УДК 004.9
DOI: 10.31891/2219-9365-2021-68-2-13

МАНЗЮК Е. А.
Хмельницький національний університет

КОМПЛЕКСНИЙ МЕТОД ВИЗНАЧЕННЯ ВІДПОВІДНОСТІ ОНТОЛОГІЇ ДОВІРИ ДО ІНТЕЛЕКТУАЛЬНИХ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА СТРУКТУРОВАНОГО ДОМЕНУ

В роботі наведено результати досліджень встановлення відповідності складових довіри до інтелектуальних інформаційних систем на основі онтології та структурованого домену, який сформовано з корпусу узагальненої інформації предметного поля етичних принципів за людино центрованим підходом. Метод сформовано з об'єднанням синтаксичного, структурного та семантичного методів з метою всебічного та об'єктивного встановлення відповідності складових довіри.
Ключові слова: інтелектуальна інформаційна технологія, отологія довіри, метод встановлення відповідності.

EDUARD MANZIUK
Khmelnytskyi National University

A COMPREHENSIVE METHOD FOR DETERMINING THE ALIGNMENT OF THE TRUST ONTOLOGY TO INTELLIGENT INFORMATION TECHNOLOGY AND THE STRUCTURED DOMAIN

The paper presents the results of research on the alignment of trust components to intelligent information systems based on ontology and structured domain, which is formed from the corpus of generalized information of the subject field of ethical principles on a human-centered approach. The method is formed by combining syntactic, structural and semantic methods in order to comprehensively and objectively alignment the correspondence of the components of trust. The approach base on the generalized informativeness of documents, which are basically designed to present the appropriate level of generalization and are official documents of the subject area of state, national, supranational structures. That is, the corpus was based on documents that have been widely discussed and represent the consolidated and balanced opinion of a wide range of associations of professional opinions in the domain segment. Thus, further generalization of the elements of the corpus, i.e. individual sectors of the domain will allow to obtain an objective picture in terms of completeness, comprehensiveness and to obtain a level of representation of the domain of discourse. The methods of analysis were based on a "scoping review" in order to synthesize and reflect the available informativeness of documents, which is most suitable for complex and heterogeneous areas of research. For alignment, a combination of three different approaches to assessing the similarity of entities: syntactic, structural and semantic. Comparative evaluation is given based on how similar the entities are, and is seen as a measure of the degree of substitution of one class for another with the possibility of use. The measure of the degree of alignment is the basis of the ontology of alignment and directly affects the quality of comparison of conceptual structures. The development of ontologies of models based on human-centric principles is an important step in the development of intelligent information systems for those responsible and critical applications.
Keywords: intelligent information technology, trust otology, alignment method.

Постановка проблеми

Для забезпечення об'єктивності та всебічності представлення предметної області необхідно формувати об'ємний корпус, який містить дослідження авторських груп. Об'ємність корпуса обґрунтовується метою отримання об'єктивної картини на рівні тенденцій. Однак обробка такого корпусу є складним завданням відповідно до вказаних причин, які базуються на наявній специфіці кожного окремого документа дослідження та складністю узагальнення.

Відповідно до цього був обраний інший підхід, який базувався на узагальненій інформативності документів, які у своїй основі розроблені для представлення відповідного рівня узагальнення та становлять собою офіційні документи предметної області державних, національних, наднаціональних структур, корпоративних об'єднань, інституціональних утворень, дослідницьких організації, професійних об'єднань тощо. Тобто в основу корпусу було покладені документи, які пройшли широке коло обговорення, та представляють собою консолідовану та зважену думку широкого кола об'єднань професійних думок сегменту домену. Таким чином подальше узагальнення елементів корпусу, тобто окремих секторів домену дасть змогу отримати об'єктивну картину з погляду повноти, всебічності та отримати рівень представлення домену дискурсу.

Аналіз останніх джерел

В основу покладено дослідження та дані отримані в роботі Jobin A. (2019) [1], в якій на глобальному рівні аналізуються вимоги, технічні стандарти, етичні принципи інтелектуальної інформаційної технології (ІІТ). Концепт "довіра до ІІТ " визначений в домені етики людиноцентричного підходу. Тому для цього використаємо методи прямої відповідності та покриття повноти з використання функцій близькості та

інтерпретації принципів. На базі контентного аналізу визначені етичні принципи та документи в яких вони мають змістовний характер. Цінність вибраного корпусу полягає в тому, що він складається з документів, які є наслідком широкого обговорення та консолідованої думки. Тобто це не є думка окремого науковця, дослідника і так далі, а є даними високого рівня узагальнення. Корпус складається з документів приватних компаній, відповідних державних установ, академічних та науково-дослідних установ, міжурядових або наднаціональних організації, некомерційних організації, професійних асоціацій/наукових товариств, об'єднань приватного сектора, дослідницьких об'єднань, наукових фондів, федерації профспілок працівників, політичних партій (private companies, governmental agencies respectively, academic and research institutions, inter-governmental or supra-national organizations, non-profit organizations, professional associations/scientific societies, private sector alliances, research alliances, science foundations, federations of worker unions, political parties) [1].

Корпус базується на документах, які представленні у виді директив, керівних принципів (guidelines) та документів "м'якого права" (soft-law documents). Документах, які розглядається як квазіправовий інструмент в контексті використання з точки зору пошуку та охоплення регулюючих систем для розробки звичайних правових документів. Втілюють собою відтворення, наміри, бачення формування контексту, відточування формулювання тощо і є підґрунтям для подальшого створення формальних документів правового рівня. Документи відносяться до "сірої літератури" (gray literature) і не індексується у звичайних наукових базах та не контролюється комерційними виданнями. Є менш тиражованою, та її отримання є більш об'єктивним оскільки не визначається популярністю в системах пошуку. Така література як правило є новітньою [2, 3] та можна описати її, як представлення більш об'єктивної та загальної думки.

Методи проведення аналізу базувались на "огляді за обсягом" (scoping review) з метою синтезу та відображення наявної інформативності документів, який найбільше підходить для складних та неоднорідних областей дослідження [4, 5]. Використовувався розроблений протокол для виявлення відповідностей та адаптований до Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [6, 7]. Базувався на кращих практиках для сірої літератури, використовуючи багатоетапний огляд на базі індуктивного огляду, дедуктивної ідентифікації відповідних організації з використанням принципів пошуку та релевантності меті дослідження. Використовувались джерела багатьох мов, з певною перевагою за кількістю англомовних, в тому числі у зв'язку з тим, що документи міжнародного значення видані англійською мовою.

Метою роботи є: розробка методу визначення відповідності онтології довіри до інтелектуальних інформаційних систем та структурованого домену

Виклад основного матеріалу

Структурований домен є ширший за представлену онтологію і значною мірою онтологія повинна бути в ньому представлена. Проте стандарт ISO/IEC TR 24028 на якому розроблена онтологія може містити аспекти, які виходять за межі предмету етики ІІТ.

Порівняння концептуальних структур проводиться згідно представленого підходу. Порівнюють усі сутності онтології з усіма сутностями представленими в структурованому домені. Порівнювання відбувається згідно з припущенням закритого світу [8, 9], відповідно до якого, якщо будь-які пари сутностей, які не згадані у порівнянні, вважаються такими, які не мають зв'язків. На початковому кроці використано фільтрацію сутностей [10, 11], щоб визначити щодо яких сутностей відбувається порівняння. Для порівняння використаємо комбінацію трьох різних підходів оцінки подібності сутностей: синтаксичний, структурний та семантичний. Порівняльна оцінка дається виходячи з того наскільки подібними є сутності, і розглядається як міра ступеня заміщення одного класу іншим з можливістю використання.

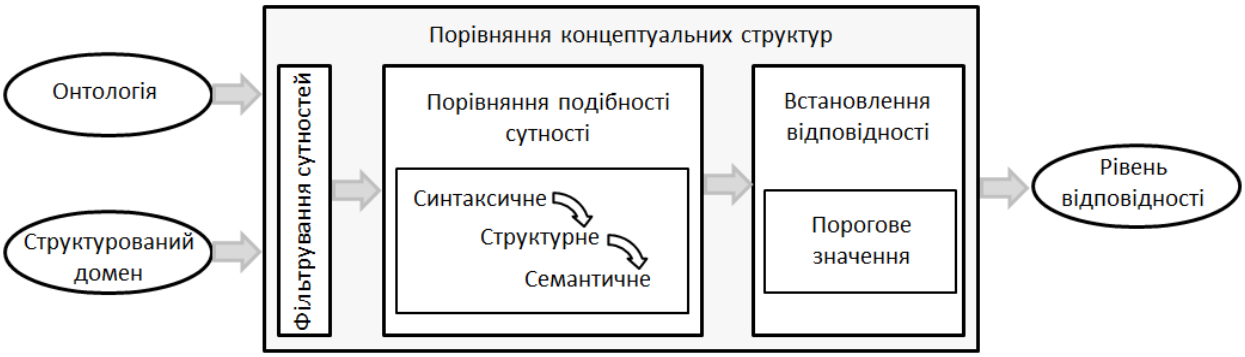

Рис. 1. Загальна структура системи порівняння концептуальних структур онтології та структурованого домену

Синтаксичне порівняння [12, 13] формується на основі порівняння позначення термінів $syntactic(\ )$. Терміни є мітками сутності, а також можуть містити коментарі, примітки до суб'єкта. Цей підхід базується на інформації отриманої з природної мови, як приклад, слова, поєднанням слів як стійких форм, речень як формі поєднання слів.

Вказані елементи представляються в онтології логічних аксіом. В певних випадках існує неоднозначність інформації слова. Відповідно досліджується лексична інформація сутностей, поєднання сенсів слів, збагачується лексична інформація та зв'язки між позначеними сутностями словами.

Структурне порівняння [14, 15] враховує структурні зв'язки інформативності сутностей $structure(\ )$. Наприклад, дві сутності з тим же суперкласом є більш близькими ніж сутності, які не мають такої спільності. Також сюди відноситься функціональні структурні зв'язки. Якщо сутність пов'язана з іншими сутностями зв'язками, при порівнянні враховуються аналогічні зв'язки з області порівняння. Цей підхід базується на поєднанні семантичної інформації такого виду: є частиною, представляє собою, підклас, підвластивість. Шляхом представлення у вигляді графу та проекцією відношень між концептами отримується представлення одиниці інформації. Суміжні концепти на рівні зв'язків семантичних відношень повною мірою розкриваються, що показує перевагу цього підходу.

Семантичне порівняння [16, 17] використовує значення міток сутностей у поєднанні з їхнім описом $semantic(\ )$. Порівнянням здійснюється без використання зовнішніх ресурсів, словників, тезаурусів тощо. Це обмежить привнесення неоднозначності зовнішнього впливу. Цей підхід базується на пояснені мови опису онтології, тобто онтології описової мови. Змістовність сутності проявляється фрагментами інформації, які містяться в онтологічних конструкціях та аксіомах. Для практичного використання використовуються техніки табличних конструкції, структурного об'єднання, використання яких визначається цілями порівняння.

Для проведення порівняння визначимо поняття "значення сутності". Значення сутності полягає у визначенні опису, згідно з яким можна дану сутність однозначно ідентифікувати. В межах домену дослідження представляється у вигляді ідентифікатора сутності $id \in I$, позначення $lb \in L$ та опису $dsc \in Dsc$. Зміст сутності описується кортежем $ent = \langle id, lb, dsc \rangle$. Множина понять "зміст сутності" належить відповідному домену.

Два елементи змісту сутності ent_1 та ent_2 співпадають, якщо вони співвідносяться з однією і тією ж сутністю предметної області з об'єктивної реальності. Ці дві сутності взаємозамінним тобто справедливе співвідношення $ent_1 \equiv ent_2$, або в межах дослідження $ent_O \equiv ent_{Ds}$. В загальній множині сутностей $\forall ent_O \in Ent_O$, $\forall ent_{Ds} \in Ent_{Ds}$, $Ent_O \cup Ent_{Ds} = Ent$, яка є об'єднанням сутностей онтології та структурованого домену, відповідно наявні повтори сутностей. Відповідно до досліджень [18–24] розрізняють два підходи з виділенням множини сутностей $Rp(Ent) = \left\{ \left(ent_{O,i}, ent_{Ds,j} \right) \middle| ent_{O,i} \in Ent_O, ent_{Ds,j} \in Ent_{Ds}, ent_{O,i} \equiv ent_{Ds,j}, i, j \in \left| Rp \right| \right\}$, тобто підмножини $Rp \subseteq Ent$, яка містить тільки елементи з повторами. Інший підхід полягає у виділенні множини зв'язків між сутностями на рівні взаємозаміни кожного елементу із двох множин $Link(Ent_O, Ent_{Ds}) = \left\{ \begin{matrix} link_{ent_{O,i}, ent_{Ds,j}} \middle| ent_{O,i} \in Ent_O, ent_{Ds,j} \in Ent_{Ds}, \\ ent_{O,i} \equiv ent_{Ds,j}, i \in \left| Ent_O \right|, j \in \left| Ent_{Ds} \right| \end{matrix} \right\}$. Співвідношення між множинами $\left| Link \right| = \left| Rp \right| / 2$.

Ефективність порівняння концептуальних структур по відношенню до онтології встановимо таким чином, що визначає наскільки розроблена онтологія відповідає стану предметної області

$$Ef_O = \frac{\left| Link \right|}{\left| Ent_O \right|} \tag{1}$$

Фільтрація сутностей множин Ent_O та Ent_{Ds} здійснимо через функцію подібності $Ent_O \times Ent_{Ds} \xrightarrow{f_{sim}} \square$, з визначеним пороговим значенням trh_{sim}. Отримаємо множину подібності на рівні зв'язків $Link_{O,filter} = \left\{ \begin{matrix} link_{ent_{O,i}, ent_{Ds,j}} \middle| \left(ent_{O,i}, ent_{Ds,j} \right) \in Ent_O \times Ent_{Ds}, ent_{O,i}, \\ i \in \left| Ent_O \right|, j \in \left| Ent_{Ds} \right|, f_{sim} \left(ent_{O,i}, ent_{Ds,j} \right) \geq trh_{sim} \end{matrix} \right\}$. Оскільки досліджуємо відповідність онтологія O по відношенню до Ent_{Ds} множину зв'язків визначаємо з базовим відношення стосовно сутностей, які визначені в онтології $Link_{O,Ds,filter} = Ent_O \text{ в}_{trh_{sim}} Ent_{Ds}$.

Множина подібності сутностей на рівні зв'язків формується за інтуїтивним припущенням про

відповідні вибраної сутності онтології певним сутностям структурованого домену.

Формування множини зв'язків $Link_{O,Ds,filter}$ здійснюється згідно з такими етапами:

1. Для кожної сутності $ent_{O,i}$, $i \in \left|Ent_O\right|$, множини сутностей онтології Ent_O вибирається сукупність сутностей структурованого домену Ent_{Ds}. Відкидаються сутності істино негативні Ent_O в $_{TN}$ Ent_{Ds}.

2. Наступним кроком є зменшення кількості зв'язків множини $Link_{O,Ds,filter}$ до знайдених ймовірних відповідностей з базовою множиною Ent_O Зменшення $Link_{O,Ds,filter}$ здійснюється відповідно до порогового значення подібності $Link_{O,Ds,filter} = \sigma_{\geq th_{sim}}\left(Link_{O,Ds,filter}\right)$.

3. На цьому кроці відбувається перехід до кроку 1 та мінімізується кількість сутностей множини онтології Ent_O, які не мають зв'язків $\min\left(Ent_O \triangleright_{TN} Ent_{Ds}\right) = \min\left(Ent_O - Ent_O \text{ в }_{TN} Ent_{Ds}\right)$. Ця умова є важливою оскільки на цьому кроці визначається кількість сутностей Ent_O для яких не було знайдено порівнювальних концептуальних структур в Ent_{Ds}

Встановлення відповідності відносно сутності онтології базується на змісту сутності. Сутності вважається відповідними, якщо належним чином співвідносяться їхні описи dsc. Найбільш поширеними є автоматизовані системи порівняння, які базуються на роботі з токенами стрінгових послідовностей для ідентифікації позначень сутностей [25, 26]. Проте в нашому випадку об'єм даних є обмеженим та визначається областю дослідження, головною метою якої є якість порівняльного аналізу співвідношення, яке виконується без використання автоматизованих систем. Ґрунтуючись на цьому базовою є множина описів $dsc \in Dsc$.

Онтологію порівняння концептуальних структур онтології довіри та структурованого домену представимо у такому вигляді

$$O_{match} = \left\langle Ent_O, Ent_{Ds}, Link_{O,Ds,filter}, \rho, \sigma_{trh_{match}} \right\rangle \quad (2)$$

Функція відбору $\sigma_{trh_{match}}$ визначає обмеження множини $Link_{O,Ds,filter}$ використовуючи встановлення відповідностей зв'язків базової множини Ent_O на основі порогового значення ступеня відповідності trh_{match}. Тип відповідності $\rho = \left\{\equiv, \phi, \text{x}\right\}$. Міра ступеня відповідності отримує базову сутність з Ent_O та присвоює кожному елементу підмножини зв'язків з $Link_{O,Ds,filter}$, в яких присутня базова сутність, відповідне число в межах $\left[0,...,1\right]$. Таким чином крайні значення визначають у випадку рівності 1 - ідентичність сутностей, або у випадку рівності 0 - відсутність значимого зв'язку між цими сутностями, тобто визначають силу зв'язку.

Міра ступеня відповідності є основою онтологія відповідності і безпосередньо впливає на якість порівняння концептуальних структур.

В системі порівняння використовується два граничних значення, граничне значення подібності trh_{sim} та граничне значення ступеня відповідності trh_{match}. Граничне значення подібності trh_{sim} використовується на етапі фільтрації для визначення подібних сутностей Ent_O на множині Ent_{Ds}. Цей етап призначений для того, щоб встановити максимальну кількість подібних сутностей з наявними зв'язками. В кінцевому випадку кожна сутність множини Ent_O повинна мати щонайменше один зв'язок $link_{Ent_O,Ent_{Ds}}$ для забезпечення максимального представлення відповідностей. Цей етап є початковим та підготовчим. Результати наступних етапів порівняння за синтаксичним, структурним та семантичним порівнянням встановлюються за ступенем відповідності trh_{match}.

Схематично представимо схему зв'язків множин сутностей.

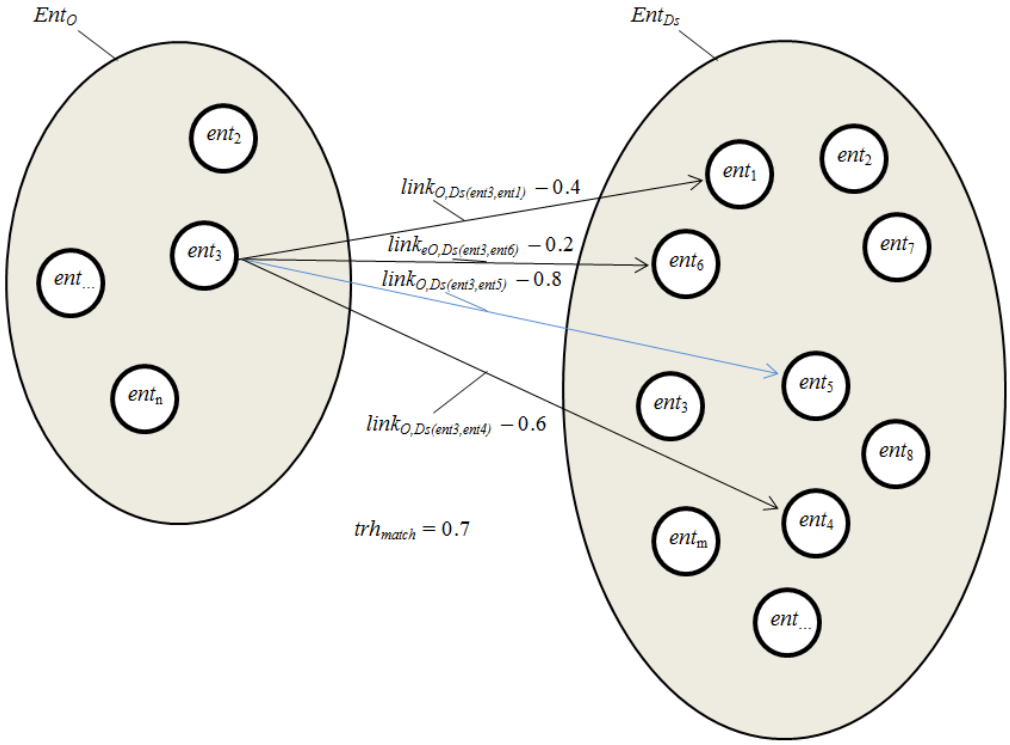

Рис. 2. Схема зв'язків множини сутностей Ent_O базової онтології довіри та множини сутностей Ent_{Ds} структурованого домену

Порівняння концептуальних структур здійснюється функцією встановлення ступеня відповідності структурованого домену онтології довіри

$$matching : Ds \to O \tag{3}$$

В зв'язку з тим, що $Dom\ O \subseteq Dom\ Ds$ і функція встановлення ступеня відповідності є функцією редукції Ds на відповідний простір O. Процес редукції полягає в послідовному вилученні на підставі онтології порівнянні концептуальних структур. Метод рекурсивного вилучення полягає у рекурсивному послідовному скороченні концептуальних структур, які задані природною мовою.

Функція редукції є поєднанням трьох функцій

$$matching(\ ) = syntactic(\ ) \circ structure(\ ) \circ semantic(\ ) \tag{4}$$

Кожна з трьох функції призначення для формалізації отримання певних аспектів інформативності, які можна отримати з концептуальних структур для їх встановлення відповідності.

Висновки

Швидкий розвиток інформаційного поля ІІТ знаходить відгук в роботі в роботі експертних груп та міжнародних стандартів - International Organization for Standardization (ISO). Формалізуються поняття, принципи, підходи, сфера використання, концепції етичності та інші. Це є важливим кроком впорядкування, визначення в сторону практичного використання експоненційного зростання інформаційного поля сфери ІІТ. Ринок технічних рішень на базі ІІТ вийшов за межі вузького використання та набув величини, яка здійснює суттєвий вплив на людину з перспективами швидкого зростання. Відповідно це потребує розробки та впровадження технічних стандартів ІІТ. Формалізація всіх етапів, принципів та підході ІІТ потребує поглибленої розробки наукових основ. Це поряд із пришвидшенням інформатизації всього суспільства подекуди має змагальний характер. Формалізація з цієї позиції є важливим напрямком систематизації інформації предметного поля. Розроблення онтологій моделей на базі людино центричних принципів є важливим етапом розробки систем ІІТ для відповідальних на критичних застосувань. Слід зазначити, що запропонована модель розроблена в рамках стандарту ISO/IEC TR 24028 та обмежена самим стандартом.

Література

1. Jobin A. The global landscape of AI ethics guidelines / A. Jobin, M. Ienca, E. Vayena // Nat Mach Intell. — 2019. — Vol. 1, № 9. — Pp. 389–399. doi: 10.1038/s42256-019-0088-2.
2. Zhou X. How to Treat the Use of Grey Literature in Software Engineering / X. Zhou // Proceedings of the International Conference on Software and System Processes. ICSSP '20. New York, NY, USA. — 2020. — Pp. 189–192. doi: 10.1145/3379177.3390305.
3. Gul S. Is grey literature really grey or a hidden glory to showcase the sleeping beauty / S. Gul, T. A. Shah, S. Ahmad, F. Gulzar, T. Shabir // Collection and Curation. — 2020. — Vol. 40, № 3. — Pp. 100–111. doi: 10.1108/CC-10-2019-0036.
4. Tong R. Professional and interprofessional identities: a scoping review / R. Tong, M. Brewer, H. Flavell, L. D. Roberts // Journal of Interprofessional Care. — 2020. — Pp. 1–9. doi: 10.1080/13561820.2020.1713063.
5. Peters L. A scoping review exploring 'opportunity' in occupational science: Possibilities for conceptual development / L. Peters, R. Galvaan // Journal of Occupational Science. — 2021. — Vol. 28, № 2. — Pp. 249–267. doi: 10.1080/14427591.2020.1832906.
6. Rethlefsen M. L. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews / M. L. Rethlefsen, S. Kirtley, S. Waffenschmidt, A. P. Ayala, D. Moher, M. J. Page, J. B. Koffel, H. Blunt, T. Brigham, S. Chang, J. Clark, A. Conway, R. Couban, S. de Kock, K. Farrah, P. Fehrmann, M. Foster, S. A. Fowler, J. Glanville, E. Harris, L. Hoffecker, J. Isojarvi, D. Kaunelis, H. Ket, P. Levay, J. Lyon, J. McGowan, M. H. Murad, J. Nicholson, V. Pannabecker, R. Paynter, R. Pinotti, A. Ross-White, M. Sampson, T. Shields, A. Stevens, A. Sutton, E. Weinfurter, K. Wright, S. Young, PRISMA-S Group // Systematic Reviews. — 2021. — Vol. 10, № 1. — Pp. 1–39. doi: 10.1186/s13643-020-01542-z.
7. Snyder H. Literature review as a research methodology: An overview and guidelines / H. Snyder // Journal of Business Research. — 2019. — Vol. 104. — Pp. 333–339. doi: 10.1016/j.jbusres.2019.07.039.
8. Schneider T. Special Issue on Ontologies and Data Management: Part I / T. Schneider, M. Šimkus // Künstl Intell. — 2020. — Vol. 34, № 3. — Pp. 287–289. doi: 10.1007/s13218-020-00682-7.
9. Schneider T. Special Issue on Ontologies and Data Management: Part II / T. Schneider, M. Šimkus // Künstl Intell. — 2020. — Vol. 34, № 4. — Pp. 439–441. doi: 10.1007/s13218-020-00693-4.
10. Ayala D. LEAPME: Learning-based Property Matching with Embeddings / D. Ayala, I. Hernández, D. Ruiz, E. Rahm // arXiv:2010.01951 [cs]. — 2020.
11. Karimi H. A learning-based ontology alignment approach using inductive logic programming / H. Karimi, A. Kamandi // Expert Systems with Applications. — 2019. — Vol. 125. — Pp. 412–424. doi: 10.1016/j.eswa.2019.02.014.
12. Laadhar A. POMap: An Effective Pairwise Ontology Matching System / A. Laadhar, F. Ghozzi, I. Megdiche Bousarsar, F. Ravat, O. Teste, F. Gargouri // IC3K 2017 : 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. Funchal, Portugal. — 2017. — Pp. 161–168.
13. Karadeniz İ. Linking entities through an ontology using word embeddings and syntactic re-ranking / İ. Karadeniz, A. Özgür // BMC Bioinformatics. — 2019. — Vol. 20, № 1. — Pp. 156. doi: 10.1186/s12859-019-2678-8.
14. Eine B. Ontology-Based Big Data Management / B. Eine, M. Jurisch, W. Quint // Systems. — 2017. — Vol. 5, № 3. — Pp. 1–45. doi: 10.3390/systems5030045.
15. Zhang C. MetaGO: Predicting Gene Ontology of Non-homologous Proteins Through Low-Resolution Protein Structure Prediction and Protein–Protein Network Mapping / C. Zhang, W. Zheng, P. L. Freddolino, Y. Zhang // Journal of Molecular Biology. — 2018. — Vol. 430, № 15. — Pp. 2256–2265. doi: 10.1016/j.jmb.2018.03.004.
16. Kulmanov M. Semantic similarity and machine learning with ontologies / M. Kulmanov, F. Z. Smaili, X. Gao, R. Hoehndorf // Brief Bioinform. — 2021. — Vol. 22, № 4. doi: 10.1093/bib/bbaa199.
17. Yu G. Gene Ontology Semantic Similarity Analysis Using GOSemSim / G. Yu // Stem Cell Transcriptional Networks: Methods and Protocols / Kidder B. L. — New York, NY : Springer US, 2020. — Pp. 207–215.
18. Hand D. A note on using the F-measure for evaluating record linkage algorithms / D. Hand, P. Christen // Stat Comput. — 2018. — Vol. 28, № 3. — Pp. 539–547. doi: 10.1007/s11222-017-9746-6.
19. Li B.-H. A Survey on Blocking Technology of Entity Resolution / B.-H. Li, Y. Liu, A.-M. Zhang, W.-H. Wang, S. Wan // J. Comput. Sci. Technol. — 2020. — Vol. 35, № 4. — Pp. 769–793. doi: 10.1007/s11390-020-0350-4.
20. Christophides V. An Overview of End-to-End Entity Resolution for Big Data / V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, K. Stefanidis // ACM Comput. Surv. — 2020. — Vol. 53, № 6. — Pp. 127:1-127:42. doi: 10.1145/3418896.
21. Brunner U. Entity matching with transformer architectures - a step forward in data integration / U. Brunner, K. Stockinger // International Conference on Extending Database Technology, Copenhagen, 30 March-2 April 2020. 2020. doi: 10.21256/zhaw-19637.
22. Papadakis G. The Four Generations of Entity Resolution / G. Papadakis, E. Ioannou, E. Thanos, T. Palpanas // Synthesis Lectures on Data Management. — 2021. — Vol. 16, № 2. — Pp. 1–170. doi: 10.2200/S01067ED1V01Y202012DTM064.
23. Weikum G. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases / G. Weikum, L. Dong, S. Razniewski, F. Suchanek // arXiv:2009.11564 [cs]. — 2021.
24. Chen Z. Towards Interpretable and Learnable Risk Analysis for Entity Resolution / Z. Chen, Q. Chen, B. Hou, Z. Li, G. Li // Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. SIGMOD '20. New York, NY, USA. — 2020. — Pp. 1165–1180. doi: 10.1145/3318464.3380572.
25. Medvet E. Interactive example-based finding of text items / E. Medvet, A. Bartoli, A. De Lorenzo, F. Tarlao // Expert Systems with Applications. — 2020. — Vol. 154. — Pp. 113403. doi: 10.1016/j.eswa.2020.113403.
26. Christiani T. Scalable and Robust Set Similarity Join / T. Christiani, R. Pagh, J. Sivertsen // 2018 IEEE 34th International Conference on Data Engineering (ICDE). — 2018. — Pp. 1240–1243. doi: 10.1109/ICDE.2018.00120.