УДК 004.94 https://doi.org/10.15407/pp2020.04.022

О. Захарова

ОСНОВНІ АСПЕКТИ СЕМАНТИЧНОГО АНОТУВАННЯ ВЕЛИКИХ ДАНИХ

Семантичні анотації, у силу своєї структурованості – невід'ємне складове ефективного вирішення задач великих даних. Але, сама проблема визначення семантичних анотацій є досить не тривіальною. Ручне анотування є не прийнятним для великих даних з огляду на їх розмір та різнорідність, а також трудомісткість та вартісність самого процесу, задача повністю автоматичного анотування для великих даних поки що не має вирішення. Тобто вирішення задачі семантичного анотування вимагає сучасних змішаних підходів, які б забезпечували вирішення основних задач анотування: виявлення та витягнення сутностей та відношень з контенту будь-якого типу та визначення семантичних анотацій за основі існуючих джерел знань (словників, онтологій, тощо). Отримані анотації повинні бути точними та забезпечувати подальшу можливість вирішення прикладних задач з анотованими даними. Слід зазначити, що контенти великих даних є дуже різноманітними, як наслідок, дуже різняться їх властивості, що підлягають анотуванню. Це вимагає різних метаданих для опису даних та обумовлює наявність великої кількості різних стандартів метаданих для даних різних типів чи форматів представлення. Але, для ефективного вирішення задачі анотування треба мати узагальнену характеристику типів метаданих, в межах якої розглядати їх специфіку. Визначення загальної класифікації метаданих, спільних аспектів та підходів до семантичного анотування контенту великих даних за їх допомогою і є метою даної роботи.

Ключові слова: великі дані, анотування великих даних, класифікація метаданих, семантичні анотації, процес анотування, кероване машинне навчання, некероване машинне навчання, витягнення сутностей, витягнення відношень, домени анотування, онтологічна модель анотування, засоби онтологічного анотування, ручне анотування, автоматичне анотування, напівавтоматичне семантичне анотування, анотатор, аспекти семантичного анотування

Вступ

У поточному стані концентрації даних, існує дивовижний ресурс інформації всіляких типів, що може використовуватись з різними цілями, наприклад, для вивчення музичних інструментів або програмування, та застосунками в різних прикладних галузях. Але, існує інший шар інформації, що доступна та передається через блоги, твіти, статті, журнали. Наприклад, веб, що містить інформацію різних типів та форматів, включаючи текст, рисунки, відео та аудіо, є комунікаційним середовищем, яке дозволяє людині зрозуміти контент і контекст, а також встановити їх відношення/ зв'язати один з іншим засоби масової інформації (ЗМІ). Незважаючи на те, що комп'ютери чудово передають цю інформацію зацікавленим користувачам, системи погано розуміють саму мову представлення інформації. Тому, вирішення задач, що використовують контент великих даних, та задач самих великих даних вимагає структурованого семантичного опису цих даних.

Одним з найбільш розповсюджених підходів до семантизації є визначення семантичних анотацій. Анотації – це назви, атрибути, описи, коментарі, приєднані до документу або частини документу будь-якого формату. Вони надають додаткову інформацію (метадані) про цей, існуючий фрагмент даних. Чіткого, формального визначення терміну «семантична анотація» насьогодні не існує. Відповідно до визначення Ontotex (2016) [1] "семантичне анотування є процесом приєднання додаткової семантичної інформації (метаданих) до різних концептів контенту (сутностей)». Даний тип метаданих забезпечує інформацію як про клас сутності, так і про її екземпляри. Слід зазначити, що семантичні анотації можуть бути застосовані до будь-якого типу контенту: веб-сторінок, звичайних документів, полів у базі даних тощо. Здобуття знань може здійснюватися на основі визначення більш складних залежностей – аналізу відношень між сутностями, опису події чи ситуації тощо. Семантичне анотування забезпечує підтримку розширеного пошуку (на основі концептів), міркування про веб-ресурси та інфор-

маційну візуалізацію на основі онтологій. Окрім цього, анотації використовуються для конвертації синтаксичних структур у структури знань. Іншими словами, семантичне анотування полягає у тому, щоб генерувати специфічні метадані та схеми використання, що забезпечує нові методи доступу до інформації та розширює існуючі.

Семантичні анотації можуть використовуватись для вирішення цілої низки задач великих даних, включаючи інформаційний пошук на основі семантик, категоризацію та композицію документів, перехід від неструктурованого контенту до релевантних знань, візуалізацію інформації на основі онтологій. Якщо документ (чи будь-яка частина деякого контенту, наприклад, відео) є семантично анотованим, він стає джерелом інформації, яку простіше інтерпретувати, комбінувати та повторно використовувати автоматизованим чином. Великим даним притаманна різнорідність джерел та різноманітність форматів їх представлення. Зрозуміло, що міркуючи про набори метаданих, що описують великі дані певних типів, треба враховувати, що кожний з них має власні характеристики. Метою даної роботи є визначити спільні аспекти та підходи до семантичного анотування контенту великих даних за допомогою метаданих.

Домени та моделі семантичного анотування

Можна класифікувати 4 моделі семантичної анотації [2]: теги, атрибути, відношення та онтології. Теги займають нижчий рівень та відповідають найпростішій формі анотування з точки зору користувача; водночас, як онтології находяться на верхньому рівні та відповідають найскладнішій формі з точки зору користувача.

Теги: елемент анотація тегу є призначеним ресурсу ключовим словом або виразом, що описує певну властивість ресурсу. Прикладами тегів можуть бути назви місць, де зроблене фото, імена осіб на фото або тема статті.

Атрибути: елемент анотація атрибуту є парою двох елементів: назва атрибута та його значення. Назва атрибуту означає властивість анотованого ресурсу (наприклад, “Країна”, “день народження”), а значення атрибуту – відповідне значення (наприклад, “Туніс”, “1909”).

Відношення: елемент анотація відношення є парою двох компонент: назва відношення та пов’язаний ресурс. Анотований ресурс зв’язаний з відношенням його назвою. Іншими словами, модель анотації відношення є розширенням моделі анотації атрибуту до домену ресурсу, дозволяючи користувачеві з’єднувати ці ресурси. Наприклад, посилання в одній науковій праці на інший документ є анотуванням відношення, що визначає зв’язок між цими документами.

Онтології: онтологічна модель описує метадані, які співставляють ресурс або його частину з деякими описами його властивостей та характеристик відповідно до формальної концептуальної моделі (онтології). Онтології є корисними для витягнення знань про домен та специфікації загальноприйнятого розуміння домену (що може повторно використовуватися і бути спільним для спільнот та застосунків). Побудова онтології може бути реалізована визначенням концептів, екземплярів концептів, властивостей концептів та екземплярів, обмежень на ці властивості, відношень між концептами та відношень між екземплярами. Користувач, що використовує онтологічну модель анотації, може описати та з’єднати існуючі ресурси шляхом структурування ресурсів (концептів або екземплярів) та визначення обмежень між відношеннями та властивостями.

Використання онтологій як словників для визначення метаданих

Метадані можна класифікувати як залежні та незалежні від контексту [3]. Серед метаданих, що залежать від контексту, можна виділити прямі явні (що згадуються у контенті) та неявні (що виводяться з контексту), та не прямі (зовнішні метадані – посилання на контекст через URL у контенті). Така класифікація (рис. 1) є дуже суттєвою для подальшого визначення процесу анотування.

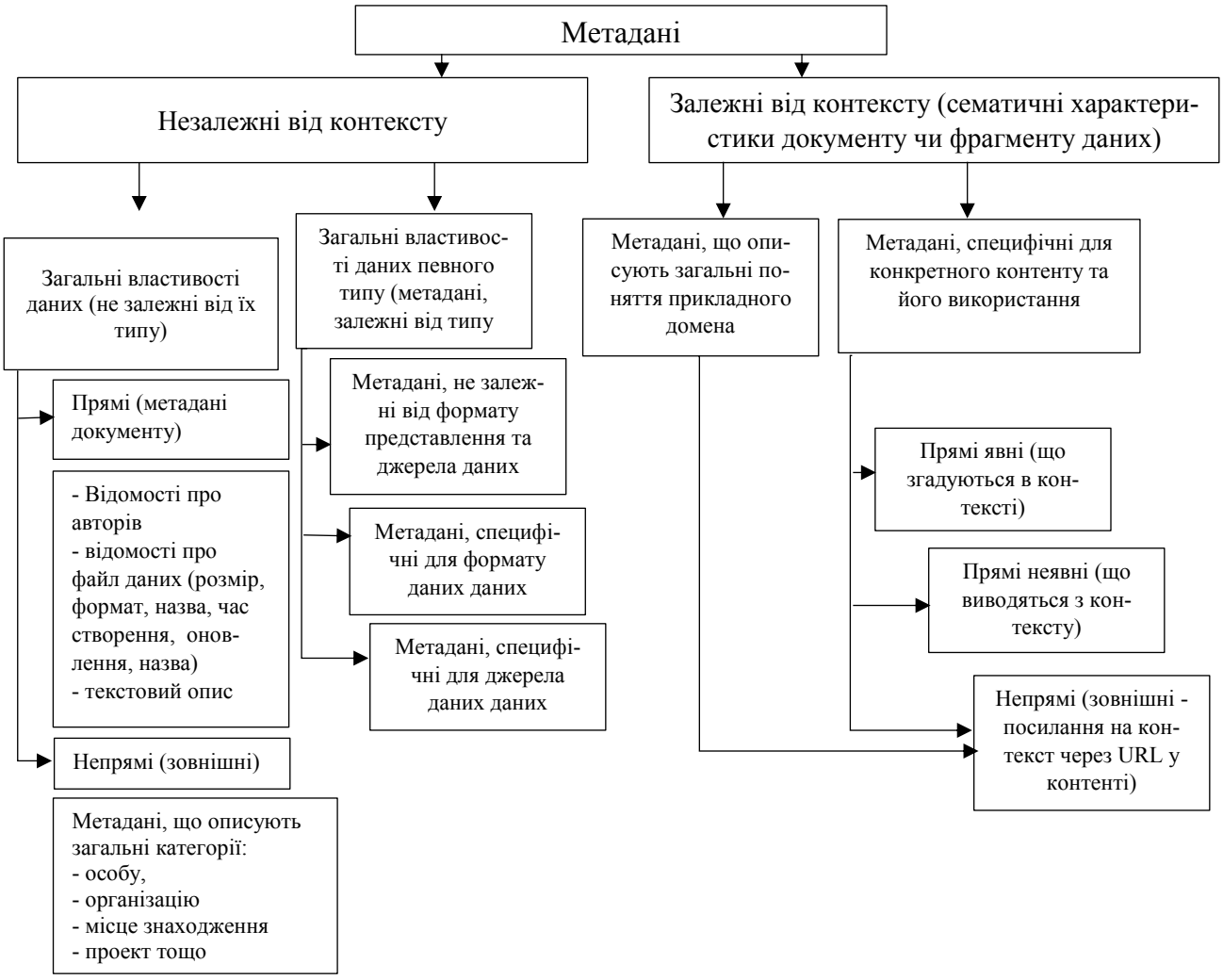

Рис. 1. Класифікація метаданих великих даних

Для створення семантичних метаданих можуть використовуватися стандарти метаданих, тезауриси та контрольовані словники (наприклад, MeSH [4], TGN [5] тощо), а також полегшені або повноцінні онтології. Стандарти, контрольовані словники та тезауруси, на відміну від повноцінної онтології, не є повністю формальними, наприклад, відношення між термінами, які вони включають, не мають явних семантик. Як правило, вони використовуються для забезпечення погодженої термінології у конкретних доменах.

Слід зазначити, що відповідно до наведеної класифікації метаданих можна виділити два головних напрямки анотування великих даних: анотування загальних характеристик документа та анотування прикладного контенту. Метадані для анотування загальних властивостей об'єктів великих даних є контекстно-незалежними і для їх визначення розроблено чимало спеціальних стандартів. Більшість з них орієнтовано на анотування даних певних типів: відео, зображень, аудіо тощо, та враховують їх специфічні характеристики. Також існують стандарти, що дозволяють описувати загальні характеристики документів будь-якого типу. Наприклад, добре відомий стандарт Dublin Core, який широко використовується для визначення характеристик електронних документів. Цей стандарт специфікує множину наперед визначених характеристик документів, таких як автор, дата створення, опис, формат тощо, які можуть бути застосовані для документу будь-якого формату.

Інша група – більш специфічні онтології, що призначені для вирішення

певних задач/підзадач опису даних. Наприклад, полегшена онтологія FOAF (Friend of a Friend) [6], яка має за мету створення анотованої мережі домашніх сторінок людей, груп, компаній тощо. Відповідно, онтологія містить такі концепти як Агент, Особа, Організація, Група, Проект, Документ, Зображення тощо, та де-які базові характеристики, що описують екземпляри цих класів. Подібне призначення мають й онтології OntoWeb [7], KnowledgeWeb [8] та онтологія опису публікацій, які описують осіб, організації, проекти, публікації тощо. Всі перелічені групи онтологій дозволяють описувати певні загальні характеристики документів в цілому, або документів певних форматів або призначення. Вони дозволяють автоматично визначити велику кількість технічних характеристик документа та деяку загальну його семантику, але не охоплюють деталей його контенту (можливо, лише деякі ключові слова та описи природньою мовою в таких характеристиках, як тема або опис). Звісно, це сприяє вирішенню певних задач, але справжній семантичний опис, який викриває сутність та особливості контенту даних, вимагає визначення різних типів контекстно-залежних метаданих, що не можливо без застосування прикладних онтологій домену. Це можуть бути загальні прикладні онтології, що призначені для анотування документів або вирішення проблем у деякому широкому домені. Наприклад, онтології Esperonto Cultural Tour [9] та Fund Finder дозволяють анотувати документи в прикладних доменах культура та фінансування, відповідно. Такі онтології, як правило, визначають загальні характеристики для обраної прикладної області. А можуть бути більш специфічні прикладні онтології, які дозволяють точніше визначити семантику даних, але їх застосування у повністю автоматизованому режимі є проблематичним з огляду на те, що є досить специфічним для конкретного об'єкта даних та задачі, що вирішується.

В цілому всі підходи анотування, що використовують такі додаткові джерела знань, як словники та онтології, відносяться до групи методів, що засновані на онтологіях. А процес створення метаданих з використанням онтологій як словників називається онтологічним анотуванням.

Використання онтологій дозволяє зв'язати фрагменти даних з формальними концептами, що забезпечує підтримку інтероперабельності та гарантує автоматичні міркування на базі структури, яка лежить в основі цих онтологій.

Анотації на основі онтологій, зазвичай, містять три типи інформації [2]:

1) *екземпляри концепта* пов'язують частину документа з одним або декількома концептами в онтології. Так, наприклад, «Інформація про рейс» представляє екземпляр сутності *Рейс* та має ім'я *AA7615_Feb08_2003*, хоча екземпляри концепту не завжди мають ім'я.

2) *значення атрибутів* пов'язують екземпляр концепта з частиною документа, що є значенням одного з його атрибутів. Так, «*Американські авіалінії*» може бути значенням атрибуту *companyName*.

3) *екземпляри відношень* зв'язують два екземпляри концептів конкретного домену. Наприклад, рейс *AA7615_Feb08_ 2003* та місцезнаходження *Мадрид* можуть бути зв'язані відношенням *departurePlace*.

Слід зазначити, що існує два типи відношень між концептами: зв'язки між концептами однієї онтології, або різних онтологій, таким чином, створюючи відображення між джерелами знань. Такі відображення є основою інтеграції множини онтологій, що створюють всебічну базу знань для семантичного опису даних. Визначення зв'язків між поняттями з різних джерел є надзвичайно корисним для взаємодії неоднорідних даних, але досить складним завданням. Системи анотування можуть використовувати спеціальні бібліотеки сервісів, що відповідають за відображення онтологій, що дозволяють використовувати вже опубліковані онтологічні відображення. Так, наприклад, спеціальні веб-сервіси NCBO [10] надають доступ до мільйонів онтологічних відображень, опублікованих у BioPortal. Це дозволяє веб-сервісу NCBO Анотатор

автоматично «тегувати» текст за допомогою термінів з онтологій BioPortal, а вебсервісу NCBO індексування ресурсів надавати доступ до онтологічного індексу публічних онлайн-ресурсів даних. Сукупність таких сервісів, разом зі спеціальними віджетами NCBO, (рис. 2) забезпечує можливість реалізації процесів візуалізації онтологій, анотування та інтеграції даних у галузі біомедицини.

Повноцінні онтології також дозволяють перевіряти обмеження на допустимі значення та їх відношення у відповідних анотаціях. Значення елементів анотацій, зазвичай, є посиланнями на екземпляри в онтології.

Рис. 2. Сервіси NCBO для анотування на основі BioPortal онтологій

Основні аспекти семантичного анотування великих даних

Концепти, що представляють контент будь-якої прикладної області, можна розділити на дві основні категорії: концепти, що представляють процеси, та концепти, які описують фізичні об'єкти, що приймають участь в цих процесах.

До основних категорій анотування відносять сутності (екземпляри фізичних об'єктів), події (екземпляри концептів, що представляють процеси) та відношення між сутностями та/або подіями.

Таким чином, при анотуванні будь-якого контенту, необхідно, перш за все, виділити події та сутності, для яких доцільно створити метадані, визначити ці концепти або їх атрибути, а потім, індексувати їх, класифікувати шляхом визначення зв'язків з онтологією прикладного домену та визначити зв'язки цих сутностей у базі даних семантичного графу.

Як основні аспекти анотування можна виділити [11]:

- *Ідентифікація тексту*. Текст витягується з будь-яких інформаційних джерел, в тому числі не текстових – відео, аудіо, pdf – файли тощо.
- *Розділення тексту на процеси та фізичні сутності.* Розпізнавання іменованих об'єктів.
- *Витягування основних сутностей та їх ідентифікація* (визначення типу сутності, її зв'язку з визначенням прикладного домену, наприклад, URI на об'єкт прикладної онтології). На цьому етапі необхідно знайти та класифікувати елементи тексту за попередньо визначеними категоріями.
- *Класифікація та ідентифікація відношень між визначеними сутностями, аргументів відношень, та визначення їх зв'язків із зовнішніми чи внутрішніми знаннями домену*.
- *Індексація та зберігання семантизованих даних в базі даних семантичного графа*. (Усі розпізнані та збагачені машинно-читаємими метаданими дані зберігаються у базі даних семантичного графа для подальших посилань та використання.)

На рис. 3 показано перелічені аспекти на прикладі анотування тексту, а саме: витягнення основних сутностей (Рим та Римська імперія) з текстового фрагменту, їх ідентифікація, встановлення зв'язків з прикладним доменом, та визначення їх місця у семантичному графі.

Рис. 3. Приклад анотування текстового фрагменту даних

Методології процесу семантичного анотування

Процес анотування має базуватися на наявному загальному теоретичному апараті: методах машинного навчання, статистичного навчання, обробки текстів природньою мовою тощо. Результати анотування великих даних мають бути придатними та уможливлювати вирішення конкретних прикладних задач з цими великими даними, як, наприклад, семантичний пошук чи доступ до даних. Сам процес анотування має вирішувати задачі виявлення та витягування концептів та відношень. Задача анотування полягає в описі визначених сутностей та відношень відповідно до онтології.

У роботі [12] наведена досить вдала загальна трирівнева архітектура процесу семантичного анотування, що визначає його основні задачі (рис. 4).

Анотування може здійснюватися в ручному, автоматичному чи напівавтоматичному режимі [1].

Ручне анотування – це методологія, яка перетворює інформаційні ресурси у взаємопов'язані структури знань шляхом додавання метаданих до деякого рівня документу. Процес ручної анотації є вартісним та трудомістким, і часто не враховує існування різних точок зору на джерела даних, які можуть представлятися різними онтологіями. Насьогодні, з'являється чимало спеціальних засобів для полегшення процесу ручного анотування.

Так, Protégé [13] дозволяє використовувати для анотування обраного фрагменту тексту екземпляри класів онтологій. Інший приклад, визначення еквівалентних об'єктів у різних місцях документу чи документів, шляхом співвідношення цих об'єктів з однією сутністю реального світу. Таке анотування еквівалентностей особливо важливе у медичних застосунках. Такий інструмент, як Semantator [14], дозволяє визначати інший важливий тип анотацій, а саме, дозволяє користувачам обрати два екземпляри та створити відношення між ними, або додати їх до списку кандидатів для побудови відношень. Після чого можна обрати будь-які властивості об'єкта онтології та визначити зміст цього нового відношення.

Очевидно, що, коли мова йде про великі дані, застосування методів ручного анотування стає неможливим, оскільки робота експерта вимагає багато часу та зусиль. Анотація, щоб забезпечити інтенсивні знання, та для розуміння контенту, має бути точною, але вона водночас має бути максимально автоматичною. Тобто, ідеальним рішенням було б автоматичне семантичне анотування, але задача повністю автоматичного створення семантичних анотацій також не має вирішення. Найефективніші семантичні метадані, створені за допомогою інструментів автоматичного анотування чи тегування, будуються на базі різних алгоритмів машинного навчання, які потребують навчальних виборок.

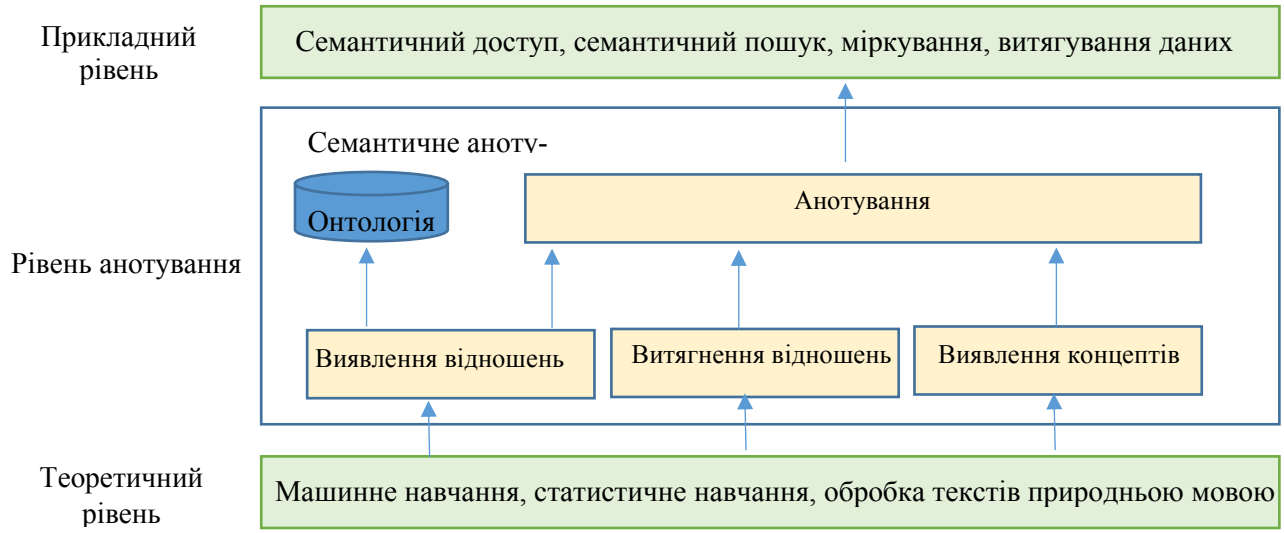

Рис. 4. Загальна архітектура процесу семантичного анотування

Серед методів, що використовуються для автоматичного анотування, можна виділити.

Методи керованого машинного навчання, що складаються з двох етапів: анотування та навчання. На етапі анотування треба визначити сутності та семантичні відношення між цими сутностями для заданого контенту. Задача етапу навчання полягає у вивченні моделі чи моделей, які використовуються на етапі анотування. Для вивчення моделей, вхідні дані часто розглядаються як послідовність деяких одиниць, наприклад, текстовий документ можна розглядати як послідовність слів або рядків тексту. Даний метод потребує розмічених даних.

Метод некерованого машинного навчання намагається створити метадані без розмічених даних. При цьому, для отримання даних з Інтернету можуть бути використані узагальнені зразки.

Слід зазначити, що в силу різноманітності та різнорідності великих даних не можливо досягти ефективного анотування при використанні якоїсь певної однієї категорії методів. Дані з різними характеристиками вимагають різних підходів для вирішення задачі. Так, для веб-сторінок, що побудовані на основі шаблонів та генерують дані з баз даних, можуть бути ефективними методи на основі правил, але вони є не прийнятними для повнотекстових фрагментів даних. Припущення, що документи мають схожу структуру або подібний текст, які досить активно використовуються існуючими підходами машинного навчання, здаються не реалістичними, враховуючи гетерогенну природу веб, та не можуть застосовуватися у випадку великих даних та сучасних типів контентів. Анотування великих даних вимагає використання комбінацій різноманіття методів та підходів для вирішення кожної з задач анотування, та динамічного прийняття рішень щодо використання тих чи інших методів та забезпечення можливості їх використання. Так, у [12] автори наводять де-які методи та моделі, що можуть використовуватися для виявлення та витягнення сутностей та відношень при автоматичному семантичному анотуванні за допомогою підходів керованого машинного навчання, а саме: витягнення сутностей на основі правил, класифікації, послідовного розмічення даних, альтернативних умовних випадкових полів, не лінійні випадкові поля Маркова тощо. Слід зазначити, що кожна з перелічених категорій містить досить широкий спектр методів та моделей (рис. 5). З більш детальним їх описом можна ознайомитись у [12].

Насьогодні, найбільш реалістичними (та використовуваними у сучасних системах) є підходи напівавтоматичного або змішаного анотування, що комбінують процеси ручного та автоматичного анотування. Вони вимагають втручання людини на певних рівнях процесу. Зазвичай, процес, що легко анотується, анотується

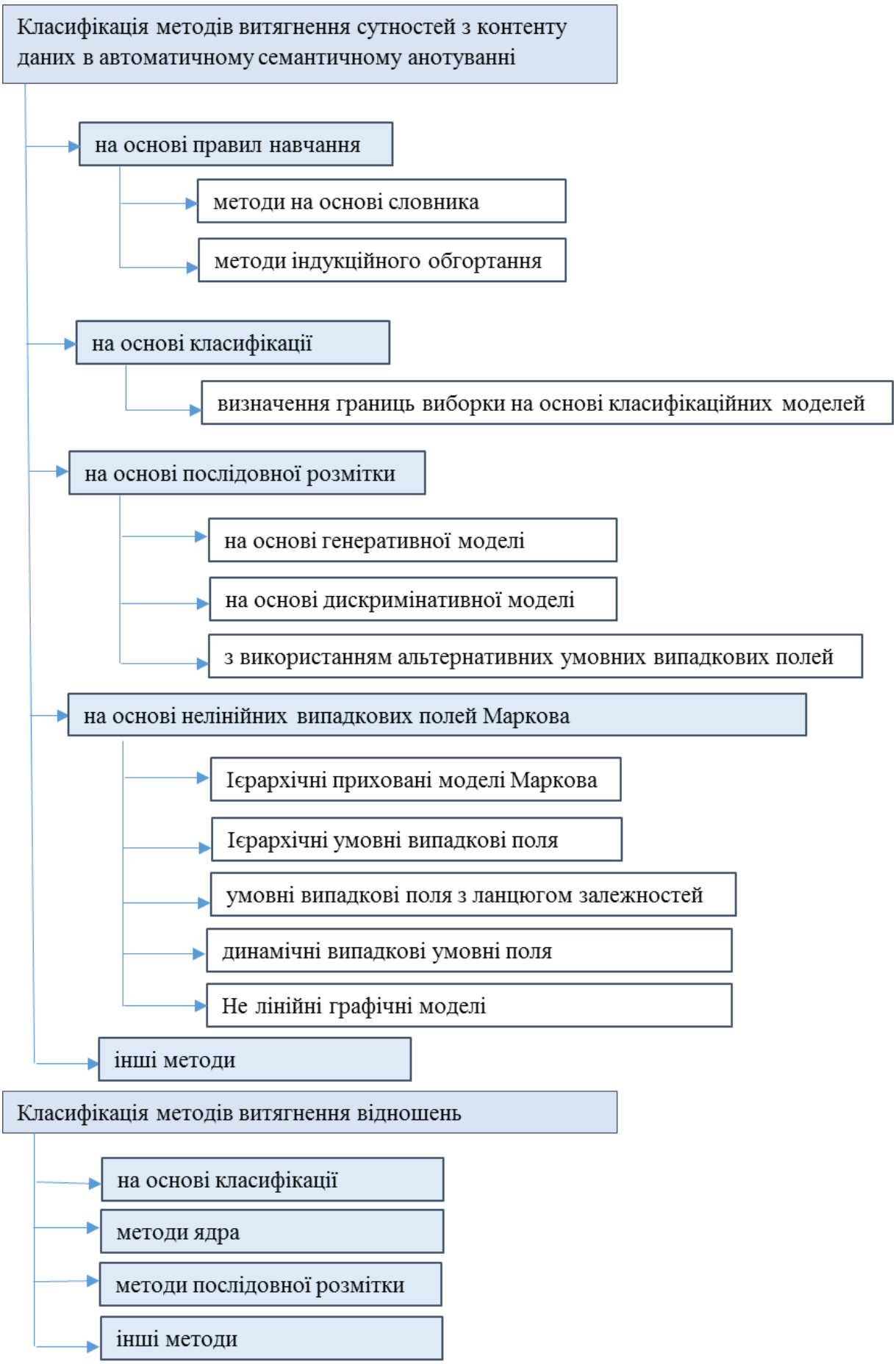

Рис. 5. Категорії методів анотування

автоматично, але існують деякі складні контенти, де втручання людини є необхідним для надання більш змістовної анотації. Ця категорія систем анотування вирізняється архітектурою, методами та засобами витягнення інформації, обсягами ручної праці, продуктивністю, організацією зберігання тощо. Прикладами інструментів напівавтоматичного анотування є GATE [15], NCBO анотатор та cTAKES [16]. NCBO анотатор та cTAKES використовують напівавтоматичне анотування додатково до Semantator.

Фокусом даної роботи є використання онтологій для різних підходів, етапів та категорій процесу анотування, хоча моделі семантичного анотування включають й простіші форми.

Засоби анотування на основі онтологій. Більшість інструментів анотування (анотаторів) на основі онтологій з'явились разом з виникненням семантичного вебу.

Метою їх розробки було забезпечення вставки розмітки на основі онтологій до веб сторінок та подальшої її підтримки. Анотатори, спочатку, були задумані як засоби для полегшення процесу ручного анотування, тобто ручного додавання анотації на веб сторінки. Згодом більшість з них перетворилися на більш повноцінне середовище, яке використовує методи вилучення інформації (IE) та машинного навчання (ML) для забезпечення напівавтоматичного процесу анотування документів. Але засоби, що полегшують процес ручного визначення анотацій, також не втрачають своєї актуальності.

Так, OntoMat-Annotizer [17, 18] та SHOE Knowledge Annotator [19] є прикладами досить широко використовуваних інструментів ручного анотування. OntoMat-Annotizer – розширюваний Java застосунок, який дозволяє створювати OWL анотації. Він включає браузер онтології для дослідження концептів та екземплярів онтології та HTML браузер для відображення документів та їх анотованих частин.

Дозволяє перетягувати частини тексту до анотацій, що створюються. Користувач може визначати екземпляри концептів, з атрибутами та екземплярами зв'язків. OntoMat-Annotizer завантажує OWL онтології.

Анотації, створені за його допомогою, зберігаються в OWL як окремі файли або як вбудовані в анотовані HTML документи та можуть використовуватися широким спектром застосунків.

SHOE Knowledge Annotator призначений для створення ручних анотацій в HTML сторінках за допомогою мови SHOE. Створені анотації можуть посилатися на концепти та відношення однієї або декількох онтологій, що реалізовані в SHOE. Але процес анотування, що забезпечується переліченими засобами, є повністю ручним, і тому використання наведених інструментів для анотування великих даних не є доцільним.

Більш розвиненим є плагін-вкладка редактора онтологій Protege ONTO-H, що дозволяє створювати анотації RTF документів. Даний плагін інтегрований до редактора Protege та має можливість використовувати багато його властивостей, наприклад, браузер онтологій. Окрім функцій ручного анотування (drag&drop), даний редактор забезпечує можливість анотування частин тексту за допомогою розпізнавання іменованих сутностей, анотацій, які вже існують з тими самими іменами або іменами-синонімами і т. і., тобто надає можливість «керованого», а не повністю ручного анотування.

Розширюваний Java-застосунок MnM [20] інтегрує веб-браузер і переглядач онтології та призначений як для ручного, так й для напівавтоматичного та автоматичного анотування. Він має можливість завантажувати онтології, які зберігаються на сервері WebOnto або у файлах, або в URL-адресах на будь-якій мові онтологій: RDF (S), OWL та OCML. Анотації, які створені за допомогою цього інструменту, можна використовувати для заповнення існуючих онтологій або

приєднання до існуючого документа (в XML форматі, де тег-імена – це назви концептів, їх атрибутів та зв'язків).

Для автоматичного анотування MnM використовує механізми витягнення даних для виявлення в документі екземплярів концептів. Ці механізми навчальними на наборах анотованих текстових або html-документів, та готовий, натренований модуль генерує правила для витягнення інформації з інших документів, виявлення екземплярів концептів, значень атрибутів, екземплярів зв'язків. Потім користувачі можуть, за необхідності, відредагувати анотації, що автоматично додані модулем. MnM зберігає екземпляри в різних форматах (OCML, RDF, OWL, XML), а анотації, що ним генеруються, можуть використовуватися в різних середовищах.

Застосунок UBOT AeroSWARM [21], що розроблений як частина UBOT (UML Based Ontology Toolset) проекта, автоматично генерує RDF анотації з текстових документів. Анотатор AeroSWARM доступний в двох версіях: як веб-форма та як окремий застосунок. У веб-версії користувач відсилає текстовий файл, а AeroSWARM повертає RDF анотації для цього тексту, які створюються відповідно до OWL версій OpenCyc [22], SUMO [23] та AeroSWARM. Функція автоматичного анотування AeroSWARM підтримується системою обміну тексту AeroText, яка аналізує тест природньою мовою та витягує з нього елементи, які відповідають онтології, що використовується. Правила витягнення, які використовуються за замовченням системою AeroText, можуть бути зміненими. AeroSWARM генерує екземпляри понять (власні іменники, загальні іменники, кількісні значення валют тощо), значення атрибутів та екземпляри властивостей (наприклад, особа належить організації тощо). Оскільки AeroSWARM забезпечує анотації в RDF форматі, вони можуть використовуватись будь-яким інструментом, що підтримує RDF. Таким чином, AeroSWARM може використовуватися як сервіс автоматичного анотування для забезпечення RDF анотацій в он-лайн режимі.

Висновки

Проведені дослідження були спрямовані на виявлення загальних характеристик як самих великих даних, так і процесів їх семантизації. Це дозволило визначити узагальнену класифікацію метаданих великих даних, що є, насьогодні, найпоширенішим інструментом визначення семантичних описів великих даних, та основні аспекти та категорії процесу анотування для контенту великих даних взагалі, не прив'язуючись до специфіки конкретних типів чи форматів. Онтології є потужним та ефективним засобом семантизації. Тому, при визначені основних аспектів процесу анотування за основу приймаються онтологічні підходи. Ефективність використання онтологій обумовлюється не лише їх розвиненими властивостями семантичного опису прикладного домена, а й можливостями, які надають онтологічні мови та відповідний апарат міркування щодо встановлення подібності сутностей, екземплярів, визначення ступеня їх відповідності, класифікації сутностей та відношень контента відповідно до таксономії онтології тощо. Методологія анотування обов'язково повинна базуватися на існуючому теоретичному апараті, охоплювати вирішення задач анотування контенту та визначення семантичних анотацій, що, в свою чергу, забезпечуватиме розв'язування прикладних задач з даним контентом, як, наприклад, семантичний пошук, витягування даних, міркування тощо. Дослідження існуючих підходів дозволив визначити основні групи методів (теоретичного апарату), що є найбільш ефективними сьогодні, та проаналізувати наявні та використовувані засоби для ручного та автоматичного створення анотацій.

Література

1. https://www.ontotext.com/services/semantic-data-modeling/
2. Thabet Slimani, Taif University, Taif, Saudia Arabia, "Semantic Annotation: The Mainstay of Semantic Web". *International Journal of Computer Applications Technology and*

Research. 2013. Vol. 2. Issue 6. P. 763–770. ISSN: 2319–8656
3. http://oa.upm.es/5638/2/IJMSO_Corcho_FinalVersionPrintedInJournal.pdf
4. http://www.nlm.nih.gov/mesh/meshhome.html
5. http://www.getty.edu/research/tools/vocabulary/tgn/index.html
6. http://www.foaf-project.org/
7. http://www.ontoweb.org/
8. http://knowledgeweb.semanticweb.org/
9. http://www.esperonto.net
10. https://pubmed.ncbi.nlm.nih.gov/23734708/
11. Phesto Enock Mwakyusa. Semantic Annotation and Big Data Techniques for Patent Information Processing. Master's Thesis in Information Technology, October 10, 2017.
12. Tang, Jie, Duo Zhang, Limin Yao, and Yi Li, "Automatic Semantic Annotation Using Machine Learning". IGI Global 1:1. doi: 10.4018/978-1-60566-028-8.ch006, 2009.
13. https://studme.org/235608/informatika/proektirovanie_ontologiy_srede_protege
14. Song D., Chute C.G., Tao C. 2011. Semantator: a semi-automatic semantic annotation tool for clinical narratives. In 10th International SemanticWeb Conference (ISWC2011).
15. Cunningham H., Maynard D., Bontcheva K., Tablan V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL '02), Philadelphia.
16. Savova G.K., Masanz J.J., Ogren P.V., Zheng J., Sohn S., Kipper-Schuler K.C., Chute C.G. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association.* 2010. 17(5). P. 507–513.
17. http://annotation.semanticweb.org/
18. Handschuh S., Staab S. and Maedche A. (2001) 'CREAM – creating relational metadata with a componentbased, ontology-driven annotation framework', in Gil, Y., Musen, M. and Shavlik, J. (Eds.): First International Conference on Knowledge Capture (KCAP'01), ACM Press, Victoria, Canadá, 1-58113-380-4. New York. P. 76–83.
19. http://www.cs.umd.edu/projects/plus/SHOE/KnowledgeAnnotator.html.
20. Vargas-Vera M., Motta E., Domingue J., Lanzoni M., Stutt A. and Ciravegna F. (2002) 'MnM: ontology driven semi-automatic and automatic support for semantic markup', in Gómez-Pérez, A. and Benjamins, V.R. (Eds.): 13th International Conference on Knowledge Engineering and Management (EKAW 2002), Springer Verlag. P. 379–391.
21. Kogut P. and Holmes W. (2001) 'AeroDAML: applying information extraction to genérate daml annotation from web pages', in Handschuh S., Dieng R. and Staab S. (Eds): KCAP'01 Workshop on Semantic Markup and Annotation, Victoria, Canadá.
22. http://www.cyc.com/2003/04/01/cyc
23. http://reliant.teknowledge.com/DAML/SUMO.owl

References

1. https://www.ontotext.com/services/semantic-data-modeling/
2. Thabet Slimani, Taif University, Taif, Saudia Arabia, "Semantic Annotation: The Mainstay of Semantic Web". *International Journal of Computer Applications Technology and Research.* 2013. Vol. 2. Issue 6. P. 763–770. ISSN: 2319–8656
3. http://oa.upm.es/5638/2/IJMSO_Corcho_FinalVersionPrintedInJournal.pdf
4. http://www.nlm.nih.gov/mesh/meshhome.html
5. http://www.getty.edu/research/tools/vocabulary/tgn/index.html
6. http://www.foaf-project.org/
7. http://www.ontoweb.org/
8. http://knowledgeweb.semanticweb.org/
9. http://www.esperonto.net
10. https://pubmed.ncbi.nlm.nih.gov/23734708/
11. Phesto Enock Mwakyusa. Semantic Annotation and Big Data Techniques for Patent Information Processing. Master's Thesis in Information Technology, October 10, 2017.
12. Tang, Jie, Duo Zhang, Limin Yao, and Yi Li, "Automatic Semantic Annotation Using Machine Learning". IGI Global 1:1. doi: 10.4018/978-1-60566-028-8.ch006, 2009.
13. https://studme.org/235608/informatika/proektirovanie_ontologiy_srede_protege
14. Song D., Chute C.G., Tao C. 2011. Semantator: a semi-automatic semantic annotation tool for clinical narratives. In 10th

International SemanticWeb Conference (ISWC2011).
15. Cunningham H., Maynard D., Bontcheva K., Tablan V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL ’02), Philadelphia.
16. Savova G.K., Masanz J.J., Ogren P.V., Zheng J., Sohn S., Kipper-Schuler K.C., Chute C.G. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association.* 2010. 17(5). P. 507–513.
17. http://annotation.semanticweb.org/
18. Handschuh S., Staab S. and Maedche A. (2001) 'CREAM – creating relational metadata with a componentbased, ontology-driven annotation framework', in Gil, Y., Musen, M. and Shavlik, J. (Eds.): First International Conference on Knowledge Capture (KCAP'01), ACM Press, Victoria, Canadá, 1-58113-380-4. New York. P. 76–83.
19. http://www.cs.umd.edu/projects/plus/SHOE/KnowledgeAnnotator.html.
20. Vargas-Vera M., Motta E., Domingue J., Lanzoni M., Stutt A. and Ciravegna F. (2002) 'MnM: ontology driven semi-automatic and automatic support for semantic markup', in Gómez-Pérez, A. and Benjamins, V.R. (Eds.): 13th International Conference on Knowledge Engineering and Management (EKAW 2002), Springer Verlag P. 379–391.
21. Kogut P. and Holmes W. (2001) 'AeroDAML: applying information extraction to genérate daml annotation from web pages', in Handschuh S., Dieng R. and Staab S. (Eds): KCAP'01 Workshop on Semantic Markup and Annotation, Victoria, Canadá.
22. http://www.cyc.com/2003/04/01/cyc
23. http://reliant.teknowledge.com/DAML/SUMO.owl

Одержано 04.11.2020

Про автора:

Захарова Ольга Вікторівна,
кандидат технічних наук,
старший науковий співробітник.
Кількість наукових публікацій в українських виданнях – 29.
http://orcid.org/0000-0002-9579-2973.

Місце роботи автора:

Інститут програмних систем
НАН України,
проспект Академіка Глушкова, 40.
Тел.: 526 5139.

E-mail: ozakharova68@gmail.com.