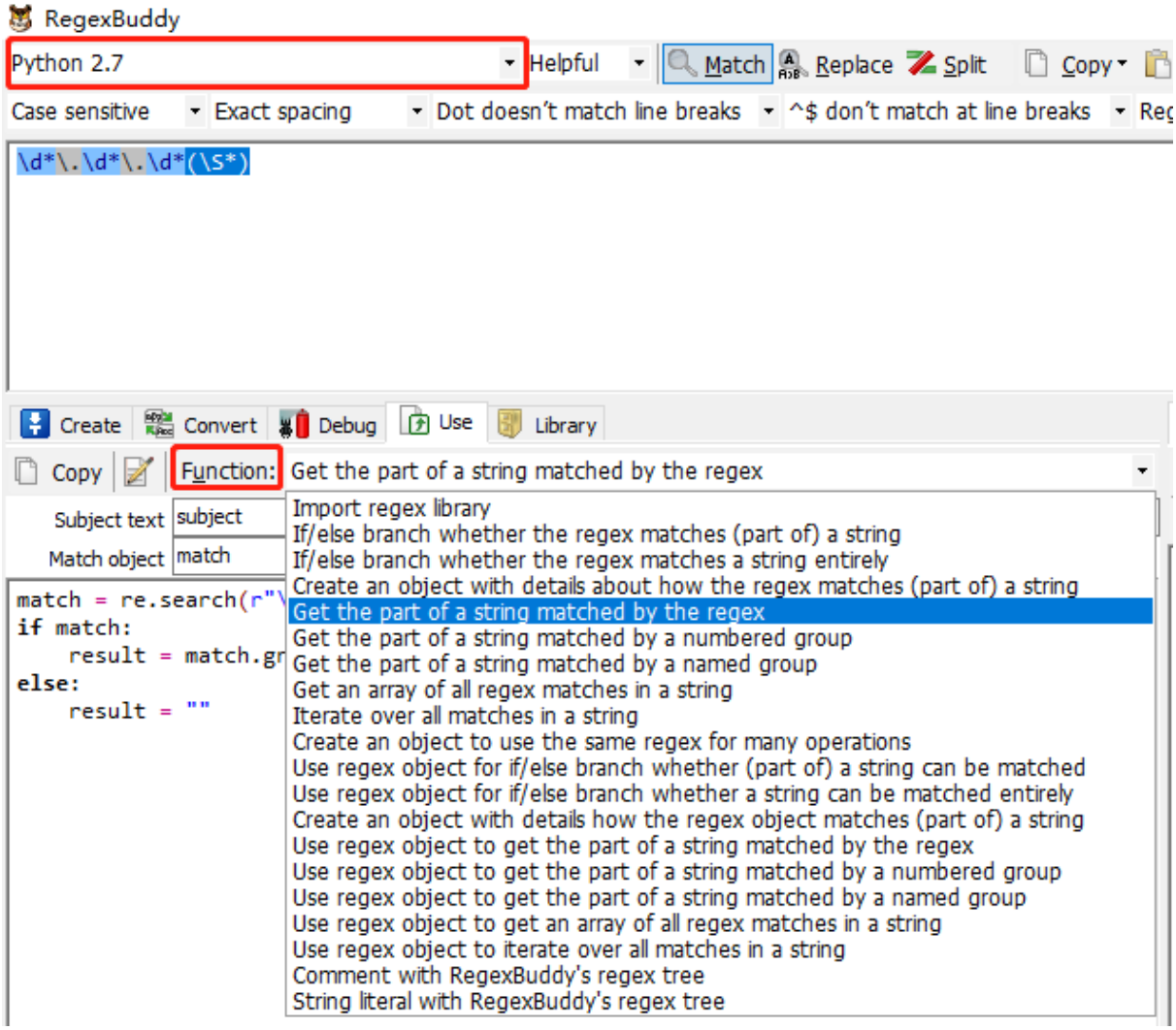

Figure 2. The export of regular expressions in programming language

2.3. Knowledge base on quality issues prevention

After knowledge extraction, it is necessary to set up a database to store and manage the quality problem prevention data. After comparing the common relational databases in the market, MySQL database was chosen for this study, which is a free and open source relational database from Oracle Corporation and is a cross-platform small and medium-sized database[16]. Although there is still a gap between MySQL and other large databases in some of its functions, its rich scenario applicability makes it still widely used among developers. Under the client, the second tier architecture contains most of the core services of the MySQL database functions, including querying , caching , optimization and so on[17]. It also implements some cross-storage engine features such as stored procedures, views, triggers, ect. In addition, the third tier of the architecture is the storage engine, which is responsible for storing and extracting data from the database. MySQL provides several storage leads, and choose the

most suitable storage engine according to the specific needs of the project, and InnoDB is chosen by default.

The common similarity measurements or algorithms and their characteristics are shown in Table 1. Considering the nature of the quality issue texts addressed in this paper, which typically range from 100 to 200 words, they fall into the category of short texts. Additionally, the distribution of text vectors is relatively scattered. The application requirement is to measure the similarity between these texts, rather than their absolute values. Therefore, cosine similarity will be used subsequently to calculate the text similarity. Cosine similarity, a fundamental and widely-used method in text similarity calculation, evaluates similarity based on the cosine of the angle between two vectors, X and Y. These vectors, representing text, are n-dimensional where n is the number of items in the sample space. The cosine value ranges from -1 to 1, indicating varying degrees of similarity: a value close to 1 suggests a small angle and high similarity, while values near 0 or -1 indicate orthogonality or opposition, respectively, reflecting less similarity. To apply this method, texts are first vectorized; then, their similarity is calculated. Texts with similarities above a certain threshold are identified, allowing the original text to be recognized after processing. Cosine similarity stands out for its solid theoretical basis, simplicity, and proven high accuracy in feature similarity analysis, making it a preferred choice in many fields.

Names	Related Meaning	Advantages	Disadvantages
Euclidean distance calculation	Measurement of the absolute distance between points in a multidimensional space	Suitable for dense and continuous data calculations	Ensure that the scales of each dimension are at the same level
Cosine similarity calculation	The cosine of the Angle between two vectors is used as a measure of the difference between two individuals	More directly compare directional differences between vectors	Not sensitive to absolute values
BM25 algorithm	Based on probabilistic retrieval model, a given query is sorted according to its relevance to the text	Effectively avoid the effect of increasing the frequency of long text target words	Focus only on the correlation between a given query and the target text

Table 1. The common similarity measurement algorithms and their characteristics

3. Illustrative examples

3.1. Construction quality problems related text preprocessing

In this detailed and comprehensive study, we delve into the intricate realm of construction quality management, leveraging a wealth of explicit and implicit knowledge embedded in specialized texts. At the core of our investigation are key excerpts from the authoritative "Manual for the Prevention and Control of Common Construction Quality Problems (Fourth Edition)," a Chinese manual that provides exhaustive guidelines on the construction of crucial structural elements, including steel processing and installation, concrete works, and cast-in-place reinforced concrete structures. These guidelines are pivotal in preventing and addressing quality issues in construction projects.

To augment our research, we incorporated quality rectification notification sheets from a supervisory unit, covering the period from September 2021 to March 2022. This inclusion of real-world data adds a layer of practicality to our analysis, enabling us to connect theoretical knowledge with on-ground realities.

The research process began with the meticulous preprocessing of this base corpus, transforming it into an actionable data source primed for in-depth analysis. A key step in this process was the utilization of OCR technology. By importing the relevant quality text data files into an advanced OCR editor, we leveraged the technology to automatically recognize text areas within the documents. This efficient method facilitated an effortless transition from image-based texts to editable and analyzable formats.

The OCR software we used was particularly adept at identifying and highlighting characters with low recognition confidence. This feature proved invaluable, as it allowed users to easily spot potential errors in the OCR process and make necessary corrections. Users could conveniently right-click on text with low confidence to correct it using the software's tools, or directly edit the recognized text through the "Recognition - Verify Text" function. This functionality extended to viewing low-confidence characters, adding new words to a user-defined dictionary, and more.

Given the complexity of the technical terminology in construction, choosing the right lexical tool for dictionary construction was critical. After extensive comparisons, considering factors like usability, stability, and open-source availability, we selected Jieba as our lexical tool. Jieba's superiority lay in its enhanced accuracy in recognizing technical terms, especially when compared to direct OCR recognition.

Our analysis was grounded in the comparison of various lexical tools, focusing specifically on their performance with the excerpted corpus from the "General Specification for Concrete Structures" under their default lexical modes. This methodical approach ensured the objectivity and reliability of our comparisons, leading to a well-informed choice of Jieba for our domain-specific dictionary construction.

In conclusion, this study represents a significant step forward in the field of construction quality management, blending cutting-edge OCR technology with deep domain expertise to offer actionable insights into the prevention and rectification of quality issues in construction projects.

3.2. Building a Knowledge Base for Quality Problem Prevention

This study employs entity-relationship (E-R) diagrams to meticulously outline the data storage architecture, thereby presenting a more vivid and intuitive representation of the associative relationships between various data elements. This approach is instrumental in laying a solid foundation for developing the system's database. By meticulously analyzing and designing the entity-relationship diagram, it aligns with the unique data storage characteristics pertinent to this research.

The study introduces two critical data tables, pivotal to the system's database:

1. Quality Problems Knowledge Sheet: This table serves as a repository for a vast array of information regarding quality issues. It meticulously catalogues each quality problem by its unique number and name, along with detailed descriptions of its causes and preventive measures. Additionally, it includes information about related or 'coupling' quality problems. This sheet is not just a collection of data; it's a comprehensive knowledge base that forms the cornerstone of the database, providing insights and facilitating effective management of quality-related issues.

2. Quality Record Sheet: The quality record table is a crucial component designed to chronicle individual quality problem records. It plays a vital role in supporting subsequent analyses, particularly in the coupling of quality problems. Key fields in this table include the quality problem number, the problem's name, the site of occurrence, and the quality record number. This sheet is not just a

record-keeping tool; it's an analytical instrument that aids in pinpointing patterns, identifying recurrent issues, and suggesting areas for improvement.

These sheets form an integrated framework within the system's database, enabling a more holistic and dynamic approach to quality management. The E-R diagrams not only simplify the understanding of these complex relationships but also enhance the efficiency and accuracy of data management, paving the way for more informed decision-making and strategic planning in quality control processes.

4. Conclusions

The aim of this research is to create a comprehensive knowledge base addressing engineering quality issues. This is to mitigate the challenges of subjectivity, instability, and the overlooked interconnections among these issues in preventing engineering quality problems. Utilizing Optical Character Recognition (OCR) technology and regular expressions, the study preprocesses textual data to extract pertinent knowledge in the field of engineering quality. Consequently, an initial knowledge base dedicated to preventing engineering quality issues is established. Following this, algorithms are employed to scrutinize the root causes of these quality issues and document them. This process uncovers and records the interrelated nature of these problems, enriching the knowledge base with new data for future reference and use. The study then validates the effectiveness of the algorithm and the relevance of the data through empirical tests and practical application. By providing systematic technical and data support, this research significantly contributes to the enhancement of construction quality problem prevention, offering new perspectives for future investigations in this area.

The research presented in this paper makes significant strides in constructing a knowledge base for engineering quality issues using OCR technology and regular expressions. However, there are areas that could be improved upon or explored further in future studies. One notable limitation is the reliance on OCR technology, which, despite advancements, may still struggle with accurately recognizing and processing highly technical or poorly scanned texts. This could lead to gaps or inaccuracies in the extracted data, potentially affecting the overall reliability of the knowledge base. Additionally, while regular expressions are powerful for text processing, they require predefined patterns, which might not capture the nuanced or evolving language used in construction documents. This could limit the ability to fully understand and document the complexity of construction quality issues.

Future research directions could include exploring machine learning algorithms for more dynamic and adaptable text processing, capable of evolving with the language used in construction documents. Integrating artificial intelligence could improve the accuracy of text recognition and extraction, particularly in dealing with complex and technical language. Furthermore, expanding the scope of data sources, such as including more diverse types of construction documents and languages, could enhance the comprehensiveness of the knowledge base. Finally, there is room for developing more sophisticated algorithms to analyze the interconnections between different quality issues, which could provide deeper insights into the causal relationships and potential preventative measures in the field of construction quality management.

ACKNOWLEGEMENTS

The work described in this paper was supported by the National Natural Science Foundation of China (Grant No. 52078101).

REFERENCES

[1]X. W. Zhou and Y. S. Wang, "Understanding competency requirements in the context of AEC industry informatization: policy insights from China," *Eng. Constr. Archit. Manag.*, 2023 Aug 2023, doi: 10.1108/ecam-11-2022-1080.

[2]Q. Wang and J. W. Shang, "Analysis of the quality improvement path of supply chain management under the background of Industry 4.0," *International Journal of Technology Management*, vol. 91, no. 1-2, pp. 1-18, 2023, doi: 10.1504/ijtm.2023.127854.

[3]O. Ardon *et al.*, "Quality Management System in Clinical Digital Pathology Operations at a Tertiary Cancer Center," *Laboratory Investigation*, vol. 103, no. 11, Nov 2023, Art no. 100246, doi: 10.1016/j.labinv.2023.100246.

[4]J. Willems, I. Bablok, E. Farin-Glattacker, and T. Langer, "Barriers and facilitating factors of care coordination for children with spinal muscular atrophy type I and II from the caregivers' perspective: an interview study," *Orphanet Journal of Rare Diseases*, vol. 18, no. 1, Jun 2023, Art no. 136, doi: 10.1186/s13023-023-02739-w.

[5]G. Lv, Y. N. Sun, F. D. Nian, M. F. Zhu, W. L. Tang, and Z. Z. Hu, "COME: Clip-OCR and Master ObjEct for text image captioning," *Image and Vision Computing*, vol. 136, Aug 2023, Art no. 104751, doi: 10.1016/j.imavis.2023.104751.

[6]Q. D. Nguyen, N. M. Phan, P. Krömer, and D. A. Le, "An Efficient Unsupervised Approach for OCR Error Correction of Vietnamese OCR Text," *Ieee Access*, vol. 11, pp. 58406-58421, 2023, doi: 10.1109/access.2023.3283340.

[7]P. Jain, K. Taneja, and H. Taneja, "Which OCR toolset is good and why? A comparative study," *Kuwait Journal of Science*, vol. 48, no. 2, 2021, doi: 10.48129/kjs.v48i2.9589.

[8]H. Elshahaby and M. Rashwan, "An end to end system for subtitle text extraction from movie videos," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 4, pp. 1853-1865, Apr 2022, doi: 10.1007/s12652-021-02951-1.

[9]P. Thierfelder, Z. G. Cai, S. T. Huang, and H. Lin, "The Chinese lexicon of deaf readers: A database of character decisions and a comparison between deaf and hearing readers," *Behavior Research Methods*, 2023 Dec 2023, doi: 10.3758/s13428-023-02305-z.

[10]N. Chida and T. Terauchi, "On Lookaheads in Regular Expressions with Backreferences," *Ieice Transactions on Information and Systems*, vol. E106D, no. 5, pp. 959-975, May 2023, doi: 10.1587/transinf.2022EDP7098.

[11]S. Broda, A. Machiavelo, N. Moreira, and R. Reis, "Location automata for regular expressions with shuffle and intersection," *Information and Computation*, vol. 295, Dec 2023, Art no. 104917, doi: 10.1016/j.ic.2022.104917.

[12]X. W. Sun *et al.*, "Efficient regular expression matching over hybrid dictionary-based compressed data," *Journal of Network and Computer Applications*, vol. 215, Jun 2023, Art no. 103635, doi: 10.1016/j.jnca.2023.103635.

[13]F. Parolini and A. Miné, "Sound static analysis of regular expressions for vulnerabilities to denial of service attacks," *Science of Computer Programming*, vol. 229, Jul 2023, Art no. 102960, doi: 10.1016/j.scico.2023.102960.

[14]D. Conficconi, E. del Sozzo, F. Carloni, A. Comodi, A. Scolari, and M. D. Santambrogio, "An Energy-Efficient Domain-Specific Architecture for Regular Expressions," *Ieee Transactions on Emerging Topics in Computing*, vol. 11, no. 1, pp. 3-17, Jan 2023, doi: 10.1109/tetc.2022.3157948.

[15]J. Doleschal, B. Kimelfeld, and W. Martens, "THE COMPLEXITY OF AGGREGATES OVER EXTRACTIONS BY REGULAR EXPRESSIONS *," *Logical Methods in Computer Science*, vol. 19, no. 3, 2023, Art no. 12, doi: 10.46298/lmcs-19(3:12)2023.

[16]K. Gopi *et al.*, "Developing a MySQL Database for the Provenance of Black Tiger Prawns (Penaeus monodon)," *Foods*, vol. 12, no. 14, Jul 2023, Art no. 2677, doi: 10.3390/foods12142677.

[17]P. Yin and J. Cheng, "A MySQL-Based Software System of Urban Land Planning Database of Shanghai in China," *Cmes-Computer Modeling in Engineering & Sciences*, vol. 135, no. 3, pp. 2387-2405, 2023, doi: 10.32604/cmes.2023.023666.