УДК 681.3 http://doi.org/10.15407/pp2023.04.039

Ю.В. Рогушина

ТРИВИМІРНА МОДЕЛЬ СЕМАНТИЧНОГО ПОШУКУ: ЗАПИТИ, РЕСУРСИ ТА РЕЗУЛЬТАТИ

Запропонована тривимірна модель семантичного пошуку , що аналізує пошукові запити, інформаційні ресурси та результати пошуку, пропонується як додатковий інструмент опису та співставлення інформаційно-пошукових систем (ІПС), що використовують різноманітні елементи штучного інтелекту та менеджменту знань для більш ефективного та пертинентного задоволення інформаційних потреб користувачів. Потрібно відмітити, що значення параметрів, які аналізує ця модель, не є взаємовиключними, тобто та сама ІПС може підтримувати кілька варіантів пошуку. Крім того, засоби подання запитів та ресурсів не завжди є порівнюваними. Проаналізовано існуючі підходи до семантизації пошукових запитів та використання зовнішніх джерел знань для їх виконання.

Модель дозволяє виявляти ІПС, для яких перетинаються тріади "запит-ІР-результат", та виконувати їх порівняння саме на цих підкласах пошукових задач. Це дозволяє визначати, які алгоритми пошуку виявляються більш пертинентними для конкретних задач користувачів і на основі цього обирати такі сервіси як джерело інформації для подальшої обробки. Важливою особливістю запропонованої моделі є використання лише тих характеристик ІПС, які можуть бути проаналізовані користувачами.

Ключові слова: семантичний пошук, онтологія, пошуковий запит.

Вступ

Пошук інформації (у локальних і глобальних мережах, на окремому комп'ютері) сьогодні є одним із найпоширеніших завдань, що входить до складу різних застосовних систем. Швидке зростання обсягу інформації, яку потрібно обробляти, та ускладнення її структури зумовлюють усе більшу потребу в розвитку засобів знаходження відомостей, необхідних користувачам для виконання їхніх завдань.

Семантичний пошук (СП) – термін, який використовують для позначення набору методів, призначених для покращення пошуку в документіах або у базі знань. На відміну від традиційних методів пошуку, зосереджених на ранжуванні документів на основі набору ключових слів (як у запиті користувача, так і в індексованому контенті), методи СП спрямовані на те, щоб враховувати контекст і семантику як запиту користувача, так і тих ресурсів, в яких здійснюється пошук, за допомогою використання засобів обробки природної мови, технологій Semantic Web та методів машинного навчання для отримання більш релевантних результатів.

Система семантичного пошуку (ССП) – це інформаційна система, що забезпечує пошук і розпізнавання інформаційних об'єктів (ІО) різних типів із використанням знань для зіставлення запиту з наявними інформаційними ресурсами на семантичному рівні [1]. ССП можна розглядати як певну інтелектуальну надбудову над традиційними *інформаційно-пошуковими системами* (ІПС) як загального призначення, так і спеціалізованими.

Актуальність проблеми семантизації пошуку

Багато дослідників звертають увагу на різні аспекти розвитку СП та критерії їх оцінювання [2] . Деякі з них докладно аналізують окремі аспекти СП: наприклад, зосереджуються на ПМ-запитах до баз знань або онтологій RDF [3].

Класичні інформаційно-пошукові системи (ІПС) отримують вхідні дані у вигляді запиту користувача, що подається як список ключових слів, а як вихідні дані генерують впорядкований список документів, релевантних цим ключовим словам.

Пошук у природномовних ресурсах

Методи обробки ПМ застосовуються для розуміння семантики запиту або документів, для розпізнавання частини мови (Part-Of-Speech – POS) у текстовому контенті (наприклад, визначити граматичні теги,

такі як іменник, сполучник або дієслово до окремих слів). Такий аналіз дає точні результати для повних, правильно сформованих речень [4], але набагато складніший для коротких текстів, таких як запити [5]. Теги POS можна використовувати також для таких задач, як: 1) розрізнення текстових ключових слів, наприклад, для розпізнавання іменованих об'єктів (Named-Entity Recognition – NER), де завдання полягає в тому, щоб визначити, які слова відповідають екземплярам об'єктів реального світу; 2) розділення співпосилання (co-reference resolution), тобто виявлення всіх ключових слова, що посилаються на ту саму сутність у тексті. Синтаксичний аналіз речень (parsing) виводить аналіз ПМ на наступний рівень, охоплюючи загальну структуру речень (як правило, через дерево аналізу залежностей).

Методи обробки ПМ часто поєднуються з лексичними базами знань для того, щоб ідентифікувати об'єкти в текстовому запиті або контенті та зіставити їх із відповідними об'єктами у базі знань (або іншому зовнішньому джерелі інформації про предметну область, такі як вікіресурси) для покращення результатів пошуку. Наприклад, WordNet використовується [6] для усунення неоднозначності слів ПМ. Аналіз структури Вікіпедії дозволяє краще оцінювати відповідність об'єктів пошуковому запиту [7] або для ідентифікації сутностей під час пошуку у колекціях документів [8] , тоді як [9] обробляють згадки сутності у Вікіпедії разом з іншими характеристиками документів.

Концептуально подібні підходи використовуються й у контексті Semantic Web – замість вікіресурсу аналізуються більш структуровані інформаційні джерела, такі як онтології, для отримання інформації щодо структури та екземплярів інформаційних об'єктів для кращого розуміння контексту запитів або сутностей, описаних в ПМ-документах. Наприклад, [10] пропонують використовувати онтології для інтерпретації ПМ-запитів, перетворюючи на основі дескриптивної логіки набір ключових слів на кон'юнкцію понять. Інші дослідника [11] використовують об'єкти та семантичні зв'язки з бази знань DBpedia для групування результатів пошукової системи у більш значущі групи.

Методи машинного навчання часто використовуються у СП для визначення семантики слів або сутностей у документах на основі гіпотези семантичної подібності слів, які зустрічаються в подібних контекстах. Ранні підходи в цій царині пов'язані з побудовою багатовимірних матриць, де кожне слово представлене розрідженим вектором у просторі великої розмірності. Зараз використовуються методи аналізу близькості слів на основі щільних векторних [12], які визначають максимальну ймовірність спільної появи слів у певному контексті за допомогою нейронної мережі. Це дозволяє генерувати векторні представлення слів із великих текстових корпусів для машинного навчання.

Пошук у базах знань

Деякі підходи до СП спрямовані на обробку декларативних баз знань (таких як онтології або графи знань), а не на колекцій ПМ-документів. Такі бази знань можуть бути подані багатьма різними способами, але наразі більшість існуючих реалізацій базується на стандартах Semantic Web – RDF та OWL. Відомі приклади таких баз знань – Google Knowledge Graph [13] , Dbpedia [14] та Wikidata [15] . Користувачі можуть будувати запити до таких джерел інформації як до традиційних ІПС (наприклад, як набір ключових слів) або структуровано (наприклад, SPARQL-запити) .

У багатьох дослідженнях, пов'язаних із проєктом Semantic Web [16, 17], аналізується формальний підхід до виконання структурованих запитів до онтологій, де дескриптивна логіка використовується для ранжирування результатів пошуку. Для покращення пошукових запитів, орієнтованих на онтології, можуть застосовуватися мета-онтології: як-от, у [18] WordNet використовується для зіставлення елементів онтології з лексичними об'єктами.

У роботах [19] та [20] розглянуто пошук у ресурсах спеціалізованих інформаційних об'єктів, які характеризуються

типами або відношеннями на основі природномовних запитів або ключових слів.

Гібридні підходи до пошуку використовують як текстовий, так і структурований контент: ІПС доповнює запити з ключових слів шляхом дослідження онтологічного графу [21], пошук використовує текстові метадані про об'єкти в структурованому сховищі [22].

Пошук на основі векторного представлення об'єктів адаптується для забезпечення пошуку у базах знань для аналізу RDF-графів у базах знань. У [23] підходи до вбудовування баз знань поділяються на дві основні групи: методи, засновані на перекладі, котрі інтерпретують зв'язки в базі знань як вектор трансляції між двома об'єктами, пов'язаними відношенням, і моделі семантичної відповідності, які використовують функції оцінки подібності об'єктів.

Приклади семантичних ІПС

Практично всі провідні промислові ІПС, такі як Google або Bing, так чи інакше реалізують семантичний пошук, але зазвичай не оприлюднюють детально методи, які використовують. Здебільшого вони підтримують пошук у масивах документів, що значно різняться рівнем структурованості та якістю метаданих, а бази знань використовують для вдосконалення запитів та для кращого фільтрування та сортування списку або повернутих документів у цьому контексті. Деякі семантичні ІПС (наприклад, Swoogle [24]) спеціалізуються на пошуку саме у базах знань, здебільшого поданих на основі стандартів Semantic Web – RDF та OWL, але використовують для цього традиційні пошукові алгоритми. Наприклад, SWSE [25] здійснює семантичний пошук у наборах RDF на основі їхніх метаданих та впорядковує результати пошуку з використанням алгоритму PageRank на графі відношень між URI та їхніми джерелами у тріплетах RDF.

SemSearch [26] здійснює пошук у ресурсах Semantic Web, перетворюючи запити користувача за ключовими словами на формальні запити. Sindice [27] підримує пошук у напівструктурованих даних великого обсягу як за ключовими словами та URI, так і структуровані запити.

Пошукова система Watson [28] здійснює пошук в онтологіях, виконуючи пошукові запити за ключовими словами та SPARQL-запити.

Напрямки семантизації пошуку

Наведений вище огляд показує, що методи семантичного пошуку можна розділити на дві основні групи залежно від цільового контенту [29] :

- методи підвищення релевантності класичних пошукових систем, де запит складається з тексту природною мовою (ПМ) – наприклад, списку ключових слів, а результати є ранжованим списком документів – наприклад, веб-сторінок або докумнетів;
- методи пошуку частково структурованих даних (зокрема, інформаційних об'єктів певної структури або RDF-трійок) у базі знань (наприклад, в онтології, семантично розміченому Wiki-ресурсі або графі знань) за запитом користувача, який може подаватися у формі ПМ-тексту або декларативної мови запитів, як-от, SPARQL.

Для обох груп використовується широкий спектр методів, таких як обробка природної мови для кращого розуміння запиту та контенту даних, технології Semantic Web для керування процесом пошуку з використанням декларативних баз знань, таких як онтології, а також машинного навчання.

Складові Інформаційно-пошукових систем

У найбільш загальному вигляді пошук складається з трьох складових: 1. запиту користувача q, що відображає його інформаційну потребу; 2. масиву даних I, в яких здійснюється пошук; 3. результатів пошуку R – тієї інформації, яку отримує користувач внаслідок виконання пошукової процедури. В такому випадку пошук можна розглядати як функцію $R = S(q, I)$, таку, що $R \subseteq I$.

Семантичний пошук є одним з підтипів інформаційного пошуку $R_{sem} \subseteq R$, який має ті самі складові, але доповнюється використанням зовнішніх джерел знань

К та методами їх застосування у пошуковому процесі: $R_{sem} = S_{sem}(q, I, K)$.

Методи пошуку S значною мірою залежать саме від цих трьох складових. У відкритому інформаційному просторі є можливість контролювати тільки q та S, тому дослідження та порівняння алгоритмів пошуку виконуються на спеціально сформованих тестових наборах I. Системи пошуку досить складно порівнювати саме через те, що вони значно різняться за всіма цими складовими. Тому доцільно визначати координати кожної застосовної системи у такому тривимірному просторі. Але для цього доцільно впорядкувати певним чином типові варіанти значень всіх цих параметрів відповідно до їхньої складності. Ця задача не є тривіальною через те, що серед них зустрічається багато непорівнюваних значень, і тому така класифікація є нечіткою. Тож виникає потреба проаналізувати можливості семантизації кожної із цих трьох складових та визначити, як саме вони впливають на методи пошуку.

Постановка задачі

Через велике розмаїття моделей, методів та засобів пошуку інформації, що ускладнюється внаслідок семантизації пошукових процедур, виникає проблема співставлення та вибору тих пошукових сервісів, що відповідають потребам користувачів застосовних систем. Цей вибір має враховувати як особливості ресурсів, серед яких планується здійснювати пошук, так і способи подання інформаційних потреб користувачів. Тому недостатньо аналізувати лише методи співставлення запитів з наявними джерелами інформації. Виникає потреба більш точно описувати властивості таких складових пошуку, як запити, результати запитів та інформаційні ресурси. Для цього пропонується тривимірна модель семантичного пошуку, яка базується на аналізі цих трьох складових та доповнює класифікацію систем семантичного пошуку. Для того, щоб обирати ІПС, що пертинентна певній задачі, необхідно визначити, які значення можуть мати ці параметри, та встановити часткове впорядкування цих значень там, де це можливо, відповідно до їхньої відповідності проблемі семантизації пошуку.

Пошукові запити та їх семанитизація

Запит користувача – це формалізований опис інформації, доступ до якої він прагне отримати. Цей опис може містити ключові слова, пов'язані логічними операторами; документ-зразок; тип документа і його тему за класифікатором; списки рекомендованих чи заборонених користувачем інформаційних джерел; обмеження стосовно часу або обсягу пошуку тощо. Деякі ІПС дають змогу також вводити такі параметри шуканого ІР, як час створення, обсяг, мова подання тощо. У більш складних або спеціалізованих пошукових механізмах користувач може вказувати тип інформаційного об'єкта, відомості про який він прагне отримати з наявних природномовних ІР (приміром, Web-сервіс, дані про особу чи організацію). Найпростішим варіантом запиту є непорожній набір послідовностей символів. Найчастіше це набір слів ПМ або чисел, але у більш узагальненому випадку можуть застосовуватися будь-які послідовності символів, що не потребують додаткової інтерпретації змісту (наприклад, пошук масок вірусів у файлах). Якщо ж використовуються саме слова певної мови – природної або формальної, то запит може уточнюватися та доповнюватися на основі знань щодо цієї мови.

Ускладнення запиту збільшує час його обробки, але використання елементів штучного інтелекту та менеджменту знань для побудови запитів дозволяють значно підвищити пертинентність його результатів. Тому дослідники в сфері СП значну увагу приділяють класифікації засобів семантизації пошукових запитів та доцільності їх застосування для різних задач.

Запити з ключових слів та їх розширення

Традиційні підходи до розширення запиту (query expansion – QE) спираються на інтеграцію неструктурованого корпусу та імовірнісних правил для виділення термінів – кандидатів для роз-

ширення. Ці методи не враховують семантику пошукового запиту, що призводить до неефективного пошуку інформації. Семантичні підходи до QE долають це обмеження, завдяки чому пошуковий запит розширюється значущими термінами, які відповідають потребам користувача. Ці підходи застосовують різні моделі та стратегії до різних структури знань – лінгвістичні методи, методи на основі онтології тощо. Таксономія таких методів, верхній рівень якої наведено на рис.1, пропонується в [30].

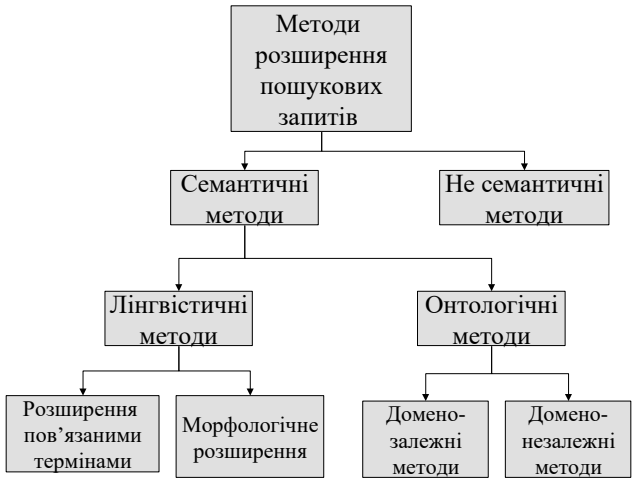

Рис.1. Таксономія методів розширення пошукових запитів

Основна задача ІПС полягає в тому, щоб відібрати документи, які містять потрібні користувачам відомості, та здобути з них ці відомості. Для цього ІПС обчислюють подібність між пошуковим запитом і документами та отримують список документів, розташованих у порядку зменшення подібності. Отриманий список документів іноді завеликий і містить багато нерелевантних документів. Основною проблемою, яка виникає під час пошуку, є невідповідність терміносистем користувачів та авторів документів: терміни, які автор використав для опису поняття в документі, відрізняються для користувачів. Основними причинами цього є вживання в ПМ слів зі схожим значенням (синонімія) та багатозначних слів (полісемія).

Ця проблема невідповідності ще більше посилюється короткими запитами, які користувачі застосовують для пошуку у Web: більшість таких запитів містить не більше двох-трьох слів [31], а цього недостатньо для автоматичного розв'язання неоднозначності.

Розширення запиту (QE) спрямоване на збільшення набору слів у запиті, і семантичні підходи мають суттєві переваги перед ручними та статистичними методами , оскільки вони розширюють кожен пошуковий запит значущими поняттями, які беруться зі структури знань (створеної вручну або автоматично) для представлення пошукового запиту. Нехай запит q містить непорожню множину ключових слів: $q = \{k_1,...,k_i \mid i \geq 1\}$, тоді розширення запиту – це перетворення його на набір $q_{ex} = \{k_1,...,k_m \mid m \geq i\}$. Залежно від того, з яких джерел здобуваються додаткові терміни та як саме здійснюється це здобуття, можна поділити методи розширення запитів на несемантичні (ручні або статистичні) та семантичні.

В свою чергу, існуючі семантичні підходи поділяють на лінгвістичні, онтологічні та змішані (гібридні). У лінгвістичних підходах значення слів виводяться з лінгвістичних баз знань – словників, тезаурусів тощо, які містять синоніми, гіпоніми та інші можливі зв'язки слів між поняттями, що відповідають словам з пошукового запиту, і є термінами розширення. Підходи, що базуються на онтологіях, використовують змістовні відношення між поняттями, що входять до запиту, щоб отримати значущі слова для розширення запиту. Змішаний підхід поєднує особливості лінгвістичного та онтологічного підходів: терміни розширення запиту можуть бути здобуті з баз знань різних типів.

Визначення розширення запиту на основі онтології. Нехай запит q містить непорожню множину ключових слів: $q = \{k_1,...,k_i \mid i \geq 1\}$, таких, що пов'язані з непорожнім набором понять $c = \{c_1,...,c_j \mid j \geq 1\}$ онтології O: $sem(O,q) = c$. Розширення запиту – це перетворення його на набір $q_{ex} = \{k_1,...,k_m \mid m \geq i\}$, таке, що це переформулювання запиту зберігає його семантику, тобто $sem(O,q_{ex}) = c$.

Таке визначення передбачає існування принаймні одного поняття для кожного ключового слова запиту . Крім того, кількість понять може не дорівнювати кількості ключових слів. Розширення запиту не означає розширення понять, передбачених у запиті (тобто мета запиту залишається незмінною), а саме розширення набору ключових слів через включення термінів, більш релевантних вже обраним поняттям, щоб ціль запиту стала більш конкретною та зрозумілою для ІПС.

Згідно з наведеним вище визначенням, можуть виникати два особливі випадки пошукового запиту: 1) всі ключові слова запиту стосуються одного поняття онтології, і тоді розширення запиту здійснюється на основі аналізу цього поняття в онтології [33]. ; 2) всі ключові слова запиту відповідають різним поняттям, тобто ключові слова можна вважати незалежними одне від одного.

Сутність e складається з непорожньої множини атомарних сутностей e^a: $e = \{e^a_1,...,e^a_n\}, n \geq 1$, де кожна атомарна сутність e^a відображає певний елемент інформації, який не можна розділити на інші сутності в обраному контексті. Тож, інтерпретація атомарності сутностей залежить від контексту. Наприклад, ім'я людини в одному контексті може розглядатися як атомарна сутність, а в іншому контексті розкладатися на ім'я та прізвище.

Документ D з точки зору пошуку – це набір сутностей, $D = \{e_1,...,e_k\}, k \geq 1$. В такому розумінні сам документ не є сутністю, але містить набір сутностей. Предметна область (ПрО) – це підмножина світу, що характеризується певною множиною знань, яка може описуватися через корпус ПрО $K = \{D_1,...,D_m\}, m \geq 1$. Кожна сутність ПрО може міститися в документах більш ніж один раз.

Експерти в певній ПрО використовують у побудові запитів власні знання і відомі їм терміни, а для запитів в областях за межами сфери їхньої компетенції такі досвідчені користувачі широко використовували тезауруси та інші зовнішні джерела знань для знаходження термінів. Знання ПрО впливає на поведінку користувачів та забезпечує більш ефективні стратегії вибору термінів, коротші запити [34] і зменшення помилок у тактиці пошуку [35].

Але користувачі-початківці частіше використовують лише свої обмежені знання про область пошуку і рідко звертаються до інших джерел знань, хоча низький рівень їхніх знань потребує більше змін початкового запиту для отримання потрібної інформації. Тому саме вони потребують автоматизованих засобів уточнення та вдосконалення їхніх запитів з використанням знань ПрО, зробивши встановлення зв'язків між запитами та документами більш коректним.

Багато досліджень показує, що онтології дозволяють подолати розрив між термінами запиту та документами, використовуючи семантику ПрО. Онтології та тезауруси, які можуть розглядатися як окремі випадки онтологій зі спрощеною формалізацією, можуть використовуватися і для розширення запиту як джерело релевантних термінів, і у обробці його результатів для усунення неоднозначності та обчислення подібності між запитами та документами.

Онтологія ПрО може розглядатися як комбінація інтенсіональних і екстенсіональних знань. Інтенсіональні знання про домен (TBox) подібні до схеми бази даних та формалізують структуру об'єктів ПрО як набір аксіом, а екстенсіональні знання (ABox) відображають відомості про екземпляри об'єктів. Інтенсіональне знання виражається у TBox.

Основна мета розширення запиту полягає в тому, щоб обчислити терміни, які відповідають намірам користувача, але не містяться в його запиті, і додати їх до початкового пошукового запиту. Традиційні підходи використовують статистичний аналіз вмісту корпусу текстів для знаходження термінів-кандидатів. Тому такі підходи добре працюють тільки тоді, коли доступний великий корпус, а вміст цього корпусу релевантний ПрО пошукового запиту. Семантичні підходи не мають таких обмежень, оскільки вони базуються на незалежних від корпусу зовнішніх джерелах знань (наприклад, лексичному тезаурусі або онтології ПрО).

Онтологія містить знання про структуру понять ПрО, тобто являє собою потенційне джерело відомостей щодо семантично пов'язаних термінів. Семантичне розширення запиту забезпечує інтерпретацію пошукового запиту, використовуючи інформацію про структуру понять. Терміни розширення отримують на основі визначення кількісних оцінок семантичної подібності між початковими термінами пошукового запиту та іншими поняттями ПрО: запит доповнюється тими термінами, які найближче до термінів у запиті користувача.

Структура знань може бути пов'язаною з ПрО (тобто описувати класифікацію та структуру об'єктів певної області) або загальною (наприклад, Cyc і EuroWordNet). Поняття, відношення між поняттями та властивості понять становлять словник структури знань, тим самим фіксуючи набір семантично значущих термінів для розширення запитів. Тому ефективність такого розширення значною мірою залежить як від якості словникового запасу (його точності, повноти, актуального представлення знань), так і пертинентності обраної структури знань інформаційним потребам користувача та відповідності рівня її узагальненості та складності.

Прості таксономічні зв'язки , такі як гіпернімія (гіпернім (hypernym) – "Has-A" – слово з широким значенням, під яке підпадають більш конкретні слова, такі як, "тварина" – це гіпернім слова "собака") і гіпонімія (гіпонім (hyponym) – "Is-A" – слово з конкретнішим значенням, ніж більш загальний термін, наприклад, "пацюк" – це гіпонім до слова "тварина"), дозволяють переходити таксономію вгору і вниз для більш загальних категорій і підкатегорій відповідно. Використання таких зв'язків для розширення запитів забезпечує отримання більш загальних або більш конкретних понять для термінів пошукового запиту зі структури знань. Однак вибір відповідної ієрархічної відстані (наприклад, два або більше рівнів від вихідного поняття) для отримання понять-кандидатів розширення зі структури знань залишається досить складною проблемою.

Інший тип підходів до розширення запитів зосереджується на нетаксономічних відношеннях структури знань, таких як синонімія, тропонімія, антонімія, відношення "частина-ціле", семантична роль, залежність, типове розташування, причинно-наслідкові відношення тощо [36], забезпечуючи структурне представлення змісту слів. Наприклад, у [37] запропоновано метод структурних семантичних взаємозв'язків (structural semantic interconnections – SSI), який створює структурні специфікації можливих значень для кожного слова в контексті та вибирає найкращу гіпотезу відповідно до граматики, що описує зв'язки між змістовними специфікаціями. Метод може застосовуватися до проблем семантичного усунення неоднозначності, таких як автоматична побудова онтології, семантичне розширення запитів та усунення неоднозначності слів у глосарію.

Однорівневі відношення, такі як синоніми, антоніми та зв'язки пов'язаних понять, є ефективними для усунення неоднозначності у значеннях термінів запиту та можуть бути легко отримані з лінгвістичних джерел знань та онтологій ПрО (наприклад, словників, тезаурусів або WordNet).

Значення слів, що описані в лінгвістичних базах знань (наприклад, у WordNet), широко використовуються багатьма дослідниками для усунення неоднозначності початкових термінів пошукового запиту [38] : слова, які здобуті зі зв'язків глосарію WordNet, є кращими кандидатами для розширення запитів, ніж слова вищого або нижчого рівня таксономії.

Кожну категорію можна далі розділити на підкатегорії відповідно до ключових характеристик. Лінгвістичні підходи базуються на інформації про властивості природної мови для створення термінів розширення. До них належать морфологічні підходи, методи на основі синонімії.

Підходи морфологічного розширення використовують морфологічні форми слів запиту (наприклад, основу, частину мови та форми слова) для створення функцій розширення. Експерименти з використанням корпусів різних мов продемонстрували, що розширення запитів морфологічними варіантами термінів за-

питу (автоматично здобутих з документів) дає задовільну продуктивність пошуку [39].

Підходи до розширення з використанням пов'язаних термінів використовують синонімію та інші типи семантично пов'язаних слів природної мови для розширення пошукового запиту. Джерелами таких знань є словники та тезауруси.

Наприклад, найбільш відома лексична база даних WordNet [6] об'єднує функції словника та тезауруса. Вона класифікує слова ПМ на іменники, прикметники, дієслова та прислівники, а також групує слова, які мають однакове значення, в набори, що називаються синсетами. Кожен синсет має семантичні зв'язки з іншими (наприклад, зв'язки гіпонімів і меронімів). Саме синсети надають інформацію для розширення запитів: синсети, найбільш подібні до ключових слів запиту, додаються до цього запиту.

В онтологічних підходах до розширення запитів поняття з онтології додаються до початкових запитів. Для цього можуть використовуватися як онтології верхнього рівня (доменно-незалежні), так і онтології окремих ПрО, а також більш специфічні онтології різної виразності – онтології задач, користувачів тощо. Для здобуття понять-кандидатів можуть використовуватися запити мовою SPARQL. Якщо пошук здійснюється в певній ПрО, то доцільно застосовувати релевантні онтології. У більш узагальнених випадках використовують доменно-незалежні онтології, такі як OpenCyc [40], YAGO [41], DBpedia [42] і UNIpedia [43]. Особливий інтерес становлять онтології, пов'язані із Wiki-технологіями, тому що користувачам легше сприймати їхню структуру та обсяг.

Наприклад, у [44] запропонована модель збагачення семантичного запиту з використанням онтологій Wikipedia та Dbpedia для отримання термінів для розширення, що семантично споріднені з ключовими словами запиту. Але використання таких доменно-незалежних онтологій призводить до двох проблем: 1. загальні онтології містять неоднозначні терміни, що мають різні значення у різних ПрО; 2. такі онтології зазвичай не містять спеціалізовані властивості та специфічні терміни окремих ПрО.

Звичайні методи семантичного розширення запитів не використовують контекст пошуку окремого користувача (тобто профіль користувача чи історію пошуку), необхідний для визначення правильного контексту запиту користувача. Але визначення контексту пошуку запиту користувача важливо з двох причин: 1) однакові пошукові запити різних користувачів можуть мати різні цілі; 2) інформаційні потреби одного користувача можуть з часом змінюватися. Таким чином, потрібно персоніфікувати розширення запитів та розробити засоби відбору актуального контексту.

Окрім профілю користувача та історії його запитів, джерелами для збору персоналізованої інформації можуть бути його профілі та поведінка у соціальних мережах, такі як Twitter, Facebook і Linkedln [45]. Але потрібно враховувати, що здебільшого така інформація є закритою та охороняється законами про персональні дані.

Мультионтологічний підхід, що полягає у використанні кількох онтологій до розширення запитів, є ефективним інструментом для пошуку на перетині кількох ПрО [46]. Але його застосування значно ускладнює необхідність узгодження та вирівнювання таких онтологій.

Тож, запити з ключових слів (незалежно від того, були ці ключові слова введені самим користувачем, чи отримані завдяки різноманітним методам розширення запитів – у тому числі й семантичним) з точки зору ІПС обробляються однаково.

Природномовні запити

Багато сучасних ІПС забезпечують користувачам можливість формулювати запити природною мовою. Обробка природномовних пошукових запитів здебільшого стосується перетворення ПМ-конструкцій у структуровані запити з використанням методів морфологічного, лінгвістичного та семантичного аналізу [47], що знаходяться поза сферою даного дослідження.

Обробка запитів ПМ передбачає такі функції, як видалення стоп-слів, морфологічний пошук (відображення слів запиту у базову форму), розпізнавання частин мови.

Багато ІПС, що підтримують ПМ-запити, використовують онтології для уточнення та співставлення елементів запиту з поняттями відповідної області [48]. Якщо користувач ставить запитання, тобто вводить набір слів, що починаються з прислівника (“який”, “коли”, “як” тощо), які потрібно інтерпретувати у структурований запит, побудувавши відповідну логічну форму: наприклад, перетворити “хто” на “категорія:персоналія”. Якщо є наявні знання щодо ПрО, тоді може бути виконане подальше перетворення, що відповідає специфіці ПрО. Наприклад, перетворити “категорія:персоналія” на “категорія:працівник” або “категорія:пацієнт”. Далі логічна форма перетворюється на вираз відповідної ІПМ. Інформаційними ресурсами, які використовуються для відповідей на запити, можуть бути зовнішні або внутрішні бази знань та онтології. У найбільш узагальненому вигляді обробка ПМ-запитів з використанням онтологій у пошукових системах наведена на рис.2.

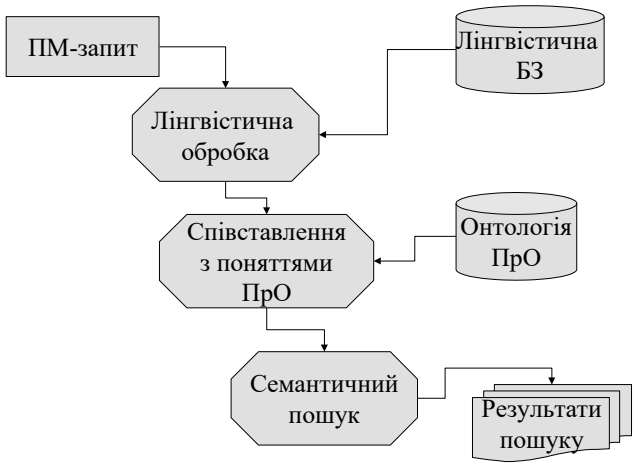

Рис.2. Обробка ПМ-запитів з використанням онтологій

Приклади ІПС, що обробляють ПМ-запити – SemanticWeb Search Engine (SWSE) [49] та Orakel [50]. Google також використовує підходи на основі ПМ для обробки запитів.

Структуровані запити

Більш складним варіантом пошукових запитів є структуровані запити, які дозволяють користувачам формально описувати умови до відомостей, які вони хочуть знайти. Багато традиційних ІПС підтримують такі прості елементи структурування, як кон’юнкція та диз’юнкція ключових слів [51]

Для побудови таких запитів використовуються спеціальні формальні мови – *інформаційно-пошукові мови* (ІПМ) – спеціальні формалізовані штучні мови, створені для відображення інформаційної потреби користувача у такій формі, що забезпечує її співставлення з інформацією про наявні ІР. Залежно від методу побудови системи пошуку ІПМ поділяють на класифікаційні та дескрипторні [52] (рис.3).

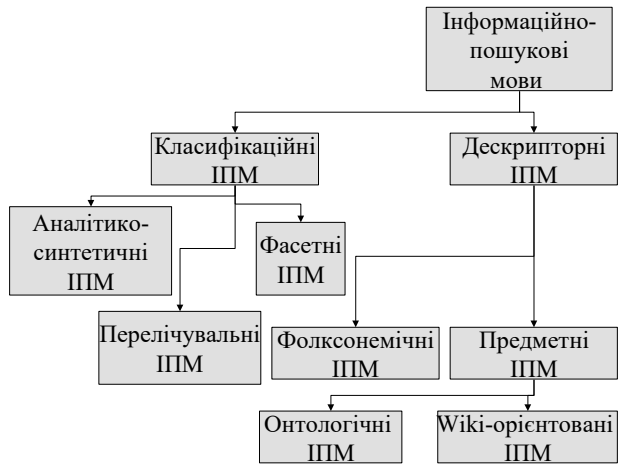

Рис.3. Класифікація ІПМ

Класифікаційні ІПМ порівнюють об’єкти за наборами ознак (які можуть бути пов’язані родо-видовими відношеннями), щоб віднести кожен об’єкт до певного класу. До таких ІПМ належать перелічувальні, аналітико-синтетичні та фасетні мови.

Перелічувальні ІПМ використовують ієрархічні набори ознак пронумерованих класів. На верхньому рівні такі класифікації містять найбільш загальну ознаку. Приклади – десяткова класифікація Дьюї, бібліотечні класифікатори.

Фасетні ІПМ містять сукупності фасетів («фасетна формула» пошукового запиту), які описують комбінації спільних оз-

нак об'єктів. Приклад – класифікація двокрапкою Шіалі Рамамріта Ранганатаном.

Аналітико-синтетичні ІПМ поділяють об'єкти на класи за незалежними ознаками від загальних до більш конкретних. Приклад – Універсальна десяткова класифікація (УДК), де перший фасет є головним, а інші – допоміжні [53] .

Дескрипторні ІПМ описують запит та об'єкти за допомогою ключових слів. Ключове слово, що виражає найзагальніше, головне значення, за допомогою якого можна точно описати зміст документу або запиту, називається дескриптором. Упорядковані в алфавітному порядку дескриптори та їхні синоніми утворюють дескрипторний словник, тоді як більш складні зв'язки між дескрипторами та їхніми значеннями відображає інформаційно-пошуковий тезаурус – структурований словник, що формалізує семантичні відношення (такі як відношення еквівалентності, ієрархічні та асоціативні) між термінами ПМ. Це дозволяє встановлювати змістовні зв'язки між тими ключовими словами запиту, що не є дескрипторами, та власне дескрипторами.

Серед ІПМ, що використовуються у веб-орієнтованих системах, виділяють такі підкатегорії [54]:

- класифікаційні ІПМ: пошук виконується на основі певної класифікаційної системи, каталогів або таксономії (наприклад, пошук на основі категорій у Вікіпедії);
- предметні ІПМ: пошук виконується за допомогою ключових слів або певних предметних рубрик (наприклад, семантичний пошук у ресурсах на основі Semantic MediaWiki [55]);
- дескрипторні ІПМ: пошук виконується за допомогою дескрипторів;
- фолксонемічні ІПМ: у пошуку використовуються різноманітні типи фолксономій, що візуалізуються як хмари тегів, глосаріїв та онтологій.

У багатьох електронних бібліотеках та інформаційно-аналітичних порталах підтримуються одночасно кілька видів пошукових сервісів, що базуються на різних типах ІПМ.

У запитах до структурованих та слабо структурованих ІР можуть застосовуватися складніші ІПМ. Основним питанням у кожній ІПМ є складність оцінки запиту і, зокрема, вплив кожного компонента мови на цю складність. Чим більш розгалуженою є структура запиту – тим більша точність і тим кращий результат пошуку. З іншого боку, ускладнення структури запиту призводить до двох негативних наслідків: ускладнення самого процесу побудови запиту для користувачів та зростання обсягу метаданих, які потрібно зберігати та обробляти по кожному документу.

Структуровані запити передбачають використання знань щодо структури об'єктів пошуку. Для того, щоб задавати у запиті певні умови щодо властивостей (як формальних, так і семантичних) цілого документу або його фрагментів, потрібно визначити назви цих елементів, тобто використовувати певну схему метаданих або структуру того об'єкта, інформацію про який потрібно знайти.

Існує багато мов запитів для пошуку в RDF, такі як DQL, N3QL, R-DEVICE, RDFQ, RDQ, RDQL, SeRQL і т.д., але найпоширенішою є SPARQL – стандарт W3C, який, на відміну від SQL з неоднозначною граматикою і семантикою, має чітку структуру і більшу виразність [55]. Основна частина запиту на SPARQL – шаблон, що описує підграф, який потрібно знайти в графі RDF. Цей шаблон представляється у вигляді набору трійок з перемінними. На сьогодні SPARQL є однією з найбільш виразних мов обробки даних. Крім мови запитів, стандарт SPARQL регламентує протокол взаємодії з базою даних і формат результату, що є великим кроком вперед порівняно із SQL. Наприклад, для пошуку в онтологіях використовують запити мовою SPARQL [56]. Це мова, розроблена для моделі даних RDF. Використання твердження SPARQL як стандартної мови запитів для RDF дозволяє багатьом сховищам даних стати точками доступу SPARQL, у такий спосіб забезпечуючи гнучкий обмін даними між системами. Ця мова є елементом стеку технологій Semantic Web, що підтримує витяг значень зі структурованих і напівструктурованих даних, дослідження відношень між даними та складні об'єднання розрізнених баз даних в одному запиті.

Поширеним прикладом частково структурованих ІР є різноманітні семантичні розширення Wiki-ресурсів, в яких елементи контенту явно пов'язуються за допомогою розмітки з поняттями певної ПрО. Пошук у таких ІР підтримується відповідними ІПМ, що враховують засоби та можливості такої семантизації даних. Виразність ІПМ, що використовується у семантичних Wiki, значно менша за виразність SPARQL, тому що виразність засобів подання знань у Wiki-ресурсах також значно поступається виразності RDF та OWL.

KiWi (http://www.kiwi-project.eu/) – це семантичне розширення Wiki-технології з додатковими можливостями здобуття інформації, персоналізації, логічного виведення та створення запитів. Основними одиницями інформації в KiWi є елементи контенту (Content Items), що розширюють концепцію Wiki-сторінок і можуть бути вкладеними. Кожен такий елемент однозначно ідентифікований своїм URI, може містити фрагменти тексту або мультимедіа, посилання та теги [57] . KWQL – це мова запитів на основі правил, яка поєднує характеристики пошуку за ключовими словами з характеристиками веб-запитів для уможлення різноманітних запитів у KiWi. Мова дозволяє створювати комбіновані запити щодо текстового вмісту, метаданих, структури документа та формальних семантичних анотацій. Запити KWQL варіюються від елементарних і відносно неспецифічних до вибору складних і повністю визначених метаданих.

Для пошуку у Wiki-ресурсах, що семантизовані на основі Semantic MediaWiki, використовується проста, але потужна мова запитів SMW-QL [58]. Мова запитів SMW-QL дозволяє фільтрувати сторінки за заданими критеріями і виводити як результати запиту тільки потрібну інформацію, а не весь текст Wiki-сторінки. Якщо сторінки, з яких отримуються потрібні дані, будуть змінюватися, то результати запитів також будуть автоматично оновлюватися, забезпечуючи несуперечність і погодженість даних.

Найчастіше використовуються вбудовані запити, сполучені з функцією ask, яка має три основні параметри: перший параметр задає умови щодо набору категорій та значень семантичних властивостей сторінок; другий параметр визначає, які саме значення семантичних властивостей цих сторінок потрібні користувачу, а третій параметр вказує форму подання результатів. Таким чином користувач може отримати не тільки перелік документів, а саме потрібні елементи їхнього контенту [59].

Тож, на основі аналізу ІПМ, можна виділити наступні типи запитів, що обробляються в ІПС:

- набори ключових слів (що безпосередньо вводяться користувачами чи будуються на основі таких запитів за допомогою різних методів розширення);
- ПМ-запити, в яких значення має також порядок слів та їхня форма (в ІПС такі запити також перетворюються на набори ключових слів, але вибір ПМ та методи перетворення значною мірою впливають на результати пошуку);
- структуровані запити, в яких явно описані логічні відношення (диз'юнкція, кон'юнкція, заперечення тощо) між термінами та умови щодо властивостей.

Відповідно ІПС за функціоналом обробки пошукових запитів можна класифікувати за наявністю наступних сервісів:

- співставлення набору ключових слів (довільних послідовностей символів) з наявними ІР – мінімальний функціонал ІПС;
- розширення набору ключових слів за допомогою зовнішніх та внутрішніх джерел знань;
- перетворення ПМ-запитів на набори ключових слів (видалення роздільників, ключових слів, виправлення орфографічних помилок, перетворення слів на нормальну форму);
- обробка структурованих запитів, де ключові слова пов'язані логічними відношеннями та обмеженнями;
- перетворення ПМ-запитів на структуровані запити на основі морфологічного, лінгвістичного та семантичного аналізу.

Інформаційні ресурси, серед яких здійснюється пошук

Інформаційні ресурси, серед яких здійснюється пошук різними ІПС, значно різняться [60]:

- моделями подання інформації;
- рівнем структурованості контенту;
- ступенем розподіленості;
- обсягом.

Пошук може здійснюватися на окремому носії, на певному сайті або порталі, у локальній мережі, у базі знань, у відкритому середовищі Web тощо. Інформація, серед якої здійснюється пошук, може бути однорідною або гетерогенною. Метадані, що характеризують наявну інформацію, можуть бути уніфіковані або різнорідні та потребувати інтеграції й узгодження. Чим більше попередніх умов накладено на структуру та подання відомостей в ІР, тим складніші та точніші пошукові запити можна будувати з використанням цих вимог.

Залежно від рівня структурованості, ІР поділяють на:

- структуровані;
- слабо структуровані;
- неструктуровані.

ІР можуть розглядатися як неструктуровані, якщо вони містять певні структурні елементи, але ці елементи не можуть бути використані для мети пошуку [55].

Найбільш розповсюдженою моделлю збереження структурованих даних з кінця 70-х років 20 ст. є реляційна модель, а стандартом на їхню обробку – мова SQL. Однак для НСД ця модель неефективна.

Існує велика кількість ІПС, що спеціалізуються на пошуку певних типів ІР зі специфічними метаописами (відео, музика, мапи, книги тощо) або на пошуку у певних ПрО (наприклад, товари в електронних магазинах). Крім того, у багатьох інформаційно-аналітичних системах використовуються спеціалізовані сервіси, що підтримують пошук різноманітних складних інформаційних об’єктів – подорожей, навчальних курсів [61].

Якщо пошук здійснюється в ІР великого обсягу або таких, що швидко змінюються, то це потребує застосування масштабованих технологій подання даних та відповідних методів пошуку в них. Наприклад, для задачі, що виходять за рамки реляційної моделі, прийнято використовувати моделі даних класу NoSQL, такі як документо-орієнтовані, об'єктні та графові БД. Такі БД мають певні обмеження на операції, що підтримуються традиційними БД. Наприклад, великі розподілені БД повністю відмовляються від транзакцій, що забезпечує підвищення продуктивності за рахунок використання паралелізму. Інший клас задач, які важко розв'язувати на реляційній моделі, – це задачі на сильно зв'язаних даних (графові задачі). Для них сьогодні найбільше поширення мають RDF-сховища, які використовують стандарти W3C для мови RDF (Resource Description Framework) і запити SPARQL.

Ще однією важливою умовою пошуку є відкритість даних. Наявність доступу до даних є основою їх повторного використання.

Багато вимог щодо підтримки доступності та ефективного пошуку інформації відображено в FAIR [62] – принципах керування даними (а саме – знаходжуваності, доступності, інтероперабельності та повторного використання) без утручання користувача, що були розроблені для формування цифрової інфраструктури трансферу наукових даних. Згідно FAIR, функції пошуку, здобуття і представлення даних реалізують не користувачі, а інформаційна система. Водночас мова йде не тільки про власне дані і метадані, а й про алгоритми та інструменти керування ними. Щоб використовувати дані, їх необхідно спочатку знайти там, де вони зберігаються. Метадані та дані повинні бути легко доступними як для людей, так і для комп’ютерів, і тому вимоги FAIR чітко характеризують ті властивості ІР, що мають забезпечити їх знаходження та автоматизовану обробку метаданих.

Значна частка існуючих ІР, що розроблялися незалежно до цих принципів, але з урахуванням можливостей пошуку, відповідають вимогам FAIR. Наприклад, наведений в [63] аналіз виразних властивостей середовища Semantic MediaWiki свідчить про те, що семантичні Wiki-

ресурси, які будуються в цьому середовищі, відповідають вимогам до відкритих даних великого обсягу.

Результати пошуку

Серед типів пошуку виокремлюють:

- адресний пошук, коли результатом структурованого запиту є посилання (адреси, імена) документів, файлів, вебсайтів тощо;
- документальний пошук, коли результат запиту – це або сам документ, або додаткові метадані про нього;
- фактографічний пошук, коли результатом пошуку є певна інформація, здобута з доступних ІР.

Залежно від того, як задаються умови пошуку, результати пошуку можуть обмежуватися певною кількістю знайдених об'єктів або певною межею, що визначає рівень релевантності запиту та тих об'єктів, з якими цей запит співставляється.

Усі ці типи пошуку можуть бути результатом обробки як набору ключових слів, так і структурованого запиту, але в першому випадку потрібно окремо вказувати, що саме має бути результатом запиту.

ІР, серед яких здійснюється пошук, можуть значно різнитися (містити текст, зображення, відео, програмний код, структуровані дані тощо) та супроводжуватися різними видами метаданих (що характеризують документ в цілому або також ї його складові), і саме це є причиною того, що й результати пошуку можуть вказувати на певні документи або знаходити окремі елементи цих документів, що відповідають запиту. Крім того, запити можуть явно визначати умови щодо того, яку інформацію потрібно надати користувачу.Тому іноді структурно прості результати пошуку можуть бути результатом семантичного пошуку та обробки запитів зі складною структурою.

Отже, за конкретизацією результати запитів можна класифікувати наступним чином:

- бінарні ("так-ні") відповіді щодо наявності потрібної інформації у наявних ресурсах (наприклад, чи наявні документи, що містять рядок символів "абв" або чи існує сайт "abc.org");
- кількісні (скільки документів містять рядок символів "абв" або скільки разів у поточному документі зустрічається цей рядок);
- посилання на документи (URL, імена файлів, вікісторінки тощо), які відповідають умовам запиту;
- посилання на інформаційні об'єкти (наприклад, класи онтології або елементи документів), які відповідають умовам запиту;
- обрані відповідно до умов запиту значення властивостей знайдених об'єктів (документів або їхніх елементів), які визначені користувачем;
- більш складні результати обробки таких знайдених значень властивостей (наприклад, сума отриманих значень або графік).

Досить часто підсистеми семантичного пошуку підтримують усі ці варіанти надання результатів пошуку (наприклад, пошук у семантизованих вікіресурсах), але наявність такої класифікації може значно спростити аналіз придатності конкретного технологічного середовища для задач користувача

Висновки та перспективи

Основною метою цього дослідження було визначення тенденцій розвитку сервісів семантичного пошуку, що можуть бути застосовані для підтримки функціювання інформаційно-аналітичних порталів, які базуються на вікітехнологіях [64]. Аналіз існуючих підходів до семантизації пошуку та результатів їх застосування дозволяє виокремити перспективні напрямки впровадження елементів онтологічного аналізу у розробку таких систем.

Запропонована тривимірна модель семантичного пошуку пропонується як додатковий інструмент опису та співставлення пошукових систем, що використовують різноманітні елементи штучного інтелекту та менеджменту знань для більш ефективного та пертинентного задоволення інформаційних потреб користувачів. Як такі методи виконання співставлення між

запитами та ІР в цій моделі не аналізуються, тому що це співставлення є наступним кроком виконання пошуку. Потрібно зауважити, що значення параметрів, які аналізує ця модель, не є взаємозаперечними, тобто та сама ІПС може підтримувати кілька варіантів пошуку. Крім того, засоби подання запитів та ресурсів не завжди є порівнюваними (наприклад, структуровані дані, що описуються різними схемами метаданих, можуть бути орієнтовані на різні типи задач та відображати різні аспекти даних, але водночас одна схема не є повнішою, але виразнішою за іншу). Так само, різні способи фільтрації та подання результатів мають відповідати різним потребам і не завжди можуть порівнюватися за виразністю (наприклад, можливість візуалізації отриманих значень у вигляді графіка не може бути порівняна з можливістю виконання логічних або арифметичних операцій над цими значеннями). Але наявність самих критеріїв порівняння та розширений набір їхніх параметрів надає більш зручний апарат для вибору відповідної ІПС. Отже, така модель дозволяє виявляти ІПС, для яких перетинаються тріади "запит-ІР-результат" та виконувати їх порівняння саме на цих підкласах пошукових задач. Це дозволяє визначати, які алгоритми пошуку виявляються більш пертинентними для конкретних задач користувачів і на основі цього обирати такі сервіси як джерело інформації для подальшої обробки.

Важливою особливістю запропонованої моделі є те, що вона використовує лише ті характеристики ІПС, які можуть бути проаналізовані користувачами (алгоритми, які використовуються в ІПС для співставлення запитів та ресурсів, доступні тільки розробникам цих систем, а їхні наявні описи – приміром, в наукових публікаціях або в документації – можуть значно відрізнятися від використаних у поточній версії програмної реалізації).

References

1. Rogushina, J. (2015) The Web semantic ontology-based search: development of models, tools and methods – Melitopol, 291 p. (in Ukrainian)
2. Bast, H., Buchhold, B, Haussmann, E. (2016) Semantic search on text and knowledge bases. Foundations and Trends in Information Retrieval 10(2-3): 119-271.
3. Mangold, C. (2007) A survey and classification of semantic search approaches. Metadata Semantic Ontologies 2(l):23-34.
4. Manning, C. (2011) Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? Gelbukh AF (Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, 171-189.
5. Hua, W., Wang, Z., Wang, H, Zheng, K, Zhou, X (2015) Short text understanding through lexical-semantic analysis. In: 2015 IEEE 31st International Conference on Data Engineering, 495-506.
6. Fellbaum, C. (2010). WordNet. In: Theory and applications of ontology: computer applications, 231-243.
7. Pehcevski, J., Vercoustre, A., Thom, J. (2008) Exploiting locality of Wikipedia links in entity ranking. In: Advances in Information Retrieval, Springer Berlin Heidelberg, , 258-269.
8. Kaptein, R., Serdyukov, P., de Vries A., Kamps, J. (2010) Entity ranking using wikipedia as a pivot. In: Proc. of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, 69-78.
9. Schuhmacher, M., Dietz, L., Ponzetto S (2015) Ranking entities for web queries through text and knowledge. In: Proc. of the 24th ACM International on Conference on Information and Knowledge Management, 1461-1470.
10. Tran, T., Cimiano, P., Rudolph, S., Studer, R. (2007) Ontology-based interpretation of keywords for semantic search. In: Proc. of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference, ISWC'07/ASWC'07, 523-536.
11. Schuhmacher, M., Ponzetto, S.P. (2013) Exploiting dbpedia for web search results clustering. In: Proc. of the 2013 Workshop on Automated Knowledge Base Construction, ACM, DOI 10.1145/2509558. 2509574.
12. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013) Efficient estimation of

word representations in vector space. arXiv preprint arXiv:1301.3781.
13. Zou, X. (2020). A survey on application of knowledge graph. In: Journal of Physics: Conference Series Vol. 1487, No. 1, 012-016.
14. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Bizer, C. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic web, 6(2), 167-195.
15. Vrandečić, D., Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.
16. Horrocks, I., Tessaris, S. (2002) Querying the semantic web: A formal approach. In: Horrocks I., Hendler J. (eds) The Semantic Web, ISWC 2002, 177-191
17. Stojanovic, N., Studer, R., Stojanovic, L. (2003). An approach for the ranking of query results in the semantic web. In: The Semantic Web-ISWC 2003: Second International Semantic Web Conference, . Proc. 2, 500-516.
18. Maedche, A., Motik, B., Stojanovic, L., Studer, R., Volz, R. (2003). An infrastructure for searching, reusing and evolving distributed ontologies. In: Proc. of the 12th international conference on World Wide Web, 439-448).
19. Tonon, A., Demartini, G., Cudrd-Mauroux, P. (2012) Combining inverted indices and structured search for ad-hoc object retrieval. In: Proc. of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12,125-134, DOI 10.1145/2348283
20. Pound, J., Mika, P., Zaragoza, H. (2010). Ad-hoc object retrieval in the web of data. In: Proc. of the 19th international conference on World Wide Web, 771-780.
21. Rocha, C., Schwabe, D., Aragao, M. P. (2004). A hybrid approach for searching in the semantic web. In Proc. of the 13th international conference on World Wide Web, 374-383).
22. Zhang, L., Yu, Y., Zhou, J., Lin, C., & Yang, Y. (2005). An enhanced model for searching in semantic portals. In Proc. of the 14th international conference on World Wide Web, 453-462).
23. Wang, Q., Mao, Z., Wang, B., Guo, L. (2017) Knowledge graph embedding: A survey of approaches and applications. In: IEEE Transactions on Knowledge and Data Engineering 29(12):2724-2743,
24. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Sachs, J. (2004). Swoogle: a search and metadata engine for the semantic web. In: Proc. of the thirteenth ACM international conference on Information and knowledge management, 652-659.
25. Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A., Decker, S. (2011) Searching and browsing linked data with swse: The semantic web search engine. In: Web Semantics: Science, Services and Agents on the World Wide Web 9(4):365-401.
26. Lei, Y., Uren, V.S., Motta, E. (2006) Sem-search: A search engine for the semantic web. In: Managing Knowledge in a World of Networks, 15th International Conference EKAW-2006, 238-245.
27. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhom, H., Tummarello, G. (2008) Sindice.com: a document-oriented lookup index for open linked data. In: IJMSO 3(1):37-52.
28. d'Aquin, M., Motta, E. (2011) Watson, more than a semantic web search engine. In: Semantic web 2(l):55-63.
29. Cudré-Mauroux, P. (2019). Semantic Search. https://exascale.info/assets/pdf/cudre2018abigdata.pdf.
30. Raza, M. A., Mokhtar, R., Ahmad, N., Pasha, M., Pasha, U. (2019). A taxonomy and survey of semantic approaches for query expansion. In: IEEE Access, 7, 17823-17833.
31. Gabrilovich, E., Broder, A., Fontoura, M., Joshi, A., Josifovski, V., Riedel, L., Zhang, T. (2009). Classifying search queries using the web as a source of knowledge. In: ACM Transactions on the Web (TWEB), 3(2), 1-28.
32. Wu, J., Ilyas, I., Weddell, G. (2011). A study of ontology-based query expansion. In: Technical report CS-2011–04. https://cs.uwaterloo.ca/research/tr/2011/CS-2011-04.pdf.
33. Qiu, Y., & Frei, H. P. (1993). Concept based query expansion. In: Proc. of the 16th annual international ACM SIGIR

conference on Research and development in information retrieval, 160-169.
34. Duggan, G. B., Payne, S. J. (2008). Knowledge in the head and on the web: Using topic expertise to aid search. In: Proc. of the SIGCHI conference on Human factors in computing systems, 39-48.
35. Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. In: Journal of the american society for information science and technology, 55(3), 246-258.
36. Loukachevitch, N. V., Dobrov, B. V. (2004). Development of Ontologies with Minimal Set of Conceptual Relations. In: LREC.
37. Navigli, R., Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated web sites. In: Computational Linguistics, 30(2), 151-179.
38. Liu, S., Liu, F., Yu, C., Meng, W. (2004). An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In: Proc. of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 266-272.
39. Moreau, F., Claveau, V., Sébillot, P. (2007). Automatic morphological query expansion using analogy-based machine learning. In: Advances in Information Retrieval: 29th European Conference on IR Research, ECIR 2007, Proc. 29, 222-233).
40. Best, B. J., Gerhart, N., Lebiere, C. (2010). Extracting the ontological structure of OpenCyc for reuse and portability of cognitive models. In: Proc. of the 17th Conference on Behavioral Representation in Modeling and Simulation.
41. Suchanek, F. M., Kasneci, G., Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. In: Journal of Web Semantics, 6(3), 203-217.
42. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Bizer, C. (2015). Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia. In: Semantic web, 6(2), 167-195.
43. Kalender, M., Dang, J., Uskudarli, S. (2010). Unipedia: A unified ontological knowledge platform for semantic content tagging and search. In: 2010 IEEE Fourth International Conference on Semantic Computing, 293-298.
44. Aggarwal, N., Buitelaar, P. (2012,). Query Expansion Using Wikipedia and Dbpedia. In: CLEF (Online Working Notes/Labs/Workshop).
45. Zhou, D., Wu, X., Zhao, W., Lawless, S., Liu, J. (2017). Query expansion with enriched user profiles for personalized search utilizing folksonomy data. In: IEEE Transactions on Knowledge and Data Engineering, 29(7), 1536-1548.
46. Ray, S. K., Singh, S., Joshi, B. P. (2009). Exploring multiple ontologies and WordNet framework to expand query for question answering system. In: Proc. of the First International Conference on Intelligent Human Computer Interaction: (IHCI 2009), 296-305).
47. Deutch, D., Frost, N., & Gilad, A. (2017). Provenance for natural language queries. In: Proc. of the VLDB Endowment, 10(5), 577-588.
48. Unni, M., Baskaran, K. (2011). Overview of approaches to semantic web search. In: International Journal of Computer Science and Communication (IJCSC), 2, 345-349.
49. Sudeepthi, G., Anuradha, G., Babu, M. S. P. (2012). A survey on semantic web search engine. In: International Journal of Computer Science Issues (IJCSI), 9(2), 241-245.
50. Cimiano, P., Haase, P., Heizmann, J., Mantel, M., Studer, R. (2008). Towards portable natural language interfaces to knowledge bases– The case of the ORAKEL system. In: Data & Knowledge Engineering, 65(2), 325-354.
51. Croft, W. B., Turtle, H. R., Lewis, D. D. (1991). The use of phrases and structured queries in information retrieval. In: Proc. of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, 32-45.
52. Teletska, A. O., Zagnitko, A. P., Nadutenko, M. V. (2018). Classification of information search languages. History, philosophy, law, 120. (in Ukrainian)
53. Chowdhury G. G. (2010) Information Retrieval, 3rd edition. London: Facet Publishing, 488 p.

54. Serbin, O. (2008). Representation of information search languages in web-oriented systems. In: Scientific works of the V.I. Vernadskyi National Library of Ukraine, (20), 176-184. (in Ukrainian)
55. Rogushina, J. V. (2019). Means and methods of the unstructured data analysis. In: Problems in programming, (1), 57-77.
56. Pérez, J., Arenas, M., Gutierrez, C. (2009). Semantics and complexity of SPARQL. In: ACM Transactions on Database Systems (TODS), 34(3), 1-45.
57. Weiand, K., Hartl, A., Hausmann, S., Bry, F., Furche, T. (2012). Keyword-Based Search over Semantic Data. Semantic Search over the Web, 159-192.
58. Bao, J., Ding, L., Hendler, J. (2008). Knowledge representation and query in semantic MediaWiki: a formal study. Tetherless World Constellation (RPI) Technical Report. DOI 10.1.1.187.4263.
59. Rogushina, J., Priyma, S., Strokan, O. (2017) Creating and Use of Semantic Wiki Resources: A Study Guide. – Melitopol, 169 p. (in Ukrainian)
60. Rogushina, J., Grishanova, I. (2022) Semantic Information Resources with a Complex Structure: Knowledge Representation, Scaling and Search Problems. In: UkrPROG, CEUR Vol-3501, 158-171.
61. Pryima, S., Rogushina, J., Strokan, O. (2018). Use of semantic technologies in the process of recognizing the outcomes of non-formal and informal learning. In: CEUR Workshop Proceedings, 226-235
62. The FAIR Guiding Principles for scientific data management and stewardship. Available from: https://www.nature.com/articles/sdata2016118.
63. Rogushina, J., Grishanova, I. (2022). Study of principles, models and methods of FAIR paradigm of scientific data management for analysis for BIG data metadata. In: Problems in programming, (4), 26-35.
64. Rogushina, J. (2023). Development of intelligent information analytical webportals based on semantic Wiki technologies: problems and challenges. In: Problems in programming, (3), 66-80.

Одержано: 23.11.2023

Про авторів:

Рогушина Юлія Віталіївна,
Кандидат фіз.-мат.наук, с.н.с.
Інституту програмних систем
НАН України,
публікації в українських виданнях – 200,
публікації в іноземних журналах – 40.
Індекс Хірша: Scopus – 5,
Google Scholar – 20.
ORCIDhttp://orcid.org/0000-0001-7958-2557.

Місце роботи авторів:

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