БАЗИ ДАНИХ, БАЗИ ЗНАНЬ ТА ІНЖЕНЕРІЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

УДК 681.3

Рогушина Ю.В., Гладун А.Я.
Інститут програмних систем НАН України

МЕТОДИКА РОЗРОБКИ ТЕРМІНОЛОГІЇ ІНФОРМАЦІЙНИХ РЕСУРСІВ ЯК БАЗИСУ ФОРМУВАННЯ ОНТОЛОГІЙ ТА ТЕЗАУРУСІВ ДЛЯ СЕМАНТИЧНОГО ПОШУКУ

Сьогодні виникає потреба в засобах та методах створення тезаурусів та онтологій – інтероперабельних моделей подання знань предметної області, що використовуються розподіленими Web-застосуваннями. Щоб коректно визначити відношення між різними термінами тезаурусу, пропонується використовувати елементи онтологічного аналізу та основи мереологічного підходу. Крім того, в роботі пропонуються застосовні правила створення визначень термінів тезаурусу, які дозволяють формувати більш якісні знання про предметну область. Розглянуто використання такого тезаурусу для семантичного пошуку.

Сегодня существует потребность в средствах и методах создания тезаурусов и онтологий – интероперабельных моделей представления знаний предметной области, которые используются распределенными Web-приложениями. Чтобы корректно определить отношения между разными терминами тезауруса, предлагается использовать элементы онтологического анализа и основы мереологического подхода. Кроме того, предлагаются прикладные правила создания определений терминов тезауруса, которые позволяют формировать более качественные знания о предметной области. Рассмотрено использования такого тезауруса для семантического поиска.

Now a great necessity in means and methods for creation of ontology and thesauri – interoperable models of domain knowledge representation that is used in distributed Web-applications – is exists. For correct definition of relations between the different thesaurus terms we propose to use some elements of ontological analysis and basic foundations of mereology. In addition some applied rules of thesaurus terms creation are proposed to build the knowledge with higher quality. Use of this thesaurus for semantic search is considered.

Ключові слова: *програмне забезпечення, онтологія, тезаурус, семантичний пошук.*

Вступ

Сучасні напрями розвитку інформаційних технологій (ІТ) пов'язані зі створенням систем, що базуються на знаннях щодо цікавлячої користувача предметної області (ПрО). У дослідженнях в сфері розподіленого менеджменту знань термін "*онтологія*" застосовують для явної опису системи знань певної ПрО або інформаційного ресурсу. Саме онтології забезпечують спільний словник певної сфери діяльності та визначають (з різними рівнями формалізації) значення термінів і відношення між ними [1]. У найбільш загальному випадку онтологія являє собою угоду про спільне використання понять, що містить засоби представлення предметних знань і домовленості про методи розуміння.

Термін "*онтологія*" запозичений з філософії, де він позначав частину метафізики – навчання про всім сущому, про його найбільш узагальнені філософські категорії (буття, субстанція, причина, дія, явище тощо), зараз він активно застосовується в інформатиці і штучному інтелекті для формального і декларативного визначення концептуалізації певного розділу знань. Часто онтологією називають базу знань

спеціального вигляду, яку можна розділяти, відчужувати і самостійно використовувати. Онтології дозволяють представити поняття так, що вони стають придатними для машинної обробки.

Незалежно від виду онтології, до неї необхідно включити словник термінів і деякі специфікації їхніх значень, що дозволяє обмежити інтерпретацію цих термінів і відбити їхню взаємодію. Слід відмітити, що при такому підході поняття онтології перетинається з уже давно прийнятим в інформатиці і лінгвістиці поняттям *тезауруса*. Онтологія – це база знань, що описує факти, що передбачаються завжди істинними в рамках певного співтовариства на основі загальноприйнятого значення тезауруса.

Онтологія може використовуватися як посередник: між користувачем і інформаційною системою або між членами співтовариства, наприклад, між користувачами деякого корпоративного сховища даних.

Наявність онтології певного інформаційного ресурсу (ІР) дозволяє автоматизувати обробку семантики такого ІР (приміром, шукати в Інтернеті саме ті ІР, що допомагають користувачеві розв'язати певну задачу). Побудова онтології є складною задачею, яка не може виконуватися повністю автоматично, але, використовуючи певні правила та технологічні прийоми, можна полегшити та пришвидшити цей процес.

Онтологічний підхід до подання знань предметної області

Онтологія – угода про спільне використання понять (термінів), що містить засоби подання предметних знань і домовленості про методи логічного виведення (міркувань). Це формалізований опис погляду на світ у конкретній сфері інтересів, що складається з набору термінів і правил використання цих термінів, що обмежують їхнє значення в рамках конкретної ПрО [2].

Знання в онтологіях формалізують, використовуючи класи, відношення, функції, аксіоми та екземпляри. У різних джерелах пропонуються різні формальні моделі представлення онтологій. В усіх цих моделях присутні:

- множина термінів (понять, концептів), що може підрозділятися на множину класів і множину екземплярів;
- множина відношень між поняттями, у якій можуть явно виділятися відношення «клас-підклас», ієрархічні (таксономічні) відношення і відношення синонімії (подоби), а також функції – спеціальний випадок відношень, для яких n-й елемент відношення однозначно визначається n-1 попередніми елементами;
- аксіоми і функції інтерпретації понять і відношень. Формальна модель онтології – це впорядкована трійка [3] $O = \langle T, R, F \rangle$, де

- T – скінчена множина термінів ПрО, що описує онтологія О;
- R – скінчена множина відношень між термінами заданої ПрО;
- F – скінчена множина функцій інтерпретації, заданих на термінах і/чи відношеннях онтології О.

Дана модель носить загальний характер, у той час як на практиці користаються більш точними моделями. Наприклад, у [4] онтологія визначається як структура

$O = \langle C, \leq_c, R, \sigma_R, \leq_R, A, \sigma_A, T \rangle$, що складається з:

- чотирьох множин C, R, A і T, що не перетинаються, а елементи яких називають відповідно ідентифікаторами концептів, ідентифікаторами відношень, ідентифікаторами атрибутів і типами даних;
- структури часткового упорядкування $\leq_c$ над C з верхнім елементом root_C, що називають ієрархією концептів чи таксономією;
- функції $\sigma_R : R \to C^+$, що називають ідентифікатором (сигнатурою) відношення;
- часткового упорядкування $\leq_R$ над R, що називають ієрархією відношень, де $r_1 \leq_R r_2$ означає, що $|\sigma_R(r_1)| = |\sigma_R(r_1)|$ і $\pi_i(\sigma_R(r_1)) \leq_R \pi_i(\sigma_R(r_2))$ для $\forall i: 1 \leq i \leq |\sigma_R(r_1)|$ (тут $\pi_i(t)$ – це i-й компонент кортежу t);
- функції $\sigma_A : A \to C \times T$, що називають ідентифікатором (сигнатурою) атрибута;
- множини типів даних T (наприклад, рядок, ціле).

Для структури $\leq$ виконуються наступні умови:

- рефлексивність;
- антисиметричність;
- транзитивність;
- наявність верхнього елемента;
- супремум.

У [5] онтологія визначається як кортеж $O = \langle C, I, R, T, V, \leq, \perp, \in, = \rangle$, такий, що:

- C – множина класів;
- I – множина екземплярів;
- R – множина відношень;

• T – множина типів даних;

• V – множина значень (множини C, I, R, T, V попарно не перетинаються);

• $\leq$ – відношення на $(C\times C)\cup(R\times R)\cup(T\times T)$, що називається спеціалізацією;

• $\perp$ – відношення на $(C\times C)\cup(R\times R)\cup(T\times T)$, що називається виключенням;

• $\in$ – відношення на $(I\times C)\cup(V\times T)$, що називається реалізацією (створенням екземпляра);

•= – відношення на $I\times P\times\cup(I\cup V)$, що називається присвоюванням.

Така онтологія може бути перетворена в граф, вузли якого є типами.

Семантика – це розділ науки, що описує відношення мовних виражень до об'єктів, що вони позначають і їхній змісту [6]. Сучасна логічна семантика базується на роботах Г.Фреге, Б.Рассела, С.Лесневського, Р.Карнапа, А.Тарського, А.Черча. Семантику онтологічних мов звичайно представляють через теорію моделей. Зокрема, вона визначає функцію інтерпретації, що відображає кожен елемент онтології на множина D, що називають областю інтерпретації.

Інтерпретацією онтології $O=\langle C,I,R,T,V,\leq,\perp,\in,=\rangle$ є пара $\langle I,D\rangle$, у якій D – область інтерпретації, а I – функція інтерпретації, така, що:

- $\forall c\in C,I(C)\subseteq D$;
- $\forall r\in R,I(r)\subseteq D\times(D\cup V)$;
- $\forall i\in I,I(i)\subseteq D$;
- $\forall t\in T,I(t)\subseteq V$;
- $\forall v\in V,I(v)\subseteq V$.

Про твердження, виражене онтологічною мовою, говорять, що воно задовольняється інтерпретацією, якщо інтерпретація погоджується з цим твердженням.

Для онтології $O=\langle C,I,R,T,V,\leq,\perp,\in,=\rangle$ її моделлю є інтерпретація $m=\langle I,D\rangle$, що задовольняє всім твердженням онтології о: $\forall\sigma\in o,m|=\sigma$.

При створенні онтологій найбільшу складність становить формування множини F, тому що цей процес вимагає застосування спеціальних навичок з ПрО інженерії знань і формальної логіки. У той же час стосовно трудомісткості основна робота з формування онтологій припадає на формування множини X, але слід зазначити, що ця робота доступна більшості фахівців довільної ПрО. Складніше визначити множину відношень R, які треба використовувати для моделювання знань. У [7] виділені найбільш загальні онтологічні відношення в реальних доменах – зв'язки еквівалентності, таксономічний, структурний, залежності, топологічний, причинно-наслідковий, функціональний, хронологічний, подоби, умовний і цільовий. Таксономія – це окремий випадок онтології, в якій присутні тільки ієрархічні зв'язки одного типу.

Одним з найпоширеніших відношень в онтології є відношення іменування. Воно є фундаментальним, тобто на його базі може бути побудована формальна система, що дозволяє виражати основні математичні поняття. Існує чотири фундаментальних відношення: приналежності (теорія множин ZF і NF), між функцією, її аргументом і результатом (теорія множин фон Неймана), іменування (онтологія Лесьневського) і "частина-ціле" (мереологія) [8].

Мереологія – це формальна теорія про частини і зв'язані з ними поняття, розроблена С.Лесьневським [9]. Об'єктом мереології є дослідження відношення "частина-ціле". Мереологія – частина тріади дедуктивних теорій, що включає також прототетику й онтологію (у Лесьневського онтологія розглядається тільки як система з єдиним відношенням «є» – "*is_a*"). Відношення "*частина-ціле*" є винятково важливим тому, що воно утворює основу поняття *системи,* яке є центральним у сучасному науковому пізнанні. Система являє собою структурне з'єднання своїх елементів. Її базовою формальною характеристикою є те, що елементи не просто входять у систему, а входять у неї в результаті взаємодії з іншими елементами.

Інші аксіоми мереології описують взаємозв'язки між системою і її елементами. Приклад – аксіома системи і частин: якщо елемент зв'язаний в одну сторону, то він зв'язаний і в іншу. Мереологія виходить за межі вивчення часткових відношень між елементами спільних систем. Вона також займається тими об'єктами, частини яких релевантні цілому. Такі об'єкти ідентифікуються як екземпляри. Серед мереологічних відношень можна виділити сім різних класів, і взагалі, транзитивність не прийнята серед екземплярів різних класів:

1. компонент-об'єкт: сторінка-книга;
2. член-колекція (наприклад, дерево-ліс);
3. частина-маса (наприклад, шматок-хліб);

4. матеріал-об'єкт (наприклад, алюміній-літак);

5. властивість-діяльність (наприклад, бачити-читати);

6. стадія-процес (наприклад, заварювання-готування чаю);

7. місцевість-область (наприклад, Закарпаття-Україна).

Більшість досліджень відношення "частина-ціле" присвячені вивченню частин, але можна також ідентифікувати різні типи цілого відповідно до таких властивостей: 1. чи віддільна частина від цілого (мелодія-пісня або вагон-потяг); 2. чи є частини просторовими або часовими (кімната-квартира або зима-рік); 3. чи відіграє частина певну функціональну роль відносно цілого (двигун-автомобіль); 4. чи є частини неподільними (атом-молекула).

Знання цих теоретичних принципів допомагає більш точно визначити мереологічні відношення, що вводяться до онтології. Визначивши тип відношення за такою класифікацією, можна більш чітко визначити, чи можна використовувати для визначення зв’язків між поняттями одне або різні відношення.

Постановка задачі

Встановивши певні застосовні правила, за якими потрібно створювати описи понять онтологій та тезаурусів, можна підвищити якість онтологій, що створюються, та забезпечити їх більшу інтероперабельність. Щоб коректно визначити відношення між різними термінами тезаурусу, пропонується використовувати основи мереологічного підходу для більш чіткого вибору відношень типу “частина-ціле”.

Формування тезауруса предметної області, що цікавить користувача

Окремим випадком онтології, який простіше формувати та обробляти, є *тезаурус* – повний систематизований набір даних про будь-яку область знань, що дає змогу людині чи комп’ютеру в ній орієнтуватися [10]. Можна досліджувати як тезауруси окремих фахівців, так і тезауруси ПрО знань.

Формальна модель тезауруса – $Ts = \langle T, R \rangle$, де:

- T – скінчена множина термінів,
- R – скінчена множина відношень між цими термінами.

Тезаурус можна розглядати як семантичну мережу, у вузлах якої знаходяться терміни, пов'язані відношеннями з обмеженого набору R. Основними технологічними фазами створення тезауруса, докладніше розглянутими в [11], є:

- виділення лексичних одиниць, тобто формування словника (глосарія) T;
- розробка набору семантичних зв'язків;
- актуалізація зв'язків – установлення зв'язків між термінами.

При цьому дуже важливо сформулювати принципи, за якими буде здійснюватися кожна процедура. Для першого пункту визначальними є два аспекти – джерело лексичних одиниць та критерій їх добування. При розробці набору семантичних відношень можна знаходити їх у тексті, що описує дану ПрО (намагатися вичленувати й уніфікувати ті відношення, що існують в текстах між термінами) або безпосередньо аналізувати знання. На практиці звичайно використовують поєднання обох методик. Для актуалізації семантичних зв'язків між термінами тезауруса можна використовувати знання експертів, а також документи, призначені як для фіксації структури знань (словники, класифікатори тощо), так і самі знання, що відображають ПрО (реферати, статті, монографії тощо).

Термін – це слово або словесний комплекс, що співвідноситься з поняттям певної організованої сфери пізнання (науки, техніки) та вступає в системні відношення з іншими словами і словесними комплексами і утворює разом з ними в будь-якому окремому випадку чи у певний час замкнуту систему, що відрізняється високою інформативністю, однозначністю, точністю й експресивною нейтральністю.

Для створення тезауруса можна скористатися методологією розробки онтологічних моделей – стандарт IDEF5 [12] сімейства IDEF, згідно з якою побудова тезауруса ПрО складається з п'яти основних дій:

- *вивчення і систематизація початкових умов* – мети і контексту розробки тезауруса, визначення меж ПрО, яка цікавить користувача;
- *збирання і накопичення даних* – підбір ІР, що відносяться до цієї ПрО;
- *аналіз даних* – вивчення відібраних ІР, формування словника термінів ПрО, що містяться у відібраних ІР;
- *початкова розробка тезауруса* – встановлення зв'язків між термінами ПрО [12], з якої потім витягуються базові терміни ПрО;
- *уточнення і затвердження тезауруса* – аналіз користувачем отриманого тезауруса та його коректування.

Під час формування тезауруса доцільно враховувати наступні рекомендації, які стосуються побудови визначень даних і метаданих та враховують вимоги, розроблені

підкомітетом зі стандартизації ПК-6 „Телекомунікації та обмін інформацією між системами” з урахуванням ISO/IEC 11179.

Практичні рекомендації стосовно визначення термінів тезаурусу

Визначення термінів тезауруса може здійснюватися в автоматичному режимі (шляхом аналізу повнотекстових документів та інших інформаційних джерел), шляхом вилучення з інших баз знань (тезаурусів, онтологій тощо) або надаватися безпосередньо експертом ПрО.

Формальні вимоги до визначень термінів тезауруса:

a) визначення має бути **викладене в однині.** Виняток становлять поняття, які самі є множинними. Наприклад, *“номер статті”:* добре визначення – "номер посилання, що ідентифікує статтю"; погане визначення – "номер посилання для ідентифікації статей". У поганому визначенні використовується слово “статті”, що може бути формою множини і це можна зрозуміти так, ніби один номер може посилатися на кілька статей.

b) визначення повинне пояснювати**, чим є** наведене поняття**, а не тільки чим воно не є.** Наприклад, “розмір вартості фрахтування”: добре визначення – "розмір витрат, які несе вантажовідправник для переміщення товарів з одного місця до іншого"; погане визначення – "розмір витрат, що не належать до витрат на пакування, документальне оформлення, завантаження, розвантаження та страхування". У поганому прикладі не вказано, що входить до поняття елемента даних.

c) визначення повинне **мати вигляд описової фрази або речення.** Речення необхідне для формування точного визначення, яке містить важливі характеристики поняття. Просте наведення одного або кількох синонімів не є достатнім. Наприклад, “ім’я агента”: добре визначення – "назва сторони, яка уповноважена діяти від імені іншої сторони"; погане визначення – "представник ". “Представник” є синонімом імені елемента даних, який не може бути адекватним визначенням.

d) визначення повинне **містити лише широко відомі скорочення.** Розуміння значення скорочення, зокрема абревіатур та ініціалів, зазвичай обмежується певним середовищем. В іншому середовищі ті ж самі скорочення можуть викликати неправильне розуміння або непорозуміння. Таким чином, для запобігання неоднозначності, у визначеннях використовуються тільки повні слова без скорочень. Наприклад, “прилад для вимірювання щільності”: добре визначення – "прилад, який використовується для вимірювання концентрації рідини, в одиницях виміру маси до одиниці об’єму (м.д.о.) (тобто фунтів на кубічний фут; кілограмів на кубічний метр)"; погане визначення "прилад, який використовується для вимірювання концентрації рідини в термінах м.д.о. (тобто фунтів на кубічний фут; кілограмів на кубічний метр)". Проте м.д.о не є загальновідомим скороченням і його значення може бути незрозумілим для деяких користувачів. Скорочення має бути наведене повними словами.

e) визначення повинне **бути викладене без використання визначень інших даних або базових понять.** Визначення термінів має наводитись у відповідному глосарії. Якщо потрібне інше визначення, воно має додаватись як примітка після тексту первинного визначення або як окремий запис у словнику. Пов’язані визначення можна отримати за допомогою атрибутів посилання (перехресних посилань). Наприклад: “код типу зразка”: добре визначення – "код, який ідентифікує тип зразка"; погане визначення – "код, який ідентифікує тип обраного зразка. Зразок – це мала частка, вилучена для проведення експериментів. Він може бути як єдиним зразком для тестування, так і сурогатним зразком для контролю якості. Зразок для контролю якості – це сурогатний зразок, обраний для перевірки результатів тестування єдиних зразків". Погане визначення містить два додаткових визначення –“зразка” та “зразка для контролю якості”.

Семантичні вимоги до визначень термінів тезауруса:

a) визначення має **відображати суттєвий зміст поняття.** Усі первинні характеристики поняття, мають бути відображені у визначенні з відповідним рівнем специфічності залежно від контексту. При цьому необхідно запобігати пояснення неважливих параметрів. Рівень деталізації залежить від потреб користувача системи та середовища.

Наприклад, “номер послідовності завантаження вантажу” (визначений контекст: будь-яка форма транспортування): добре визначення – «номер, що вказує на послідовність, в якій здійснюється завантаження до транспортного засобу або елемента транспортного середовища»; погане визначення – “номер, який відображає послідовність, в якій здійснюється завантаження до вантажівки” (у визначеному контексті вантажі можуть транспортуватись різними транспортними засобами:, вантажівками, кораблями, вантажними потягами і не обмежене лише вантажівками).

Інший приклад: “сума за рахунком-фактурою”: добре визначення – “загальна сума, яку потрібно сплатити за рахунком-фактурою“; погане визначення – “загальна сума вартості всіх елементів, зазначених в рахунку-фактурі, включаючи усі відрахування, зокрема знижки та дисконти, та додаткові платежі, зокрема страхові, транспортні та накладні витрати тощо“. У поганому визначенні міститься зайва інформація.

b) визначення має бути **точним та однозначним.** Визначення має бути достатньо зрозумілим, щоб забезпечити його однозначну інтерпретацію. Наприклад, “дата отримання вантажу”: добре визначення – «дата, на яку вантаж передається отримувачу»; погане визначення – «дата, на яку здійснюється доставка вантажу». У поганому визначенні не роз’яснюється, що таке “доставка”. Під “доставкою” можна зрозуміти як момент розвантаження товару в певному місті, так і факт передачі товару кінцевому отримувачу. Не виключено, що кінцевий отримувач ніколи не отримає вантаж або його передача може здійснитися через кілька днів після розвантаження.

c) визначення має **бути коротким.** Слід запобігати використання додаткових фраз описового характеру, подібних до “для забезпечення використання цього реєстру метаданих”, “терміни, що мають бути описані”. Наприклад, “ім’я набору символів”: – добре визначення “ім’я, що присвоюється набору фонетичних або ідеографічних символів, в які зашифровані дані”; погане визначення – “ім’я, що присвоюється набору фонетичних або ідеографічних символів, в яких зашифровані дані для забезпечення використання цього реєстру метаданих або, якщо говорити про загальний вжиток, спроможність системного обладнання і програмного забезпечення обробляти дані, зашифровані одним або декількома шифрами”. У поганому визначенні всі фрази після виразу “… в яких зашифровані дані” є зайвими.

d) визначення повинне **мати можливість використовуватися окремо.** Зміст поняття має бути наочним у визначенні. Для розуміння поняття не повинні бути потрібні додаткові роз’яснення. Наприклад, “назва міста розміщення школи”: добре визначення – “назва міста, де знаходиться школа”; погане визначення – “див. сайт школи”. Погане визначення не є самостійним, оскільки необхідно звернутися до додаткового джерела.

e) визначення повинне **бути поданим без використання пояснювальної інформації, функціонального використання або процедурної інформації.** Пояснення не слід включати до визначень, тому що вони містять зайву інформацію. У разі потреби такі пояснення можуть бути розміщені в інших атрибутах метаданих. Припустимо додати кілька прикладів після визначення. Наприклад, “мітка поля даних”: добре визначення – “ідентифікація поля в індексі, тезаурусі, базі даних тощо”; погане визначення – “ідентифікація поля в індексі, тезаурусі, базі даних тощо, яка застосовується для таких елементів інформації, як примітки, колонки в таблицях”. У поганому визначенні містяться примітки, що стосуються функціонального використання. Якщо інформація, що починається зі слів “яка застосовується…” є необхідною, то вона має бути розміщена в іншому атрибуті.

f) визначення повинне **запобігати циклічних посилань**

Два поняття не повинні бути розкриті одне через одне. Визначення одного поняття не може використовувати інше поняття як своє визначення, тому що це може призвести до ситуації, коли поняття визначається через інше поняття, яке, у свою чергу, визначається через перше поняття. Наприклад, два елементи даних з поганими визначеннями – “ідентифікаційний номер працівника – номер, що призначається працівнику; “працівник – людина, яка має відповідний ідентифікаційний номер працівника”. Визначення посилаються одне на одне, але в жодному з них не наведено зміст поняття.

g) визначення повинні **використовувати однакову термінологію та логічну структуру для пов’язаних визначень**

Для близьких або пов’язаних визначень має використовуватись одна й та ж сама термінологія та синтаксис. Наприклад, “дата відправлення товарів – дата, в яку товари були відправлені даній стороні”, “дата отримання товарів – дата, в яку товари були отримані даною стороною”. Використання єдиної термінології значно спрощує розуміння.

Мови подання тезаурусів

Для інтероперабельного представлення тезаурусів, що забезпечує їх повторне використання в різних застосуваннях, доцільно використовувати створені на сьогодні мови представлення онтологій.

Мови подання онтологій, які існують на цей момент, можуть бути класифіковані як мови на основі XML або не-XML-орієнтовані. До мов, які основані на XML, відносяться такі мови, як SHOE, XOL, OIL, RDF, DAML+OIL, OWL [13] тощо. До не-XML мов відносяться Cycl, Ontolingua, уніфікована мова моделювання UML.

XML-орієнтовані мови отримали більш широкий розвиток для Semantic Web, оскільки XML найбільш добре підходить для рішення питання інтероперабельності, що є одним з основних вимог Semantic Web. RDF(S) є найбільш простою та найбільш масштабованою XML-мовою подання онтологій. Вона базується на обчисленні предикатів і використовує для визначення семантики трійки суб'єкт, предикат і об'єкт. Однак, вона не достатньо виразна і може використовуватися тільки для вказівки концептів та бінарних відношень між ними.

Мова OWL розширює RDF(S) і забезпечує конструкції для вираження понять, відношень, потужності, анотацій та конкретизації понять і т.д. Окрім того, OWL підтримується великою кількістю редакторів онтологій і вирішувачів. Все це дозволяє називати OWL найбільш підходящою мовою для подання онтологічних знань.

Дескриптивні логіки (DL – description logics) – сімейство мов представлення знань, що дозволяють описувати поняття предметної області в недвозначному, формалізованому виді.

Вони виникли як розширення фреймів і семантичних мереж механізмами формальної логіки. Зараз DL використовуються в Semantic Web для побудови онтологій.

Для того, щоб задати яку-небудь DL, необхідно задати її синтаксис і семантику. Синтаксис описує, які вирази (концепти, ролі, аксіоми і т.п.) вважаються правильно побудованими в даній логіці. Семантика вказує, як інтерпретувати ці вирази, тобто додає їм формальний зміст [14].

DL – це сімейство мов представлення знань, що дозволяють описувати поняття предметної області у формалізованому виді. Кожна логіка DL є логікою першого порядку, але не навпаки.

DL визначають формальну мову для понять і відношень (ролей) разом з теорією доказу. DL є мовою виразу тверджень про факти (про їхню істинність) і запитів до них, включаючи виконуваність (satisfiability) і включення (subsumption).

Формальна семантика DL може використовуватися для того, щоб автоматично виконувати доказу на основі БЗ. Такі автоматичні докази дозволяють знаходити відповіді на такі запити, як [15]:

- виконуваність понять (satisfiability) – чи може існувати деяке поняття C;
- включення (subsumption) – чи є деяке поняття C варіантом поняття D;
- погодженість (consistency) – чи є вся БЗ погодженою;
- перевірка екземпляра (instance checking) – чи є деяке твердження правдивим.

ALC (Attributive Language with Complements) – це підмножина DL, на якому базується OWL, причому для дуже багатьох реальних онтологій цілком достатньо ALC. Логіка ALC досить проста, але містить багаті ключові можливості OWL.

Синтаксис ALC містить алфавіт (набір базових символів), що складається з трьох компонентів:

- набір імен базових класів (NC) і два спеціальних класи (універсальний клас top і порожній клас bottom);
- набір імен властивостей (NR);
- набір імен об'єктів (NI).

ALC дозволяє описувати складні поняття за допомогою наступних конструкторів класів, де C і D – довільні класи:

- перетинання класів $C \cap D$;
- об'єднання класів $C \cup D$;
- доповнення класу $\neg C$;
- універсальне обмеження властивості $\forall R.C$;
- екзистенціальне обмеження властивості $\exists R.C$.

Дескриптивну логіка можна використовувати для управління онтологіями, відображаючи поняття і відносини онтології у твердження логіки DL. Тоді для виконання доказів треба використовувати екземпляри понять, що представлені в онтології. DL застосовується при створенні БЗ, доказу її логічної погодженості й одержання відповідей на запити. Доказ виконуваності БЗ (її логічної несуперечності) обґрунтовує правильність онтології.

Використання онтологій та тезаурусів у семантичному пошуку

Розглянута вище методи побудови тезаурусу дозволяють користувачеві більш коректно формувати тезаурус, що характеризує його поточні інформаційні потреби. Такий тезаурус може стати основою бази знань, яка використовується для персоніфікації семантичного пошуку. Для цього за природномовними документами, які користувач вважає релевантними цікавлячій його ПрО, будуються їх тезауруси та поєднуються в тезаурус ПрО. Потім цей тезаурус співставляється з тезаурусами знайдених нових ІР, щоб оцінити їх релевантність ПрО (співставлення тезаурусів є значно простішою операцією порівняно зі співставленням онтологій).

Розглянемо це на прикладі системи МАІПС.

Мультиагентна інформаційно-пошукова система (МАІПС) з розвинутими засобами інтелектуалізації її поведінки, що детально описана в [16, 17], надає користувачеві високо релевантні результати пошуку, які досягається завдяки:

- орієнтованості системи на користувачів, які мають в мережі сталі інформаційні інтереси та потребують постійного надходження відповідної інформації (функціонально МАІПС спрямована на виконання складних багаторазових запитів в досить вузьких областях, пов'язаних з професійними або науковими інтересами користувачів). Запити таких користувачів можуть повторюватися від сеансу до сеансу або змінюватися, але предметні області пошуку, в яких користувачі є експертами, практично не змінюються.
- Застосуванню інтегральних програмних агентів, створених з використанням розроблених моделей, які здатні діяти в динамічному середовищі, навчаючись на власному досвіді.
- Використанню бази знань (БЗ) предметної області (ПрО), яка подана у вигляді онтології.

Наукова новизна МАІПС полягає в інтегрованому використанні онтологічного подання знань, агентної парадигми та технологій Semantic Web для пошуку інформації на семантичному рівні [18].

Основні технології та методи, інтегровані в МАІПС:

1. онтології та тезауруси ПрО;
2. теоретико-множинні операції над тезаурусам;
3. генерація тезаурусів за природномовними текстами;
4. використання технологій Web 2.0 (хмар тегів – для візуалізації пошукових тезаурусів; соціальних сервісів – для взаємодії між користувачами);
5. застосування технологій Semantic Web;
6. оригінальні алгоритми впорядкування інформаційних ресурсів, знайдених системою, з урахуванням ваги онтологічних термінів;
7. використання критеріїв оцінки читабельності тексту для пошуку інформації, що відповідає персональним потребам користувача;
8. використання методів індуктивного виведення для узагальнення досвіду роботи МАІПС;
9. застосування мультиагентного підходу до створення моделі інтелектуальної інформаційно-пошукової системи та представлення компонентів системи як інтелектуальних BDI-агентів для формалізації поведінки системи в цілому;
10. використання парадигми інтелектуальних Web-сервісів для опису функцій агентів системи, що дозволяє їх інтероперабельне багаторазове використання.

Процес виконання запиту є послідовністю наступних кроків.

1. В результаті виконання інформаційного запиту користувача до Q за ключовими словами формується множина I. $I = \bigcup_{i=1}^{n} I_j$, де I_j – результат пошуку в ІР Q_j. Якщо є метаінформація про відповідний ІР (наприклад, у форматі RDF або MPEG7), то пошук здійснюється з її урахуванням. На жаль, більшість ІПС, що здійснюють пошук за ключовими словами, включають у I багато непотрібної користувачу інформації – повтори, не релевантні і застарілі посилання, а також посилання на документи, уже відомі користувачу. Щоб позбавити користувача від необхідності переглядати вручну всі ці документи, потрібно здійснити їх фільтрацію, використовуючи відомості про попередні запити цього користувача і сферу його інформаційних інтересів.
2. Якщо множина I не порожня, то виконується її упорядкування за URL-адресами посилань. Інакше – завершення роботи.
3. Якщо отримана на кроці 2 підмножина I не порожня, то відфільтровуються посилання-“дзеркала”. Інакше – завершення роботи.
4. Відфільтровуються застарілі посилання.
5. Якщо отримана на кроці 3 підмножина I не порожня, то здійснюється перевірка за БД користувача, чи одержував він раніше кожне з посилань (якщо одержував, тоді рішення про те, чи залишати це посилання, залежить від того, як у минулому користувач обробив це посилання, а також від інших його інструкцій). Інакше – завершення роботи.
6. Якщо сформована на кроці 5 підмножина I не порожня, то виконується перевірка на відповідність документів $i_j, j = \overline{0,k}$ з множини $I`, I` \subseteq I$ контексту пошуку. Інакше – завершення роботи.

Пошук може здійснюватися у два етапи. В результаті виконання першого етапу формується непорожня множина слів (чи словосполучень)

$W = \{w_1,...,w_m\}$, кожне з яких може мати свою позитивну або негативну вагу $v_k, k = \overline{1,m}$. Потім для кожного документа $i_j, j = \overline{0,k}$ з множини $I`, I` \subseteq I$ формується коефіцієнт відповідності контексту пошуку

$$s_j,\ j = \overline{0,k}, s_j = \sum_{k=1}^{m} v_k * f(i_j, w_k),$$

де $f(i_j, w_k) = \begin{cases} 1, & w_k \in i_j \\ 0, & w_k \notin i_j \end{cases}$.

Чим вище цей коефіцієнт, тим, імовірно, вище релевантність документа запиту користувача. У деяких випадках корисно використовувати більш складну формулу розрахунку коефіцієнта відповідності контексту пошуку

$$s`_j,\ j = \overline{0,k}, s`_j = \sum_{k=1}^{m} v_k * f(i_j, w_k) * t_k,$$

де $t_k, k = \overline{1,m}$ – кількість входжень терміна $w_k, k = \overline{1,m}$ в документ $i_j, j = \overline{0,k}$.

Користувач може звертатися до онтологій, створених іншими користувачами – переглядати їх, задавати за ними контекст пошуку, копіювати з них потрібні фрагменти, але не має права змінювати їх. ІПС має передбачати пошук онтологій, що містять уведені користувачем терміни, а також пошук онтологій, схожих на обрану користувачем онтологію. Це дозволяє створювати групи користувачів із загальними інформаційними інтересами і запобігти дублюванню у виконанні однакових багаторазових запитів різних користувачів. Адекватним засобом подання таких онтологій є мова OWL.

Якщо онтологія ПрО характеризує сферу інформаційних потреб користувача досить загально, то тезаурус дозволяє специфікувати її більш точно. Цей тезаурус використовується МАІПС для відбору найбільш пертинентних інформаційних ресурсів.

Іноді користувача цікавить ПрО, що відображається підмножиною певної онтології або є перетином різних ПрО, відображених різними онтологіями. Щоб відобразити в тезаурусі таку ПрО, користувач може виконувати різні операції (об'єднання, перетинання тощо) над тезаурусами, створеними за однією або кількома онтологіями.

При редагуванні тезаурусу можна вводити вагу різних термінів, що позначають їх важливість для пошуку (як позитивну, так і негативну), – цілі числа від –9 до +9. Ця інформація дозволяє відображати тезаурус у вигляді хмари тегів (червоним кольором відмічені терміни з негативною вагою, синім – з позитивною, розмір шрифту відображає числові значення ваги).

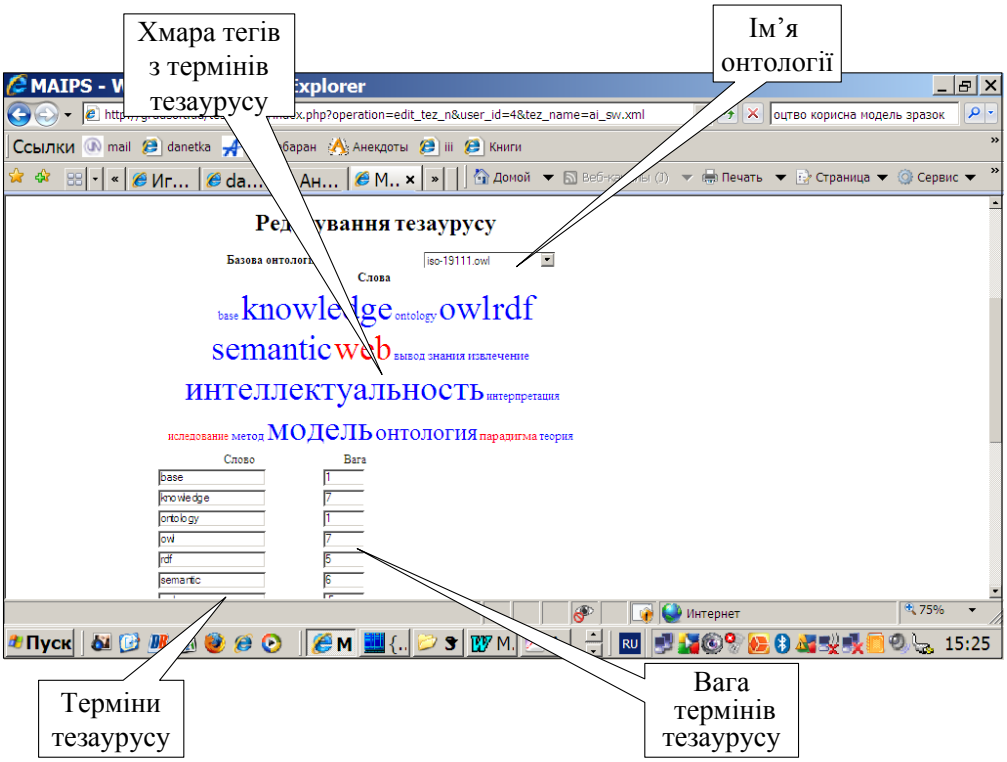

Рис.1. Редагування тезаурусу з можливістю вводу семантичної ваги окремих термінів

Результати виконання запиту надаються користувачеві у вигляді таблиці, яка містить гіперпосилання на знайдені ІР, їх короткі описи та рейтинг, визначений пошуковою системою за аналізом цих описів. У заголовку таблиці містяться відомості про зовнішню пошукову систему, за допомогою якої отримані ці результати, та описані основні параметри запиту.

Можна виділити два аспекти в навчанні МАІПС: пошук знань щодо контенту ІР, доступ до яких має система; пошук закономірностей, пов'язаних з поведінкою користувачів системи. Ще один варіант індуктивного узагальнення досвіду взаємодії користувача з ІПС – внесення уточнень в онтологію ПрО. Детальніше алгоритми, розроблені для цього, описані в [19, 20]. У цьому випадку результуючим параметром навчальної вибірки є той термін, зв'язки якого з іншими термінами онтології користувач хоче уточнити, а прикладами навчальної вибірки – тільки ті ІР, що задовольняють інформаційні потреби користувача. Ті терміни онтології, що ввійшли в побудоване по такої навчальної вибірки дерево рішень і зв'язані з результатом "Термін зустрічається в ІР часто" галузями, також зв'язаними зі значеннями "Термін зустрічається в ІР часто", повинні бути й у користувальницькій онтології бути зв'язані з цим терміном (семантику зв'язку має визначити користувач).

Проведені експерименти дозволяють показати переваги семантичного пошуку в порівнянні з традиційними пошуковими машинами. Порівнювалися результати пошуку за тими самими запитами у Google та їх упорядкування у МАІПС. Через те, що пошук здійснювався на одній індексній базі (Google), повнота та абсолютна точність пошуку співпадають, але значно відрізняється умовна точність пошуку для 10, 20 та 30 документів. Практично МАІПС автоматизує ту роботу, яку зазвичай виконує користувач, проглядаючи сторінки з результатами пошуку.

Висновки

Сьогодні для опису певної предметної ПрО та створення інтелектуальних систем та мереж велике значення має розробка нових алгоритмів та методик формування тезаурусів та онтологій. Для адекватного відображення інтероперабельних моделей подання знань в роботі запропоновано методику, яка використовує мереологічний підхід для створення визначень термінів тезаурусу, що дозволило сформувати більш якісні знання, які потім можуть використовуватися різноманітними інтелектуальними застосуваннями (приміром, для семантичного пошуку)

Список використаних джерел

1. Gruber T. R. A translation approach to portable ontology specifications // Knowledge Acquisition, 1993. – V. 5. – P. 199 – 220.

2. Клещев А. С., Артемьева И.Л. Отношения между онтологиями предметных областей. Ч. 1. Онтологии, представляющие одну и ту же концептуализацию. Упрощение онтологии // Информационный анализ. – Вып.1. – С. 2. – 2002. – С. 4 – 9.

3. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – Спб.: Питер. – 2001. – 382 с.

4. Cimiano P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. – Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006. – 347 p.

5. Euzenat J., Shvaiko P. Ontology matching. – Berlin: Springer-Verlag Heidelberg, 2007. – 332 p.

6. Смирнова Е.Д., Таванец П.В. Семантика в логике // Логическая семантика и модальная логика. – М.: Наука, 1967. – С. 3 – 53.

7. Gómez-Pérez A., Moreno A., Pazos J., Sierra-Alonso A.. Knowledge Maps: An essential technique for conceptualisation // Data & Knowledge Engineering. – 2000. – V. 33(2). – P. 169 – 190.

8. Непейвода Н.Н. Мереология. – http://www.logic.ru/Russian/events/ifras/nnn.pdf.

9. Лесьневский С. Об основаниях математики [Электронный ресурс] – Режим доступа: http://www.philosophy.ru/library/logic/lesnewxi.html.

10. Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология [Электронный ресурс] – Режим доступа: http://www.artint.ru/articles/narin/teon.htm.

11. Гладун А.Я., Рогушина Ю.В. Основи методології формування тезаурусів з використанням онтологічного та мереологічного аналізу // Искусственный интеллект. – 2008. – №5. – С. 112 – 124.

12. IDEF5 – Ontology Description Capture Method [Electronic resource] – Access mode: http://www.idef.com/IDEF5.html.

13. OWL 2 Web Ontology Language Document Overview. W3C. 2009 [Electronic resource] – Access mode: http://www.w3.org/TR/owl2-overview/.

14. Золин Е. Дескрипционная логика [Электронный ресурс] – Режим доступа: http://lpcs.math.msu.su/~zolin/dl/.

15. Тузовский А.Ф. Работа с онтологической моделью организации на основе дескриптивной логики // Известия Томского политехнического университета. – Т. 309. – № 7. – 2006 [Электронный ресурс] – Режим доступа: http://www.duskyrobin.com/tpu/2006-070030.pdf.

16. Рогушина Ю.В., Гришанова І.Ю. Літературний твір наукового характеру "Модель мультиагентної інформаційно-пошукової системи "МАІПС"("Модель МАІПС") // Свідоцтво про реєстрацію авторського права на твір №32068.

17. Гришанова І.Ю., Рогушина Ю.В. Комп`ютерна програма "Мультиагентна інформаційно-пошукова система "МАІПС"("МАІПС") // Свідоцтво про реєстрацію авторського права на твір № №32015.

18. Рогушина Ю.В. Семантический поиск как составляющая управления знаниями в Semantic Web // Материалы международной научно-технической конференции OSTIS-2012. – Минск: БГУИР, 2012. – С. 239 – 244.

19. Рогушина Ю.В., Гришанова І.Ю. Индуктивное извлечение знаний об информационных потребностях пользователей на основе онтологической модели предметной области поиска // Міжнародний семінар з індуктивного моделювання МСІМ-2005. Збірник праць. – К.: Міжнародн.наук.-навч. центр інформаційних технологій та систем НАН та МОН України, 2005. – С. 263 – 269.

20. Рогушина Ю.В., Гришанова І.Ю. Индуктивное извлечение знаний об информационных потребностях пользователей на основе онтологической модели предметной области поиска // Міжнародн.семінар з індуктивного моделювання МСІМ-2005. Збірник праць. – К.: Міжнародн.наук.-навч.центр інформаційних технологій та систем НАН та МОН України, 2005. – С. 263 – 269.

Відомості про авторів:

Рогушина Юлія Віталіївна – к. ф.-м. наук, старший науковий співробітник Інституту програмних систем Національної Академії Наук України. Наукові інтереси: онтологічний аналіз на базі технологій Semantic Web, інтелектуальний інформаційний пошук, методи індуктивного вилучення знання, дослідження інтелектуальних інформаційних систем і поведінку програмних агентів.

E-mail: ladamandraka2010@gmail.com

Гладун Анатолій Ясонович – старший науковий співробітник Міжнародного науково-навчального центру інформаційних технологій і систем (Національна Академія Наук України і Міністерство науки і освіти України). Наукові інтереси: розпізнавання та інтерпретацію інтелектуальних інформаційних об'єктів на базі технологій Semantic Web, семантичну інтероперабельність в середовищі інтелектуальних програмних агентів і мультиагентних систем; інтелектуалізацію систем управління в бездротових мережах NGN, сервіс-орієнтовані архітектури та GRID-технології.

E-mail: glanat@yahoo.com