УДК 004.89

В.В. Литвин
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

АВТОМАТИЗАЦІЯ ПРОЦЕСУ РОЗВИТКУ БАЗОВОЇ ОНТОЛОГІЇ НА ОСНОВІ АНАЛІЗУ ТЕКСТОВИХ РЕСУРСІВ

Розглянуто автоматизацію побудови онтології на основі аналізу текстових ресурсів. Розвиток онтології починається з деякої базової онтології, яка задає досліджувану предметну область. Розроблено алгоритм розвитку такої онтології.

Ключові слова: онтологія, концепт, відношення, інтерпретація.

In the paper the automation of ontology development on the basis of textual resources analysis is considered. Ontology development starts from a basic ontology which predefines the subject area under research. The algorithm of such ontology is developed.

Keywords: ontology, concept, relation, interpretation.

Постановка проблеми у загальному вигляді

Для задоволення своїх інформаційних потреб кожний інтернет-користувач періодично відвідує сайти професійних співтовариств, підписується і переглядає тематичні розсилання та RSS-повідомлення, шукає у пошукових системах невідомі терміни. Отже, у кожної людини вибудована своя система інтеграції знань у цікавій для нього предметній області. Однак задачі користувачів потребують більшої систематизації та механізму інтеграції розподілених і різнорідних знань у цілісну картину предметної області.

Зазначимо, що оригінальна специфікація WWW розроблялася саме для розв’язування задачі інтеграції наукових матеріалів. Очевидно, що для ефективної інтеграції даних якоїсь предметної області з різних інтернет-джерел необхідно працювати із семантикою веб-ресурсів. У цьому зв'язку актуальним є використання різних технологій Semantic Web [1].

В Інтернеті використовують мови представлення даних, які базуються на XML. Відповідно до проекту Semantic Web для представлення даних консорціум W3 розробив мову RDF (Resource Definition Framework – Середовище Опису Ресурсу). RDF надає можливість запису триплетів, трійок даних – суб'єкт – предикат – об'єкт. Об'єкт і суб'єкт відповідають вузлам графа, а предикат або властивість – напрямленій дузі графа. Дуга спрямована від суб'єкта до об'єкта. Кожний з елементів триплета називають RDF-ресурсом й ідентифікують за допомогою URI ідентифікаторів.

Платформа RDF активно використовується для представлення різних даних, зокрема RSS 3.0, агрегатори новин збирають інформацію у форматі RDF.

Аналіз останніх досліджень та публікацій

Для машинного представлення різних предметних областей в Інтернет і використовуються онтології і словники. Онтологія – специфікація концептуалізації [2] або явний, формальний опис предметної області. Як і в об’єктно-орієнтованому описі, онтологія складається з класів та їхніх екземплярів. Класи та екземпляри володіють властивостями, на властивості можуть накладатися логічні обмеження.

Пошукова система SWOOGLE сьогодні проіндексувала понад 10 тисяч онтологій і словників, доступних у мережі Інтернет. Онтології використовують наукові співтовариства – для опису термінології [3], в електронній комерції – для опису товарів і послуг тощо. Завдяки своїй популярності онтології почали використовуватися як бази знань локальних інтелектуальних систем.

Для опису онтологій створено мови RDFS (RDF Schema – RDF Схема) і OWL (Ontology Web Language – Мова Мережних Онтологій). Як базові елементи ці мови використовують RDF-ресурси. RDFS застосовується для запису словників, а OWL – онтологій. Мережеві онтології надають більші можливості порівняно з RDF словниками, наприклад, логічні операції над класами і логічні обмеження властивостей.

Інтелектуальні системи на основі онтологій показали на практиці свою ефективність [1]. Робота з розроблення онтологій є дуже наукомісткою, вимагає експертних знань у досліджуваній предметній області, значних витрат часу та матеріальних ресурсів, тому актуальним є завдання автоматизації процесу побудови онтології. Для цього пропонується використовувати текстовий зміст масиву ресурсів описового характеру визначеної тематики. Базовою є задача розроблення алгоритму автоматичної побудови семантичної карти ресурсу за допомогою аналізу його текстової інформації. Семантичною картою ресурсу є відображення контенту ресурсу в концептуалізацію його змісту, подану у вигляді OWL онтології.

Формування цілей

Розробити математичну модель розвитку онтології на основі базової онтології предметної області та алгоритми автоматичного наповнення цієї онтології новими поняттями, зв'язками та аксіомами на основі аналізу тестових ресурсів, що описують відповідну предметну область.

Основний матеріал

Для того щоб вручну побудувати повну зв'язану онтологію для певної предметної області (ПО) необхідно затратити достатньо багато часу та ресурсів. Причина цього полягає в тому, що такі онтології повинні містити десятки тисяч елементів, щоб бути придатними для розв'язування широкого кола прикладних задач, які виникають у цих предметних областях (зокрема аналізу наукових текстів заданої предметної області). Тобто ручна побудова онтології людиною-оператором – це довгий рутинний процес, який, до того ж, вимагає ґрунтовних знань предметної області та розуміння принципів побудови онтологій.

Зрозуміло, що повністю автоматизувати процес побудови онтологій неможливо – базові терміни і поняття повинні бути введені людиною-експертом. Однак процес подальшої побудови онтології можна організувати у вигляді навчання на основі текстів заданої предметної області, упорядкованих за зростанням складності опрацювання. Міра складності опрацювання тексту може ґрунтуватись на різних критеріях, наприклад, за кількістю невідомих термінів, які трапляються у тексті, або за допомогою топологічного порядку дерева наукових праць, які посилаються одна на одну.

Огляд відомих підходів та проектів автоматизації побудови онтологій

Сфера автоматизації побудови онтологій викликає дуже великий інтерес, тому не дивно, що у світі є багато напрацювань у цьому напрямку. Розглянемо декілька з них.

«Додаток до індуктивного концептуального аналізу для конструювання онтологій предметних областей» – новий підхід, запропонований Hele-Mai Haav (Institute of Cybernetics at TUT, Таллін, Естонія). Оснований на автоматичному створенні онтологій предметних областей в процесі опрацювання природної мови та формального концептуального аналізу текстових даних.

«Ontosophie» – напівавтоматична система для створення онтологій з текстів, розроблена David Celjuska та Maria Vargas-Vera з Knoweledge Media Institute (Великобританія). Система дає змогу будувати онтологію із неструктурованого тексту, самонавчаючись в процесі побудови онтології і використовувати отримані знання для аналізу нових текстів. Система складена з трьох основних компонентів: Marmot – процесор природної мови, Crystal – інструмент розширення словника, Badger – інструмент для виділення інформації. Частиною процесу є користувач, який підтверджує, відхиляє чи модифікує запропоновані системою концепти, зв'язки і об'єкти. Система показала непогані результати при аналізі великої кількості наукових текстів.

«On-To-Knowledge» – керована контентом система керування знаннями на основі онтологій. Ця система призначена для аналізу і керування вмістом великих інформаційних мереж рівня корпорацій. Такі мережі можуть містити мільйони неструктурованих документів, які розміщені на

фізично віддалених один від одного серверах та мають різні формати. Завдяки використанню онтології, що будується в процесі аналізу та індексації документів у мережі, вдалося побудувати систему пошуку, яка основана на інтелектуальному аналізі вмісту документів, а не на примітивному порівнянні тексту з пошуковим запитом. Система сумісна зі всіма широковживаними форматами документів, мережевими протоколами, підтримує імпорт онтологій з форматів OWL, RDF та OIL.

Christian Blaschke та Alfonso Valencia із Universidad Autónoma de Madrid (Мадрид, Іспанія) розробили метод автоматичної генерації класифікації функції генних продуктів на основі бібліографічної інформації. Отримана класифікація призначена для допомоги експертам під час побудови чи перевірки онтологій людьми-експертами.

Mehrnoush Shamsfard та Ahmad Abdollahzadeh Barforoush розробили систему побудови онтологій на основі аналізу текстів природної мови та невеликої базової онтології, яка містить базові поняття та зв'язки. Особливість цієї системи в тому, що вона аналізує тексти перською мовою.

В IBM T.J. Watson Research Center (Нью-Йорк, США) розробляється метод побудови онтологій на основі наукових запитів із використанням технологій аналізу тексту. Цей метод формує онтологічні концепти і зв'язки на основі аналізу результатів пошуку текстів предметної області у мережі Інтернет. Метод дає змогу не тільки будувати онтології з нуля, але й розширювати вже наявні. Однією із особливостей є персоналізація онтологій за рахунок підтримки користувацьких сесій.

«Artequakt project» – система, розроблена в Univercity of Southampton (Великобританія), яка автоматично будує базу знань про митців на основі аналізу даних з мережі Інтернет. Архітектура системи містить три основні компоненти – виділення знань, керування інформацією та побудова біографічних даних. Для ілюстрації ефективності роботи системи автори пропонують ознайомитися із біографіями всесвітньовідомих художників, побудованими їх системою.

Функціональна модель автоматичної побудови онтологій

Наша ідея, що лежить в основі автоматичної побудови онтологій, полягає в тому, що опрацьовані тексти зі знаннями предметної області використовуються для отримання даних для доповнення наявної онтології. Водночас проміжна онтологія використовується для опрацювання текстів ПО. У результаті отримуємо рекурсивний процес, який можна вважати самонавчанням системи (рис. 1). Процес навчання може бути як автоматичним, так і автоматизованим, із допомогою вчителя. У ході навчання системи необхідність у вчителі зникне і процес стане повністю автоматичним. Початкова онтологія із базовими поняттями предметної області та загальновживаними термінами повинна бути задана апріорі.

Для визначення нових елементів, які можуть бути додані в онтологію, можна використовувати різноманітні методи, переважно основані на евристиках, що враховують наявні в онтології елементи. Використавши різні методи, отримуємо набір можливих модифікацій онтологій, серед яких потрібно вибрати правильні. Вибір здійснюється вчителем або автоматично, на основі попереднього навчання.

Евристики, що беруть участь у визначенні нових елементів, можуть мати вигляд як продукційних правил, так і основуватись на алгоритмах розпізнавання образів, намагаючись доповнити ділянки онтологій пропущеними елементами на основі наявних шаблонів. Щоб далі продовжити дослідження автоматичної побудови онтологій, необхідно проаналізувати алгоритм побудови онтології загалом.

Основні кроки в процесі побудови онтології є доволі очевидними. Переважно онтологія будується згідно з такими етапами:

1) збирання знань про предметну область;

2) об'єднання отриманих інформаційних ресурсів в єдину, узгоджену та достатньо повну систему термінів і понять, які використовуються для опису предметної області;

3) створення базової онтології ПО;

4) розроблення загальної концептуальної структури предметної області. Цей етап передбачає визначення основних концептів предметної області, їх властивостей, зв'язків між концептами, створення абстрактних класів для підтримки наслідування властивостей і зв'язків, посилання чи

включення допоміжних онтологій, віднесення екземплярів за концептами. Цей етап нині практично не піддається автоматизації, всі дії повинна здійснювати людина;

5) збереження отриманої онтології як базової для подальшого розширення;

6) додавання концептів, зв'язків та об'єктів до рівня деталізації, необхідного для забезпечення вимог, які ставлять перед онтологією, щоб використати її для розв'язування задач предметної області;

7) перевірка результатів роботи онтології;

8) перегляд синтаксичних, логічних та семантичних несумісностей між елементами онтологій. Під час цієї перевірки може відбуватися автоматичне виділення додаткових абстрактних концептів на основі зв'язків та властивостей наявних;

9) перевірка онтології експертами предметної області та розгортання її в середовищі, де вона буде використовуватися.

В описаному вище алгоритмі кроки 6–8 можуть бути об'єднані в єдиний процес, що передбачає автоматичне доповнення онтології.

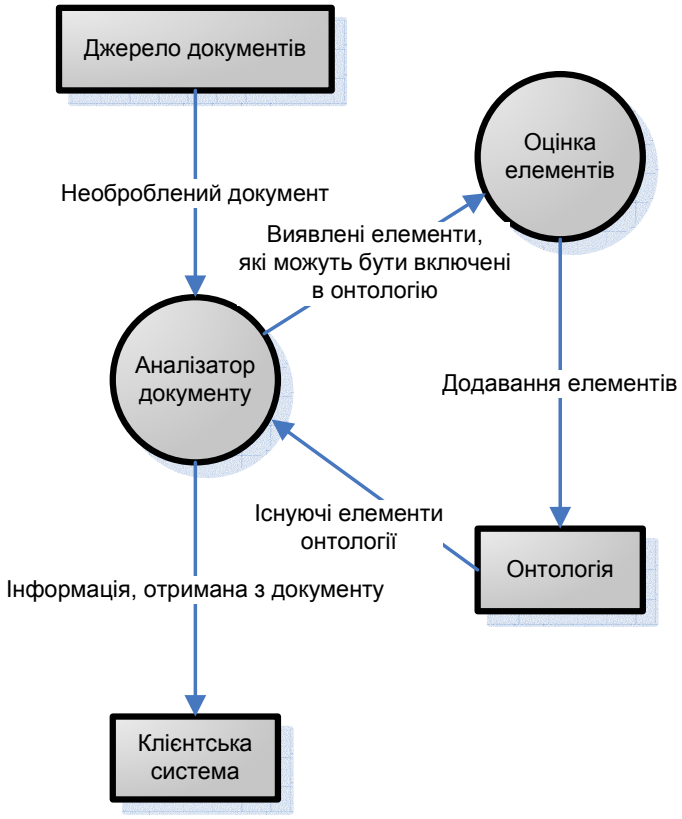

Рис. 1. Діаграма потоків даних при автоматичній побудові онтології в процесі аналізу документів

Математична модель розвитку онтології

Вище вже наголошувалося, що поняття онтології припускає визначення і використання взаємозв'язаної та взаємоузгодженої сукупності трьох компонентів: таксономії термінів, визначень термінів і правил їх опрацювання. Враховуючи це, введемо визначення поняття моделі онтології.

Під *формальною моделлю онтології* O розумітимемо впорядковану трійку такого вигляду:

$$O = \langle C, R, F \rangle,$$

де C – скінченна множина концептів (понять, термінів) предметної області, яку задає онтологія O; R – скінченна множина відношень між концептами (поняттями, термінами) заданої предметної області; F – скінченна множина функцій інтерпретації (аксіоматизація), заданих на концептах чи відношеннях онтології O.

Зазначимо, що природним обмеженням, що накладається на множину C, є його скінченність і непорожність. Інша річ з компонентами F і R у визначенні онтології O. Зрозуміло, що і в цьому випадку F і R мають бути скінченними множинами.

Введемо позначення:

$C = \{C_1, C_2, ..., C_n\}$ – множина концептів (понять, термінів) предметної області;

$R = \{r_1, r_2, ..., r_m\}$ – множина відношень між концептами;

$F = \left\{ f_1^C, f_2^C, ..., f_n^C, f_1^R, ..., f_m^R \right\}$ – множина функцій інтерпретації концептів та відношень.

Позначимо множину властивостей концептів $V = \{v_1, v_2, ..., v_s\}$. Тоді поняття відношення можна записати як відображення із C в C, зважене V:

$$R : C \xrightarrow{V} C$$

Тоді відношення r_i являє собою триплет $r_i = \left\langle C_{i_1}, v_{i_j}, C_{i_2} \right\rangle$.

C_{i_1} – домен (область визначення) відношення r_i.

C_{i_2} – множина значень відношення r_i.

На орграфі, який задає онтологію O, таке відношення задається у вигляді напрямленої дуги від концепту C_{i_1} до концепту C_{i_2}.

Окремо виділимо відношення $\text{IS-A} = \left\langle C_{i_1}, \text{IS-A}, C_{i_2} \right\rangle$, $\text{PART-OF} = \left\langle C_{i_1}, \text{PART-OF}, C_{i_2} \right\rangle$.

Множина класів (концептів) ділиться на дві підмножини $C^1 = \left\{ C_1^1, C_2^1, ..., C_k^1 \right\}$ – множина первинних понять та $C^2 = \left\{ C_1^2, C_2^2, ..., C_{n-k}^2 \right\}$ – множина визначених понять.

Очевидно: $C = C^1 \cup C^2$, $C^1 \cap C^2 = \varnothing$.

На орграфі визначені класи позначатимемо штрихованими вершинами.

Визначений клас C_i^2 означає, що для нього побудований набір аксіом A_i, які інтерпретують це поняття. Тобто можна задати однозначну відповідність

$$C_i^2 \leftrightarrow A_i = \left\{ a_{i_1}, a_{i_2}, ..., a_{i_l} \right\} \quad \text{і} \quad C^1 \leftrightarrow A = \varnothing .$$

Тоді

$$f_j^C = \begin{cases} Comment, & C_j \in C^1 \\ A_j, & C_j \in C^2 \end{cases}$$

З погляду побудови онтології істотно, де міститься концепт в ієрархії понять, а також його місцезнаходження (область визначення, множина значень) у відношенні (напрям стрілки орграфа). Тому визначимо клас

$$C = \left\langle N, R^X, R^Y, S, D, A, Ob \right\rangle,$$

де N – ім'я концепту, R^X – множина відношень, в яких клас C є доменом (областю визначення); R^Y – множина відношень, в яких клас C є множиною значень; S – суперкласи C; D – підкласи C, A – аксіоми визначення C, Ob – екземпляри C.

Нехай задана базова онтологія $O_{base} = \left\langle C_b, R_b, F_b \right\rangle$, яку ми автоматизовано доповнюватимемо. Вважаємо, що базова онтологія є правильною, тобто немає суперечних аксіом і визначені всі необхідні відношення між концептами базової онтології.

Отже, в процесі розвитку базової онтології O_{base} необхідно будувати триплети r_i та нові поняття C, які задаються сімкою величин. Наведемо розроблений нами алгоритм розвитку базової онтології на основі аналізу природно-мовних текстів:

1) із тексту виділяються семантичні одиниці із прив'язкою до відповідних їм елементів у онтології;

2) серед взаємопов'язаних семантичних одиниць виділяють підмножини тих, які утворюють певні семантичні шаблони, котрі можуть утворити нові елементи для онтології. Приклади таких шаблонів наведено на рис. 2;

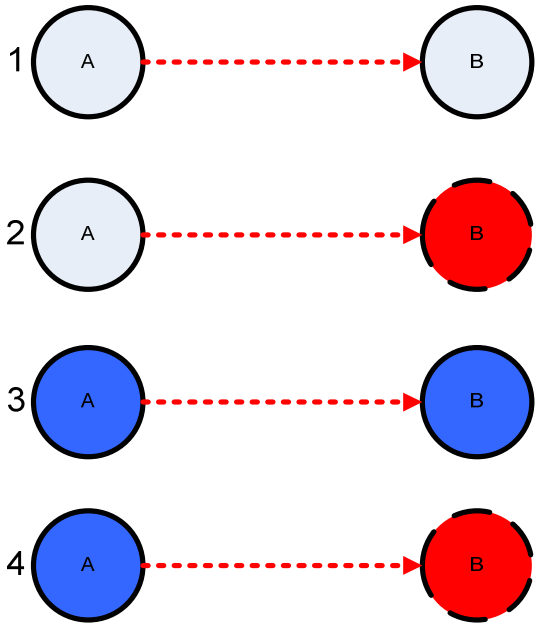

Рис. 2. Семантичні шаблони:
блакитний колір – існуючі елементи онтології,
які належать предметній області, синій – літерали
(існуючі елементи, що не належать предметній області),
червоний – неіснуючі елементи, які можуть бути додані в онтологію

3) семантичні шаблони додаються у масив, по якому після опрацювання текстового документа здійснюється серія проходів. Протягом кожного проходу шаблон розглядається на можливість додавання у онтологію. Якщо такий шаблон дозволений для додавання політикою побудови онтології, то він поміщається в чергу для розгляду адміністратором або автоматично додається залежно від ступеня довіри до типу шаблону, встановленого політикою побудови. Проходи здійснюються доти, доки не перестануть додаватися нові елементи або ж фіксовану кількість разів, встановлену політикою побудови;

4) черга шаблонів являє собою орієнтований ациклічний граф пропозицій вставляння нових елементів в онтологію. Адміністратор розглядає пропозиції з верхнього рівня; якщо пропозицію відхилено, то автоматично відхиляються всі пропозиції з нижніх рівнів, які стали можливими завдяки додаванню скасованої в чергу. Якщо пропозиція прийнята адміністратором, то йому надаються до розгляду наступні пропозиції поточного рівня, якщо таких не залишилося, то відбувається перехід до наступного рівня. Роль адміністратора може виконувати евристичний алгоритм додавання, залежно від політики побудови онтології;

5) будь-які дії в онтології логуються у базі даних, підтримується транзакційність і можливість відхилити зміни, починаючи від певного моменту.

Висновки

Розроблено математичну модель розвитку базової онтології як процес побудови нових концептів предметної області та відношень між ними. Для побудови цих елементів онтології

здійснюється аналіз природно-мовних текстів, які належать до заданої предметної області. У роботі не розглядається процес побудови аксіом визначених концептів та перевірка отриманої онтології на відсутність суперечностей, повноту. Ці задачі є предметом подальших досліджень, однак з деякими підходами для їх розв'язання можна ознайомитись у нашій роботі [1].

1. Досин Д.Г. Інтелектуальні системи, базовані на онтологіях // Д.Г. Досин, В.В. Литвин, Ю.В. Нікольський, В.В. Пасічник. – Львів: Цивілізація, 2009. – 414 с. 2. Даревич Р.Р. Метод автоматичного визначення інформаційної ваги понять в онтології бази знань / Р.Р. Даревич, Д.Г. Досин, В.В. Литвин // Відбір та обробка інформації. – 2005. – Вип. 22(98). – С.105–111. 3. Даревич Р.Р. Застосування інформаційних технологій для координації наукових досліджень // Р.Р. Даревич, Д.Г. Досин, В.В. Литвин, Л.С. Мельничок. – Львів: СПОЛОМ, 2008. – 240 с. 4. Рассел С. Искусственный интеллект / С. Рассел, П. Норвиг. – М., СПб., К.: Вильямс, 2006. – 1408 с. 5. Даревич Р.Р. Оцінка подібності текстових документів на основі визначення інформаційної ваги елементів бази знань / Р.Р. Даревич, Д.Г. Досин, В.В. Литвин, З.Т. Назарчук // Искусственный интеллект. – Донецк. – № 3. – 2006. – С.500–509.

УДК 004.652.4+004.827

О.А. Лозицький, В.В. Пасічник
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

КОМП'ЮТЕРНІ ЗАСОБИ ОСВІТНІХ ПРОЦЕСІВ ДЛЯ ЛЮДЕЙ З ВАДАМИ ЗОРУ. АНАЛІТИЧНИЙ ОГЛЯД

Розглянуто проблематику стану комп'ютерного інформаційного забезпечення людей з вадами зору, наводиться аналіз світових розроблень та досліджень за такою тематикою. Описано різні проекти, націлені на створення віртуальних навчальних середовищ, пристосованих для людей з вадами зору та незрячих. У сучасних електронних навчальних підручниках може використовуватись контент з вставками рисунків, діаграм, формул тощо.

Ключові слова: шрифт Брайля, незрячий, аудіокнига, DAISY, ІТ, аудіоматеріали.

The problems of computer information providing of blind and visually impaired people is examined in this article. The main world developments and researches are shown in this article. This paper describes a few projects aimed to design virtual educative applications specifically for the visually impaired people. DAISY books can be enriched with different media content like images, diagrams and equalizations.

Keywords: Brail, visual impaired, audio book, DAISY, talking book.

1. Постановка проблеми

Соціальна адаптація – це процес взаємодії особи із соціальним середовищем; вона полягає в засвоєнні норм і цінностей оточення у процесі соціалізації.

Адаптація незрячого у студентському колективі зрячих є одним з видів його адаптації до довкілля. Тому незрячий повинен усвідомлювати, що успішно адаптуватися він зможе, якщо набуде необхідних вмінь і навиків не тоді, коли вже опиняється в колективі зрячих, а раніше.

Проблеми навчання і доступу до електронних і друкованих видань інвалідів – незрячих, людей з різними формами порушення зору – потребують невідкладного вирішення.