УДК 681.3.06

О. Захарова, В. Міненко

ТЕХНОЛОГІЯ ОПТИМАЛЬНОГО ВИБОРУ ВАРІАНТІВ НА ОСНОВІ СЕМАНТИЧНОГО АНАЛІЗУ ІНФОРМАЦІЙНИХ ОБ'ЄКТІВ БІЗНЕС-ПРОЦЕСУ

У роботі розглядається проблема використання онтологій для вирішення задач прийняття рішення на основі проведення семантичного аналізу та визначення відповідності предметних областей інформаційних об'єктів. Пропонуються інтелектуальні методи оптимального підбору варіантів на базі порівняння онтологій відповідних предметних областей. Запропоновані методи розглядаються на прикладі вирішення задач бізнес процесу публікації наукової праці, а саме: вибору видання, вибору рецензента, визначення коду УДК.

Вступ

В сучасному світі накопичена велика кількість інформації, що представлена в електронному вигляді. Це дає можливості її інтелектуальної автоматизованої обробки, тобто обробки інформації не лише як структур даних, а й охоплюючи рівень їх семантики. На сьогоднішній день досить ефективним засобом явного представлення інформаційних елементів є онтологічні описи предметних областей. Використання онтологій як джерел даних для прикладної системи дозволяє більш ефективно обробляти складну та різноманітну інформацію.

В даній роботі розглядатиметься проблема використання онтологій для вирішення задач прийняття рішення на основі проведення семантичного аналізу та визначення відповідності предметних областей інформаційних об'єктів. Тобто мова йде про вибір найбільш прийнятних пар пов'язаних семантичними відносинами сутностей на основі аналізу їх предметних областей. Як приклад розглядається процес подання наукової праці для публікації, де підхід, що пропонується, може бути використаний для вирішення цілої низки задач, а саме:

- вибору найбільш прийнятного видання;
- вибору рецензента;
- визначення для наукової праці коду УДК за описом її предметної області.

Метод залучення онтологій для встановлення відповідності предметних областей як критерію прийняття рішення в задачах оптимального вибору варіантів

Нехай X та Y – два об'єкти реального світу, що пов'язані певними відносинами. Кожний екземпляр цих об'єктів має власну предметну область (ПО). Необхідно визначити ступінь відповідності їх предметних областей, що далі може бути критерієм для оптимального вибору пари y_j з множини екземплярів об'єкта Y для кожного екземпляра x_i об'єкта X.

Слід зазначити, що в даному випадку ми вирішуємо односторонню задачу, тобто для кожного x_i необхідно обрати один найбільш підходящий y_j. При чому, один і той самий y_j може виявитися найприйнятнішим для декількох x_i. Теоретично може існувати ситуація, коли жоден екземпляр y_j не підходить x_i, тобто множина екземплярів об'єкта Y, що аналізується, просто не містить варіанта для екземпляра x_i об'єкта X, в цьому випадку для екземпляра x_i на наявній множині екземплярів Y задача не має рішення – ПО інформаційного об'єкта Y не відповідає даному екземпляру.

У випадку повної відповідності предметних областей x_i та y_j, предметна область x_i буде підмножиною предметної

області y_j, тобто, якщо оперувати поняттями алгебри онтологій, онтологія ПО x_i є перетином онтологій ПО x_i та y_j.

В решті випадків має місце часткова відповідність предметних областей x_i та y_j, тобто їх онтології мають перетин, але він не співпадає з онтологією ПО x_i. У такому разі, відповідність ПО визначається коефіцієнтом відповідності.

В загальному випадку, коефіцієнт відповідності предметних областей має виражати зворотне відношення загальної кількості понять, що описують предметну область екземпляра x_i, до кількості цих понять, що належать до предметної області.

Тобто, якщо предметна область x_i описується множиною понять $T_1, T_2, ..., T_n$ та k_s $(s = 1,...,n)$ – коефіцієнт, що є ознакою належності поняття T_s до предметної області y_j, такий що:

$k_s = 1$, якщо поняття T_s належить до предметної області y_j, та

$k_s = 0$, якщо поняття T_s не належить до предметної області y_j, то

μ_{ij} – коефіцієнт відповідності предметних областей x_i та y_j можна виразити наступним чином:

$$\mu_{ij} = \sum_{s=1}^{n} \frac{k_s}{n}.$$

Як результат обирається пара з найбільшим значенням коефіцієнта відповідності предметних областей. Очевидно, що коефіцієнт приймає своє найкраще значення $(\mu_{ij} = 1)$ у випадку повної відповідності.

Слід зазначити, що існуючі на цей час методи встановлення належності сутності до певної ПО шляхом порівняння її текстового опису з онтологією ПО обумовлюють однозначну відповідь тільки у випадку, коли сутність не належить ПО, а онтологія ПО є повною.

Крім того, методи порівняння двох онтологічних описів дають точну адекватну оцінку лише за умови однорідності цих описів.

Уніфікація онтологій предметних областей

Онтологія ПО кожного з екземплярів може мати своє представлення, та у загальному випадку ці онтології є гетерогенними. Якщо мова йде про встановлення відповідності шляхом пошуку одного концепта в онтології, що характеризує її контент в цілому, як, наприклад, може відбуватися при вирішенні задачі вибору видання для наукової праці за кодом УДК, то ця проблема не є критичною. Але, у загальному випадку, необхідно визначити відповідність ПО, проаналізувавши всю множину понять, що її описують, та зв'язки між ними.

В такому разі для проведення аналізу цих ПО необхідна їх попередня уніфікація, та, перш за все, приведення їх до однієї форми представлення. Зважаючи на те, що будь-яка онтологія, незалежно від вигляду, може бути представлена орієнтованим графом, в якості форми представлення доцільно обрати графічну модель [1]. Для погодження онтологічних контекстів може бути використаний один з підходів, що описані в [2].

Визначення відповідності для онтологій ПО, що представлені графами

Далі розглянемо, як трансформується алгоритм визначення відповідності ПО інформаційних об'єктів, у разі використання графічної моделі для представлення їх онтологій. В даному випадку, визначення відповідності ПО інформаційних об'єктів зводиться до порівняння відповідних онтологічних графів. У випадку повної відповідності онтологічний граф ПО одного інформаційного об'єкта є підграфом онтологічного графа ПО об'єкта, який шукається до пари, та коефіцієнт відповідності їх ПО дорівнює 1.

Якщо онтологічні графи предметної області екземплярів інформаційних об'єктів x_i та y_j мають перетин, але граф перетину не співпадає повністю з

онтологічним графом ПО екземпляра x_i, то має місце часткова відповідність предметних областей, яку можна виразити коефіцієнтом відповідності.

Коефіцієнт відповідності ПО для ПО x_i та y_j, у загальному випадку:

$$\mu_{ij} = \frac{\left(\sum_{s=1}^{n_i} k_s\right) * \left(\sum_{q=1}^{m_i} k_q\right)}{n_i * m_i},$$

де n_i – кількість вузлів в онтологічному графі ПО екземпляра x_i,

m_i – кількість дуг в онтологічному графі ПО екземпляра x_i;

$k_s = 1$, якщо для вузла N_s онтологічного графа ПО x_i існує відповідний вузол в онтологічному графі ПО y_j, та $k_s = 0$ – в іншому випадку;

$d_q = 1$, якщо для дуги E_q онтологічного графа ПО x_i існує відповідна дуга в онтологічному графі ПО y_j, та $d_q = 0$ – в іншому випадку.

Очевидно, що найбільш відповідним вибором буде пара з максимальним значенням коефіцієнта відповідності.

Методи вирішення задачі вибору видання для публікації наукової праці

Далі розглядається приклад використання запропонованих алгоритмів в межах загальної технології, що пропонується для формалізації та автоматизації бізнес процесу подання наукової праці до друку, перш за все, для вирішення задачі вибору автором найбільш прийнятного видання для публікації своєї нової праці. Нині це робиться автором самотужки на основі власних знань та уподобань щодо відомих йому видань. Очевидно, що такий вибір є досить обмеженим, а результат складно вважати оптимальним. При використанні автоматизованих процедур вибору основними критеріями мають бути:

– ступінь відповідності ПО видання;
– рейтинг видання;
– ступінь, в якій видання може задовільнити вимоги, що висуваються автором, наприклад, щодо терміну публікації наукової праці.

Для визначення ступеня відповідності ПО наукової праці та видання можуть бути застосовані підходи, що запропоновані вище.

Слід зазначити, що тут доцільно врахувати два варіанти вибору видання:

1) за кодом УДК, якщо він визначений для наукової праці;

2) за відповідністю онтологічних описів ПО множиною понять, що її характеризують, та зв'язків між ними (ключові слова – вироджений випадок онтології ПО).

В першому випадку інтегрована онтологія ПО видань, що аналізуються, має вигляд (рис. 1).

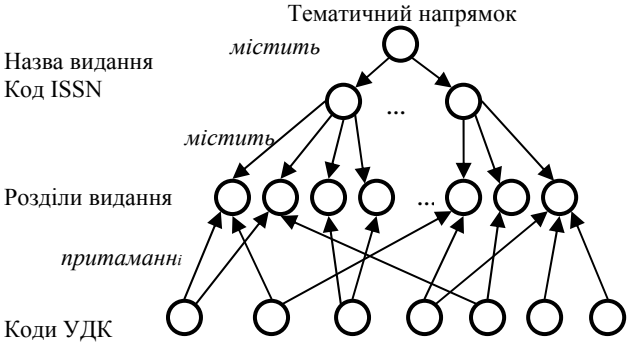

Рис. 1. Структура інтегрованої онтології

ПО наукової праці описується кодом десятинної класифікації, що їй присвоєний. Задача встановлення відповідності полягає у знаходженні в ієрархії інтегрованої онтології вказаного коду УДК, та визначення всіх можливих кодів ISSN, що пов'язані з цим кодом через рівень *Розділів видання*. Слід зазначити, що під відповідністю кодів УДК в даному випадку розуміється не обов'язково точна їх рівність. Якщо код УДК наукової праці визначає підрозділ розділу, що визначається кодом УДК, який знайдено в ієрархії інтегрованої онтології, та цей вузол не має нащадків, то ці коди вважаються відповідними. Іншими словами, якщо видання публікує наукові праці за деяким тематичним розділом, то воно публікує

наукові праці з будь-якого підрозділу цього розділу. Зважаючи на технологію формування коду УДК [3], відповідність таких кодів може бути визначена шляхом аналізу рядків кодів. Для встановлення факту відповідності, рядок коду УДК наукової праці має бути підрядком, починаючи з першого символу рядка коду УДК, за яким видання приймає до публікації наукові праці.

Якщо відповідних кодів ISSN знайдено декілька, то подальший вибір здійснюється з урахуванням виконання інших вимог автора та аналізу рейтингів видань.

У другому випадку, для визначення коефіцієнта відповідності ієрархія інтегрованої онтології має бути розширена ієрархіями понять або ключових слів, що характеризують кожний з тематичних підрозділів. Це може бути зроблено відповідно до стислого опису підрозділів. З іншого боку, ПО наукової праці також описується онтологією, що містить поняття, які описують контент наукової праці.

Відповідність цих предметних областей у загальному випадку розглядається по гілкам ієрархії інтегрованої онтології. Кожна гілка відповідає коду ISSN видання. Тобто відповідність предметних областей наукової праці та видання визначається коефіцієнтом відповідності ПО, який розраховується для кожного видання, що визначається кодом ISSN, та є присутнім у базі доступних видань, тобто є елементом множини, на якій здійснюється пошук.

Коефіцієнт відповідності ПО має виражати зворотне відношення кількості понять, що описують ПО наукової праці, до кількості цих понять, що належать до ПО видання.

Тобто, якщо ПО наукової праці описується множиною понять $T_1, T_2, ..., T_n$, та k_j $(j = 1,...,n)$ – коефіцієнт, що є ознакою належності поняття T_j до ПО видання, такий що:

$k_j = 1$, якщо поняття T_i належить до ПО видання, та

$k_j = 0$, якщо поняття T_i не належить до ПО видання, то

μ_i – коефіцієнт відповідності ПО наукової праці до ПО видання P_i можна виразити наступним чином:

$$\mu_{ij} = \frac{\sum_{j=1}^{n} k_j}{n}.$$

Найбільш прийнятним є видання з максимальним значенням коефіцієнта відповідності ПО. Очевидно, що коефіцієнт приймає своє найкраще значення ($\mu = 1$), якщо всі поняття ПО наукової праці належать до ПО видання – повна відповідність.

У випадку представлення онтологій ПО орієнтованими графами, визначення їх відповідності зводиться до порівняння відповідних онтологічних графів. У випадку повної відповідності онтологічний граф наукової праці є підграфом онтологічного графа ПО видання, та коефіцієнт відповідності ПО дорівнює 1.

Коефіцієнт відповідності ПО для видання P_i у загальному випадку має вигляд:

$$\mu_{ij} = \frac{\sum_{j=1}^{n} k_j * \sum_{l=1}^{m} d_l}{n * m},$$

де n – кількість вузлів в онтологічному графі ПО наукової праці, m – кількість дуг в онтологічному графі.

$k_j = 1$, якщо для вузла N_i онтологічного графа ПО наукової праці існує відповідний вузол в онтологічному графі ПО видання, та $k_j = 0$ – в іншому випадку.

$d_l = 1$, якщо для дуги E_i онтологічного графа ПО наукової праці існує відповідна дуга в онтологічному графі ПО видання, та $d_l = 0$ – в іншому випадку.

Задача визначення рейтингу має вирішуватися за допомогою методів багатокритеріального оцінювання на основі аналізу множини критеріїв. В якості таких критеріїв можна розглядати: чи є це видання признане ВАК, міжнародне чи внутрішнє, термін існування, кількість праць, що були в ньому надруковані за весь термін існування, за останні декілька років, кількість праць, що опубліковані за

даною тематикою. Окрім того, досить достовірною оцінкою є індекс цитування наукових праць, що були надруковані в даному виданні.

Під індексом цитування видання (і(вид)) [4] в даному випадку розуміється відношення кількості посилань на наукові праці, що опубліковані у виданні, до загальної кількості опублікованих в ньому наукових праць у певному науковому напрямку за певний період часу:

$$i(\text{вид})td = K_n / K(S_t),$$

де $t = \langle T_1, T_2 \rangle$, $0 < T_1 < T_2$;

$$d = \{U \in M(УДК)\};$$

K_n – число посилань на наукові праці, що надруковані у виданні, за період часу t;

$K(S_t)$ – загальна кількість опублікованих наукових праць за період часу t.

Для оцінювання за критеріями, що не мають точного кількісного вираження, необхідне залучення незалежних експертів.

Ступінь виконання вимог автора щодо терміну публікації наукової праці визначається прогнозованим часом очікування T, що має задовольняти умові:

$$T \le d_{\max} - d_{\text{register}},$$

де d_{register} – дата подання наукової праці,

$d_{\max}$ – максимальний термін публікації, визначений автором.

Час очікування публікації науковою працею залежить від наступних показників:

1) обсяг черги – кількість наукових праць, що очікують публікації;

2) періодичність виходу видання;

3) обсяг вибірки видання (стала величина для всіх екземплярів видання);

4) рейтинг наукової праці.

Рейтинг наукової праці $r(S_t)$ визначається як медіана [5] множини індексів цитування [6] авторів статті, тобто

$$r(S_t) = \mu(i(\text{авт}), \text{авт} \in A(S_t)),$$

де $i(\text{авт})$ – індекс цитування автора статті, $i \ge 0$;

$A(S_t)$ – множина авторів статті.

Під індексом цитування автора (i(авт)) [7] в даному випадку розуміється відношення кількості посилань на статті автора до загальної кількості опублікованих статей автора в певному науковому напрямку за певний період часу:

$$i(\text{авт})td = K_n / K(S_t),$$

де $t = \langle T_1, T_2 \rangle$, $0 < T_1 < T_2$;

$$d = \{U \in M(УДК)\};$$

K_n – число посилань на статті автора за період t;

$K(S_t)$ – загальна кількість статей автора за період t.

Інші варіанти використання запропонованого алгоритму для формалізації процесу подання наукової праці для публікації

Якщо розглядати процес подання наукової праці в цілому, то наведені методи встановлення відповідності ПО можуть бути використані для вирішення принаймні ще двох задач бізнес процесу:

1) визначення для наукової праці коду УДК за описом її предметної області;

2) вибір рецензента для наукової праці.

Визначення для наукової праці коду УДК за описом її предметної області.

Пропонується підхід до визначення коду УДК наукової праці шляхом порівняння множини понять наукової праці з множиною понять, що характеризують кожний код десятинної класифікації.

Тобто для вирішення цієї задачі, перш за все, необхідно мати онтологію, що описуватиме універсальну десятинну класифікацію зі стислим описом кожного підрозділу поняттями відповідної ПО. За наявності такої онтології, алгоритм визначення УДК буде аналогічним

алгоритму, що був застосований при порівнянні предметних областей наукової праці та видання. Для наукової праці обиратиметься УДК з найбільшим коефіцієнтом відповідності β_i $(i=1,...,n)$, де

$$\beta_i = \sum_{j=1}^{m} k_{ij},$$

та n – число вузлів $N_i (i = 1,...,n)$, що представляють коди УДК найнижчого рівня ієрархії в онтологічному графі універсальної десятинної класифікації,

m – кількість понять E_j, що описують ПО наукової праці,

$k_{ij} = 1$, якщо поняття E_j належить до підграфу понять вузла N_i, та $k_{ij} = 0$, в іншому випадку.

Вибір рецензента для наукової праці

Множина можливих рецензентів для наукової праці може формуватися із залученням авторів, що публікувалися у відповідній ПО. При чому онтологія ПО автора або рецензента має розширюватися по мірі публікації нових праць. Так, у разі використання графічної моделі представлення, вихідною онтологією ПО автора/рецензента є онтологічний граф його першої опублікованої наукової праці. Поточною онтологією його ПО є онтологічний граф, що описує всю його наукову діяльність на поточний момент, починаючи з першої статті до останньої. Формування ПО автора/рецензента виконується шляхом покрокового об'єднання без перетину онтологічних графів [1, 2, 8] всіх його наукових праць.

При здійсненні вибору, перш за все, необхідно вилучити з множини, що аналізується, автора та співавторів даної наукової праці.

Для оцінювання відповідності для решти рецензентів та вибору найкращого може бути застосований підхід, що запропонований для вибору видання, – тобто встановлення ступеня відповідності ПО наукової роботи та рецензента. Повна відповідність рецензента науковій праці означає, що онтологічний граф ПО наукової праці є підграфом онтологічного графа ПО рецензента. Коефіцієнт відповідності ПО рецензента та наукової роботи, що подана на публікацію, може бути визначений за формулою (2), де n – кількість вузлів в онтологічному графі ПО наукової праці, m – кількість дуг в онтологічному графі наукової праці; $k_j = 1$, якщо для вузла N_i онтологічного графа ПО наукової праці існує відповідний вузол в онтологічному графі ПО рецензента, та $k_j = 0$ – в іншому випадку; $d_l = 1$, якщо для дуги E_i онтологічного графа ПО наукової праці існує відповідна дуга в онтологічному графі ПО рецензента, та $d_l = 0$ – в іншому випадку.

Слід зазначити, що для уточнення та покращення вибору, до рецензентів також може бути застосована система рейтингів.

Зважаючи на те, що рецензенти обираються з множини авторів, що мають наукові праці в заданій тематиці, рейтинг рецензента може визначатися як його індекс цитування як автора (*i*(авт)) [4]. Окрім цього, при призначенні рецензента можуть бути враховані наступні критерії:

- його науковий ступінь;
- термін опублікування останньої праці за цією тематикою;
- завантаженість рецензента (кількість наукових праць, що були передані йому на рецензію, та очікують своєї черги).

Для формування результуючого коефіцієнта рецензента доцільно використати методи багатокритеріального оцінювання [9].

1. *Prasenjit Mitra, Gio Wiederhold, Martin Kersten.* A Graph-Oriented Model for Articulation of Ontology Interdependencies. - Technical Report, 1999, http://www.dit.unitn.it~p2p/RelatedWork/Matching/graphmodel.pdf
2. *Захарова О.В.* Основні принципи побудови онтологічного граф-орієнтованого опису прикладної області// Проблеми програмування. – 2010 – № 4. – С. 51–59.

3. *Універсальна* десяткова класифікація - http://www.nbuv.gov.ua/texts/libdoc/udc.htm#0
4. *CiteSeer^{x}beta*: http://citeseerx.ist.pru.edu/
5. *Феллер В.* Введение в теорию вероятностей и её приложения. Том 1. – М.: Мир, 1967. – 498 с.
6. *Плескач В.Л., Рогушина Ю.В.* Агентні технології: Монографія. – К.: нац. торг.-екон. ун-т, 2005. – 338 с.
7. *Александров Д., Костров А., Макаров Р., Хорошева Е.* Методы и модели информационного менеджмента. – М.: Финансы и статистика, 2007.
8. http://www.obitko.com/tutorials/ontologies-semantic-web/operations-on-ontologies.html
9. *Міненко В.Д.* Формування редакційного портфеля для автоматизованої системи видавництва // Системний аналіз та інформаційні технології. – 2009.

Одержано 27.09.2011

Про авторів:

Захарова Ольга Вікторівна,
кандидат технічних наук,
науковий співробітник,

Міненко Валерій Дмитрович,
інженер-програміст.

Місце роботи авторів:

Інститут програмних систем
НАН України,
проспект Академіка Глушкова, 40.
Тел.: 526 5139,
e-mail: ozakharova68@gmail.com

Компанія Artofbytes
e-mail: valmin@ukr.net