Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 2 (24)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

УДК 001.89:004.5:004.91:81`3
DOI: https://doi.org/10.30837/ITSSI.2023.24.027

С. Гайко

ОНТОЛОГО-КЕРОВАНІ ЗАСОБИ ОБРОБЛЕННЯ ТА ПОДАННЯ ВЕЛИКИХ МАСИВІВ НЕСТРУКТУРОВАНИХ ТЕКСТІВ

Предметом дослідження статті є методи онтолого-керованого оброблення та подання неструктурованих текстів у глобальному середовищі. **Мета роботи** – підвищення ефективності автоматичного пошуку, класифікації та вибору необхідної інформації, що міститься в електронних виданнях у неструктурованій формі шляхом розроблення моделі, методу й засобів автоматизованого оброблення та трансдисциплінарного подання текстових документів, створених українською, російською та англійською мовами. Відповідно до мети визначено такі **завдання**: розглянути моделі оброблення та подання неструктурованих текстів; виявити переваги інформаційних систем з онтолого-керованою архітектурою; розробити концептуальну модель і засоби автоматизованого оброблення та трансдисциплінарного подання текстових документів; удосконалити метод рекурсивної редукції; за допомогою розроблених методів і засобів обробити великий масив неструктурованих текстів (зокрема інформаційну базу знань наукової продукції (ІБЗ НП), навчальні програми, а також навчально-дослідницькі проєкти НЦ "Мала академія наук України"); подати їх у вигляді, що дасть змогу виявляти якість і повноту наявної в базах системи знань і, отже, експериментально підтвердити ефективність розроблених засобів. Проведені дослідження ґрунтуються на таких **методах**: системологічної класифікації, алгебро-логічний та аксіоматичний методи, метод рекурсивної редукції. Здобуто такі **результати**: описано наявні моделі оброблення та подання неструктурованих текстів; удосконалено технологічні аспекти онтолого-керованого підходу до оброблення та подання великих масивів мережних документів; розроблено технологію формування онтологій документів на основі репрезентації їх таксономій, зв'язків між їх контекстами та множинного подання, що забезпечує ефективний пошук інформації в неструктурованих текстах. **Висновки**: обґрунтовано переваги онтолого-керованих засобів оброблення та подання неструктурованих текстів; подальшого розвитку набув метод рекурсивної редукції шляхом побудови перетворення, яке приводить до предикативної форми онтологічно задані описи інформації; викладено ідеї ефективного пошуку, оброблення, класифікації та вибору необхідної інформації.

Ключові слова: неструктуровані тексти; онтологія; рекурсивна редукція; таксономія; оброблення текстів; подання інформації.

Вступ

Найпоширенішою формою подання знань і досі залишаються неструктуровані тексти. Переважна більшість видань – від наукових статей до художніх творів – створюються людьми природною мовою та в такому неструктурованому вигляді зберігаються в надвеликій кількості в глобальному цифровому середовищі. Проте обробити та використати подібні інформаційні масиви людям без машинних засобів не під силу.

Для вирішення окресленої проблеми експертами розроблено цілі класи інформаційних систем, зокрема: повнотекстові бази даних та інтелектуальний пошук; системи автоматичного індексування й рубрикації; системи автоматичного анотування й реферування; інформаційно-пошукові системи; системи машинного перекладу; системи класу "Штучний інтелект" (Текст → База знань); системи генерації тексту (База знань → Текст) [1]. Перелічені системи призначені для вирішення вузькоспеціалізованих завдань. Однак у сучасних умовах усе частіше виникає потреба у вирішенні комплексних завдань, пов'язаних із пошуком, обробленням, структуризацією та інтеграцією гетерогенної, розподіленої, неповної інформації.

Саме тому виникає необхідність у створенні та розвитку систем, що здатні обробляти великі масиви неструктурованих даних (зокрема природномовних документів) і цим надавати технологічну підтримку фахівцям різних галузей в ефективному та конструктивному застосуванні накопичених людством знань.

Аналіз проблеми й наявних методів

Аналізу методів оброблення та подання природномовних текстів (ПМТ) присвячені роботи таких дослідників: K. Jones [2], E. Liddy [3], D. Jurafsky [4], О. Барковська [5], S. Praveena [6], С. Субботін [7] та ін. Пропонуємо стислий огляд найбільш поширених підходів до оброблення та подання ПМТ.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 2 (24)

Обчислювальний підхід. Першими моделями, що розроблялися для завдань автоматичного перекладу, були: ієрархія граматик Хомського, граматики Вудса, граматики залежностей, імовірнісні автомати, моделі "Зміст – Текст" тощо. Вивчення подібних моделей сприяло створенню розвиненої теорії формальних мов.

Лінгвістичний підхід має чотири рівні. Перший полягає у виділенні окремих елементів тексту / документа, наприклад, розділів, абзаців, речень і под. Другий – у виявленні морфологічних характеристик окремого слова. Третій рівень відповідає за визначення синтаксичної залежності слів у реченнях. Останній рівень пов'язаний зі змістовим розумінням тексту, що передбачає розробки у сфері штучного інтелекту.

Статистичний підхід. В основі зазначеного підходу лежить припущення, що зміст тексту може бути визначено за найуживанішими словами. Основним завданням статистичного підходу є виявлення кількості повторень конкретного слова в тексті.

Імовірнісний підхід. Поданий такими моделями, як *N*-грами, системи, основані на деревах рішень, та ймовірнісні узагальнення контекстно-вільних граматик. *N*-грамна модель побудована на послідовності з *n* елементів: речень, слів, букв, звуків тощо. Модель дає змогу розрахувати ймовірність появи будь-якого елемента в тексті.

Символічний підхід. Цей підхід глибоко аналізує лінгвістичні явища та ґрунтується на явному поданні знань, що здійснюється шляхом використання добре досліджених схем подання знань і алгоритмів, які працюють із ними. Джерелом знання про мову можуть бути словники, формули й правила, розроблені людьми.

Коннективістський підхід. Зазначений метод оброблення природної мови відповідає за оброблення загальних моделей із використанням конкретних прикладів мовних явищ. Найбільш значуща відмінність коннективістського підходу від інших статистичних методів полягає в поєднанні статистичних знань і різних теорій уявлень, що дають змогу працювати з логічними висновками та трансформацією логічних формул.

Аналіз природномовних текстів неминуче пов'язаний із необхідністю подавати певним способом результати такого аналізу. До класичних методів подання знань належать: формальні логічні моделі, продукційні моделі, семантичні мережі, фрейми, нейронні мережі.

Формальні логічні моделі передбачають, що вся інформація, необхідна для вирішення прикладних завдань, розглядається як сукупність фактів і тверджень, що подаються як формули в деякій логіці. Знання відображаються сукупністю таких формул, а набуття нових знань зводиться до реалізації процедур логічного висновку.

Продукційні моделі (або моделі, основані на правилах) описують процедурні знання та дають змогу подати їх у вигляді пропозицій типу "якщо (умова) …, то (дія)". Механізм, реалізований як засіб висновку в продукційних системах, називається машиною логічного висновку й виконує функції пошуку в базі правил, послідовно здійснює операції над знаннями та отримує висновки.

Семантичні мережі. Це графічні системи позначень для подання знань у шаблонах пов'язаних вузлів і дуг. Більш формально: це орієнтований граф, вершинами якого є поняття, а дугами – відношення між ними. За умови індуктивного виведення на основі фактів і відношень між ними, що існують у навколишньому світі, люди використовують асоціативні зв'язки.

Фрейми. Фреймова модель є систематизованою психологічною моделлю пам'яті людини та її свідомості. Фрейм – структура даних для подання деякого концептуального об'єкта. Це мінімальна структура інформації, необхідна для подання класу об'єктів, явищ або процесів.

Нейронні мережі – один із напрямів штучного інтелекту, мета якого змоделювати аналітичні механізми, що здійснюються людським мозком. Завдання, що вирішує типова нейромережа, – класифікація, передбачення та розпізнавання. Нейромережі здатні самостійно навчатися й розвиватися, будуючи свій досвід на помилках.

Домінування окремих методів умовно поділяють на такі періоди:

I етап (кінець 1940-х – кінець 1960-х рр.) – акцентування уваги на створенні систем машинного перекладу;

II етап (кінець 1960-х – кінець 1970-х рр.) – домінування теорій штучного інтелекту;

III етап (кінець 1970-х – кінець 1980-х рр.) – розроблення обчислювальних граматик і логічного програмування;

IV етап (1990-і роки) – застосування статистичного підходу;

V етап (2000-і рр. – сьогодення) домінування експліцитних методів семантичного аналізу текстової

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 2 (24)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

інформації (алгоритми онтологічного семантичного аналізу) та методів латентно-семантичного аналізу.

Інформаційні системи з онтолого-керованою архітектурою (або знання-орієнтовані ІС) реалізують інтегровану інформаційну технологію, що передбачає комп'ютерне оброблення природномовних об'єктів, заданих лінгвістичним корпусом текстів, які описують деяку проблемну галузь, та вилучення предметно-орієнтованих знань з метою їх формально-логічного подання та автоматизованого оброблення. Зазначені системи набули особливої переваги для проведення складних міждисциплінарних наукових досліджень.

Попри широку популярність онтологічного інжинірингу серед фахівців з комп'ютерних наук, і, як наслідок, відчутну кількість фундаментальних і прикладних досліджень у цій галузі, вона залишається актуальною та перспективною [8, 9]. Особливо варто наголосити про дефіцит якісних методів і засобів онтолого-керованого оброблення природномовної інформації для україномовного сегменту.

Метою роботи є підвищення ефективності автоматичного пошуку, класифікації та вибору необхідної інформації, що міститься в електронних виданнях у неструктурованій формі, шляхом розроблення моделі, методу й засобів автоматизованого оброблення та трансдисциплінарного подання текстових документів, створених українською, російською та англійською мовами.

Вирішення завдання

Матеріали й методи

Технологічною базою для реалізації поставлених завдань застосовується web-орієнтований сервіс Когнітивна інформаційна технологія (КІТ) "Поліедр" [10], оскільки ця платформа дає змогу здійснювати такі інформаційні процеси:

– лінгвістично-семантичний аналіз мережних інформаційних ресурсів, що мають значну кількість міждисциплінарних відношень, та створення на основі використання різних інформаційних технологій і стандартів;

– трансдисциплінарну інтеграцію з іншими мережними інформаційними системами та web-орієнтованими застосунками;

– таксономізацію наративів довільних документів і відображення їх структури, зокрема міжконтекстних зв'язків;

– створення онтологічних інтерактивних документів;

– виявлення латентної інформації в інформаційних ресурсах, що аналізуються;

– глибинне й машинне навчання (*Deep Learning, Machine Learning*);

– підтримку форматів і протоколів *Semantic Web*;

– опрацювання великих даних (*Big Data*).

"Поліедр" має значну кількість застосунків і сервісів, об'єднаних у багатофункціональні модулі різного призначення. Базовими застосунками є такі: переглядач онтологій, редактор онтологій, пошукова призма, ранжування альтернатив, онтологічне автоматизоване робоче місце. Базовими сервісами є: індексатор (призначений для повнотекстового пошуку інформації у великих масивах структурованих і неструктурованих документів), агрегована таблиця (призначений для зберігання великих масивів структурованої інформації), конспект (призначений для структуризації неструктурованих документів шляхом виділення з них термінів), рекурсивний редуктор (призначений для структуризації слабкоструктурованої та неструктурованої інформації за допомогою спеціалізованих правил, заданих у форматі λ-виразів) тощо.

Усі модулі можуть додатково модифікуватися шляхом розроблення та підключення онтологічних шаблонів подання, унаслідок чого користувач із правами адміністрування може створювати нові, більш спеціалізовані застосунки, спрямовані на виконання конкретних завдань.

Для вирішення завдань, сформульованих у цій статті, створено спеціалізований застосунок "система трансдисциплінарного подання інформаційних ресурсів різних стилів", що ґрунтується на засадах трансдисциплінарності [11–13] та гіпотезі про ймовірність розвинення сервісу "рекурсивний редуктор" шляхом доповнення її класифікатором стилів видань і створенням бази еталонних видань різних стилів для ефективнішого оброблення різнорідної інформації [14].

Концептуальна модель автоматизованого оброблення та подання неструктурованих текстів

Запропонована концептуальна модель є описом технології автоматизованого оброблення різностильових документів глобального середовища й подання їх у формі, придатній для подальшого їх перетворення в активний формат інтерактивних баз знань.

Структура моделі має три складники:

– модель онтолого-керованої ідентифікації інформації;

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 2 (24)

– модель класифікації природномовного тексту за стилем;

– модель трансдисциплінарної інтеграції інформаційних ресурсів.

Розглянемо кожний із трьох складників.

Модель онтолого-керованої ідентифікації інформації

Для ідентифікації інформації може бути використаний метод рекурсивної редукції [15]. У цьому разі основним недоліком методу є те, що правила, застосовані для задання шаблонів ідентифікації, подані в предикативній формі, не зручній для користувача. Ефективність процесу створення бази правил можна значно підвищити, побудувавши перетворення, що приводило б до предикативної форми онтологічно задані описи інформації, яку необхідно ідентифікувати. У такому разі можна говорити про процес онтолого-керованої рекурсивної редукції.

Онтологія, що задаватиме такий опис, має максимально повно описувати предметну галузь (ПГ), тобто бути активною онтологією такого вигляду [16]:

$$O = \langle X, R, F, A, (D, R_S) \rangle, \quad (1)$$

де X – кінцева множина концептів (понять) заданої ПГ;

R – кінцева множина семантично значущих відношень між концептами ПГ;

$F : X \times R$ – кінцева множина функцій інтерпретації (дій), заданих на концептах і/або відношеннях;

A – кінцева множина аксіом, що використовуються для запису завжди істинних висловлювань (визначень і обмежень) у термінах операціональної тематики ПГ;

D – кінцева множина додаткових визначень понять у термінах операціональної тематики ПГ;

R_S – кінцева множина обмежень, що визначають сферу дії понятійних структур в операціональному середовищі людини.

Однак користувачам зручно редагувати стандартні онтології вигляду (2). Для забезпечення максимальної ефективності процесу створення правил потрібно інтерпретувати множину атрибутів стандартної онтології як елементи активної онтології (1) [17].

$$O_{St} = \langle X, R, F(A'), A(A'), D(A'), R_S(A') \rangle. \quad (2)$$

Для цього необхідно побудувати перетворення інтерпретації атрибутів:

$$O_S \xrightarrow{G_{O_t}} O_{St}. \quad (3)$$

Активна онтологія вигляду (1) може бути перетворена в набір правил для рекурсивного редуктора за допомогою спеціалізованого перетворення:

$$R_s(A') \xrightarrow{G_{rd}} f_{ap}^{g}, \quad (4)$$

$$F(A') \xrightarrow{G_{rd}} f_{tr}^{g}. \quad (5)$$

Складник формування функції застосовності (4) є достатньо простим і зводиться до інтерпретації відповідних атрибутів як предикативних виразів, після чого метод рекурсивної редукції може використовуватися без змін. Складник формування функції перетворення (5) потребує модифікації безпосередньо методу для того, щоб динамічно враховувати задані онтологією формати результату.

Насамперед функція перетворення f_{tr}^{g} призначена для формування об'єктів і відношень у результуючій онтології (1). Однак у завданні оброблення текстів різних стилів додатково потрібна класифікація документів. Для коректної класифікації необхідно, крім власне назв об’єктів, виділяти і їх контексти. Для кожного з компонентів оператора редукції функції перетворення, що містяться у відповідних правилах, матимуть певну структуру.

Функція створення об’єкта використовується в правилах, що визначають оператор ідентифікації об’єкта F_x, і має вигляд

$$f_{tr}^{g}(l_1,...,l_n) = X(N(l_1,...,l_n)), \quad (6)$$

де X – операція створення об’єкта із заданою назвою.

Функція створення зв’язків є більш складною (7). Ці функції застосовуються в правилах, що визначають оператор ідентифікації відношень. Завдання цієї функції полягає у визначенні двох об’єктів і створенні зв’язку між ними.

$$f_{tr}^{g}(l_1,...,l_n) = R(X(N(l_1,...,l_m)), X(N(l_{m+1},...l_n))), \quad (7)$$

де X – операція створення (або виділення) об’єкта із заданою назвою;

R – операція створення зв’язку між двома заданими об’єктами;

m – індекс, який вказує межу між іменами об’єктів у вхідній послідовності лексеми.

Функція створення атрибутів має вигляд (8). Цей тип функції часто потребує спеціалізованих процедур для визначення об’єкта, до якого належить атрибут.

$$f_{tr}^{g}(l_1,...,l_n) = A(N(l_1,...,l_m), N(l_m,...,l_n)), \quad (8)$$

де A – операція створення атрибута за його ім’ям та значенням;

m – індекс, що вказує межу між іменем та значенням.

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 2 (24)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Модель класифікації природномовного тексту за стилем

Для вирішення завдання класифікації текстів за стилями може використовуватися механізм зростання пірамідальної мережі (ЗПМ) [18]. Цей спосіб забезпечує формування понять, ієрархічне упорядкування та класифікацію вхідної інформації на основі виявлення існування об'єднувальної властивості, що є загальною для двох понять.

Для здійснення класифікації необхідний власне опис стилів – у вигляді онтологічного реєстру. Для побудови онтологічного реєстру стилів експертним чином був оброблений масив наукових, навчальних, законодавчих, відомчих, публіцистичних і частково художніх документів. Унаслідок цього була побудована таксономічна структура функціональних стилів мови, а також визначені атрибути об'єктів, що належать до основних стилів.

Такий реєстр може бути поданий як активна онтологія за схемою (1), і в подальшому застосовуватися як база правил для методу онтолого-керованої рекурсивної редукції.

Об'єктами цього класифікатора є типові види документів, що належать до відповідного стилю. Об'єкти мають атрибути, що й описують особливості оформлення тексту конкретних документів.

Отже, можна задати особливості текстової розмітки типових документів, таких як закони, стандарти, положення, інструкції, наукові звіти тощо.

ЗПМ може формуватися на основі цього реєстру, результатів роботи методу онтолого-керованої рекурсивної редукції:

$$O \bigcup O_S \xrightarrow{G_T} T_S, \tag{9}$$

де O – онтологія, що містить результати методу рекурсивної редукції;

O_S – онтологічний реєстр стилів;

T_S – класифікаційна таксономія, що має структуру:

$$T_S = \langle X_T, R_T \rangle, \tag{10}$$

де X_T – множина об'єктів, що належать таксономії;

R_T – множина зв'язків, що містяться в таксономії.

Множина X_T складається з об'єктів X_S початкової онтології O_S, а також множини об'єктів, що є класифікаційними властивостями відповідних стилів (побудованих на основі атрибутів A'_S об'єктів X_S):

$$X_T = X_S \bigcup X\left(A'_S\right). \tag{11}$$

Множина R_T містить взаємозв'язки між об'єктами X_T, що є стилями, і об'єктами $X\left(A'_S\right)$, які є класифікаційними характеристиками об'єктів:

$$R_T = R\left(X_S, X\left(A'_S\right)\right). \tag{12}$$

Для таксономії T_S можна побудувати ЗПМ, яка буде унівалентною до неї:

$$T_S \xrightarrow{G_\Psi} \Psi_S, T_S \cong \Psi_S. \tag{13}$$

Модель трансдисциплінарної інтеграції інформаційних ресурсів

Принципи трансдисциплінарної інтеграції політематичних інформаційних ресурсів глобального середовища забезпечують процеси консолідованої взаємодії складних інформаційних систем – коректного використання мережних інформаційних ресурсів, що створені за різними стандартами, мають неоднакові формати й обробляються різними інформаційними системами, які не мають спільних інтерфейсів.

Для реалізації такої інтеграції пропонується застосування онтологічного шаблону подання разом із розширенням онтологій контекстами та використанням до них функцій пошуку й контекстної зв'язності [19].

Онтологічний шаблон подання – це вид онтології, призначений для визначення додаткових модулів натуральної системи. Ці модулі можуть бути включені в структуру системи, змінюючи її поведінку відповідно до вимог завдання.

Необхідно, щоб узагальнена інформаційна модель натуральної системи, яка підтримує онтологічні шаблони подання, мала такий вигляд:

$$\Pi_{SN} = \sum_{i=0}^{n} \Pi^i_{SN} \bigcup G_T\left(O_D\right), \tag{14}$$

де Π^i_{SN} – стандартні модулі, що забезпечують основні функції оброблення та відображення даних;

G_T – перетворення інтерпретації онтологічного шаблону подання;

O_D – онтологічний шаблон подання.

Основним стандартним модулем такої натуральної системи є системний контролер, що спостерігає за взаємодією між іншими модулями, оскільки вони виконують цільову функцію системи. Системний контролер забезпечує перетворення інтеграції функцій окремих модулів.

$$G_C : \bigcup_{i=0}^{n} S_i \bigcup S_T\left(O_D\right) \to \breve{f}, \tag{15}$$

де G_C – перетворення інтеграції;

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 2 (24)

S_i – функції стандартних модулів Π^i_{SN};

$S_T(O_D)$ – функції, отримані внаслідок інтерпретації онтологічного шаблону подання.

Інтерпретація поділяється на два паралельні процеси – створення функцій оброблення даних і функцій відображення даних.

$$S_T(O_D) = \sum_{x \in X_D} G_Q(x) \bigcup \sum_{x \in X_D} G_D(x), \quad (16)$$

де X_D – сукупність об'єктів, що належать до онтології O_D;

G_Q і G_D – перетворення інтерпретації, призначені для створення функцій оброблення даних і подання відповідно.

Завдання трансдисциплінарної інтеграції інформації вирішується шляхом перетворення розширення контекстами об'єктів онтології, здобутої внаслідок роботи методу рекурсивної редукції.

$$W^T \to X_t(W^T), \quad (17)$$

де W^T – контекст;

X_t – концепт, що міститься в контексті.

У подальшому такі онтології використовуються як вхідні дані для функції пошуку.

$$Q_S(H, x) = \left\{ H \left\langle \left\{ V(l) \times V(x) \right\} \right\rangle \right\}, \quad (18)$$

де H – індекс, отриманий унаслідок індексації масиву наративів мережних документів за допомогою спеціалізованої функції індексації Q_H;

$V(l), V(x)$ – ідентифікатори лексеми l і об'єкта x, що належать онтології O.

Функція пошуку дає змогу формувати зв'язки між контекстами всіх лексичних одиниць безлічі мережних документів. Ця операція є основною для побудови функції контекстної зв'язності:

$$Q_C(L_x, C) = \bigcup_{O \in C} \left(Q_S\left(Q_H(O), L_x\right) \right), \quad (19)$$

де C – множина онтологій O, що описують одну або кілька ПГ;

L_x – текстове подання контексту лексичної одиниці x, що належить таксономії T.

Порядок взаємодії розроблених підсистем у межах спеціалізованого застосунку

Для реалізації спеціалізованого застосунку "система трансдисциплінарного подання інформаційних ресурсів різних стилів" створено такі модулі (підсистеми):

– модуль інтерпретації онтології стилів;

– модуль формування ЗПМ;

– модуль класифікації тексту за стилем.

Взаємодія підсистем відбувається в певній послідовності.

1) Під час запуску спеціалізованого застосунку онтологія стилів витягується зі сховища та за допомогою модуля інтерпретації онтології стилів перетворюється в базу правил редукції.

2) Запускається редукція ПМТ.

3) На основі результатів редукції формується ЗПМ у відповідному модулі.

4) Створена ЗПМ подається до модуля класифікації тексту за стилем, де і відбувається класифікація тексту на основі операцій порівняння.

5) Отримана класифікація подається на вхід рекурсивного редуктора, що дає змогу повторити редукцію з більшою точністю та дістати кінцевий результат – інформаційну онтологію.

6) Сформована онтологія відображається інтерфейсом КІТ "Поліедр" у вигляді інтерактивного документа.

7) Елементи отриманого інтерактивного документа можуть бути пошуковим запитом, виконавши який за допомогою індексації, система надає користувачеві трансдисциплінарно інтегрований масив інформації.

Тобто, основний процес системи оброблення та трансдисциплінарного подання ІР різних стилів полягає в розширенні функціональності та підвищенні ефективності рекурсивної редукції шляхом застосування дворівневої схеми оброблення (з додатковою класифікацією вхідних текстів за стилем), що значно підвищує точність вихідного результату та зменшує кількість помилок оброблення.

Приклад застосування системи автоматизованого оброблення та подання великих масивів неструктурованих текстів

Великим масивом неструктурованих текстів обрано інформаційну базу знань наукової продукції (ІБЗ НП), навчальні програми, а також навчально-дослідницькі проєкти НЦ "Мала академія наук України". Трансдисциплінарне подання цих документів допоможе:

– синхронізувати інформаційні ресурси, що відображають науково-технічну продукцію (НТП) з навчальними програмами для визначення якості та повноти наявної в ІБЗ НП системи знань;

– синхронізувати інформаційні ресурси, що відображають науково-технічну продукцію (НТП)

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 2 (24)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

з навчально-дослідницькими проєктами НЦ "МАНУ" для визначення якості цих проєктів.

В обох випадках синхронізація здійснюється між ресурсами в межах ІБЗ НП, що означає необхідність внесення в неї програм і проєктів відповідно. Оскільки обидва типи документів є слабкоструктурованими, витриманими в різних стилях, до них спочатку застосовуються технологічні засоби структуризації.

Приклад навчальної програми, що обробляється, показано на рис. 1. Така програма містить значну кількість інформації, що загалом зводиться до переліку ключових слів. Ці слова необхідно виділити з тексту програми, класифікувати й налаштувати процеси автоматизованої оцінки наявних в ІБЗ НП інформаційних ресурсів на її основі.

Компетентнісний потенціал навчального предмета

1. Спілкування державною (і рідною у разі відмінності) мовами	***Уміння*:** усно й письмово тлумачити біологічні поняття, факти, явища, закони, теорії; описувати (усно чи письмово) експеримент, послуговуючись багатим арсеналом мовних засобів — термінами, поняттями тощо; обговорювати проблеми біологічного змісту. ***Ставлення:*** усвідомлення значущості здобутків біологічної науки, зокрема пошанування досягнень українських учених; прагнення до розвитку української біологічної термінологічної лексики. ***Навчальні ресурси:*** навчальні, науково-популярні, художні тексти про природу, дослідницькі проекти в галузі біології, усні / письмові презентації їх результатів

Рис. 1. Приклад фрагмента навчальної програми

Оскільки в межах поставленого завдання структура самої програми не має значення, це дає змогу обробляти їх за спрощеною схемою – з допомогою лексикографічних модулів. Відповідний дескриптор структуризації показано на рис. 2.

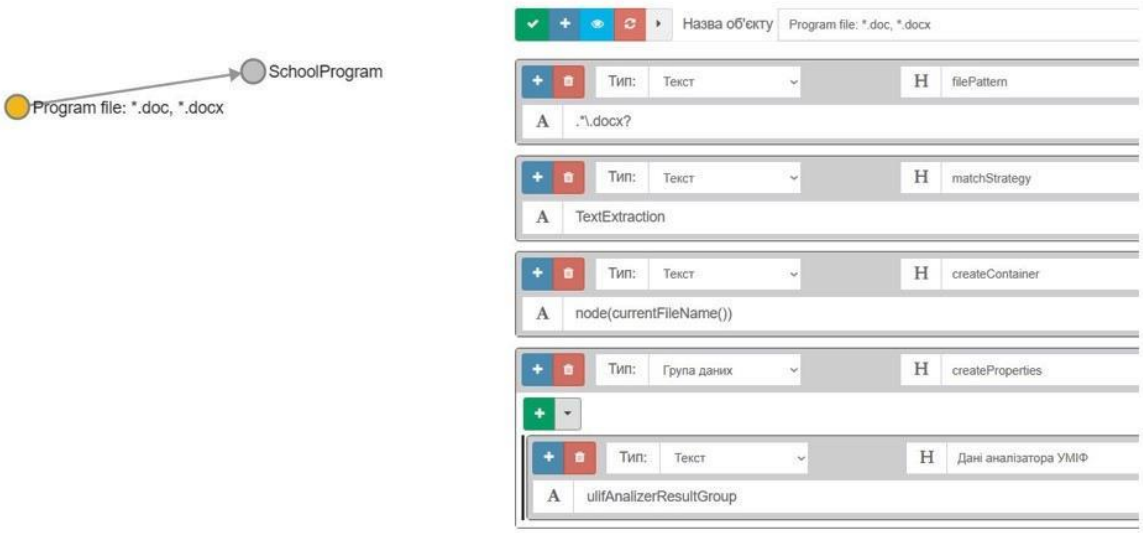

Рис. 2. Онтологічний дескриптор структуризації навчальної програми

Наведений варіант дескриптора дає змогу робити одну інформаційну онтологію на основі масиву програм. Унаслідок утворюється онтологія (рис. 3), що містить самі програми як об'єкти, а отримані з них терміни – як атрибути. Експерт може використовувати такі онтології разом із наявними механізмами повнотекстового пошуку та встановлювати ступінь відповідності матеріалів програми.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 2 (24)

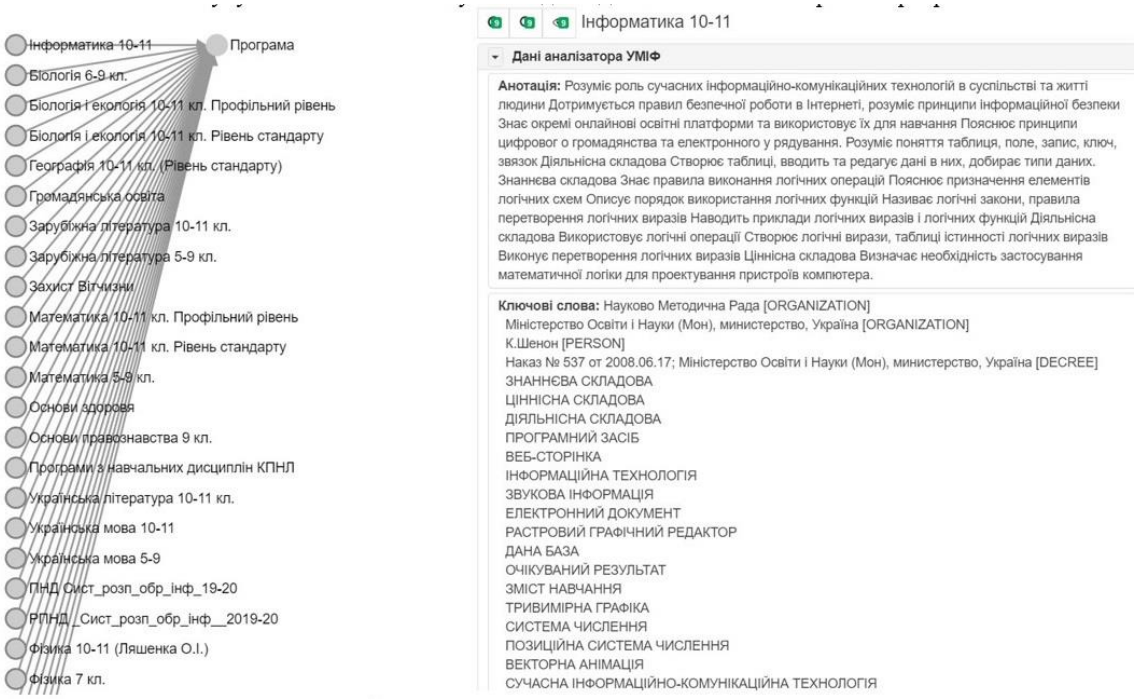

Рис. 3. Результат структуризації масиву навчальних програм

За результатами такого аналізу доповнюються або модифікуються наявні в ІБЗ НП інформаційні ресурси.

Для навчально-дослідницьких проєктів НЦ "МАНУ" використовується дещо інший підхід, що враховує структуру документа. Крім власне ключових слів, що також виділяються з таких проєктів, аналізується і структура безпосередньо документа та його складові частини.

У процесі створення онтології (рис. 4) до атрибутів об'єкта додається власне вихідний текст учнівського проєкту, розбитий на атрибути типу "Група" відповідно до вкладеності розділів. Таким чином під час аналізу роботи експерт може переглядати не тільки ключові слова, але й контекст, у якому ці вони вживаються в роботі. Такий підхід забезпечує більш повне відображення роботи.

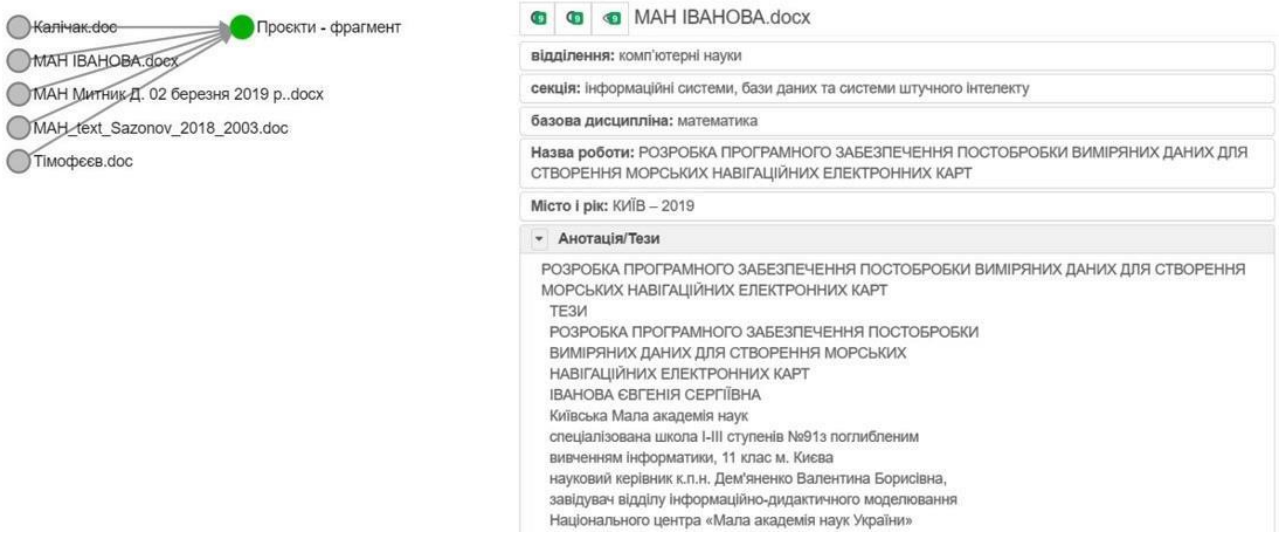

Рис. 4. Приклад обробки навчально-дослідницького проєкту

Зазначені вище термінологічні онтології не є кінцевим результатом. Вони є першим кроком до ефективного подання знань про НТП. Другим кроком є формування онтології пошукового запиту, що може здійснюватися в кілька способів. Найпоширеніший з них – оброблення термінологічної онтології

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 2 (24)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

експертно (рис. 5). Експерт відбирає з наявного списку ті терміни, що вважає релевантними своєму завданню, і вносить їх в електронну таблицю.

Сформована онтологія (рис. 6) відображає бачення експерта на структуру його завдання. Залежно від структури завдання може мати певну ієрархію або бути однорівневим (як показано на рисунку).

Така онтологія використовується для інтегрованого подання інформації про НТП у форматі онтологічної пошукової призми (рис. 7). Кожна грань такої призми є певним елементом ПГ у межах завдання експерта, а елементи на грані призми є інформаційними ресурсами, релевантними цьому елементу.

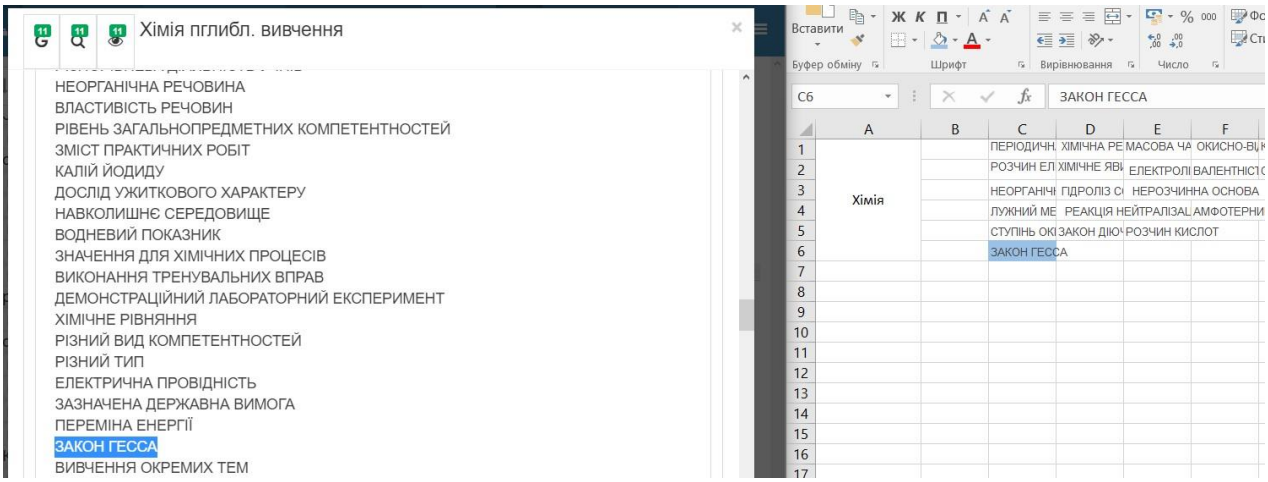

Рис. 5. Формування онтології пошукового запиту

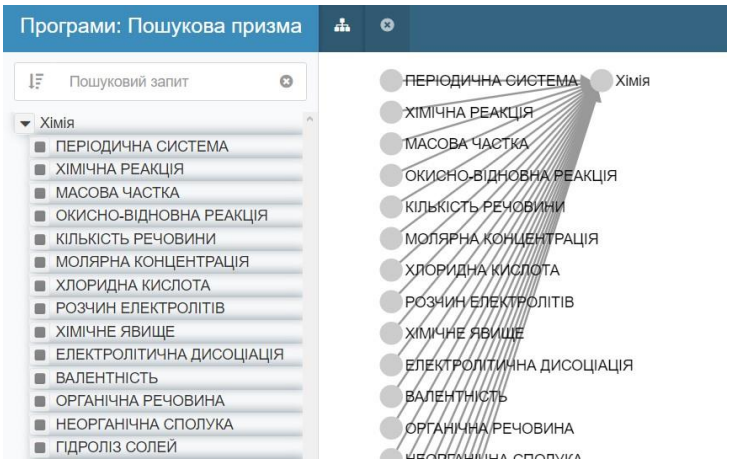

Рис. 6. Онтологія пошукового запиту

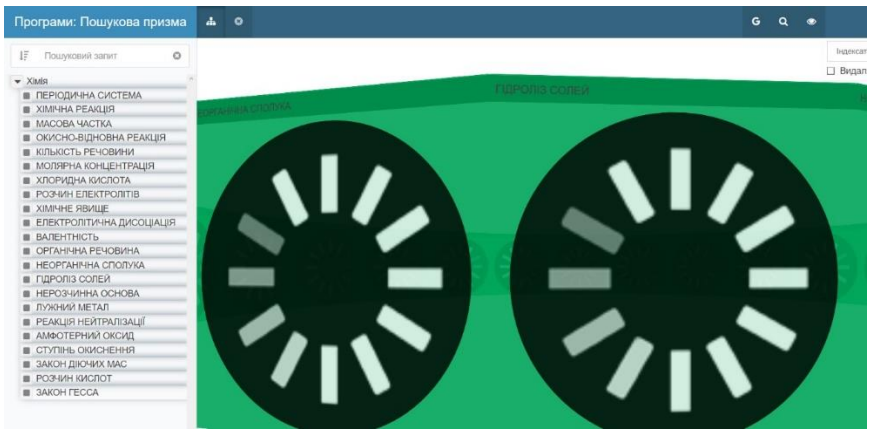

Рис. 7. Онтологічна пошукова призма НТП

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 2 (24)

Висновки

Отже, у статті вирішена актуальна науково-технічна проблема систематизації великих масивів неструктурованих текстів у глобальному середовищі шляхом розроблення моделі, методу й засобів їх трансдисциплінарного подання та інтеграції.

Обґрунтовано переваги онтолого-керованих засобів оброблення та подання неструктурованих текстів.

Набув подальшого розвитку метод рекурсивної редукції шляхом побудови перетворення, що приводить до предикативної форми онтологічно задані описи інформації.

Викладені ідеї ефективного пошуку, оброблення, класифікації та вибору необхідної інформації.

Перспективи подальшого розвитку

Розглядається можливість поширення здобутих результатів проєктування засобів онтолого-керованого трансдисциплінарного подання великих масивів неструктурованої інформації для оптимізації роботи експертів за необхідності вилучення із множини текстових документів знань, релевантних до заданої предметної галузі та здійснення їх системно-онтологічної структуризації. Структуроване подання інформації є важливим етапом для подальшої реалізації функцій систем підтримки прийняття рішень. Тому знання-орієнтовані експертні бібліотеки і репозиторії галузевих документів можуть створюватися у сфері державного управління, промисловості, науки, освіти, засобів масової інформації тощо.

Список літератури

1. Павленко П. М., Філоненко С. Ф., Бабіч К. С. та ін. Інформаційні системи і технології. Київ : НАУ, 2013. 324 с.

2. Jones K. S. Natural language processing: a historical review. 1994. URL: https://aclanthology.org/www.mt-archive.info/Zampolli-1994-Sparck-Jones.pdf (дата звернення: 10.05.2021)

3. Liddy E. D. Natural Language Processing. *Encyclopedia of Library and Information Science*. NY : Marcel Decker, Inc. 2001.

4. Jurafsky D., Martin G. H. Speech and Language Processing. Prentice Hall. 2000. URL: https://web.stanford.edu/~jurafsky/slp3 (дата звернення: 10.05.2021)

5. Barkovska O., Khomych V., Nastenko O. Дослідження методів обробки та аналізу тексту при організації електронних сховищ інформаційних об'єктів. *Сучасний стан наукових досліджень та технологій в промисловості*. 2022. № 1 (19). С. 5–12. DOI: https://doi.org/https://doi.org/10.30837/ITSSI.2022.19.005

6. Praveena S., Justus S. A Study on Knowledge Representation Models. *European Journal of Molecular and Clinical Medicine*. 2020. вип. 7. № 4. С. 2446–2452.

7. Субботін С. О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень. Запоріжжя: ЗНТУ, 2008. 341 с.

8. Yang L., Cormican K., Yu M. Ontology-based systems engineering: A state-of-the-art review. *Computers in Industry*. 2019. вип. 111. С. 148–171. DOI: https://doi.org/https://doi.org/10.1016/j.compind.2019.05.003

9. Басюк Т. М., Досин Д. Г., Литвин В. В. *Онтологічний інжиніринг*. Львів : Вид-во Львівської політехніки. 2017. 224 с.

10. Інструкція користувача КІТ "Поліедр". 2020. URL: https://storage.ulif.org.ua/storage/instructions/polyhedron_instruction.pdf (дата звернення: 10.05.2021)

11. Lawrence M., Williams S., Nanz P., Renn O. Characteristics, potentials, and challenges of transdisciplinary research. *One Earth*. 2022. вип. 5. № 1. С. 44–61. DOI: https://doi.org/https://doi.org/10.1016/j.oneear.2021.12.010

12. Renn O. Transdisciplinarity: Synthesis towards a modular approach. *Futures*. 2021. вип. 130. С. 1–18. DOI: https://doi.org/https://doi.org/10.1016/j.futures.2021.102744

13. Гончар А. В., Стрижак О. Є., Беркман Л. Н. Трансдисциплінарна консолідація інформаційних середовищ. *Зв'язок*. 2021. № 1(149). С. 3–10. DOI: https://doi.org/10.31673/2412-9070.2021.010310

14. Гайко С. І., Приходнюк В. В. Підхід до автоматизованої структуризації освітніх ресурсів на основі методу рекурсивної редукції. *Наукові записки Малої академії наук України*. 2021. № 1 (20). С. 28–38. DOI: https://doi.org/http://doi.org/10.51707/2618-0529-2021-20-03

15. Приходнюк В. В. Технологічні засоби трансдисциплінарного представлення геопросторової інформації: дис. канд. техн. наук. Інститут телекомунікацій і глобального інформаційного простору. 2017. 267 с.

16. Стрижак О. Є., Приходнюк В. В., Гайко С. І., Шаповалов В. Б. Відображення мережевої інформації у вигляді інтерактивних документів. Трансдисциплінарний підхід. *Математичне моделювання в економіці*. 2018. № 3. С. 87–100.

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 2 (24)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

17. Гайко С., Приходнюк В. Средства трансдисциплинарного представления информационных ресурсов разных стилей. *Information Models and Analysis*. Sofia : ITHEA, 2020. вип. 9. № 1. С. 78–99.

18. Величко В. Ю. Алгоритм побудови зростаючих пірамідальних мереж у паралельному обчислювальному середовищі. *Комп'ютерні засоби мережі та системи*. 2011. № 10. С. 50–57.

19. Dovgyi S., Stryzhak O. Transdisciplinary Fundamentals of Information-Analytical Activity. *Advances in Information and Communication Technology and Systems. MCT 2019. Lecture Notes in Networks and Systems*. 2019. вип. 152. DOI: https://doi.org/https://doi.org/10.1007/978-3-030-58359-0_7

References

1. Pavlenko, P. M., Filonenko, S. F., Babich, K. S. and others. (2013), *Information systems and technologies*, Kyiv, NAU, 324 p.

2. Jones, K. S. (1994), "Natural language processing: a historical review", available at: https://aclanthology.org/www.mt-archive.info/Zampolli-1994-Sparck-Jones.pdf (last accessed: 10.05.2021)

3. Liddy, E. D. (2001), "Natural Language Processing", *Library and Information Science*, New York, P. 15–25. DOI: https://doi.org/10.1145/234173.234180

4. Jurafsky, D., Martin, G. H. (2000), "Speech and Language Processing", Prentice Hall, available at: https://web.stanford.edu/~jurafsky/slp3/ (last accessed: 11.01.2023).

5. Barkovska, O., Khomych, V., Nastenko, O. (2022), "Study of text processing and analysis methods in the organization of electronic storage of information objects", *Innovative technologies and scientific solutions for industries*, No 1 (19), P. 5–12. DOI: https://doi.org/https://doi.org/10.30837/ITSSI.2022.19.005

6. Praveena, S., Justus, S. (2020), "A Study on Knowledge Representation Models", *European Journal of Molecular and Clinical Medicine*, Vol 7, No. 4, P. 2446–2452.

7. Subbotin, S. O. (2008), *Presentation and processing of knowledge in artificial intelligence and decision support systems*, Zaporizhzhia, 108 p.

8. Yang, L., Cormican, K., Yu, M. (2019), "Ontology-based systems engineering: A state-of-the-art review", *Computers in Industry*, No. 111, P. 148–171. DOI: https://doi.org/https://doi.org/10.1016/j.compind.2019.05.003

9. Basyuk, T. M., Dosyn, D. G., Lytvyn, V. V. (2017), *Ontological engineering*, Lviv, 224 p.

10. "KIT "Polyhedron" user manual", (2020), available at: https://storage.ulif.org.ua/storage/instructions/polyhedron_instruction.pdf (last accessed: 12.08.2021)

11. Lawrence, M., Williams, S., Nanz, P., Renn, O. (2022), "Characteristics, potentials, and challenges of transdisciplinary research", *One Earth*, Vol. 5, No. 1, P. 44–61. DOI: https://doi.org/https://doi.org/10.1016/j.oneear.2021.12.010

12. Renn, O. (2021), "Transdisciplinarity: Synthesis towards a modular approach", *Futures*, Vol. 130, P. 1–18. DOI: https://doi.org/https://doi.org/10.1016/j.futures.2021.102744

13. Gonchar, A. V., Stryzhak, O. E., Berkman, L. N. (2021), "Transdisciplinary consolidation of information environments" ["Transdy`scy`plinarna konsolidaciya informacijny`x seredovy`shh"], *Connection [Zv`yazok]*, No. 1(149), P. 3–10. DOI: https://doi.org/10.31673/2412-9070.2021.010310

14. Haiko, S. I., Prykhodnyuk, V. V. (2021), "Approach to the automated structuring of educational resources based on the method of recursive reduction", *Scientific notes of the Junior Academy of Sciences of Ukraine*, No. 1 (20), P. 28–38. DOI: https://doi.org/http://doi.org/10.51707/2618-0529-2021-20-03

15. Prikhodniuk, V. V. (2017), *Technological means of transdisciplinary presentation of geospatial information [Texnologichni zasoby` transdy`scy`plinarnogo predstavlennya geoprostorovoyi informaciyi]*, Ph.D. thesis, Institute of Telecommunications and Global Information Space, 267 p.

16. Stryzhak, O. E., Prykhodnyuk, V. V., Haiko, S. I., Shapovalov, V. B. (2018), "Display of network information in the form of interactive documents. Transdisciplinary approach" ["Vidobrazhennya merezhevoyi informaciyi u vy`glyadi interakty`vny`x dokumentiv. Transdy`scy`plinarny`j pidxid"], *Mathematical modeling in economics*, No. 3, P. 87–100.

17. Gaiko, S., Prikhodnyuk, V. (2020), "Means of transdisciplinary presentation of information resources of different styles" ["Sredstva transdistsiplinarnogo predstavleniya informatsionnykh resursov raznykh stiley"], *Information Models and Analysis*, Sofia, ITHEA, Vol. 9, No. 1, P. 78–99.

18. Velichko, V. Yu. (2011), "Algorithm for building growing pyramidal networks in a parallel computing environment" ["Algory`tm pobudovy` zrostayuchy`x piramidal`ny`x merezh u paralel`nomu obchy`clyuval`nomu seredovy`shhi"], *Network and system computer tools*, No. 10. P. 50–57.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2023. No. 2 (24)*

19. Dovgyi, S., Stryzhak, O. (2019), "Transdisciplinary Fundamentals of Information-Analytical Activity", *Lecture Notes in Networks and Systems*, Vol. 152. DOI: https://doi.org/https://doi.org/10.1007/978-3-030-58359-0_7

Received 16.05.2023

Відомості про авторів / About the Authors

Гайко Світлана Іванівна – Інститут телекомунікацій та глобального інформаційного простору НАН України, молодший науковий співробітник, Київ, Україна; e-mail: svitgai@i.ua; ORCID ID: https://orcid.org/0000-0002-3564-475X

Haiko Svitlana – Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine, Junior Researcher, Kyiv, Ukraine.

ONTOLOGY-DRIVEN MEANS FOR PROCESSING AND PRESENTATION OF LARGE ARRAYS OF UNSTRUCTURED TEXTS

The **subject** of the article's research is methods of ontology-driven processing and presentation of unstructured texts in a global environment. The **goal** of the work is to improve the efficiency of automatic search, classification and selection of the necessary information contained in electronic publications in an unstructured form by developing a model, method and means of automated processing and transdisciplinary presentation of text documents created in Ukrainian, Russian and English languages. In accordance with the goal, the following **tasks** were set: to conduct an overview of models of processing and presentation of unstructured texts, to identify the advantages of information systems with an ontology-driven architecture, to develop a conceptual model and means for automated processing and transdisciplinary presentation of text documents, improve the method of recursive reduction, with the help of developed method and means to process a large array of unstructured texts (in particular information base of knowledge of scientific products (IBK SP), educational programs, as well as educational and research projects of the National Center "Junior Academy of Sciences of Ukraine"), to present them in a form that allows to reveal the quality and completeness of the knowledge system available in the databases and, thus, experimentally to confirm the effectiveness of the developed means. The conducted research is based on the following **methods**: systemological classification, algebraic-logical and axiomatic methods, the method of recursive reduction. The following **results** were obtained: the existing models of processing and presentation of unstructured texts were described, the technological aspects of the ontology-driven approach to the processing and presentation of large arrays of network documents were improved, the technology of forming ontologies of documents based on the representation of their taxonomies, connections between their contexts and multiple representations was developed, which provides effective search for information in unstructured texts. **Conclusions**: The advantages of ontology-driven means for processing and presentation unstructured texts are substantiated. The method of recursive reduction by constructing a transformation, which leads to the predicative form of ontologically given descriptions of information, gained further development. The idea of effective search, processing, classification and selection of the necessary information has gained further development.

Keywords: unstructured texts; ontology; recursive reduction; taxonomy; texts processing; presentation of information.

Бібліографічні описи / Bibliographic descriptions

Гайко С. І. Онтолого-керовані засоби оброблення та подання великих масивів неструктурованих текстів. *Сучасний стан наукових досліджень та технологій в промисловості*. 2023. № 2 (24). С. 27–38. DOI: https://doi.org/10.30837/ITSSI.2023.24.027

Haiko, S. (2023), "Ontology-driven means for processing and presentation of large arrays of unstructured texts", *Innovative Technologies and Scientific Solutions for Industries*, No. 2 (24), P. 27–38. DOI: https://doi.org/10.30837/ITSSI.2023.24.027