УДК 004.89

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

ПОСТАНОВКА ЗАДАЧІ ОЦІНЮВАННЯ НОВИЗНИ ОНТОЛОГІЧНИХ ЗНАНЬ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ

Розглянуто формалізацію процесу оцінювання новизни онтологічних знань з метою побудови ефективних процедур розв’язування задач інтелектуальною системою.
Ключові слова: онтологія, оцінка новизни знань, інтелектуальна система, база знань.

In the paper the formalization of the evaluation process innovation ontological knowledge to build effective procedures for solving problems of intelligent systems..
Keywords: ontology, knowledge innovation, intelligent system, knowledge base.

Постановка проблеми у загальному вигляді

Основною компонентою інтелектуальних систем (ІнтС) є база знань (БЗ), що формується відповідно до предметної області (ПО), на яку зорієнтоване функціонування цієї системи [1]. Призначення цієї компоненти полягає у зберіганні, впорядкуванні та керуванні інформацією про ПО та задачі, які в ній виникають. Найважливіший параметр БЗ – якість та повнота знань про ПО, яку вона задає. Якість БЗ залежить від структури та формату знань, способу їх подання. Для широкого впровадження будь-якої технології чи методики необхідний чіткий і аргументований стандарт. Традиційні методи інженерії знань (отримання знань від експерта, інтелектуальний аналіз даних, машинне навчання тощо) не ґрунтуються на системі вивірених та загальноприйнятих стандартів, тому побудовані на їхній основі бази знань з часом втрачають свою функціональність через низьку ефективність їх функціонування. Як стандарт інженерії знань використовують онтологічний інжиніринг, у результаті застосування якого отримують онтологію бази знань. Онтологія – це детальна формалізація деякої області знань, подана за допомогою концептуальної схеми. Така схема складається з ієрархічної структури понять, зв’язків між ними, теорем та обмежень, які є прийняті у певній ПО.

Використання онтологій у складі БЗ ІнтС допомагає вирішити низку проблем методологічного та технологічного характеру, які виникають під час розроблення таких систем. Зокрема для України характерні проблеми відсутності концептуальної цілісності й узгодженості окремих прийомів та методів інженерії знань; нестачі кваліфікованих фахівців у цій галузі; жорсткості розроблених програмних засобів та їх низькій адаптивній здатності; складності впровадження ІнтС, що зумовлено психологічними аспектами.

Ефективність онтології напряму залежить від новизни знань, які до неї додаються. Постає завдання оцінювання знань, які пропонується додавати в онтологію БЗ ІнтС, тобто побудови деякої міри знань, подібної до відомої міри інформації, яку ввів Шеннон [2]. Однак, на відміну від інформації, для знань необхідно визначити ефект від її використання. Отже, метою цієї роботи є розроблення формальної постановки задачі оцінювання ефективності нових знань.

Аналіз останніх досліджень та публікацій

Види кількісного оцінювання інформації можна поділити на три групи: синтаксичні, семантичні та прагматичні.

1. Синтаксичне вимірювання кількості інформації ґрунтується на взаємозв’язку між інформацією і ентропією – мірою невизначеності. Якщо апріорно ситуація характеризується ентропією H_0,

а після одержання повідомлення S ентропія зменшилася до H_1, то кількість інформації, що міститься у повідомленні S (позначення I_S), визначається різницею

$$I_S = H_0 - H_1. \quad (1)$$

Отже, кількість інформації вимірюється зменшенням невизначеності ситуації у результаті отримання повідомлення S. Що більшим є число можливих повідомлень, то більшою була апріорна невизначеність H_0 і то більшу кількість інформації одержує адресат, коли ця невизначеність знижується.

Величина ентропії на один елемент повідомлення визначається за формулою К. Шеннона:

$$H = -\sum_{i=1}^{m} p_i \log p_i, \quad (2)$$

де m – кількість можливих станів елементу повідомлення; p_i – ймовірність того, що елемент перебуває в i-му стані.

2. Семантичні міри інформації. Ця категорія мір призначена для вимірювання змісту I_{sem} повідомлень, які одержує адресат інформації. Оскільки зміст повідомлень кожний конкретний їх адресат сприймає по-різному (залежно від об'єму інформації, якою він володіє), то саме "запас знань" адресата і покладено в основу семантичних мір інформації. Отже, семантичні міри враховують не тільки зміст повідомлень, але і те, який об'єм нової інформації несе повідомлення його адресату порівняно з тим, що він вже знав.

Методи кількісного вимірювання змісту інформації ще достатньо не розроблено. Найбільшого визнання здобула запропонована Ю.І. Шнейдером тезаурусна міра, у якій семантична ємність Ісем інформації пов'язується із здатністю її одержувача сприймати (асимілювати) відомості, що надходять, а це залежить від об'єму знань адресата інформації, або, як прийнято говорити, від його тезауруса.

Максимального значення семантичність інформації набуває за погодженості тезаурусів, коли інформація, що надходить, по-перше, є зрозумілою адресату і, по-друге, несе йому невідомі раніше (відсутні у його тезаурусі) відомості.

Отже, кількість семантичної інформації у відомостях (тобто кількість нових знань, що їх отримує адресат) є величиною відносною: одне й те саме повідомлення може мати зміст для компетентного і бути беззмістовним (семантичним шумом) для некомпетентного адресата. Водночас інформація, яка хоча є зрозумілою, але відомою компетентному адресатові, є для нього також семантичним шумом. Іншими словами, відомості, які не є новими, несуть нульову кількість інформації.

Отже, зміст інформації можна вимірювати ступенем зміни тезауруса адресата інформації під впливом отриманого повідомлення.

У зв'язку з цим дуже важливими є задачі моделювання тезаурусів різних інформаційних систем, розроблення взаємоузгоджених тезаурусів у системі "людина-машина", вирішення питань про те, яким повинен бути тезаурус у людей, що займаються тією чи іншою проблемною сферою (економікою, фінансами, математикою, програмуванням, інженерним проектуванням, медициною, біологією тощо).

3. Прагматичні міри інформації. Ці міри інформації відображають цінність інформації, її корисність і доцільність для досягнення поставленої мети.

Прагматичні міри, як і семантичні, є відносними, і вибір міри зумовлюється особливостями використання певної інформації у певній системі.

Як правило, цінність інформації залежить від того, ким і з якою метою вона використовується. Наприклад, повідомлення, що завтра буде значна хмарність, має різну цінність для торговельної організації і для авіації.

Відомі два основні методи визначення кількісної міри цінності інформації. Якщо мета напевне є досяжною і до того ж декількома можливими шляхами, то зручною є міра цінності, запропонована Р.Л. Стратоновичем. Вона полягає в оцінюванні умовних "штрафів" (або витрат часу, засобів, грошей та ін.) і вимірюється зменшенням витрат у результаті досягнення цілей. Так,

наприклад, кількісною мірою цінності інформації, яка стосується поїздки у місто (розклад руху транспорту, вартість квитків тощо), може слугувати зекономлений час і (або) гроші.

Якщо, навпаки, досягнення мети малоймовірне, то зручніше користуватися критерієм, який запропонували Н.М. Бонгарт і А.А. Харкевич. Мірою цінності при цьому є логарифм відношення ймовірностей досягнення мети до одержання інформації (P_{in}) і після цього (P_{fin}):

$$I_{\text{ц}} = \log \frac{P_{fin}}{P_{in}}. \qquad (3)$$

Що стосується оцінювання новизни знань, то у роботі [3] нами був запропонований підхід, що ґрунтується на використанні лексичного словника WordNet [4]. Однак існує проблема розроблення єдиного уніфікованого підходу до такого оцінювання.

Формування цілей

Здійснити формальну постановку задачі оцінювання ефективності нових знань.

Основний матеріал

Проблема автоматизованого формування онтологій БЗ супроводжується виникненням таких проблем:

1) зменшенням якості нарощуваного обсягу інформації (релевантності до заданої ПО);

2) втратою монотонності, тобто появою внутрішніх конфліктів, які порушують її цілісність;

3) появою неконтрольованої надлишковості.

Як відзначено у [5], онтологію формують вручну, напівавтоматизовано та автоматизовано. Якщо цей процес здійснюється автоматизовано, то необхідно оптимізувати онтологію [6].

Напрями оптимізації онтології безпосередньо залежать від критерію якості ІнтС, оскільки якість таких систем залежить від БЗ [7], а в розглянутому у цій роботі підході ядром БЗ є онтологія. Для оцінювання якості ІнтС використаємо стандарт ISO 9126 [8] для інформаційних систем, оскільки спеціалізованого стандарту лише для інтелектуальних систем немає. Цей стандарт призначено або для стандартизації виробничого процесу (аналогія з життєвим циклом інформаційних систем), або для оцінювання якості програмних засобів.

Характеристики ISO 9126 визначають напрями оптимізації онтологій, які полягають в оптимізації її структури, що динамічно формується під час експлуатації системи та оптимізації змісту онтології БЗ. Під змістом розуміємо інформаційне наповнення, яке має бути гнучким, тобто налаштовуваним під конкретну ПО і потреби користувача. Іншим підходом до оптимізації онтології є оцінювання новизни знань, яку ми плануємо додавати в онтологію порівняно з тими знаннями, які вже зберігаються в ній. Саме формальна постановка цієї задачі розглядається у цій роботі.

Базовими характеристиками якості ІнтС за ISO 9126 є:

- функціональна придатність до використання;
- коректність або достовірність;
- ресурсна економічність;
- практичність;
- супроводжуваність;
- мобільність.

Оскільки ми розглядаємо клас інтелектуальних систем, призначених для підтримки прийняття рішень, ядром БЗ яких є онтології, то ці характеристики напряму залежать від якості онтологій, оскільки, своєю чергою, якість ІнтС залежить від її БЗ.

Функціональна придатність ІнтС залежить від повноти онтологій, наскільки вона точно описує специфіку ПО та задач, які у ній виникають. Своєю чергою, повнота онтологій залежить від вміння давати правильні відповіді на запити до неї. Це залежить від вміння системою оцінювати новизну знань, яку пропонується додавати до онтології. Тому надалі розглянемо метод оцінювання новизни знань, який ґрунтується на використанні лексичного словника WordNet. Мірою якості

функціональної придатності буде середній відсоток нетривіальних (ненульових), правильних відповідей на запити до онтології. Тобто

$$\chi_1 = \frac{N_q^p}{N_q} \cdot 100\ \%,$$

де N_q – кількість всіх запитів до онтології БЗ; N_q^p – кількість правильних відповідей на запити.

Визначення функціональної придатності є однією з базових характеристик ІнтС.

Коректність, або достовірність функціонування ІнтС – це відсоток достовірно розв'язаних інтелектуальною системою задач. Це основна характеристика якості ІнтС, яка залежить не лише від якості БЗ, але й від моделі функціонування таких систем, тобто від побудованої метрики. Отже

$$\chi_2 = \frac{N_z^p}{N_z} \cdot 100\%,$$

де N_z – кількість задач, які розв'язала ІнтС; N_z^p – кількість правильно розв'язаних нею задач.

Використовуванність ресурсів (або *ресурсна економічність*) у стандартах відображається зайнятістю ресурсів центрального процесора, оперативної, зовнішньої та віртуальної пам'яті, каналів введення–виведення, терміналів і каналів зв'язку. Для покращення цієї характеристики розглянемо оптимізаційну задачу, критерієм якої буде фізичний обсяг пам'яті, яку займає онтологія. З іншого боку, очевидно, що онтологія займає найменший об'єм пам'яті, якщо в ній немає жодного поняття. Тому цей критерій необхідно скомбінувати з іншим критерієм, а саме з функціональною придатністю ІнтС. Отримуємо задачу оптимальної кількості понять онтології.

Практичність – важко формалізоване поняття, яке визначає функціональну придатність і корисність застосування ІнтС для певних користувачів. До цієї групи показників входять субхарактеристики, які з різних сторін відображають функціональну зрозумілість, зручність освоєння, системну ефективність і простоту використання ІнтС. Така придатність повинна ґрунтуватися на цілісності онтології, тобто відсутності в її тілі взаємозаперечувальних тверджень та дублювання, а також на збалансованості ПО, яка виражається у рівномірному поданні її окремих підрозділів в онтології.

Супроводжуваність ІнтС відображається зручністю і ефективністю виправлення, удосконалення або адаптації структури та змісту онтології БЗ залежно від змін у зовнішньому середовищі застосування, а також у вимогах і функціональних специфікаціях замовника. Узагальнено якість супроводжуваності ІнтС можна оцінювати як потребу ресурсів для забезпечення її функціональності та реалізації. Сукупність субхарактеристик супроводжуваності програмної системи, подана в стандарті ISO 9126, цілком застосовна для описання цієї якості інтелектуальних систем, переважно тими самими організаційно-технологічними субхарактеристиками.

Мобільність характеризується тривалістю і трудомісткістю інсталяції інформаційних продуктів, адаптації та заміщення при перенесенні на інші апаратні та операційні платформи. Інформація про процеси, що відбуваються у зовнішньому середовищі, може мати великі обсяги і трудомісткість первинного накопичення та актуалізації, що визначає необхідність її ретельного зберігання та регламентованої зміни. Критерієм мобільності є швидкодія, яка виражається часом відгуку ІнтС на зовнішнє звернення (час реакції на зміну параметрів зовнішнього середовища, до яких чутлива система).

Отже, основною характеристикою якості будь-якої ІнтС є достовірність отриманого розв'язку цією системою χ_2. Якщо ІнтС побудована на основі онтології, то вагомою є й перша характеристика χ_1.

Для підвищення ефективності вищенаведених шести характеристик необхідно розв'язати задачу оцінювання новизни онтологічних знань.

Нехай Z – задача, яка розв'язується за допомогою знань поданих у вигляді онтології.

$Z_j \subset Z$ – деякий підклас задачі Z.

Для розв'язування задачі Z використовується множина методів $M^Z = \left\{ M_1^Z, M_2^Z, ..., M_n^Z \right\}$.

Якість методів, які використовуються для розв'язування задачі Z, оцінюється множиною параметрів $P^Z = \left\{ p_1^Z, p_2^Z, ..., p_m^Z \right\}$.

Новий метод розв'язування задачі Z : M_{new}^{Z}.

Виграш цього методу порівняно з деяким іншим *i*-м методом, який використовується для розв'язування підкласу Z_j задачі Z : $M_i^{Z_j}$, задається як відображення в підмножину елементів, які являють собою пару (параметр, перевага):

$u\left(M_{new}^{Z_j}, M_i^{Z_j}\right) \to U_{i,new}^{Z_j} = \left\{\left(p_{i_l}^{Z_j}, \alpha_{i_l}\right)\right\}_{l=1}^{k_i}$, α – кількісний або якісний опис переваги.

Аналогічно програш $v\left(M_{new}^{Z_j}, M_i^{Z_j}\right) \to V_{i,new}^{Z_j} = \left\{\left(p_{i_s}^{Z_j}, \beta_{i_s}\right)\right\}_{s=1}^{k_i}$.

Функція виграшу нового методу порівняно з методом $M_i^{Z_j}$: $\varphi\left(U_{i,new}^{Z_j}\right)$, функція програшу нового методу порівняно з методом $M_i^{Z_j}$: $\psi\left(V_{i,new}^{Z_j}\right)$.

Читається як: Розроблено новий метод M_{new}^{Z} для розв'язування задачі Z, який на відміну від іншого методу M_i^{Z} дає змогу отримати виграш $\varphi\left(U_{i,new}^{Z_j}\right)$ та програш $\psi\left(V_{i,new}^{Z_j}\right)$.

Приклад. Нехай задача Z полягає у розв'язанні квадратного рівняння $ax^2 + bx + c = 0$.

Z^1 – підклас задачі, для якої $D = b^2 - 4ac > 0$, $a, b, c \in Z$, $a = 1$, $b \neq 0$, $c \neq 0$.

Z^2 – підклас задачі, для якої $D = b^2 - 4ac > 0$, $a, b, c \in Z$, $a \neq \{0,1\}$, $b \neq 0$, $c \neq 0$.

$M^Z = \left\{M_1^Z = '\textit{дискримінант}', ...\right\}$. $P^Z = \left\{p_1^Z = '\textit{час розв'язування}', ...\right\}$.

$M_{new}^{Z} = '\textit{теорема Вієта}'$. Тоді

$u\left(M_{new}^{Z_1}, M_1^{Z_1}\right) \to U_{1,new}^{Z_1} = \left\{\left(p_1^{Z_1}, 30c\right)\right\}$. $v\left(M_{new}^{Z_2}, M_1^{Z_2}\right) \to V_{1,new}^{Z_2} = \left\{\left(p_1^{Z_2}, 40c\right)\right\}$.

Тобто $\alpha_1 = 30c$, $\beta_1 = 40c$.

Функція виграшу: $\varphi = \alpha_1 = t_1 - t_{new}$ – різниця між часом розв'язування задач старим та новим методом.

Висновки

Здійснено формальну математичну постановку задачі оцінювання новизни онтологічних знань з погляду підвищення ефективності функціонування інтелектуальних систем, ядром баз знань яких є онтологія. Визначено основні характеристики інтелектуальних систем згідно із ISO 9126, що дало змогу обґрунтувати доцільність розв'язування такої задачі. Оцінювання новизни знань, на відміну від оцінювання новизни інформації, ґрунтується на функції виграшу, яку може отримати користувач системи, використовуючи ці знання. Наведено приклад, який ілюструє розроблену математичну модель задачі.

1. Литвин В.В. Бази знань інтелектуальних систем підтримки прийняття рішень / В.В. Литвин. – Львів: Видавництво Львівської політехніки, 2011. – 240 с. 2. Шеннон К. Работы по теории информации и кибернетике / К. Шеннон. – М.: 1963. – 830 с. 3. Литвин В. В. Оцінка новизни знань під час автоматичної розбудови онтологій / В. В. Литвин, А. С. Мельник, В. Я. Крайовський // Вісн. Нац. ун-ту "Львівська політехніка". "Інформаційні системи та мережі". – 2011. – № 699. – С. 343–353. 4. Miller G.A. WORDNET: A lexical database for English / G.A. Miller // Communications of ACM (11). – 1995. – P. 39–41. 5. Інтелектуальні системи, базовані на онтологіях // Д.Г. Досин, В.В. Литвин, Ю.В. Нікольський, В.В. Пасічник. – Львів: "Цивілізація", 2009. – 414 с. 6. Литвин В. В. Задачі оптимізації структури та змісту онтології та методи їх розв'язування / В. В. Литвин // Вісн. Нац. ун-ту "Львівська політехніка". "Інформаційні системи та мережі". – 2011. – № 715. – С. 189–200. 7. Гаврилова Т.А. Базы знаний интеллектуальных систем / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб.: Питер, 2001. – 384 с. 8. ISO/IEC 9126:1991. Information technology – Software product evaluation – Quality characteristics and guidelines for their use. – 1991. – 39 p.