УДК 004.855.5

АНАЛІЗ МЕТОДІВ АВТОМАТИЧНОЇ ПОБУДОВИ ОНТОЛОГІЙ

Кириленко Ганна

Вінницький національний технічний університет, Україна

Анотація

В роботі розглядаються основні теоретичні аспекти та характерні риси існуючих методів автоматичної побудови онтологій, а також проводиться їх порівняльний аналіз на основі розгляду переваг та недоліків кожного з методів.

Basic theoretical aspects and characteristics of existing methods of automated ontology building are considered in the report. Also comparative analysis of these methods based on reviewing their advantaged and disadvantages is carried out.

Вступ

Зі зростанням потоку інформації, з'явилася необхідність ефективних зручних способів її зберігання, і обробки. Величезний інтерес викликають системи, здатні без участі людини витягти потрібні відомості з тексту. Задача побудови семантичної моделі предметної області відіграє ключову роль у процесі створення інтелектуальної системи. Таку модель прийнято називати онтологією. Онтології є зручним засобом представлення та зберігання знань. Побудова онтологій вручну вимагає знань людини-експерта в конкретній предметній області, а також значних фінансових та часових затрат, тому питання автоматичної їх побудови є надзвичайно актуальним в наш час.

Огляд методів автоматизації

На сьогодні існує декілька основних методів автоматичної побудови онтологій. Серед них: представлення онтологій у вигляді кінцевого автомата, підхід на основі лексико-синтаксичних шаблонів, побудова семантичної карти ресурсу, а також побудова онтологій по колекції текстових документів. Розглянемо детальніше особливості кожного з них.

Автоматне представлення онтологій дозволяє вести операції над ними використовуючи операції на мовах і автоматах, що в свою чергу допоможе автоматизувати процес створення онтологій. При такому підході типи онтологій та їх ієрархій не деталізуються з метою підкреслити спільність операцій, що розглядаються. Алгебраїчні властивості введених операцій на онтологіях випливають з відповідних властивостей операцій алгебри регулярних мов. Це означає, що дані операції задовольняють наступним законам: комутативність і асоціативність операцій об'єднання та перетину, асоціативність множення, дистрибутивність операції множення щодо операцій об'єднання та перетину.

Дану множину операцій (у разі потреби) можна розширювати принаймні мере в двох напрямках. Одним з таких напрямків є розширення операціями на графах (додавання і видалення вершини і ребра, з'єднання графів, ізоморфне з'єднання, декартове множення тощо). Іншим напрямком є алгебра відношень. Оскільки кожна онтологія є поданням деякої сукупності відношень, то можна вводити операції реляційної алгебри.

Який з можливих напрямків буде вибрано, залежить від практичних потреб використання онтологій. Очікується, що представлені операції над онтологіями виявляться корисними при аналізі, синтезі і маніпулюванні онтологіями і онтологічними об'єктами.

Розглянемо тепер деякі проблеми, що виникають на шляху реалізації даних операцій. Перша проблема пов'язана з тим, що коректне виконання описаних вище операцій вимагає створення деякого загального глосарію предметних областей і понять, за допомогою якого можна було б однозначно ідентифікувати відповідні об'єкти. Ця проблема є не тільки проблемою на шляху реалізації введених операцій, але і в деякому сенсі спільною проблемою на шляху побудови онтологій та роботи з онтологіями. Друга проблема, що виникає при реалізації операцій, пов'язана з наявною ієрархією областей і понять. Справа в тому, що в різних онтологіях одні й ті ж поняття й об'єкти можуть перебувати на різних рівнях ієрархії і це необхідно враховувати при застосуванні операцій. У пропонованому підході ця проблема вирішується за допомогою побудови транзитивного замикання відношення досяжності на станах автоматів, що представляють дані онтології. Однак, автори не впевнені в тому, що цього замикання достатньо для вирішення проблеми. Тут, мабуть, необхідні експерименти на реальних онтологіях та їх представленнях. Третя проблема пов'язана з повнотою знань, наявних в представлених онтологіях. Ця проблема є основною в процесі специфікації та верифікації програмного і технічного забезпечення. Тут же ця проблема пов'язана з можливістю побудови повної онтолого-керованої інформаційної системи.

В основному тому графові моделі часто використовуються при роботі з моделями даних (RDF) і досить рідко при роботі з онтологіями [1].

Лексико-синтаксичні шаблони давно використовуються у комп'ютерній лінгвістиці і являють собою характерні висловлювання і конструкції певних елементів мови. Лексико-синтаксичний шаблон – це структурний зразок мовної конструкції, що відображає її лексичні та поверхнево-синтаксичні властивості. Лексико-синтаксичні шаблони дозволяють побудувати семантичну конструкцію, яка відповідає концептуальним змістом одиниці тексту. Для цього використовуються особливості мови, на якій представлений текст. Автоматична обробка тексту вимагає формалізувати в словнику характерні для певної області конструкції. Їх формалізація вимагає визначення множини лексем і граматичних форм, що входять в неї, а також виявлення необхідних синтаксичних умов, наприклад, узгодження граматичних характеристик лексем. Цю інформацію можна задекларувати у вигляді деякої декларативної структури, якою є лексико-граматичний шаблон. Проведені різними авторами дослідження показали, що використання шаблонів на великих корпусах текстів певної тематики дає в результаті досить адекватну таксономію понять даної предметної області [2].

Недоліком даного підходу є його трудомісткість. Перш за все для побудови шаблонів необхідно проводити дослідження сукупності текстів з урахуванням конкретної мови з метою врахування всіх можливих варіантів, складу та граматичних особливостей конструкцій. Побудова онтології по колекції текстових документів заснований на статистичних методах аналізу текстових документів на природній мові. Важливу роль в аналізі та формуванні формалізованого подання текстових документів і в обробці для користувача запитів грають тезауруси. Тезаурус являє собою словник основних понять мови, що позначаються окремими словами чи словосполученнями з певними семантичними зв'язками між ними. Тезаурус може бути загальним для мови, або орієнтованим на якусь предметну область. Зазвичай в тезаурусах підтримуються зв'язки, що визначають синоніми, омоніми, антоніми понять мови, зв'язки виду "ціле-частина "рідвид "використовується для"тощо.

В даний час застосовується два способи створення тезаурусів – ручний і автоматичний. Вручну, як правило створюються універсальні, незалежні від конкретної колекції документів тезауруси. Однак, розробка тезауруса вручну є досить дорогою і трудомісткою справою, що вимагає значних витрат часу. Тому на практиці часто використовують автоматичне створення тезаурусів [3].

Перевагою систем, що використовують тезаурус, є те, що він дозволяє при пошуку за ключовими словами розширювати запит, включаючи в нього синоніми заданих користувачем ключових слів і забезпечуючи тим самим більш повний пошук. Можуть бути ототожнені синоніми в документі і в запиті. Крім того попередня обробка початкових даних в представленому методі значно спрощена. Тезауруси також часто використовуються в процесі ручного або автоматичного індексування документів. Недоліком онтологій побудованих на колекції текстових документів є те, що створення тезаурусів здійснюється зазвичай для заданих колекцій текстових документів. Тому такі тезауруси призначені для роботи саме з цими колекціями.

Побудова семантичної карти ресурсу є одним із методів автоматичної побудови онтологій, що ґрунтується на аналізі текстових даних веб-ресурсу. Семантичною картою називають відображення контенту ресурсу на концептуалізацію його вмісту. Існує багато стандартів опису онтологій для побудови семантичної карти, найпопулярнішим на сьогоднішній день є OWL. Мова OWL дозволяє описувати класи і відносини між ними, властиві для веб-документів і додатків. В основі мови – уявлення дійсності в моделі даних «об'єкт-властивість». OWL придатний для опису не тільки веб-сторінок, але і будь-яких об'єктів дійсності. Мова OWL формалізує область визначення класів і властивості цих класів, визначає індивіди та призначає їх властивості, уточнює ці класи і індивіди до певного ступеня, що визначений формальною семантикою стандарту. Така модель не є предметно-орієнтованою. Для її більш ефективного застосування в конкретній галузі необхідно використовувати в онтологіях лише поняття притаманні даній області [4].

Дана методика дозволяє отримати логічні висновки, факти, які не представлені в онтології буквально, але випливає з її семантики. Це є незаперечною перевагою методу. Проте є певні перепони для його масового впровадження. По-перше – це людський фактор – люди можуть подавати некоректну інформацію, не додавати метаописи, використовувати неправильні метадані. Другим недоліком є надмірне дублювання інформації – її треба представляти у відповідному вигляді і для людини, і для комп'ютера.

Висновки

Проблема автоматичної побудови онтологій сьогодні є дуже актуальною задачею, для якої існує багато методів її вирішення. Представлені методи мають як спільні так і відмінні риси. Головною спільною рисою методів є їх орієнтованість на отримання знань з тексту для подальшого їх зберігання, аналізу та обробки. Відомо, що знання – це динамічна сутність, існуючі знання можуть давати нові. Тому важливим є робота з текстом не як з набором знаків і символів, а як з системою понять, що пов'язані між собою різноманітними зв'язками і відношеннями. Саме розуміння семантики тексту стоїть на першому плані в побудові інтелектуальних систем, а створення онтологій відіграє важливу роль в цьому процесі.

Серед представлених методів доцільно вибрати той, який краще підходить для конкретної задачі. Так, наприклад, автоматний метод відрізняється тим, що операції над графами просто виконувати, проте обмеженість станів автомату є його основним недоліком. Лексико-синтаксичні шаблони добре працюють з великими обсягами вихідного тексту для конкретної області, проте вимагають попереднього аналізу тексту та визначення основних синтаксичних конструкцій. Побудова онтологій на основі колекції текстових документів добре працює лише з цими ж колекціями. А побудова онтологій на основі семантичної карти відрізняється тим, що дає змогу вилучати з текстової інформації не лише знання, що в ній закладені, але й на основі визначених правил виводити нові знання. Даний метод дістав сьогодні велике поширення і значна частина досліджень в області автоматичної побудови онтологій направлена саме на його вдосконалення.

Список використаних джерел:

1. Крывый С.Л., Ходзинский А.Н.Автоматное представление онтологий и операций на онтологиях. Algorithmic and Mathematical Foundation of hte Artificial Intelligance. International Book Series "Information Science and Computing". ITHEA, Sofia, 2008. с. 173-178.
2. Рабчевский Е. А. Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска / Е.А. Рабчевский // Труды XI Всеросс. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – Петрозаводск, 2009.
3. Мозжерина Е. С. Автоматическое построение онтологии по коллекции текстовых документов // Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции – RCDL 2011 – Воронеж, 2011 – С. 293-298.
4. OWL Web Ontology Language Guide. – http://www.w3.org/TR/2004/REC-owl-guide-20040210