УДК 681.3 https://doi.org/10.15407/pp2022.03-04.171

ПРОБЛЕМИ МАСШТАБУВАННЯ СЕМАНТИЧНИХ ІНФОРМАЦІЙНИХ РЕСУРСІВ ЗІ СКЛАДНОЮ СТРУКТУРОЮ

Юлія Рогушина, Ірина Гришанова

Проаналізовано проблеми масштабування, що виникають у сучасних інтелектуальних інформаційних системах (ІІС), та класифіковано причини їх виникнення у розробках. ІІС інтегрують різноманітні елементи штучного інтелекту (ШІ) для здобуття релевантних знань для задач користувачів. Важливі особливості таких ІІС – використання даних зі складною структурою та орієнтація на семантичні інформаційні ресурси (ІР). Тому ми проаналізували особливості напрямків розвитку штучного інтелекту, що концентруються на даних, та їхні можливості щодо здобуття знань з Big Data. Системи організації знань (СОЗ) забезпечують моделі та методи для ефективного збереження, пошуку та використання інформації, яка обробляється Web-орієнтованими ІІС, і ми розглянули програмні реалізації таких СОЗ. Проаналізовано особливості масштабування систем, що орієнтовані на обробку семантичної інформації, та її відмінності від традиційних та великих даних. Ці особливості викликані складністю структури даних, кількістю семантичних відношень між інформаційними об'єктами в ІР та складністю семантичних запитів, які виконуються в СОЗ.
На прикладі е-ВУЕ – Wiki-порталу Великої української енциклопедії – проаналізовано ситуації, що виникають у процесі практичного впровадження семантичних інформаційних ресурсів, які мають великий обсяг, складну структуру бази знань та підтримують одночасне виконання великої кількості різноманітних запитів. На основі цього аналізу розроблено набір рекомендацій, спрямованих на забезпечення більш ефективного масштабування таких ресурсів.

Ключові слова: семантичний інформаційний ресурс, масштабування, онтологія, Wiki-технологія, метадані, семантична розмітка.

We analyze scaling problems arising in modern intelligent information systems (IISs) and classify main reasons for their occurrence in their practical solutions. IISs integrate various elements of artificial intelligence (AI) for acquisition of knowledge relevant to actual user tasks. Important properties of these IISs are use of data with complex structure and orientation on semantic information resources (IRs). Therefore we analyze main features of the Data-Centric AI and opportunities for acquiring domain knowledge in various representations from Big Data. Knowledge organization systems (KOS) provide models and methods for effective store, retrieval and use of information processed by the Web-oriented IISs, and we consider existing approaches for their software platforms.We analyse the specifics of the scaling for systems focused on the semantic information processing and its differences from traditional data and Big Data scaling. This specifics is caused by complexity of data structure, number of various semantic relations between information objects into IR and complexity of semantic queries executed by KOS.
On example of e-VUE – the Wiki-portal of the Great Ukrainian Encyclopedia – we analyze various situations that arise in process of practical development of semantic information resources with large volume and complex structure. Various ways of semantic retrieval into this information resource that use possibilities of the Semantic MediaWiki plugin are considered from the point of view of scaling aspects (such as increase of information objects, their relations and complication of their structure and characteristics). On base of this analysis we generate a set of recommendations aimed at ensuring more efficient development of such resources and their efficient functioning for practical use.

Keywords: semantic information resource, scaling, ontology, Wiki-technology, metadata, semantic markup.

Вступ

Щороку обсяг інформації, яка генерується та використовується людством, збільшується, а її структура ускладнюється та стає все більш гетерогенною. Велике значення для обробки такої інформації мають збільшення швидкості обчислювальних пристроїв та розвиток засобів збереження даних. Одним з перспективних напрямків для ефективного використання інформації є перехід від обробки даних до обробки знань, але потрібно враховувати, що як знання використовуються для аналізу великих даних, так і самі ці дані є джерелом для генерації нових знань. Із цього випливає, що обробка великих даних безпосередньо пов'язана із створенням методів та засобів обробки знань, які мають великий обсяг та складну структуру, тобто окремі елементи пов'язуються багатьма змістовними відношеннями різноманітних типів.

Сучасні інтелектуальні інформаційні системи (ІІС) використовують та генерують знання, орієнтовані на функціювання у відкритому середовищі Web та на застосування зовнішніх джерел інформації. Ефективність обробки підвищується, якщо вони отримують відомості з семантизованих інформаційних ресурсів (ІР), в яких зміст інформації описано формальними засобами, що забезпечує їх однозначну інтерпретацію.

Орієнтований на дані штучний інтелект

Технології аналізу даних швидко змінюються. Традиційні стратегії розробки програмного забезпечення замінюються сучасними підходами, орієнтованими на методи штучного інтелекту (ШІ). Перетворення «сирих» даних на структуровані потребує багато часу та застосування експертів, тому доцільно за можливості використовувати вже структуровані ІР: таке структурування дозволяє автоматизувати їхній аналіз на семантичному рівні. Тому збільшується важливість створення складних систем організації знань (СОЗ), які мають забезпечити доступ до контенту таких ІР.

Зараз багато дослідників розглядають концепцію орієнтованого на дані (датацентричного – Data-Centric) ШІ [1] замість підходів, орієнтовних на моделі. Традиційне програмне забезпечення базується на програмному коді, тоді як системи ШІ складаються з поєднання коду та даних, і саме проблеми, що стосуються даних, є зараз найбільш актуальними для розробки інтелектуальних застосунків.

Хоча переважна більшістьіснуючої інформації зберігається у цифровому форматі, але це не означає, що ці дані можна обробляти. Щоб зробити дані придатними для використання в ІІС, їх потрібно структурувати (наприклад, побудувати з них навчальні вибірки з множини прикладів).

Протягом тривалого періоду доступність даних і обчислювальна потужність були обмежені, що потребувало оптимізації коду для розвитку ШІ. Але розвиток Big Data [2] викликав потребу переходу до підходу, ще більше орієнтованого на дані . В обробці Big Data в ІІС ключове значення має аналіз метаданих, які містять інформацію не тільки про походження інформації, а й про її семантику [3].

Орієнтований на дані ШІ спрямований на те, щоб:

– спеціалісти з обробки даних краще розуміли та контролювали структуру наборів даних і те, як ці дані обробляються (наприклад, для навчання моделі) – це полегшує визначення найкращих шляхів удосконалення на основі постійного моніторингу й аналізу продуктивності моделі, а також визначення недоліків набору даних;

– зменшення витрат на побудову моделей, зменшуючи необхідний обсяг даних або здобуваючи більше цінного з неструктурованих, різноманітних джерел даних;

– спростити анотування даних за допомогою розумніших процесів аналізу;

– виявити дублювання даних, пошкоджених або низькоякісних даних на ранніх стадіях аналізу;

– забезпечити якісні розмітки даних, уникати суб'єктивного підходу до цього.

В орієнтованому на дані ШІ досить часто застосовують такі компоненти семантичних технологій, як семантичні ІР – це підмножини ІР, в яких елементи контенту явно та однозначно пов'язуються з елементами бази знань, зокрема, за допомогою семантичної розмітки. Або, якщо елементи контенту представлено на основі форматів подання знань – наприклад, RDF та OWL та окремих випадків онтологій, таких як тезауруси, таксономії тощо [4]. ІІС орієнтовані здебільшого на обробку та створення знань, а не даних: ефективність їх роботи значним чином визначається вибором методів аналізу та форм подання знань. Тому велике значення мають системи організації знань (СОЗ) – засоби, спрямовані на упорядкування інформації та підтримку управління знаннями [5]. Такі СОЗ використовуються як концептуальна інфраструктура для підтримки цього процесу і забезпечують розуміння, інтеграцію та пошук знань, підготовку даних до здобуття з них знань, виявлення зв'язків і узагальнень, прийняття рішень на їх основі. СОЗ є інструментами для опису контенту ІР і допомоги в доступі та пошуку документів та інформації [6].

Багато Web-орієнтованих ІР, що створюються в результаті колективної діяльності користувачів, базується на технологіях Web 2.0 [7], що робить їхній контент динамічнішим та актуальним. Однією з успішних платформ Web 2.0 для колаборативного створення контенту великого обсягу є Wiki-технології [8], наприклад, MediaWiki [9]. Для таких систем можуть використовуватися СОЗ на основі Wiki-онтологій, які є окремим випадком онтологій з набором обмежень на характеристики відношень, що відображають структуру знань семантизованих Wiki-ресурсів [10]. Функціювання ІІС значною мірою залежить від того, які саме дані в них використовуються. Тому важливою передумовою їх роботи є наявність методів та засобів збирання корисних даних, а також вибір адекватних та якісних ІР.

Big Data – дані, які з різних причин не можуть оброблятися такими традиційними інформаційними системами, як реляційні бази даних. Технології великих даних зараз широко застосовуються та підтримуються значною кількістю програмних рішень. Щоб Big Data стали корисними, потрібно знаходити ті їх набори, що можуть бути використані для конкретної застосовної задачі, тобто виникає необхідність у створенні та обробці мета даних для Big Data. Але такі метадані потребують використання знань з ІР великого обсягу та складної структури. Це викликає потребу масштабування не тільки даних, а й знань, що стосуються цих даних.

Ще однією проблемою, пов'язаною з даними, є гнучкість доступу та формати їх подання. Якщо система зберігання даних накладає обмеження на зміни масштабу даних та на перехід до інших інструментів обробки, то це може призвести до негативних наслідків у процесі створення та вдосконалення ІІС. Ці проблеми визначаються не тим, чи працездатна система взагалі, а тим, чи працює вона надійно, ефективно та доступно у великих масштабах.

Дані – це лише одна з проблем, з якою мають справи розробники ІІС з елементами аналітики та ШІ у масштабному виробництві. Однак вимоги щодо даних та їхньої інфраструктури найчастіше залишаються поза увагою, хоча вони здатні унеможливити практичне застосування ІІС. Саме тому в даній роботі ми аналізуємо проблеми, пов'язані із масштабуванням даних у семантичних ІР та враховують специфіку процесів обробки інформації на семантичному рівні.

Часто-густо проблеми в таких ІІС виникають у процесі переходу від проєктування та прототипування до розгортання, промислової експлуатації та розвитку ІІС або внаслідок накопичення значного обсягу інформації. Поширені причини, через які виникає така ситуація:

1. Зміни в середовищі виконання.

2. Вимоги щодо угод про рівень сервісів (service-level agreements – SLA) – кількісні та якісні характеристики наданих сервісів, такі як їхня доступність, підтримка користувачів, час виправлення несправності тощо.

3. Обробка даних більшого масштабу – Big Data, даних з більш складною структурою, інформації з різних джерел тощо.

Усі ці проблеми пов'язані зі змінами між налаштуваннями розробки та експлуатації, оскільки середовище тестування відрізняється від робочого. Наприклад, конкретна програма може відповідати вимогам SLA щодо затримки в ході ізольованого тестування під час розробки, але ця вимога не задовольняється під час роботи у робочому середовищі, де інші програми конкурують за ресурси, а до самої програми звертається велика кількість користувачів.

Убезпечення даних теж потребує масштабування. Те, як забезпечується безпека в локальному чи малому масштабі, що використовувався під час розробки, не обов'язково виявляється надійним у великому масштабі, і це може стати несподіваним для користувачів. Традиційні концепції безпеки, такі як дозволи процесу, ідентифікатори користувачів і SELin у масштабованих системах стають набагато менш ефективними. Тому виникає потреба у застосуванні нових технологій. Наприклад, SPIFFE (Secure Production Identity Framework for Everyone) – технологія з відкритим вихідним кодом, яка забезпечує способи боротьби з небезпекою для таких великих програм шляхом визначення криптографічно підтвердженого ідентифікатора робочого навантаження для захисту каналів зв'язку між процесами.

ІІС, які розробляються зараз із застосуванням аналітики, та ШІ, орієнтовані на використання великих наборів даних, що було неможливо з даними меншого масштабу. Такі великомасштабні дані можуть бути проблемою для використання програмного продукту, але розмір даних є лише одним із аспектів масштабу, який слід враховувати, щоб побудувати успішну ІІС.

Проблеми масштабування ІІС

Масштабування застосунків, що містять елементи ШІ та аналітики, має свою специфіку [11]. Найважливіші з них – це:

- комплексна (Comprehensive) стратегія даних та уніфікований доступ до даних;
- розділення проблем на рівні платформи;
- масштабованість, а не просто масштаб;
- багатофункціональний дизайн.

Масштабування сучасних ІІС потребує врахування різних аспектів, які стосуються наступних властивостей інформації: розміру самих даних; кількості об'єктів, які обробляються; складності алгоритмів обробки та кількості програмних модулів, які використовуються для аналізу інформації; джерел інформації тощо.

Масштабування *з точки зору розміру даних* має підтримувати можливість без потреби не збільшувати його: наприклад, створювати тільки необхідні копії, забезпечувати API відкритого доступу замість локальних копій. Але такий підхід викликає потребу в уніфікації подання даних, щоб не потрібно було б створювати адаптації наборів даних до різних інструментів аналітики чи машинного навчання.

Це непотрібне копіювання особливо поширене в проектах машинного навчання та штучного інтелекту, де спеціалісти з обробки даних регулярно застосовують широкий спектр специфічних інструментів, які відрізняються від засобів Big Data, якими користуються інженери обробки даних. Інструменти ШІ та машинного навчання зазвичай не мають прямого доступу до даних, що зберігаються на платформах Big Data. Результатом є поширення зайвих копій.

Іншою причиною такого непотрібного копіювання даних є інфраструктура даних, де відсутні повністю розподілені метадані. А це може призвести до перевантаження метаданими, коли кілька програм отримують доступ до великих наборів даних.

Масштабування *з точки зору кількості інформаційних об'єктів* (ІО) – файлів або інших елементів даних – стосується можливості одночасної обробки великої кількості різних об'єктів. Якщо інфраструктура даних не призначена для обробки дуже великої кількості ІО, це може викликати значне збільшення часу обробки, перевантажити платформу та навіть вивести систему з ладу. Прикладом такої ситуації є навантаження на інтернет-магазини з великою кількістю товарів, які можуть пропонувати багато версій кожного товару та містити чимало варіантів зображень для кожного продукту. Тож для кожного звертання користувача необхідно обробляти велику кількість невеликих файлів з зображеннями та описами.

Масштабування *з точки зору засобів обробки* пов'язане з тим, що архітектура та інфраструктура обробки даних не мають обмежуватися кількома програмами на одній платформі, бо інакше потрібно налаштувати новий кластер для підтримки кожного окремого застосунку.

Масштабування *з точки зору георозподілених місць* – це застосування даних з географічно розподілених джерел або запуск програм із різних локацій. Це викликає проблеми, пов'язані із отриманням великих обсягів даних поблизу їх джерела та з прийняттям рішень щодо того, яку частину цих даних надсилати до основних центрів обробки даних, а також як і звідки надавати аналітичні програми для обробки цих даних. Приміром, система фіксує дані датчиків Інтернету речей і виконує часткову обробку чи моделювання даних на місці, а частину інформації передає для аналізу та порівняння з даними з інших джерел.

На жаль, люди можуть неправильно оцінити потенційну масштабованість своїх систем або вважатимуть нормальним розробку систему, яка успішно задовольняє їхні поточні потреби, але не розрахована на зростання цих потреб. Іноді саме вибір архітектури та інфраструктури даних накладає такі обмеження, яких можна уникнути. Отримані компроміси є частиною того, що робить створення успішного широкомасштабного штучного інтелекту та аналітики важчим, ніж це має бути. Найпоширеніші помилки у цій сфері, які заважають ефективному масштабуванню інтелектуальних застосунків:

- ШІ та аналітика мають працювати в окремих системах (кластерах);

– ІТ-команду необхідно розширювати із зростанням масштабу даних і застосунків;
– створення великомасштабних проєктів потребує багато коштів;
– різні команди чи програми створюють особисті копії тих самих даних, навіть для дуже великих наборів даних;
– переміщення даних (наприклад, між локальним сховищем та хмарним) необхідно реалізовувати на рівні застосунків;
– застарілі програми не можуть працювати безпосередньо на сучасній інфраструктурі Big Data;
– платформи Big Data призначені для спеціалізованих проєктів замість того, щоб бути універсальною загальною платформою;
– для планування та налаштування архітектури та інфраструктури, потрібно заздалегідь знати остаточний масштаб даних і програм;
– для масштабування застосунків необхідно змінювати архітектуру існуючої системи.

В [12] аналізуються актуальні проблеми БД та порівнюються відмінності між великими та традиційними даними. Таке співставлення доцільно розширити, порівнюючи їх із сучасними семантичними ІР, які базуються на технологіях Semantic Web.

Таблиця 1. Порівняння характеристик традиційних, великих та семантичних даних

Компоненти	Традиційні бази дані	Big Data	Семантичні дані
Запити	SQL	Largely Abandoned SQL	SQL-подібні запити
Архітектура	Централізована	Розподілена	Розподілена з ієрархією елементів
Типи даних	Структуровані	Структуровані, частково структуровані або неструктуровані	Формально структуровані
Модель даних	Фіксована схема	Немає схеми	Різноманітні схеми (RDF, RDF-S OWL)
Відношення між даними	Відомий фіксований набір відношень без формалізованої семантики	Невідомі або невизначені, частково подані у метаданих	Розширюваний набір довільних відношень з формалізованою семантикою
Обсяг даних	Великий	Дуже великий	Відносно малий
Кількість відношень (порівняно з обсягом даних)	Мала	Дуже незначна	Значна
Інтегрованість даних	Висока	Низька	Дуже висока
Семантика	Неформалізована	Невизначена, тільки мета описи	Формальна інтероперабельна

Як показує таблиця 1, основні проблеми масштабування семантичних ІР стосуються саме обробки у різноманітних запитах великої кількості відношень між елементами даних, які формалізовані, але викликають генерацію надвеликої кількості комбінацій поєднань елементів контенту.

Таким чином, для семантичних ІР доцільно аналізувати наступні аспекти масштабування:
– загальний обсяг даних, що зберігаються в ІР;
– засоби та інфраструктуру збереження інформації (наприклад, обсяг та потужність серверу);
– кількість ІО різних типів (природномовних текстів, структурованих даних, мультимедійних ІО – аудіо, відео, зображень тощо;
– кількість відношень між ІО;
– кількість типових ІО та складність їхньої структури;
– інфраструктура метаданих щодо ІО (засоби представлення, індексації, перегляду та пошуку);
– кількість звертань користувачів до ІР;
– швидкість актуалізації бази знань ІР після внесення змін – як у метадані, так і у контент;
– кількість операцій (пошукових запитів, отриманні інших ІО) у типових звертаннях користувачів.

Постановка задачі

У розробці ІІС, що базуються на ІР великого обсягу, недостатня увага до проблем масштабування може призвести до неефективної роботи. Але, крім тих аспектів розробки, що є спільними для масштабування всіх розподілених інформаційних систем, у створенні систем, орієнтованих на аналіз інформації на семантичному рівні, необхідно обирати такі форми подання знань предметної області та засоби їхньої обробки, які, з одного боку, дозволяють досить повно відображати специфіку цієї області, а з іншого – придатні для виконання семантичного пошуку за прийнятний час за умов збільшення елементів бази знань інформаційного ресурсу та ускладнення її структури. У загальному випадку це є складною теоретичною проблемою, і у даній роботі ми аналізуємо її окремий випадок – масштабування обробки семантичних

Wiki-ресурсів, які містять велику кількість інформаційних об'єктів різних типів (таких, як енциклопедійні портали). Для цього необхідно виокремити ті фактори, які впливають на масштабованість семантичних IP, та визначити умови для забезпечення успішного розвитку такого ресурсу. У побудові практичних рекомендацій враховується практичний досвід розробки e-BYE.

e-BYE як приклад семантизованого Wiki-ресурсу зі складною структурою

Розглянемо це детальніше на прикладі e-BYE – семантизованого Wiki-ресурсу зі складною структурою, який реалізовано на технологічній платформі MediaWiki та її семантичного розширення Semantic MediaWiki. E-BYE – це портальна версія Великої української енциклопедії, яка містить відомості з багатьох галузей знань [13].

e-BYE використовує семантичні шаблони для подання типових інформаційних об'єктів (TIO). TIO – це підмножина сторінок Wiki-ресурсу, що належать до однакового набору категорій, мають однакову або подібну структуру та семантичні властивості. Створення системи TIO має базуватися на спільній роботі інженера зі знань та експертів ПрО. В e-BYE TIO пов'язуються з окремими Wiki-сторінками та базуються на виразних здатностях Wiki-середовища та його семантичного розширення. Зараз в e-BYE виокремлено 32 TIO, які характеризуються подібним набором семантичних властивостей відповідних Wiki-сторінок: персоналії, міста, країни, організації тощо. Для спрощення та уніфікації створення таких сторінок розроблено відповідні шаблони.

Рис. 1. Сторінка e-BYE на сайті Wikiapi.com

Із 2020 року e-BYE зареєстрована на сайті спільноти продуктів Semantic MediaWiki (https://wikiapiary.com/wiki/Great_Ukrainian_Encyclopedia), який показує швидкість зростання кількості сторінок, активності користувачів та кількості редагувань у цьому IP. Це підтверджує потребу в розробці та використанні методів керування розподіленими знаннями для подальшого розвитку цього ресурсу. Зараз портал стабільно розвивається (Рис.1), але збільшення його обсягу потребує знаходження масштабованих рішень для організації несуперечної структури бази знань.

Розвиток сайту відображає також Google Analytics – сервіс від компанії Google для аналізу Web-сайтів та мобільних застосувань, що надає статистичні дані щодо користувачів Web-застосунків. Він дозволяє відстежувати активність користувачів на Web-сайті, тривалість сеансу, кількість переглянутих за сеанс сторінок, кількість відмов тощо, а також інформацію про джерела трафіка. На рис.2 надано статистичні дані щодо зростання кількості користувачів e-BYE.

Ці статистичні дані вказують на потреби у масштабованому підході до подальшого розвитку порталу, який забезпечить його функціювання в умовах збільшення та ускладнення контенту та для більшої кількості відвідувачів.

Семантичний пошук в Semantic MediaWiki

Плагін Semantic MediaWiki є спеціальним розширенням технології Wiki, який надає можливість у середовищі MediaWiki вводити семантичну розмітку (тобто пов'язувати семантичними відношеннями сторінку із константами різних типів та з іншими Wiki-сторінками) та виконувати семантичні запити, в яких семантичні властивості можуть входити як до умов, так і до опису параметрів, що є результатом виконання запиту.

Семантичний пошук в Semantic MediaWiki – це вдосконалення традиційного Wiki-пошуку з використанням інформації про структурні елементи шуканого інформаційного об'єкту, про його властивості та відношення з іншими інформаційними ресурсами. Наприклад, можна шукати країну за назвою столиці, а людину – за місцем та роком народження. На відміну від традиційного пошуку, який пропонує, зокрема, Вікіпедія, у семантичному пошуку можна використовувати набір умов та враховувати не тільки категорії.

Рис. 2. Кількість користувачів e-BYE за даними Google Analytics.

Семантичний пошук може виконуватися кількома способами:

1. на спеціальній сторінці "Семантичний пошук", де параметри запитів вводяться у відповідні поля, не потребуючи від користувачів специфічних знань щодо синтаксису пошукової мови (достатньо знати, в які поля та за якими правилами вводити умови запиту та як описувати побажання щодо представлення результатів);
2. у вигляді пошукових запитів, що подаються спеціалізованою пошуковою мовою Semantic MediaWiki та вбудовуються в інші сторінки;
3. за допомогою запитів з використанням API , що потребують створення спеціального програмного коду.

За виразністю найбільш обмеженим є перший варіант, тому що користувач може використовувати тільки імена категорій та семантичних властивостей та вказувати обмеження щодо значень цих семантичних властивостей. Ефективнішим та швидшим є другий спосіб – користувач може використовувати додаткові змінні, такі як властивості поточної Wiki-сторінки, поточну дату й час, а сам пошук виконується у вбудованій базі знань Wiki-ресурсу серед структурованих даних. Третій спосіб має найбільшу виразність через можливість запрограмувати запит практично будь-якої складності, але виконання таких запитів потребує значно більше часу, тому що базується на повнотекстовому пошуку в усьому контенті IP.

У перших двох варіантах необхідною умовою виконання пошуку є наявність семантичної розмітки, а в третьому її наявність є теж бажаною, тому що надає зразки для пошуку потрібних елементів. У результаті виконання такого запиту користувач отримує перелік сторінок гасел, що відповідають введеним умовам, та ті значення їхніх семантичних властивостей, які він обирає. Для коректної побудови запитів потрібна інформація про правильні імена семантичних властивостей – їх можна отримати в результаті перегляду сторінок відповідних шаблонів (їх можна знаходити в звичайному пошуку у просторі імен «Шаблон»).

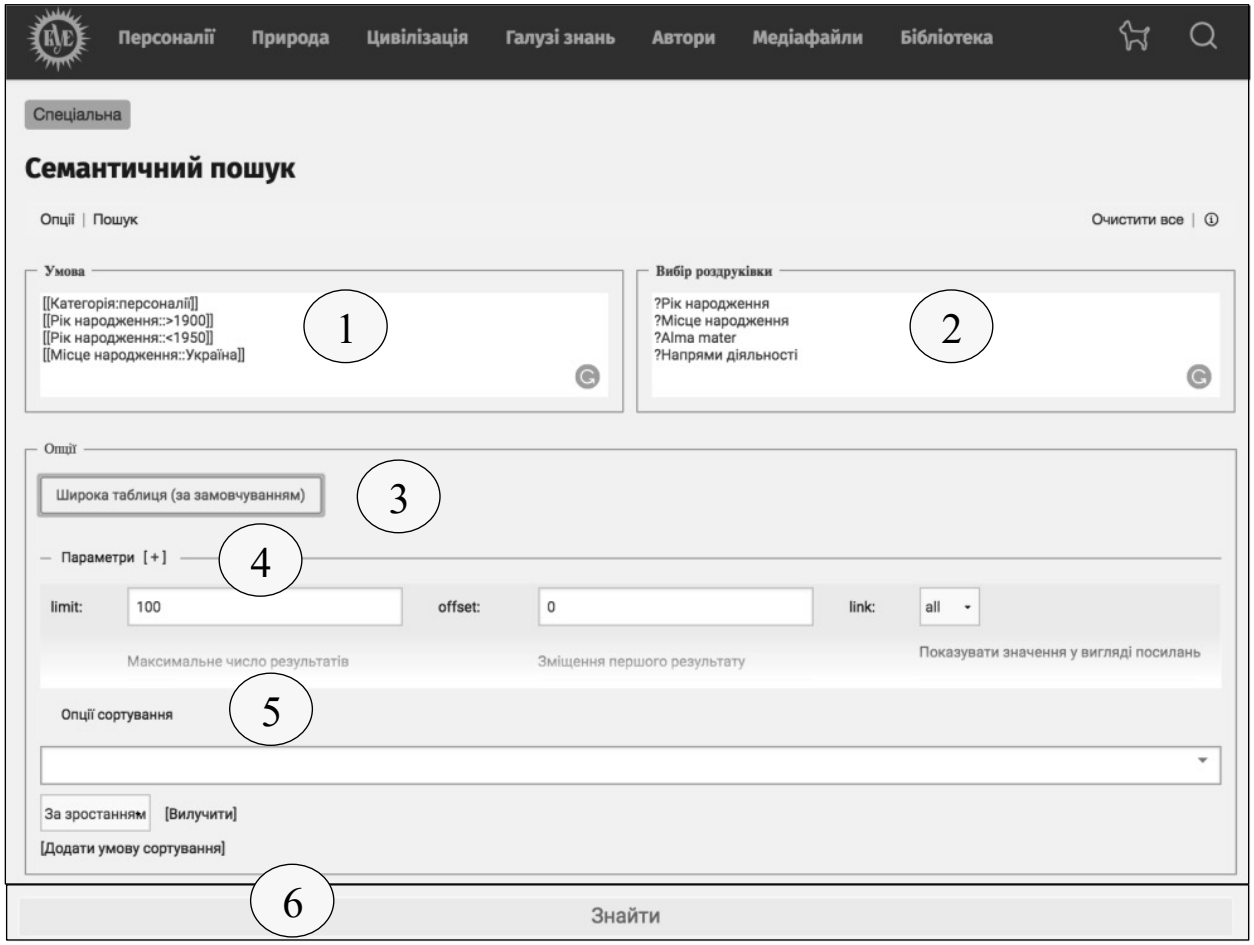

Рис. 3. Семантичний пошук у Semantic MediaWiki

Для користування першим типом семантичного пошуку доцільно надавати користувачам додаткову інструкцію та набір простих прикладів. На порталі е-ВУЕ така інформація подана на сторінці “vue.gov.ua/Пошук” (рис.3). Елементи сторінки семантичного пошуку: 1 – умови пошуку; 2 – що треба знайти; 3 – в якому вигляді виводити інформацію; 4 – скільки знайдених результатів виводити; 5 – як впорядкувати результати пошуку. Коли ці поля заповнено, потрібно натиснути кнопку «Знайти» (6).

Другий варіант пошуку розрахований на більш кваліфікованих користувачів, які володіють мовою запитів. У Semantic MediaWiki є проста, але потужна мова запитів SMW-QL, що відкриває широкі можливості для семантичного пошуку у Wiki-ресурсах. Тоді як семантичні властивості і категорії дозволяють структурувати дані у Wiki, запити потрібні для того, щоб використовувати цю інформацію: вони допомагають Wiki-користувачам і Wiki-адміністраторам комбінувати дані і візуалізувати їх. Зрозуміло, всі відвідувачі Wiki не зобов’язані вивчати цю мову запитів, і можуть навіть не знати про її існування. Однак і вони можуть відчути різницю в роботі із сайтом на Semantic MediaWiki завдяки можливості зберігання вбудованих запитів безпосередньо в тексті Wiki-статті. Мова запитів SMW-QL дозволяє, по-перше, фільтрувати сторінки за заданими критеріями, і по-друге, виводити як результати запиту тільки ту інформацію, що цікавить користувача, а не весь текст Wiki-сторінки. Найчастіше використовуються вбудовані запити, сполучені з функцією ask. Ця функція використовується так само, як і інші функції синтаксичного аналізатора MediaWiki: її виклик позначається подвійними фігурними дужками, перед іменем ставиться символ “#”, а після – двокрапка “:”. Спочатку передається сам рядок запиту, що обирає потрібну інформацію з Wiki, а потім усі параметри запиту, розділені символами вертикальної риски “|”.

Наприклад, у е-ВУЕ за допомогою таких запитів будується перелік гасел, підготовлених кожним з авторів, визначаються модератори кожної з галузей знань, будуються списки співробітників організацій та випускників навчальних закладів. Коректність виконання запитів залежить від якості оброблюваних даних, тобто від того, наскільки правильно зроблена семантична розмітка сторінок. Якісніше структурування контенту забезпечується застосуванням шаблонів типових ІО.

Слід зазначити, що перший та другий варіанти семантичного пошуку використовують ті ж способи, що забезпечує Semantic MediaWiki, але в першому варіанті запит мовою SMW-QL генерується цим плагіном (рис.2: 1 – генерується результат у вигляді, що визначається в 4, та код – в 3), а в другому – вводиться вручну користувачем. Досить часто використовують комбінований варіант – спочатку генерують запит засобами Semantic MediaWiki, а потім редагують його перед додаванням до Wiki-сторінки.

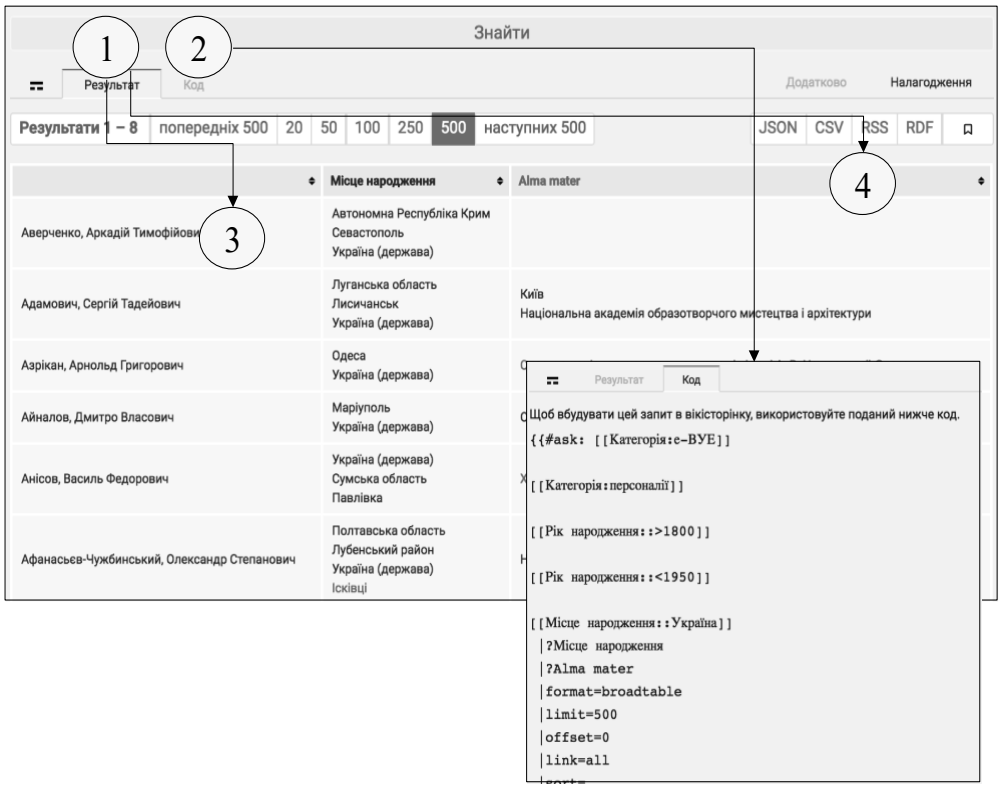

Рис. 4. Побудова коду та виконання запиту в Semantic MediaWiki

Третій варіант виконання запитів має більше можливостей щодо опису інформаційних потреб користувача, але основними елементами пошуку також є саме семантичні властивості (їхні назви) та значення.

Напрямки розвитку е-ВУЕ, що пов'язані із масштабуванням ІР

Стандартні задачі семантичної надбудови складаються з індексування сторінок для збережених раніше семантичних запитів (поданих на сторінках), пошук дублікатів, перевірка сутностей на відповідність шаблонам, типам даних, існування — індексація, збір статистики використання властивостей, пошук помилок та сутностей, що не використовуються. Наприклад, сторінка тестування API в е-ВУЕ (рис.5.1 – https://vue.gov.ua/Спеціальна:ApiSandbox#action=query&format=json&meta=siteinfo&siprop=statistics) показує кількість сторінок е-ВУЕ, статей, редагувань та користувачів), а сторінка (рис.5.2 – vue.gov.ua/Спеціальна:SemanticMediaWiki) надає відомості щодо процесу індексації даних та налаштування бази даних. Це дозволяє оцінювати обсяг ІР, задовільність або незадовільність стану індексації та ухвалювати адміністративні рішення щодо режиму індексування.

На 2.07.2022 до ВУЕ заведено 491 властивість, з яких використовується 364, і для цих властивостей визначено понад 360 тисяч значень, використовується понад 20000 вбудованих запитів. Існує можливість отримувати детальнішу інформацію та визначати, які дії потрібно виконати.

Крім того, оцінити складність структури бази знань ІР дозволяє кількість шаблонів типових ІО, їх структура (рис.6.1), кількість їх використань (рис.6.2) та кількість використань семантичних властивостей (аналізувати їх також дозволяють спеціальні сторінки), що використовуються для подання контенту.

Основними факторами, що впливають на ефективність виконання запитів, є:

– Схема бази знань, тобто система семантичних властивостей, що використовуються для структурування контенту Wiki-сторінок, яка визначає потенційну виразність таких запитів;

– Вчасна індексація змін у контенті (як у наборах семантичних відношень, так і у самих сторінках), яка, з одного боку, не повинна знижувати продуктивність роботи ІР, а з другого – забезпечувати актуальність бази даних;

– Вчасне видалення сторінок та елементів інфраструктури, що не використовуються;

– Якість виконання семантичної розмітки (відсутність помилок у назвах властивостей, семантизація посилань на інші сторінки, коректне введення значень властивостей);

– Наявність вбудованих запитів, що відповідають типовим (часто повторюваним) інформаційним потребам користувачів;

– Кількість вбудованих запитів та кількість звертань до сторінок з такими запитами;

– Зручне представлення результатів запитів;

– Розташування семантичних, вбудованих у Wiki-сторінки, запитів так, щоб результати запитів надавалися саме в тому місці ІР, де користувачам потрібна така інформація.

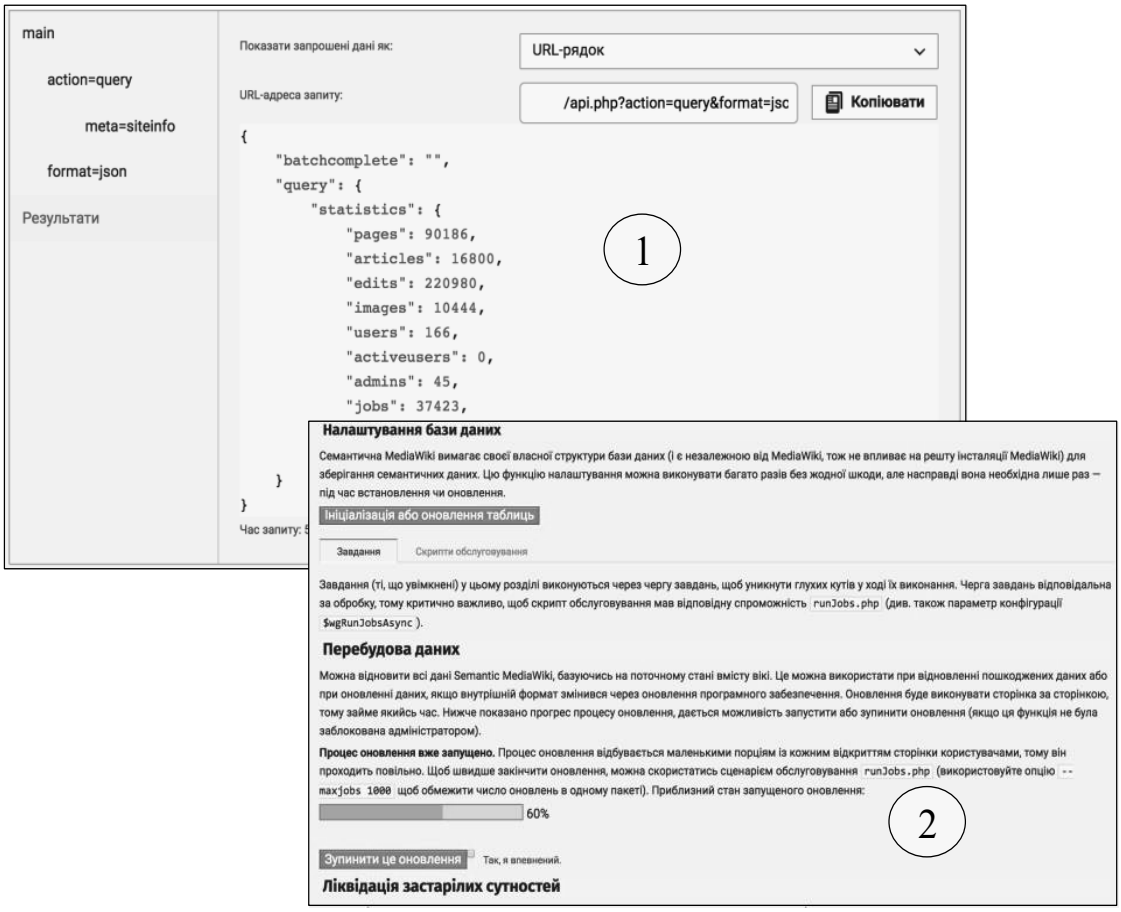

Рис. 5. Сторінка тестування API в e-BYE.

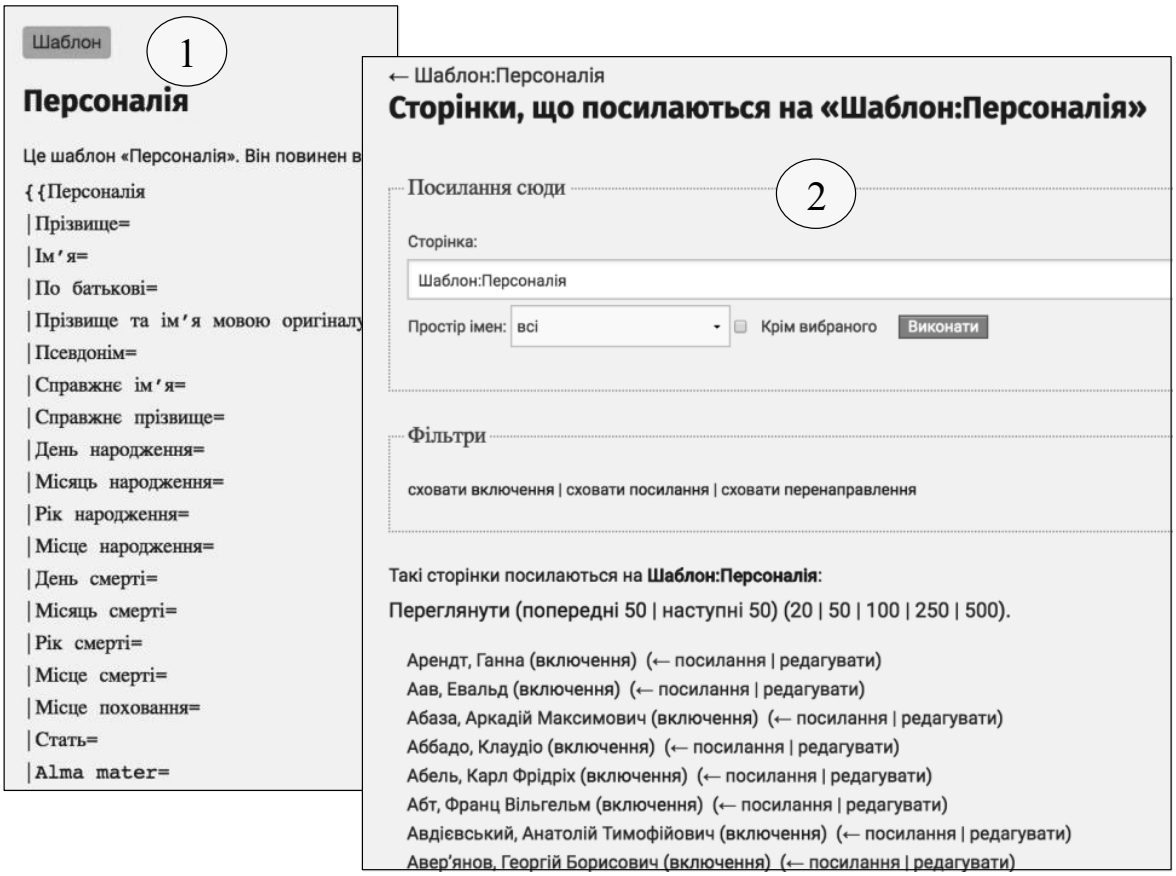

Рис. 6. Використання шаблону "Персоналія" в e-BYE .

В усіх варіантах пошуку (незалежно від того, трансформується запит у SQL-запит до бази знань ІР чи виконується повнотекстовий пошук по всьому контенту) швидкість виконання запиту залежить від його складності, тобто від кількості умов та обмежень. Тому в побудові запитів доцільно не вводити непотрібні умови (наприклад, якщо потрібно знайти освітні заклади певної країни, то недоцільно, крім категорії "Вищі навчальні заклади", вказувати категорію "Організації").

Досвід розробки та впровадження таких засобів семантичного пошуку на порталі е-ВУЕ дозволяє визначити ті особливості розробки семантичних порталів, які забезпечують його масштабування.

Умови розробки масштабованого семантичного ІР

Розглянувши особливості створення ІР на основі семантичного розширення Wiki-технологій, ми виявили, що основні фактори успішного масштабування таких ресурсів пов'язані із організацією структури семантичної розмітки Wiki-сторінок, а саме – кількістю відношень між Wiki-сторінками, коректним визначенням їх області значення та області визначення, а також із чітко формалізованими значеннями цих відношень, які забезпечують однозначне спільне розуміння сфери їх використання й запобігають дублюванню у створенні семантичних властивостей. Інші аспекти масштабування, що є універсальними для розробки систем великого обсягу, також мають специфічні характеристики, що пов'язані із технологічним середовищем Semantic MediaWiki.

Виходячи з наведеного вище аналізу тих аспектів, що впливають на можливість ефективного масштабування семантичного ІР, і враховуючи особливості організації такого ІР на технологічній основі MediaWiki та його семантичного розширення Semantic MediaWiki, доцільно дотримуватися наступних вимог:

з точки зору розміру даних:

– Контролювати розмір мультимедійних ІО, що використовуються в ІР;

– Передбачати засоби масового імпорту інформації із зовнішніх джерел у формати ІР;

з точки зору кількості ІО:

– Контролювати загальну кількість Wiki-сторінок та видаляти непотрібні, помилково створені сторінки та сторінки-дублі;

– Уніфікувати метаописи мультимедійних ІО для уникнення дублювання збережених файлів (наприклад, ті самі зображення можуть використовуватися на різних сторінках ІР);

– Розробляти шаблони ТІО, щоб уникнути збільшення кількості подібних імен семантичних властивостей та помилок у цих іменах та спростити сприйняття інформації користувачами;

– Створювати шаблони, які за допомогою запитів інтегрують контент різних сторінок ІР;

з точки зору структури бази знань:

– Формалізувати структуру бази знань ІР та інтероперабельно визначити семантику відношень між сторінками, яка використовується у семантичній розмітці ресурсу (вбудованих можливостей Semantic MediaWiki для цього недостатньо, і тому для цього доцільно застосовувати різноманітні зовнішні системи організації знань на основі онтологій);

– Визначити семантику гіперпосилань між сторінками ІР та створити відповідні семантичні властивості, явно описуючи їх область значення, область визначення та зміст;

– Розробляти шаблони для введення та подання значень семантичних властивостей ТІО, явно описати категорії сторінок, для яких вони мають використовуватися;

– Для вбудованих семантичних запитів, що викликаються на декількох різних сторінках, розробляти відповідні шаблони;

з точки зору засобів обробки:

– Визначити доцільність підключення розширень (плагінів), які розширюють функціонал обробки, і не встановлювати ті з них, в яких немає реальної потреби;

– Розробити адекватну політику індексування контенту, який враховує частоту оновлення інформації та кількість відвідувань користувачами;

– Створювати запити без надлишкових умов, аналізуючи таксономію категорій ІР;

– Мінімізувати інтегровані до ресурсу зовнішні програмні засоби (такі як лічильники відвідувань сторінок;

з точки зору місця обробки даних:

– Аналізувати кількість семантичних запитів на Wiki-сторінках та складність кожного з таких запитів;

– зменшувати кількість семантичних запитів та ІО на тих сторінках, які користувачі відвідують найчастіше (наприклад, на головній сторінці порталу недоцільно вбудовувати складні запити, які краще розміщувати на сторінках, до яких ведуть посилання з головної сторінки);

– вчасно створювати резервні копії контенту та структури ІР, забезпечувати можливість відтворення інформації;

– якщо потрібно виконувати велику кількість запитів для сторінок, які відвідують багато користувачів (наприклад, головної сторінки, сторінок категорій верхнього рівня), доцільно генерувати контент із фіксованим інтервалом часу та додавати його до контенту сторінки, а не виконувати запити окремо для кожного відвідувача.

Крім цих аспектів, для масштабування необхідно враховувати питання, що стосуються ролей користувачів та їхніх повноважень, а також інші вимоги безпеки.

Література

1. Data-Centric AI. the ultimate guide to the new ai paradigm. 2021. Available from: https://resources.kili-technology.com/dcai-ebook-2022. [Accessed: 11.07 2022].
2. Demchenko Y., De Laat C., Membrey P. Defining architecture components of the Big Data Ecosystem. In 2014 International Conference on Collaboration Technologies and Systems (CTS), 2014, P. 104-112.
3. Chen, M., Mao, S., Liu, Y. Big data: A survey. *Mobile networks and applications*, 19(2), 2014, P.171-209.
4. Рогушина Ю.В. Семантические wiki-ресурсы и их использование для построения персонифицированных онтологий. *CEUR* Workshop Proceedings 1631, 2016, P.188-195. Available from: http://ceur-ws.org/Vol-1631/188-195.pdf. [Accessed: 11.07 2022].
5. Soergel D. Knowledge organization systems: overview, 2009. Available from: www.dsoergel.com/UBLIS514DS-08.2a-1Reading4SoergelKOSOverview.pdf. [Accessed: 07 2015].
6. Hjorland B. What is knowledge organization (KO)? *KO Knowledge Organization*, 35(2-3), 2008, P.86-101. Available from: https://www.researchgate.net/profile/Birger-Hjorland/publication/277803483_What_is_Knowledge_Organization_KO/links/55d8232608aed6a199a6afce/What-is-Knowledge-Organization-KO.pdf.[Accessed: 15.07 2022].
7. Hendler J. A., Golbeck J. Metcalfe's law, Web 2.0, and the Semantic Web. *Web Sem.,* 6 (1), 2008, P.14-20.
8. Wagner C. Wiki: A technology for conversational knowledge management and group collaboration. *The Communications of the Association for Information Systems*, 2004, 13(1), P.264-289.
9. Völkel M., Krötzsch M., Vrandecic D. et al. Semantic Wikipedia. Proc. of the 15th international conference on World Wide Web, 2006, 585-594.
10. Рогушина Ю.В. Використання систем організації знань на основі онтологій у wiki-ресурсах. *Проблеми програмування*, 2022, №1, С|23-33. doi.org/10.15407/pp2022.01.23.
11. Dunning T., Friedman E. AI and Analytics at Scale. Lessons from Real-World Production Systems. 2021. O'Reilly Media. Available from: https://www.oreilly.com/library/view/ai-and-analytics/9781492094388/. [Accessed: 02.07 2022].
12. Benlachmi Y., Hsnaoui M.L. Current State and Challenges of Big Data, 2020, DOI: 10.1007/978-3-030-33103-0.
13. Andon P.I., Rogushina J.V., Grishanova I.Y. et all. Experience of Semantic Technologies Use for Development of Intelligent Web Encyclopedia. UkrPROG, *CEUR* Workshoop Proc., 2021, Vol-2866, P.246-259. Available from: http://ceur-ws.org/Vol-2866/ceur_246-259andon24.pdf. [Accessed: 22.06 2022].

References

1. Data-Centric AI. (2021). The ultimate guide to the new AI paradigm. . Available from: https://resources.kili-technology.com/dcai-ebook-2022. [Accessed: 11.07 2022].
2. DEMCHENKO Y. & DE LAAT C. (2014) Membrey P. Defining architecture components of the Big Data Ecosystem. In 2014 International Conference on Collaboration Technologies and Systems (CTS), P. 104-112.
3. CHEN M. & MAO S. & LIU Y. Big data: A survey. Mobile networks and applications, 19(2), 2014, P.171-209.
4. ROGUSHINA J. (2016) Semantic Wiki resources and their use for the construction of personalized ontologies, CEUR Workshop Proceedings 1631, P.188-195. Available from: http://ceur-ws.org/Vol-1631/188-195.pdf. [Accessed: 11.07 2022]. (in Ukrainian)
5. SOERGEL D. (2009). Knowledge organization systems: overview. Available from: www.dsoergel.com/UBLIS514DS-08.2a-1Reading4SoergelKOSOverview.pdf. [Accessed: 07 2015].
6. HJORLAND B. (2008). What is knowledge organization (KO)? KO Knowledge Organization, 35(2-3), P.86-101. Available from: https://www.researchgate.net/profile/Birger-Hjorland/publication/277803483_What_is_Knowledge_Organization_KO/links/55d8232608aed6a199a6afce/What-is-Knowledge-Organization-KO.pdf. [Accessed: 15.07 2022].
7. HENDLER J. A. & GOLBECK J. (2008). Metcalfe's law, Web 2.0, and the Semantic Web. Web Sem., 6 (1), P. 14-20.
8. WAGNER C. (2004). Wiki: A technology for conversational knowledge management and group collaboration The Communications of the Association for Information Systems, 13(1), P.264-289.
9. VÖLKEL M. & KRÖTZSCH M. & VRANDECIC D. et al. (2006). Semantic wikipedia. Proc.e of the 15th international conference on World Wide Web, P.585-594.
10. ROGUSHYNA J. (2022) Use of knowledge organization systems based on ontologies in wiki-resources. Problems on Programming , 1, P.23-33. doi.org/10.15407/pp2022.01.23. (in Ukrainian)
11. DUNNING T. & FRIEDMAN E. AI and Analytics at Scale. Lessons from Real-World Production Systems. 2021. O'Reilly Media. Available from: https://www.oreilly.com/library/view/ai-and-analytics/9781492094388/. [Accessed: 02.07 2022].
12. BENLACHMI Y. & HSNAOUI M.L. Current State and Challenges of Big Data, 2020, DOI: 10.1007/978-3-030-33103-0.
13. ANDON P.I. & ROGUSHINA J.V. & GRISHANOVA I.Y. et all. Experience of Semantic Technologies Use for Development of Intelligent Web Encyclopedia. UkrPROG, CEUR Workshoop Proc., 2021, Vol-2866, P.246-259. Available from: http://ceur-ws.org/Vol-2866/ceur_246-259andon24.pdf. [Accessed: 22.06 2022].

Одержано 03.08.2022

Про авторів:

Рогушина Юлія Віталіївна,
Канд.фіз.-мат.наук,
с.н.с Інституту програмних систем НАН України,
публікації в українських виданнях – 207,
публікації в іноземних журналах – 61,
ORCID http://orcid.org/0000-0001-7958-2557.

Гришанова Ірина Юріївна,
н.с Інституту програмних систем НАН України,
публікації в українських виданнях – 19,
публікації в іноземних журналах – 3,
ORCID http://orcid.org/0000-0003-4999-6294.
e-mail: i26031966@gmail.com

Місце роботи авторів:

Інститут програмних систем НАН України, 03181, Київ-187,
проспект Академіка Глушкова, 40,
e-mail: ladamandraka2010@gmail.com,
066 550 1999.

Прізвища та ініціали авторів і назва доповіді англійською мовою:
Rogushina Ju. V., Grishanova I. Yu.
Problems of scaling semantic information resources with a complex structure

Прізвища та ініціали авторів і назва доповіді українською мовою:
Рогушина Ю. В., Гришанова І. Ю.
Проблеми масштабування семантичних інформаційних ресурсів
зі складною структурою

Контакти для редактора: Рогушина Ю.,
старший науковий співробітник
Інституту програмних систем НАН України,
e-mail: ladamandraka2010@gmail.com,,
тел.: (38)(066) 5501999