

Інформаційні технології

УДК 004.051

Бражник Максим Русланович

студент

Інституту прикладного системного аналізу

Національного технічного університету України

«Київський політехнічний інститут імені Ігоря Сікорського»

ВИКОРИСТАННЯ СЕМАНТИЧНИХ ТА ФОРМАЛІЗОВАНИХ ЗНАНЬ В ІНТЕЛЕКТУАЛЬНІЙ ОБРОБЦІ ДАНИХ

Анотація. У статті проаналізовані підходи використання семантичного Data Mining.

Ключові слова: Семантика, структурована інформація, семантичний веб, K-means, онтологія, семантичні мови, семантичні веб-сервіси.

Актуальність теми. З появою інтернету підхід до технологій кардинально змінився. На сьогодні у всесвітній мережі зберігається велика кількість інформації. Такі масиви даних надзвичайно важко обробляти ручними способами, а з зростаннями ціни на робочу силу, це стає практично неможливо. Зараз набирають великої популярності технології структурування інформації в інтернеті для подальшої машинної обробки. До таких можна перелічити семантичний веб, структуризація за допомогою тегів тощо. Такі підходи до зберігання інформації дозволили застосовувати методи класифікації та кластеризації даних, що зможуть допомогти людині під час роботи або навіть замінити та автоматизувати весь робочий процес.

Дослідження в галузі Семантичної Мережі призвели до досить зрілих стандартів моделювання та структурування знань галузі. Сьогодні онтології

семантичної мережі стають ключовою технологією інтелектуальної обробки знань, забезпечуючи основу для обміну концептуальними моделями домену. Для цього широко використовується Інтернет-мова онтології (OWL), яка виникла фактичним стандартом для визначення семантичних онтологій Інтернету. Таким чином, семантичні веб-технології, які представляють знання про галузь, можуть створити основи для систематичного включення знань про галузь в інтелектуальне середовище видобутку даних.

Перспектива та механізм використання онтологій у семантичному аналізі даних різняться залежно від різних систем та додатків. Питання, чому онтологія корисна для сприяння процесу видобування даних, не має єдиної відповіді. Переглядаючи підходи, засновані на онтологіях, ми узагальнюємо наступні три цілі, для яких онтології були введені в семантичний аналіз даних:

- Для подолання семантичного розриву між даними, програмами, алгоритмами інтелектуального аналізу даних та результатами аналізу даних.
- Надати алгоритми аналізу даних, які керують процесом видобутку, або зменшують простір пошуку.
- Надати формальну інструкцію роботи data mining, від попередньої обробки інформації до кінцевого результату аналізу.

Питання, чому знання галузей є корисним в процесі видобутку даних, давно обговорювалось у попередніх дослідженнях семантичного видобутку даних. Дослідники стверджують, що існує розрив у знаннях між даними, алгоритмом видобутку даних та результатами видобутку на всіх етапах видобутку даних, включаючи попередню обробку, виконання алгоритму та генерацію результатів [3].

Попередня обробка даних зазвичай займається очищенням даних, їх нормалізацією та трансформацією. У більшості сценаріїв існують

семантичні прогалини на етапах попередньої обробки даних. Без урахування формальної семантики для визначення якості даних використовуються спеціальні або емпіричні методи. Наприклад, правила дефіциту та найближчого сусіда зазвичай приймаються для визначення відхилень та відсутніх значень. На етапі нормалізації та трансформації семантика даних необхідна для розуміння відносин даних. Наприклад, важливо визначити кореляцію між ознаками та атрибутами даних при нормалізації даних. Сильно корельовані атрибути можна звести до одного комбінованого атрибута. На практиці семантичні прогалини зазвичай заповнюються експертами доменів вручну. Однак онтології виявилися корисними у багатьох завданнях попередньої обробки даних.

Існує семантичний розрив між алгоритмом аналізу даних та даними. Алгоритми видобутку даних зазвичай розроблені для даних, зібраних з різних доменів та сценаріїв. Однак дані з конкретного домену зазвичай несуть специфічну для домену семантику. Алгоритми загального аналізу даних не мають можливості ідентифікувати та використовувати семантику в різних доменах та додатках. Онтології корисні для визначення семантики домену та можуть зменшити семантичну прогалину, додаючи до даних багату семантику. Семантична анотація спрямована на віднесення основного елемента інформаційних посилань до формальних семантичних описів. Такі елементи повинні складати семантику їх джерела. Семантична анотація має вирішальне значення у здійсненні семантичного видобутку даних шляхом залучення формальної семантики до даних. Анотовані дані дуже зручні для наступних етапів семантичного видобутку даних, оскільки дані просуваються до формального та структурованого формату, який пов'язує онтологічні терміни та відношення.

Багато дослідницьких зусиль присвячено подоланню семантичного розриву між результатами аналізу даних та користувачами. Результати аналізу даних можуть бути представлені онтологіями у

багатофункціональному семантичному форматі, який можна використовувати під час обміну та повторного використання. Наприклад, вилучення інформації (ІЕ) - це завдання автоматичного вилучення структурованої інформації з тексту. Результати аналізу даних / тексту - це набори структурованої інформації та знань щодо домену. За допомогою ОВІЕ (Виділення інформації на основі онтології) інформація, що видобувається, не тільки добре структурована, але й представлена предикатами в онтології, якими легко обмінюватися та використовувати повторно.

Визначення та повторне використання попередніх знань є однією з найважливіших проблем для семантичного аналізу даних. Як формальна специфікація понять та відносин, онтологія допомагає створенню формальної семантики попередніх знань. Зашифровані попередні знання можуть допомагати всім етапами процесу видобування даних і впливати на них, починаючи від попередньої обробки і закінчуючи фільтрацією та поданням результатів. Наприклад, гіперграф RDF був розроблений для збору інформації як з онтологій, так і з даних. Онтології включені в графічне представлення даних як знання для зміщення структури графіка, а також для представлення відстані між термінами та поняттями на графіку. Цей підхід перетворює гіперграф та гіперпрограми у дводольний графік, щоб представити як дані, так і онтологію в уніфікованій структурі.

Як сукупність понять і предикатів, онтологія має здатність виконувати логічні міркування і, таким чином, робити висновок про послідовність для цих предикатів. У семантичному аналізі даних можливість зробити висновок про узгодженість зазвичай представляється як обмеження. Набір обмежень, що забезпечується онтологією, має можливість виявляти суперечливі дані та результати на етапі попередньої обробки, етапі виконання алгоритму та етапі фільтрації та генерації результатів. Онтологія визначає обмеження між кількома завданнями

класифікації. Карлсон представив напівконтрольований алгоритм вилучення інформації, який поєднує в собі навчання багатьох екстракторів інформації [7]. Використовуючи онтологію як обмеження набору екстракторів, це дає більш точні результати.

Кластеризація - це задача інтелектуального аналізу даних, яка групує набір об'єктів в одному кластері, які схожі між собою. Рання робота кластеризації на основі онтології включає використання онтології в задачі кластеризації тексту для попередньої обробки даних, збагачення векторів термінів онтологічними концепціями та сприяння вимірюванню відстані семантикою онтології.

Сонг скористався онтологією на основі тезаурусу для кластеризації тексту із збагаченою концептуальною подібністю. Він запропонував генетичний алгоритм кластеризації тексту з трансформованою прихованою семантичною індексацією з використанням онтології для фіксації асоційованої семантичної подібності.

Література

1. Stumme G. Semantic Web Mining State of the art and future directions / G. Stumme, A. Hotho, 2005. 21 p.
2. Vidal M. Semantic Data Integration of Big Biomedical Data for Supporting of Personalised Medicine / Vidal Maria-Esther. 31 p.
3. Ristoski P. Semantic Web in data mining and knowledge discovery: A comprehensive survey / Ristoski Petar. Mannheim, Germany, 2015.
4. Acharyya S., Ghosh J. Context-sensitive modeling of web-surfing behaviour using concept trees, in: Proceedings of the WebKDD Workshop on Web Mining and Web Usage Analysis, 2003.
5. NASS: News annotation semantic system / Luis Angel. Boca Raton, FL, USA, 2011. 18 p.