УДК 025.3/6

Антон ВІТУШКО,
провід. інженер СІАЗ НБУВ

ЗАСТОСУВАННЯ ОНТОЛОГІЙ В АНАЛІТИЧНИХ ПІДРОЗДІЛАХ БІБЛІОТЕЧНИХ УСТАНОВ

Стаття присвячена питанням використання онтологій і створення Semantic Web в аналітичних підрозділах. Розглянуті основні етапи обробки інформації. Можливості використання онтологій під час пошуку й добування, подання й зберігання інформації. Висвітлюються такі системи керування знаннями, як корпоративна пам'ять та корпоративний портал.

Ключові слова: онтології, системи керування знаннями (СКЗ), Semantic Web, Concept-based queries, Smart Queries, XML, RDF, OWL, корпоративна пам'ять, корпоративний портал.

Процес створення та діяльності аналітичних підрозділів бібліотечних установ супроводжується зростанням ролі комп'ютерних технологій у сучасному суспільстві. Збільшення щоденного потоку інформації призводить до необхідності появи нових способів її пошуку, зберігання, подання, формалізації й систематизації, а також автоматичної обробки. На тлі цих процесів значний інтерес викликають системи, здатні без участі людини проводити певні операції з текстом.

Разом з World Wide Web з'являється його розширення, Semantic Web, в якому гіпертекстові сторінки забезпечуються додатковою розміткою, що надає відомості про семантику елементів, що включають у сторінки. Невід'ємним компонентом Semantic Web є поняття онтології, що описує зміст семантичної розмітки.

Як можуть використовуватися онтології в роботі аналітичних підрозділів бібліотечних установ? Для початку розглянемо основні етапи обробки інформації аналітичними службами. Це:

– пошук і добування інформацій з масиву даних;

– подання й зберігання інформації.

Завдання добування інформації з тексту складається з автоматичної обробки документів з метою розпізнавання й виділення релевантної інформації та подання її в структурованій формі. Практично в будь-якій предметній галузі для точного добування потрібні апріорні знання про неї – знання про поняття, об'єкти й відносини, пов'язані з цілями добування. У свою чергу витягнута з текстів інформація може нести в собі

нові знання про предметну галузь і бути корисною для подальшого добування. Тісний зв'язок між апріорною й витягнутою інформацією, а також між предметними й лінгвістичними знаннями сформував потребу в уніфікації засобів подання [1].

Уніфікація апріорних і витягнутих з текстів знань зручна тим, що дає змогу використати ті самі алгоритми й інструменти для роботи з обома типами знань. Об'єднання лінгвістичних і предметних знань в одному ресурсі, по-перше, полегшує первинне наповнення й подальшу підтримку, а по-друге, дає можливість використати предметні знання вже на етапі первинної обробки тексту правилами добування інформації. Завдяки спеціально розробленій мові запитів, правила можуть не обмежуватися словниковою інформацією, а звертатися в онтологію й базу фактів для перевірки різних умов.

Термін «онтологія» у машинному аналізі тексту використовується у вузькому значенні як синонім терміна «тезаурус» або «класифікатор» і є словником понять (концептів), кожному з яких відповідає синонімічний ряд термінів, плюс ієрархічна структура взаємозв'язків між ними типу «частина – ціле» або «загальне – частка» [2]. Такі «онтології в слабкому змісті» використовуються для формулювання запитів до пошукової машини, для автоматичної класифікації (категоризації) текстів.

Діючих прикладних програм, які належать до класу систем добування знань із тексту й, використовуючи «онтології у сильному змісті», тобто методи штучного інтелекту, здатні нетривіально переробляти витягнуті з тексту елементи знань (інтерпретувати, узагальнювати, виявляти залежності, прогнозувати й тощо), сьогодні не існує, в усякому разі для російської й української мови [3]. Таке обмежене використання онтологій зумовлене двома факторами. По-перше, слабким поширенням систем лінгвістичного аналізу тексту, здатних інтерпретувати синтаксичні відносини між словами та дійсно витягати знання як якісь нетривіальні елементи, який має внутрішню структуру, придатну для нетривіальної обробки штучним мозком. Такі системи на світовому й російському ринках тільки почали з'являтися в останні кілька років (Net Owl, Attensity, RCO Fact Extractor). По-друге, відносно низькою достовірністю тверджень, що витягають автоматично з тексту, і фактів, що зумовлено як недосконалістю алгоритмів аналізу тексту, так і якістю джерел інформації.

Інша особливість застосування онтологій у системах добування знань із тексту – необхідність мати додаткову лінгвістичну складову як для розпізнавання різних способів позначення понять (синонімічні терміни), так і для семантичної інтерпретації різноманітних мовних конструкцій у відносини між цими поняттями (синонімічні лексико-граматичні конструкції).

У результаті для систем добування знань із тексту найбільш типовою є онтологія «у слабкому змісті» з відносно бідною концептуальною, але надзвичайно багатою лінгвістичною складовою [3].

Якщо говорити про використання онтологій для подання й зберігання інформації в аналітичних підрозділах, то необхідно розглянути питання інтеграцій Semantic Web в організації.

Semantic Web організації – це реалізація глобальної концепції Semantic Web у рамках окремої організації. Семантичний веб (Semantic Web) – це концепція мережі, в якій кожний ресурс людською мовою був би позначений описом, зрозумілим комп'ютеру. В ідеальному варіанті вся інформація в Інтернеті повинна розміщуватися двома мовами: людською мовою для людини й комп'ютерною мовою, щоб «розумів» комп'ютер [4].

У нинішньому видгляді WWW – разом з пошуковими системами й іншими сервісами – уже є «мислячою мережею». Однак неважко зрозуміти, що самі веб-сторінки тут пасивні, а «мислячою частиною» є саме пошукові машини. Комп'ютерна програма нездатна, завантаживши будь-який документ (веб-сторінку або певний файл), зрозуміти його зміст. Семантичний веб дасть можливість користувачеві швидко одержувати відповіді на складні запити.

Сьогоднішні пошукові системи часто видають безліч «шумів», які не стосуються запиту, прирікаючи користувача на тривалий ручний відбір матеріалу. За допомогою технології семантичного веба забезпечується оптимізація пошуку інформації:

– за рахунок більш точного розуміння машиною понять, які цікавлять людину;

– за рахунок більш точного індексування інформації із застосуванням не тільки статистичних методів, але й аналізу семантичних зв'язків.

Наведемо приклади інтелектуального пошуку інформації:

1. «Понятійні» запити (Concept-based queries) – можливість робити запити на основі значень понять, а не на основі збігу рядків. Переваги даного методу – більш швидкий пошук потрібної інформації. На відміну від звичайного пошуку за ключовими словами, дає змогу доповнювати запит (результати) більш загальними або більш детальними поняттями стосовно пошукового ключового слова.

2. «Розумні» запити (Smart Queries) – можливість одержання не заданих явно знань зі створених моделей (шляхом логічного виводу), що дає змогу здійснювати пошук «схованої інформації» в об'єднаних схемах. Традиційні пошукові сервіси можуть зв'язувати окремі інформаційні джерела, використовуючи лексичні збіги. Реляційні БД використовують зв'язки, задані вручну в схемі даних (зв'язку між таблицями) або в коді процедур і функцій. Для виконання складного запиту над реляційною БД, необхідно заздалегідь визначити всю логіку цього запиту.

Можна виділити три ключові технологічні складові семантичного веба, що забезпечують його реалізацію:

– розширювана мова розмітки – eXtensible Markup Language, XML;

– система опису ресурсів – Resource Description Framework, RDF;

– мова мережевих онтологій – Web Ontology Language, OWL [1].

Мова мережевих онтологій дає змогу формально представляти онтології, які, у свою чергу, використовуються для опису інформації в мережі та її пошуку користувачами.

Незважаючи на очевидні переваги й перспективи глобального впровадження ідей і технологій Semantic Web, для реалізації цієї концепції в рамках всесвітньої мережі потрібен час. У процесі досліджень і розробок, проведених великими компаніями, був сформульований локальний підхід: семантичний веб організації Enterprise Semantic Web (ESW) – реалізація Semantic Web на обмеженій предметній галузі (у межах окремо взятої організації).

Enterprise Semantic Web дає можливість створити систему, що моделює сукупність бізнес-логіки, правил, процесів та інформації в організації, не прив'язаних до окремих додатків і не схованих у програмному коді. Відокремивши інформацію й знання від алгоритмів роботи додатків, розроблювачі зможуть створювати сервісні додатки, які буде використано поза залежністю від того, яку інформацію вони обробляють.

ESW не може існувати без застосування ключових технологій, рекомендованих W3C. Resource Description Framework (RDF) є новим підходом до організації даних у вигляді триплетів. Web Ontology Language (OWL) дає змогу представляти онтології, що описують значеннєву структуру даних, і створити моделі зв'язків. Ці дві специфікації, що використовуються спільно, становлять фундамент семантичного веба організації.

Під час проектування семантичного веба організації варто звернути увагу на те, що більша частина інформації всередині компанії зберігається в різних форматах, у різному ступені структурування тощо, тобто вона є занадто фрагментованою, щоб бути корисною. Першим кроком може стати впровадження RDF як базовий формат зберігання даних. RDF може описувати дані з будь-якого джерела й будь-якого формату (XML, реляційні БД, ієрархічні моделі, неструктуровані або напівструктуровані), причому реалізувати цю можливість можна без особливих проблем. Автоматизовані інструменти для розбору можуть бути використані для створення RDF триплетів з існуючих стандартних джерел. Результуючі RDF дані простої структури зберігаються й використовуються в розподілених базах даних RDF.

RDF є простим способом організації доступу до даних в уніфікованому форматі – простіше й наочніше, ніж XML структури або реляційні БД. Однак RDF не дає можливисті створювати структури, які

описують самі себе. Інакше кажучи, RDF не має достатніх семантичних можливостей, щоб зберігати зміст інформації.

Можливості RDF як базової мови доповнюються технологією OWL, що надбудовується на рівень вище (для створення ієрархії інформаційних джерел і визначення семантики зв'язків також може використатися RDF(S), але у зв'язку з тим, що OWL має ті ж можливості й додаткові зручні особливості, у даному прикладі архітектури на другому рівні розглядається саме цей стандарт).

OWL надає можливості побудови семантичної структури для організації й класифікації RDF триплетів. Онтології OWL створюються зверху вниз, виходячи з логіки бізнесу. Одночасно RDF дані генеруються знизу вгору автоматизованими інструментами.

OWL зв'язується з RDF за допомогою універсальних ідентифікаторів ресурсів (URIs). Фактично кожний індивідуальний елемент OWL може бути пов'язаний з елементом даних RDF (одним з вузлів триплету об'єкт – властивість – значення).

Знання, що здобувають під час здійснення інформаційного пошуку, вимагають подальшого нагромадження й керування у вигляді корпоративної пам'яті (КП). КП включає роботу як з явними знаннями аналітичного підрозділу у формі баз даних і електронних архівів, так і зі схованими знаннями – або фіксуючи його у формі експертних систем і БД, або забезпечуючи пошук і доступ до експертів.

Корпоративна пам'ять – це сховище інформації, що утримується в різних джерелах (бази даних, файлові системи, бази знань). Усі знання, що утримуються в корпоративній пам'яті, повинні бути описані за допомогою універсальної моделі. Останнім часом для цих цілей часто використовуються онтології. Для того щоб описати онтологію деякої предметної галузі, необхідно визначити її основні поняття (концепти), співвідношення й зв'язки між ними. Візуально онтології можуть подаватися різними способами. Наприклад, у вигляді графа або гіпертексту. Зручність використання онтологій полягає в тому, що вони дають можливість надати цілісно й одноманітно всю необхідну нам предметну галузь. Крім того, онтології дають змогу надати користувачеві тільки необхідну її частину, приховуючи все різноманіття понять і зв'язків між ними, що існує в предметній області. Це також спрощує метод фільтрації переданої користувачеві інформації.

Необхідність у КП зумовлена потребою аналітичної служби більш ефективно використовувати накопичені знання та таким явищем, як втрата досвіду. Втрата досвіду в основному пов'язана зі звільненням або виходом на пенсію висококваліфікованих співробітників, але може також відбуватися в результаті зміни кваліфікації співробітників, необхідної для реалізації нових проектів. В обох випадках досвід, придбаний під час реалізації попередніх проектів, повністю або частково втрачається.

Відновлення попереднього досвіду часто потребує істотних витрат робочого часу і як наслідок – підвищення вартості проектів. Ці обставини стимулюють організації вкладати кошти в розробку інструментів і систем, що сприяють обміну досвіду між співробітниками й зниженню залежності знань від індивідів, що володіють ними.

Характер потреб у системах корпоративної пам'яті й обсяг зусиль для їх впровадження в аналітичну діяльність організацій може залежати від розміру організацій. Серед мотивуючих факторів можна виділити такі:

1. Уникнення втрати унікальних знань фахівця після його виходу на пенсію або зміни місця роботи.

2. Збереження досвіду, отриманого у попередніх проектах, і уникнення повторення помилок.

3. Використання карти знань організацій для розробки стратегії розвитку. Регулярне впровадження інновацій має покращувати здатність служби реагувати на інформаційні зміну.

4. Поліпшення процесу поширення інформації й підвищення рівня комунікації всередині організацій.

5. Поліпшення якості процесу навчання персоналу організації.

6. Інтегрування різних інновацій.

Основні функції корпоративної пам'яті:

– збір і систематична організація інформації з різних джерел у централізоване й структуроване інформаційне сховище;

– інтеграція з існуючими автоматизованими системами збору інформації;

– забезпечення запиту потрібною інформацією.

Можна виділити два рівні корпоративної пам'яті:

1. Рівень матеріальної або явної інформації – це дані й знання, які можуть бути знайдені в документах організації у формі повідомлень, листів, статей, довідників, патентів, креслень, відео- і аудіозаписів, програмного забезпечення й тощо.

2. Рівень персональної або схованої інформації – це персональне знання, невідривно пов'язане з індивідуальним досвідом. Воно може бути передане через прямий контакт – «віч-на-віч» через процедури добування знань. Саме сховане знання – це практичне знання, що є ключовим при прийнятті рішень і керуванні технологічними процесами [5].

Кінцева мета КП полягає в тому, щоб забезпечити доступ до знання щоразу, коли це необхідно. Щоб забезпечити це, КП реалізує активний підхід до поширення знань, що засновується не на запиті користувачів, а автоматично забезпечує корисне для рішення завдання знання. Щоб запобігати інформаційному перевантаженню, цей підхід має бути поєднаний з високою вибірковою оцінкою доречності. В ідеалі система повинна діяти як інтелектуальний помічник користувачеві.

Для опису ресурсів знань та їх пошуку в КП доцільно використовувати три види онтологій:

– онтологію інформації для опису метавластивостей і доступу до інформації;

– онтологію організацій для опису контексту створення й застосування інформації;

– предметну онтологію для опису контенту інформації.

Особливу увагу під час розробки пошукових механізмів у КП пропонується приділяти саме контексту. Основне завдання КП полягає в тому, щоб виявляти інформаційну потребу протягом виконання виробничого процесу й визначати доречне знання в специфічному контексті завдання. Моделювання й використання контексту для пошуку інформації в КП створює потенціал для застосування результатів моделювання. Під час моделювання контексту розглядаються дві сторони:

– контекст потенційного застосування знань;

– контекст, в якому знання було створено.

Контекст застосування рекомендується представляти в термінах організаційної структури, процесів, функцій, які саме становлять основу онтології організації й онтологічної моделі, які створюються при моделюванні.

Навіть добре структуроване знання, якщо воно не є доступним для потенційних споживачів, мало кого цікавить. Цінність одиниці знань зі слабкою доступністю надзвичайно мала. Розрізненість, незв'язаність знань, відсутність подання про їх поточний стан стають істотною перешкодою для їх застосування в інтересах аналітичної структури.

Сьогодні знайдено універсальне рішення, що забезпечує доступ до інформації в межах систем УЗ, – це корпоративні портали. Корпоративний портал – система, що поєднує всі наявні в організації інформаційні ресурси (додатки, бази й сховища даних, аналітичні системи та ін.), і, використовуючи веб-інтерфейс, надає користувачам єдиний захищений доступ до корпоративної й зовнішньої інформації. Реалізується створенням Інтранету.

Корисні властивості, які мають в собі портал:

– він дає можливість чітко систематизувати контент і надати ефективні кошти навігації для користувачів;

– може надавати ефективні можливості пошуку, включаючи засоби інтелектуального пошуку й візуальні моделі;

– за допомогою порталу інформація доступна в будь-який день 24 години на добу з будь-якого місця з доступом в Інтернет, що дає змогу більш гнучко організовувати роботу співробітників;

– дає можливість надавати засоби керування контентом для різних груп співробітників (керування доступом);

– підтримує внутрішній корпоративний обмін знаннями й спільною роботою;

– забезпечує персоналізацію подання інформації для співробітників: персоналізація сторінок, каналів новин, оголошень.

Цінність корпоративного порталу визначається його пошуковими можливостями, які, як було показано вище, можуть бути істотно поліпшені за рахунок застосування онтології.

Отже, описані вище підходи застосування онтологій використовуються під час розробки системи інформаційного пошуку й обробки інформацій аналітичними підрозділами. Відмінною рисою цих систем інформаційного пошуку є орієнтація на явне подання знань за допомогою онтологій. Дані підходи дають можливість реалізувати інтелектуальні сервіси для пошуку й обробки електронних документів, одержаних з різних джерел, з необхідної тематики.

Список використаних джерел

1. *Овдей О. М.* Обзор инструментов инженерии онтологий [Электронный ресурс] / О. М. Овдей, Г. Ю. Проскудина. – М. : Институт развития информ. общества. – Т. 7. – Вып. 4. – 2004 / Электронный журнал, посвященный созданию и использованию электронных библиотек. – Режим доступа: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part4/op. – Загл. с экрана.

2. *Гладун А.* Онтологии в корпоративных системах / А. Гладун, Ю. Рогушина // Корпоративные системы. – 2006. – № 1. – С. 41–47.

3. *Гаврилова Т. А.* Использование онтологий в системах управления знаниями / Т. А. Гаврилова // Труды международного конгресса «Искусственный интеллект в XXI веке», Дивноморское, Россия. – М. : Физматлит, 2001. – С. 21–33.

4. *Гаврилова Т. А.* Извлечение знаний: Лингвистический аспект / Т. А Гаврилова // Корпоративные системы. – 2001. – № 10 (25). – С. 24–28.

5. *Гаврилова Т.* Онтологический подход к управлению знаниями при разработке корпоративных информационных систем / Т. Гаврилова // Новости искусственного интеллекта. – 2003. – № 2. – С. 24–30.