DOI: 10.34185/1991-7848.itmm.2026.01.043

АВТОМАТИЗОВАНЕ ВІДНОВЛЕННЯ ПРОДУКЦІЙНИХ ГРАМАТИК НА ОСНОВІ СТРУКТУРНОГО АНАЛІЗУ МАТЕМАТИЧНИХ ФОРМУЛ У ФОРМАТІ LATEX

Андрющенко В.О.[1] [ORCID], Лебеденко А.В.[2] [ORCID]

[1]Український державний університет науки і технологій,

канд. техн. наук, доц., Україна

[2]Український державний університет науки і технологій, аспірант, Україна

Анотація. *Досліджено проблему семантико-структурного аналізу математичних виразів у наукових текстах, поданих у форматі LaTeX. Проведено аналіз існуючих підходів у галузі Математичного інформаційного пошуку та виявлено їхні недоліки, пов'язані із залежністю від статичних словників або низькою інтерпретованістю. Запропоновано метод автоматизованого відновлення продукційних граматик на основі принципів конструктивно-продукційного моделювання. Розроблено алгоритм, який здійснює динамічний лексичний аналіз, побудову абстрактного синтаксичного дерева з урахуванням префіксних операторів, а також висхідне згортання дерева для генерації правил. Відмінністю підходу є динамічне виділення термінального носія і сигнатури конструкторів без попередньо заданих шаблонів. Результати є базовим етапом для створення прозорих алгоритмів кластеризації наукових документів на основі їхнього математичного апарату.*

Ключові слова: *інформаційні технології, програмне забезпечення, конструктивно-продукційне моделювання, формальні граматики, структурний аналіз.*

Вступ. Сучасні цифрові бібліотеки містять величезні обсяги документів, насичених математичним апаратом. Проте більшість існуючих систем інформаційного пошуку та обробки природної мови (NLP) стикаються з проблемою «семантичної сліпоти» щодо математичних формул, поданих у форматі LaTeX.

Аналіз останніх досліджень у галузі математичного інформаційного пошуку (MathIR) показує розвиток двох основних напрямків. Перший — використання дерев розташування символів та дерев операторів (наприклад, система Tangent-L [1]). Цей підхід є ефективним, проте він критично залежить

від жорстко заданих, статичних словників LaTeX-команд і не здатний до динамічної адаптації. Другий напрямок — застосування великих мовних моделей, таких як MathBERT [3]. Вони здатні враховувати нелінійний контекст, але функціонують як "чорні скриньки" та не надають математично строгої інтерпретації структурної подібності. З огляду на це, актуальною є задача розробки повністю прозорих структурних методів, що здатні до самостійного виведення правил побудови формул на основі теорії формальних граматик та конструктивно-продукційного моделювання (КПМ) [2].

Метод вирішення. У рамках дослідження розв'язано підзадачу автоматизованого семантико-структурного аналізу математичних виразів, поданих у форматі LaTeX. Головна проблема полягає у неможливості машинного виділення глибокої математичної логіки з "сирого" тексту розмітки без залучення громіздких ручних словників або втрати ієрархічної структури формули. Розроблений підхід вирішує задачу автоматичного відновлення формальної граматики з коду формул, перетворюючи їх на набір строгих математичних правил.

В основу роботи алгоритму покладено принципи конструктивно-продукційного моделювання (КПМ). Формула розглядається як результат роботи узагальненого конструктора [2]:

$$C = \langle M, \Sigma, \Lambda \rangle \quad (1)$$

де M — носій (елементи формул), Σ — сигнатура операцій (продукційні правила), Λ — інформаційна підтримка. Процес обробки кожної формули здійснюється у три послідовні етапи:

1. *Семантична токенізація*: спочатку текст розбивається на логічні примітиви, зберігаючи пробіли як розділювачі та класифікуючи елементи на змінні (VAR), числа (NUM), диференціали (DIFF) та оператори.
2. *Синтаксичний аналіз (побудова абстрактного синтаксичного дерева)*: Реалізовано механізм "жадібного" поглинання для префіксних операторів. Наприклад, оператор інтеграла рекурсивно сканує вираз до токена диференціала, формуючи єдиний структурний вузол.
3. *Відновлення граматики*: за допомогою методу висхідного згортання (від листків дерева до кореня) генеруються строгі продукційні правила. При цьому відбувається чітке відокремлення термінального носія (змінні, числа) від

сигнатури конструкторів (суми, дроби, рівняння), після чого локальні правила формул об'єднуються в єдину узагальнену граматику предметної області.

Розглянемо приклад відновлення продукційної граматики у табл. 1:

Таблиця 1

Приклад відновлення продукційної граматики

y = k x + b	a = (b + c) (x + y)
EQUATION → VAR OP SUM SUM → PRODUCT OP VAR PRODUCT → VAR OP VAR VAR → 'b' | 'k' | 'x' | 'y' OP → '*' | '+' | '='	EQUATION → VAR OP PRODUCT PRODUCT → PARENS OP PARENS PARENS → '(' SUM ')' SUM → VAR OP VAR VAR → 'a' | 'b' | 'c' | 'x' | 'y' OP → '*' | '=' | '+' | '(' | ')'
Відновлена продукційна граматика EQUATION → VAR OP SUM | VAR OP PRODUCT SUM → PRODUCT OP VAR | VAR OP VAR PRODUCT → VAR OP VAR PARENS → '(' SUM ')' VAR → 'a' | 'b' | 'c' | 'k' | 'x' | 'y' OP → '*' | '+' | '=' | '(' | ')'	

Розроблений підхід має дві принципові переваги над існуючими рішеннями у галузі MathIR:

По-перше, на відміну від символьних і деревоподібних методів (наприклад, системи Tangent-L), запропонований метод не потребує наперед заданих, вичерпних статичних словників LaTeX-команд. Алгоритм здатний до динамічної адаптації — він самостійно визначає арність та структуру невідомих функцій на основі синтаксичного контексту (наявності фігурних дужок чи індексів). По-друге, порівняно з сучасними нейромережевими моделями на базі трансформерів (наприклад, MathBERT), розроблений метод є концепцією "білої скриньки". Замість неінтерпретованих векторних вкладень (embeddings), алгоритм видає математично строгий, прозорий та зрозумілий людині набір продукційних правил, що дозволяє не лише групувати наукові документи, але й логічно пояснювати причину їхньої структурної подібності.

Висновки

Розроблено метод автоматизованого відновлення продукційних граматик для математичних формул. На відміну від існуючих підходів (Tangent-L, MathBERT), метод не використовує статичних словників і забезпечує повну

"білу скриньку" інтерпретації результатів. Виділення базового носія та сигнатури конструкторів ідеально узгоджується з теоретичними засадами конструктивно-продукційного моделювання. Отримані множини продукційних правил пропонується використати як ознаковий простір для TF-IDF векторизації з метою високоточної кластеризації наукових документів.

ЛІТЕРАТУРА / REFERENCE

1. Greiner-Petter R. et al. Discovering Mathematical Objects of Interest – A Study of Mathematical Notations // Proceedings of The Web Conference (WWW '20). ACM, 2020. P. 1445-1456.
2. Shynkarenko V. I., Ilman V. M. Constructive-Synthesizing Structures and Their Grammatical Interpretations. I. Generalized Formal Constructive-Synthesizing Structure // Cybernetics and Systems Analysis. — 2014. — Vol. 50, No. 5. — P. 655–662. DOI: 10.1007/s10559-014-9655-z.
3. Zhong J. et al. MathBERT: A Pre-Trained Model for Mathematical Formula Understanding // arXiv preprint arXiv:2105.00377. 2021. 12 p.

AUTOMATED RECONSTRUCTION OF PRODUCTION GRAMMARS BASED ON STRUCTURAL ANALYSIS OF MATHEMATICAL FORMULAS IN LATEX FORMAT

Vadym Andriushchenko, Anton Lebedenko

Abstract. *The problem of semantic-structural analysis of mathematical expressions in scientific texts presented in LaTeX format is investigated. The analysis of existing approaches in the field of Mathematical Information Retrieval is carried out and their shortcomings associated with dependence on static dictionaries or low interpretability are revealed. A method of automated restoration of production grammars based on the principles of constructive-production modeling is proposed. An algorithm is developed that performs dynamic lexical analysis, construction of an abstract syntactic tree taking into account prefix operators, as well as upward tree folding for rule generation. The difference of the approach is the dynamic selection of the terminal carrier and signatures of constructors without predefined templates. The results are a basic stage for creating transparent algorithms for clustering scientific documents based on their mathematical apparatus.*

Keywords: *information technology, software, structural and production modeling, formal grammars, structural analysis.*