УДК 002.53 : [510.63 : 681.3.016]

РОЗВИТОК МЕТОДІВ АВТОМАТИЧНОГО ФОРМУВАННЯ ОНТОЛОГІЙ ТА ЇХ ВИКОРИСТАННЯ В СИСТЕМАХ ШТУЧНОГО ІНТЕЛЕКТУ

С.С. Щербак, В.В. Волкова
(Харківський національний університет радіоелектроніки)

В статті розглянута проблема автоматичного формування онтологій в системах штучного інтелекту та запропоновані шляхи вирішення, які базуються на технологіях Semantic Web.

онтологія, Semantic Web, знання, штучний інтелект, вилучення знань

Вступ. Одною з головних проблем штучного інтелекту (ШІ) є проблема вилучення знань. Проблема збільшується тим, що джерелом знань для систем ШІ потенційно може виступати будь-який об'єкт матеріального світу. Одними з активно досліджуваних джерел знань є структуровані джерела знань (СДЗ), для яких може бути побудована схема даних, тобто описані типи даних і зв'язки між ними, і потенційно існує формальний спосіб одержання нових, відповідних дійсності, корисних закономірностей, що добре розуміються – знань.

Прикладами СДЗ є бази і сховища даних, різного роду користувальницькі документи у форматах MS Word і MS Excel.

Табличні джерела знань та способи формування онтологій. За проведеними дослідженнями [1] в 52% електронних документів для структурування використовувалися таблиці. Широке різноманіття табличних структур складає природну перепону для використання існуючих методів автоматичного вилучення знань з таких джерел.

СДЗ, у яких для структурування використовуються таблиці будемо називати табличними джерелами знань (ТДЗ) [2].

Вилучення знань з ТДЗ пов’язане в першу чергу з проведенням ефективного аналізу ТДЗ, який повинен забезпечувати високий рівень розуміння таблиці, як її структури, так і логічних складових, та маючий механізм адаптації до змін структури ТДЗ.

В останній час, широкий розвиток одержали онтології предметних областей (ПрО), як основа для побудови розподілених баз знань [3].

Існує багато визначень онтології, але в контексті виконаних досліджень ми притримувалися наступного визначення: онтологія – це набір

явних описів властивостей, об’єктів предметних областей та їх екземплярів.

Таким чином, для формування онтології ПрО з ТДЗ необхідно зробити вилучення вмісту ТДЗ – наборів даних, об’єднаних в деяку структуру, проаналізувати цю структуру на предмет ідентифікації об’єктів ПрО і виділення відповідних екземплярів ПрО, а потім представити результати вилучення у вигляді сумісного представлення з онтологією ПрО, наприклад, у форматі представлення знань RDF/RDFS.

У найпростіших формах під формуванням онтологій ПрО розуміється дослідження структури і змісту ТДЗ для того, щоб ідентифікувати заголовну інформацію (атрибути) і встановити відповідність між атрибутами та відповідним їм даним, тобто побудувати структурно-логічну схему ТДЗ [4]. Однак, так буває рідко, тому що найчастіше потрібно визначити взаємозв'язок між атрибутами ТДЗ для встановлення відповідності між атрибутами ТДЗ і компонентами цільової схеми моделі ПрО. У такому випадку, під формуванням онтології розуміється встановлення семантичної відповідності, тобто встановлення зв'язку між кожним елементом структури ТДЗ і об'єктами ПрО в цільовій схемі онтологічної моделі ПрО або встановлення відповідного зв'язку між об'єктами ПрО у вихідній схемі ТДЗ і зв'язку між об'єктами в цільовій схемі моделі ПрО.

Таким чином, під формування онтологій ми розуміємо ідентифікацію будь-яких іменованих і неіменованих наборів даних, атрибутивні характеристики яких відповідають онтологічній моделі обраної ПрО.

У процесі формування онтологій дані з ТДЗ перетворюються до узагальненого представлення ТДЗ(УПТДЗ), що відповідає обраній формальної моделі ПрО, яка використовується в системі Штучного інтелекту [5, 6].

Для розробки математичної моделі УПТДЗ уточнимо поняття ТДЗ:

Визначення 1. ТДЗ – це користувальницькі електронні документи, вміст яких структуровано за допомогою однієї або декількох таблиць.

Беручи до уваги, що 95% форматів електронних документів сумісні до публікації з форматом HTML, використаємо елементи комплексної табличної моделі (КТМ) HTML для розробки узагальненого представлення ТДЗ. Тому, введемо наступне визначення таблиці, як основного елемента ТДЗ.

Визначення 2. Таблиця – елемент ТДЗ, виражена в термінах КТМ HTML.

Тоді, відповідно до визначення 1, ТДЗ як множина таблиць може бути представлене таким чином:

$$T = \{TABLE_{\alpha}\}, \alpha = 1,...,\lambda, \quad (1)$$

де λ – кількість таблиць у ТДЗ.

Кожна таблиця складається з заголовка, нижньої виноски і набору даних (TBody). Заголовок і нижня виноска таблиці повинні містити описову інформацію про стовпці таблиці, а набір даних таблиці повинен містити рядки даних таблиці. Таким чином, таблиця може бути представлена наступною формулою:

$$TABLE = \{THEAD, TFOOT, TBODY_{Generic}\}, \quad (2)$$

де THEAD, TFOOT, $TBODY_{Generic}$ – заголовок, нижня виноска і множина наборів даних таблиці відповідно.

Зауваження 2.1. Згідно КТМ, у рамках однієї таблиці може бути визначено більше одного набору даних таблиці, тоді $TBODY_{Generic}$ може бути представлений наступною формулою:

$$TBODY_{Generic} = \{TBODY_{\beta}\}, \beta = 1,...,\gamma, \quad (3)$$

де β – номер TBODY у таблиці; γ – кількість TBODY у таблиці.

Зауваження 2.2. Зважаючи на те, що $|TBODY_{Generic}| > 1$ може бути тільки у випадках, коли необхідно забезпечити порціонне завантаження вмісту таблиці через Інтернет на клієнтський комп'ютер, – що не є актуальним у контексті проведених досліджень – надалі, якщо це не буде викликати незручності і невизначеності $TBODY_{Generic}$ будемо розглядати, як об'єднання окремих $TBODY_{\beta}$ яке може бути представлено наступною формулою:

$$TBODY_{Generic} = \bigcup_{\beta} TBODY_{\beta}, \ \beta = 1,...,\gamma. \quad (4)$$

Кожний $TBODY_{\beta}$ являє собою множину рядків комірок таблиці, тоді $TBODY_{\beta}$ може бути представлений наступною формулою:

$$TBODY_{\beta} = \{TR_i\}, i = 1,...,n, \quad (5)$$

де i – номер рядка TR в $|TBODY_{Generic}| > 1$; n – кількість TR у TBODY таблиці.

Заголовок таблиці являє собою множину рядків комірок, що визначають структуру набору даних (рис. 1), тоді заголовок може бути представлений наступною формулою

$$THEAD = \{TR_i\}, i = 1,...,n, \quad (6)$$

де i – номер рядка TR в THEAD; n – кількість TR у THEAD таблиці.

Нижня виноска таблиці визначається наступною формулою

$$TFOOT = \{TR_i\}, i = 1,...,n, \quad (7)$$

де i – номер рядка TR в TFOOT; n – кількість TR у TFOOT таблиці.

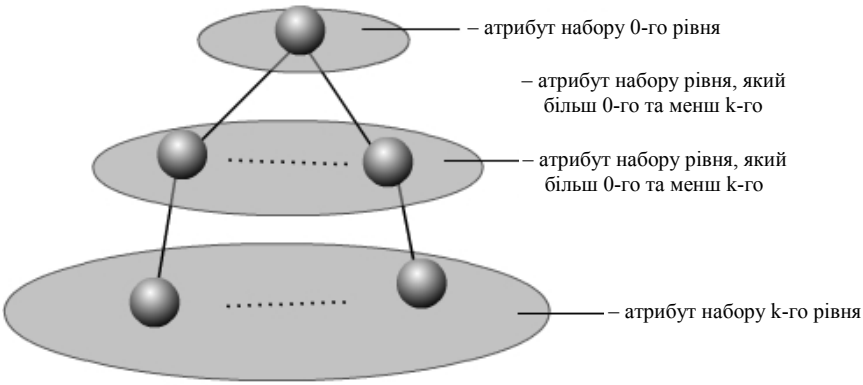

Рис.1. Структура набору даних

Рядок таблиці являє собою множину комірок (рос. “ячеек”) таблиці. Нехай комірка таблиці визначена наступною формулою:

$$\mathrm{Cell}(a,b,c,d,e,f), \tag{8}$$

де a – ім'я комірки; b – вміст комірки; з – кількість стовпців, охоплюваних коміркою; d – кількість рядків, охоплюваних коміркою; e – ідентифікатор комірки, тоді рядок таблиці може бути представлений наступною формулою; f – ідентифікатор комірки, що надає для даної комірки заголовну інформацію

$$TR_i = < id, \mathrm{Cell}_j >,\ j = 1,\ldots,m, \tag{9}$$

де j – номер комірки; m – число комірок в і-му рядку; id – ідентифікатор об’єкту ПрО, якому відповідають дані комірок рядка.

Табличні комірки бувають двох типів: TD – комірка даних і TH – заголовні комірки, тому визначимо імена комірки:

$$A = \{a_1, a_2\} \quad a_1 = "TD",\ a_2 = "TH", \tag{10}$$

де a_1, a_2 – імена комірок.

Таким чином, згідно формул (8) і (9), комірка рядка таблиці $\mathrm{Cell}_j(a,b,c,d,e,f)$ може бути визначена як

$$\mathrm{Cell}_j(a,b,c,d,e,f) = \begin{Bmatrix} TD(b,c,d,e,f), a = a_1 \in A \\ TH(b,c,d,e,f), a = a_2 \in A \end{Bmatrix}, \tag{11}$$

де j – номер комірки; a – ім'я комірки; b – вміст комірки; з – кількість стовпців, охоплюваних коміркою; d – кількість рядків, охоплюваних коміркою; e – ідентифікатор комірки; f – id комірки, що надає для даної комірки заголовну інформацію.

Таким чином, узагальнена модель ТДЗ може бути визначена формулами (1) – (11).

Запропонований підхід до формування онтологій з ТДЗ ґрунтується на використанні УМТДЗ, як основи для аналізу ТДЗ, та базових онтологій ПрО (онтологій верхнього рівня). Такий підхід припускає формування онтології відповідно до використаної базової онтології, що дозволить будувати процедури вилучення нового знання, абстрагуючи від способу і формату збереження даних кожного ТДЗ. У рамках такого підходу, процес формування онтологій зводиться до використання базових онтологій ПрО для ідентифікації заголовної інформації ТДЗ і побудови структурно-логічної схеми ТДЗ (рис. 2) шляхом порівняння з компонентами онтології ПрО, а саме термінологічними вираженнями об'єктів ПрО і їхніми атрибутами. Зважаючи на те, що не завжди можливо встановити пряму відповідність між компонентами онтології ПрО і компонентами ТДЗ необхідно додатково визначити процедури виявлення синонімів і т.п.

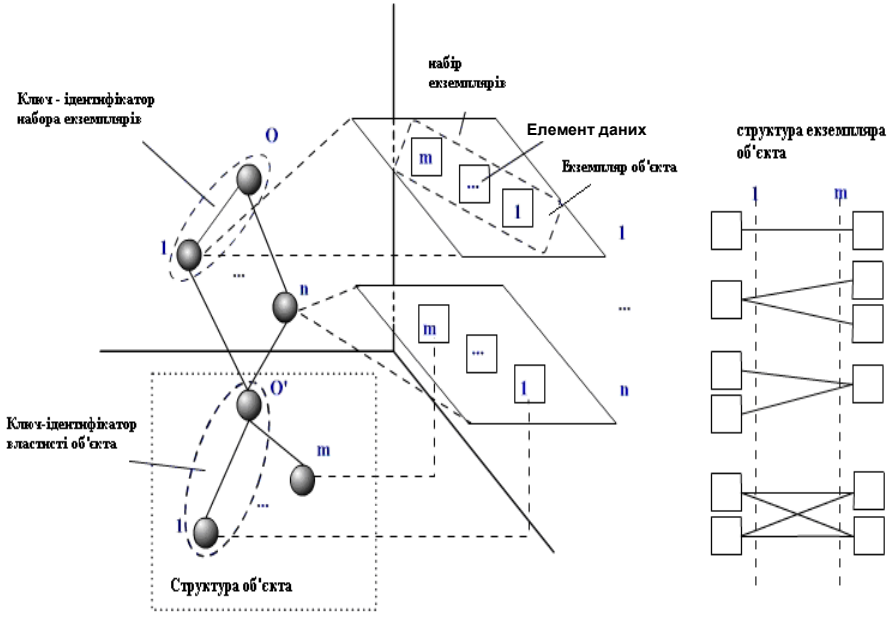

Рис. 2. Структурно-логічна схема ТДЗ

Далі за структурно-логічною схемою кожного ТДЗ, отриманому автоматичним або автоматизованим способом, у залежності від складності використовуваної структури ТДЗ, будується узагальнена модель представлення ТДЗ (УМПТДЗ), що подібно онтології представляється у виді мережі об'єктів і зв'язків між ними. Об'єкти УМПТДЗ характеризуються атрибутами. Кожному атрибутові УМПТДЗ відповідає домен, що крім множини припустимих значень цього атрибута, визначає і формат запису, прийнятий у даному ТДЗ. Після побудови УМПТДЗ можна провести за допомогою експерта її модифікацію, видаляючи частину елементів або звужуючи домени. Це дасть можливість спростити роботу з ТДЗ у випадку, коли необхідно використовувати тільки частину даних, що утримуються в ньому [5].

На наступному етапі отримана УМПТДЗ перевіряється на наявність подібних УМПТДЗ в онтології ПрО, і у випадку відсутності подібних додається в онтологію системи ШІ. Таким чином, забезпечується саморозвиток онтології ПрО. Однак, в онтологію додається тільки логічна компонента нового об’єкту ПрО без урахування типів властивостей об’єктів ПрО. При необхідності автоматичної типізації властивостей об’єктів ПрО додатково потрібно визначити процедури аналізу стовпців набору даних.

Далі, ідентифікуються конкретна реалізація УМПТДЗ, тобто ті компоненти ТДЗ, котрі, по суті, складуть реалізації об'єктів онтології ПрО та формується XML/RDF представлення онтології за допомогою технології XSLT, способом зазначеним в [7].

Взагалі, для формування XML/RDF представлення онтології можна використовувати альтернативні способи за допомогою технології GSS та інших [8].

Висновки. В статті запропоновано підхід до формування онтологій з ТДЗ оснований на використанні узагальненої моделі ТДЗ і базових онтологій ПрО, що збільшує рівень розуміння вмісту ТДЗ, як постачальника знань для систем ШІ, та забезпечує основу для побудови програмних засобів вилучення знань з структурованих за допомогою таблиць джерел.

Основна проблема і недолік запропонованого підходу в тому, що для організації формування онтології з ТДЗ, попередньо повинна бути створена модель ПрО у виді базової онтології ПрО. Але, водночас, таким чином ми забезпечуємо наявність первинних знань про зміст і можливу структуру ТДЗ, що недоступно методам синтаксичного аналізу структур [5] та може бути ефективно використано для автоматичного формування спеціалізованих баз знань.

ЛІТЕРАТУРА

1. *Щербак С.С. Автоматизована обробка табличних документів у задачах менеджменту знань // Східноєвропейський журнал передових технологій. – № 6. – С. 136-139.*
2. *Щербак С.С. Розробка засобів інтеграції інформаційних ресурсів для забезпечення високого рівня інтероперабельності корпоративних інформаційних систем // Матеріали VII Міжнародної науково технічної конференції "Системний аналіз і інформаційні технології". – К.: НТУ України "КПІ". – С. 226.*
3. *Волкова В.В., Щербак С.С. Технологія витягу семантично збагачених даних і відношень як стратегія ефективного використання інтелектуальних систем // MPZIS-2005: Тези доповідей 3-й міжнародної науково-практичної конференції. – Дніпропетровськ: ДНУ. – 2005. – С. 32-33.*
4. *Щербак С.С. Збір і перетворення різнорідних даних з різних джерел до єдиного формату в системах підтримки прийняття рішень // Автоматика-2005: Матеріали 12-й міжнародної конференції по автоматичному керуванню. – Х.: НТУ "ХПІ". – 2005. – Т. 3. – С. 140-141.*
5. *Щербак С.С. Моделі таблиць електронних документів і методи аналізу їхньої структури в парадигмі "Semantic Web" // Збірник матеріалів 8-го міжнародного молодіжного форуму "Радіоелектроніка і молодь у XXI столітті". – Х.: ХНУРЕ. – Ч. 2. – С. 207-207.*
6. *Щербак С.С. Онтологічний підхід до аналізу табличних документів / Зб. наук. праць 10-й між. наук. конференції "Теорія і техніка передачі, прийому й обробки інформації". – С. 48-49.*
7. *Щербак С.С.Інтелектуалізація обробки інформації на основі технологій Semantic Web // Системи обробки інформації. – Х.: ХВУ. – 2004. – Вип. 9 (37). – С. 224-230.*
8. *Рябова Н. В., Щербак С.С. Розвиток технологій Semantic Web: Обробка RDF-графів за допомогою XSLT // Східноєвропейський журнал передових технологій. – № 4. – С. 67-72.*

Надійшла 13.01.2006

Рецензент: доктор технічних наук, професор Є.І. Кучеренко,
Харківський національний університет радіоелектроніки.