УДК 004.942
Осейко П. – ст. гр. СП-42
Тернопільський національний технічний університет імені Івана Пулюя

РОЗРОБКА МЕТОДІВ МАШИННОГО НАВЧАННЯ СКЛАДНИХ ТОПОЛОГІЧНИХ ПРОСТОРІВ У ЗАСТОСУВАННІ ДО ЛЕКСИЧНОЇ ТАКСОНОМІЇ

Науковий керівник: к.т.н., доцент Стоянов Ю. М.

Oseiko P.
Ternopil Ivan Puluj National Technical University

DEVELOPMENT OF MACHINE LEARNING METHODS FOR COMPLEX TOPOLOGICAL SPACES IN APPLICATION TO LEXICAL TAXONOMY

Supervisor: Ph.D. in Engineering, associate professor Stoianov Ju .M.

Ключові слова: програмна платформа, моніторинг ризиків, стохастичні системи, штучні нейронні мережі, стохастичні обчислення, глибоке навчання
Keywords: software platform, risk monitoring, stochastic systems, artificial neural networks, stochastic computing, deep learning

У сучасних дослідженнях у галузі штучного інтелекту та обробки природної мови (NLP) задача автоматичного вилучення знань та побудови ієрархічних структур залишається однією з найважливіших. Традиційні підходи до моделювання відношень між сутностями (наприклад, гіпернімії) здебільшого обмежуються виявленням зв'язків між ізольованими парами елементів. Такий бінарний підхід ігнорує глобальні структурні обмеження та контекстуальну взаємозалежність даних.

Теорія претопології пропонує потужний математичний апарат для моделювання складних відношень між множинами сутностей. На відміну від класичних методів, використання такого дрібномодульного (fine-grained) моделювання дозволяє значно точніше фіксувати реальні взаємозв'язки. Однак застосування претопології на практиці стикається з серйозними обмеженнями щодо масштабованості (scalability) через високу обчислювальну складність алгоритмів. Ця робота пропонує вирішення проблеми масштабованості шляхом переформулювання задачі вилучення відношень як "моделі поширення" (propagation model) із заданими структурними обмеженнями, що дозволяє враховувати апріорні знання про очікувану структуру даних.

Основу нашої методології становить визначення оператора псевдозамикання (pseudo-closure operator), який у претопології виступає математичною моделлю концепції поширення. Ми визначаємо цей оператор як логічну комбінацію гетерогенних околів або джерел інформації.

Такий підхід дозволяє навчати моделі, які одночасно експлуатують знання, отримані за допомогою як статистичних, так і числових (наприклад, нейромережевих) підходів. Формалізація процесу поширення через оператор псевдозамикання дозволяє системі динамічно оцінювати, як певна властивість або семантичне відношення "поширюється" від однієї множини термінів до іншої в межах заданого простору.

Процес машинного навчання для такого оператора природним чином вписується у парадигму багатекземплярного навчання (Multiple Instance, MI). У цій парадигмі навчання відбувається не на рівні окремих екземплярів (інстансів), а на рівні мультимножин — так званих "мішків" (bags) екземплярів.

Незважаючи на те, що середовище MI добре підходить для концептуалізації цієї задачі, його пряме застосування для навчання претопологічного простору призводить до формування набору "мішків", розмір якого зростає експоненційно. Це робить класичні MI-алгоритми непридатними для роботи з великими масивами даних. Для подолання цієї обчислювальної перешкоди ми пропонуємо новий метод навчання — LPSMI (Learning Pretopological Space via Multiple Instances). Метод базується на використанні нижньої оцінки (low estimate) кількості "мішків", що покриваються концептом у процесі його побудови. Це дозволяє радикально скоротити простір пошуку та зробити процес навчання масштабованим і обчислювально ефективним без суттєвої втрати точності моделювання простору.

На першому етапі перевірки ефективності запропонованого підходу ми провели експериментальну валідацію через симуляцію просторових процесів перколяції. Як типовий приклад процесу поширення було обрано симуляцію розповсюдження лісових пожеж. Моделі поширення пожежі були вивчені та відтворені за допомогою нашого претопологічного підходу. Експерименти наочно продемонстрували, що запропонований MI-підхід з оптимізацією LPSMI є винятково ефективним у задачах розпізнавання та прогнозування моделей поширення в динамічних середовищах. Модель успішно виявила складні нелінійні патерни розповсюдження, які недоступні для фіксації базовими статистичними методами.

Головним практичним внеском цієї роботи є застосування розробленої претопологічної моделі до розв'язання реальної задачі NLP — навчання та побудови лексичної таксономії. Ми концептуалізуємо задачу вибудовування ієрархічних словникових зв'язків (наприклад, "транспортний засіб" -> "автомобіль" -> "позашляховик") як комплексну задачу семантичного поширення.

На базі розробленої теорії нами розроблено універсальний фреймворк для тренування моделей побудови лексичних таксономій. Цей фреймворк здатний органічно поєднувати та агрегувати результати різноманітних існуючих методів вилучення відношень, зокрема: статистичних методів (аналіз частоти сумісної зустрічальності слів, розподільна семантика); методів на основі шаблонів (pattern-based), таких як лексико-синтаксичні шаблони Херста (Hearst patterns); методів на основі векторних представлень (embedding-based), що використовують сучасні нейромовні моделі для визначення семантичної близькості в багатовимірному просторі.

Отже, застосування теорії претопології для аналізу текстових даних відкриває нові горизонти у розумінні структурних взаємозв'язків між сутностями. Запропонований метод LPSMI вирішує критичну проблему масштабованості алгоритмів претопологічного аналізу в межах парадигми багатекземплярного навчання. Перехід від парадигми "оцінки ізольованих пар" до "оцінки структурного поширення" дозволив створити потужний гібридний фреймворк. Експериментальні дані підтверджують, що агрегація гетерогенних джерел інформації через претопологічний оператор псевдозамикання суттєво підвищує якість побудови реальних лексичних таксономій, забезпечуючи більш глибоке розуміння семантичної ієрархії природної мови.