https://orcid.org/0000-0003-1845-5014

УДК 004.047

О.О. КРЯЖИЧ[*]

ВИЗНАЧЕННЯ РЕЛЕВАНТНОСТІ ІНФОРМАЦІЇ, ОТРИМАНОЇ ВІД ПОШУКОВО-ДОВІДКОВОГО СЕРВІСУ НА ВЕБ-ПЛАТФОРМІ

[*]Східноукраїнський національний університет імені Володимира Даля, м. Сєвєродонецьк, Україна

Анотація. *У статті проаналізовано можливості швидкого визначення релевантності інформації, отриманої з пошукових та довідкових сервісів, розташованих на різноманітних веб-платформах. Визначена важлива особливість для забезпечення живучості інформації в мережі Інтернет для довідкових і пошукових сервісів – забезпечення сервісу масивом із вибіркою даних для підтвердження інформації з першоджерел або механізмом переходу на посилання, яке підтвердить інформацію або забезпечить відповідь за запитом у разі збою. Запропоновано розглядати сервіс пошуку чи довідки за продукційними правилами IF-THEN структур у залежності від наявних фактів. Виведений показник ефективності формалізованої інформації через визначник рівня погрішності інформації при її обробці за умов невизначеності, який висвітлює ті дані, яких не вистачає для опису об'єкта запиту і серед яких, можливо, є вирішення системних проблем. Наведене його представлення через оціночні значення. При розробці пошукового або довідкового сервісу на веб-платформі за підходом, що пропонується, подібні сервіси будуть утримувати не просто масиви інформації, структуровані за тематикою, а сформовані масиви інформації на засадах символьних перетворень із використанням як загальних, так і окремих схем виведення інформації на запит, а також із можливістю переформулювання задач і запитів для виведення максимально повної та різної за структурою інформації за запитом. Запропоновано механізм виявлення істини та хиби при отриманні відповідей на запит від сервісу за правилами нечіткої логіки, коли розуміння множини можна замінити розумінням характеристичної функції. Наведений приклад алгоритмічної реалізації. Наведене може бути використане для оптимізації довідкових та пошукових сервісів на веб-платформах, вдосконалення роботи пошукових систем, різноманітних сервісів розподіленої обробки інформації.*

Ключові слова: *релевантність, живучість, формалізація, масив, граф, алгоритм, веб-сервіс.*

Abstract. *The article analyzes some possibilities of a quick determination of the relevance of information obtained from some search and reference services located on various web platforms. An important feature to ensure the survivability of the information on the Internet for reference and search services has been identified. This feature consists in providing services to an array of sample data to verify information obtained from primary sources or through the mechanism of following links, which confirm the information or provide an answer to the request in case of failure. It is proposed to consider a search service or some reference platforms using the production rules of if-THEN structures, depending on the available facts. An indicator of the effectiveness of the formalized information is derived through a determinant of the level of the information error when processing it under the conditions of uncertainty, providing that the determinant highlights the data that is not enough to describe the object of the request, and that there may be a solution to system problems among the data. Its representation through some estimated values is given. When developing a search or reference service on a web platform using the proposed approach, such services will hold not just arrays of information structured by topic, but formed arrays of information based on symbolic transformations using both general and separate schemes for displaying information on the request, as well as with the ability to reformulate tasks and queries with the aim of getting the most complete and different information on the request. A mechanism for detecting truth and falsity when receiving responses to the request from a service according to the rules of fuzzy logic, when understanding the set can be replaced by understanding the characteristic function, is proposed. An example of an algorithmic implementation is provided as well. All stated above can be used to optimize reference and search services on web platforms, improve the work of search engines and various distributed information processing services.*

Keywords: *relevance, survivability, formalization, array, graph, algorithm, web-service.*

DOI: 10.34121/1028-9763-2021-1-52-63

1. Вступ

У середньому нова інформація в мережі Інтернет цікавить користувачів близько шести годин [1], хоча зберігатися може нескінченно довго [2]. Тобто, інформація залишається в мережі, але у стрічках новин та в топі запитів вона вже не виникає, хоча є доступною за запитом. Використовуючи сервіс пошуку або отримання довідки, реалізованої на веб-платформі, споживач інформації часто має класичні умови [3], коли мета, обмеження та наслідки дій, зроблених на основі отриманої інформації, є невідомими. Виник, навіть, такий термін, як «веб-серфінг» (Інтернет-серфінг) – отримання інформації, пошук її на сайтах для задоволення власних запитів [4].

Пошуковою системою в Інтернеті є деякий он-лайн (мережевий) сервіс – апаратно-програмний комплекс із веб-інтерфейсом для пошуку інформації [5]. Довідковою системою є набір систематизованих файлів визначеної структури з організацією переходів гіперпосиланнями для отримання користувачем інформації з інформаційного сховища [6]. Одночасно або окремо і пошукова, і довідкова система може бути представлена в мережі Інтернет у вигляді сайту (веб-сайту) [7] в залежності від категорії і типу веб-ресурсу [8]. Більшість веб-сайтів у мережі Інтернет утримують неструктуровану і слабо систематизовану інформацію [1], що ускладнює для користувача формування висновку, чи є джерело інформації релевантним. Під релевантністю розуміється ступінь відповідності знайденої інформації потребам користувача [9].

Актуальність даної теми зростає з кожним роком. Наприклад, все більше користувачів використовують мережу Інтернет як швидкий універсальний довідник замість тривалої і кропіткої роботи з першоджерелами. Бувають непоодинокі випадки, коли автори наукових статей у посиланнях приводять Вікіпедію як першоджерело, хоча Вікіпедія є похідним елементом – обробленою інформацією, яка не завжди може бути ідентичною джерелу, на яке вона спирається. Те ж саме стосується різних пошукових та довідкових сервісів, які видають оброблену інформацію, а не доступ до першоджерел – користувач інформації не може визначитися, чи є отримана інформація релевантною, що примушує здійснювати пошук далі, поки не отримає декілька аналогічних підтверджень чи спростувань стосовно свого запиту. З аналогічним прикладом нерелевантної інформації можна іноді зіштовхнутися в Google Scholar, коли сервіс урахування цитувань показує число менше, ніж є насправді. Пов’язано це з тим, що деякі користувачі внесли перекручену інформацію про роботи та цитування, яка не ідентифікується з базовим шаблоном у Google Scholar. Система це розглядає як помилку, як невідповідність, як загублену цитату від іншого твору. І у користувача виникає задача отримання самостійного висновку про релевантність/нерелевантність такої інформації поставленого запиту.

Метою роботи є представлення можливого підходу до визначення релевантності інформації, отриманої за допомогою пошукових, довідкових та різноманітних інформаційних сервісів у мережі Інтернет відповідно до потреб користувача та ідентичності до першоджерел інформації.

Задачі роботи:

– розглянути особливості забезпечення живучості інформації в мережі Інтернет;

– проаналізувати процеси формалізації та збереження структури інформації при її перетвореннях на веб-сервісах;

– запропонувати механізм виявлення істини та хиби при отриманні відповідей на запит від сервісу та його алгоритмічну реалізацію.

2. Постановка задачі

Варто зауважити, що ідейні викладки обробки інформації з забезпеченням її релевантності можна знайти ще в роботах Дж. фон Неймана стосовно оптимізації інформаційних ресурсів. Зокрема, ним було висловлене таке положення: «Ми хотіли б ввести до машини у вигляді електронних схем тільки такі логічні структури, які або потрібні для функціонування повноцінної системи, або дуже зручні, оскільки часто використовуються» [10]. Виходячи з цього висловлювання можна зробити висновок, що для отримання релевантної інформації на запит користувача слід проаналізувати з отриманих на запит результатів основні статичні або динамічні виміри частоти тієї інформації, що найбільше зустрічається серед досліджуваних відповідей.

У цьому випадку виникне задача проведення обробки інформації, що надійшла на запит, таким чином, щоб було можливо виконати такі дії:

– перепланувати проходження завдання (для кращого завантаження наявних ресурсів, для керування черговістю вирішення задач при кластерних обчисленнях);

– визначити найбільш «вузькі місця» інформаційного сервісу, з яким ведеться робота;

– визначити живучість інформаційної технології при різних варіантах конфігурації системи, в тому числі при роботі з технологією через різні браузери, з різною операційною системою тощо;

– визначити міру сумісності від двох до n процесів;

– удосконалити існуючу систему пошуку шляхом введення нових команд і виключення тих, що використовуються зрідка;

– оптимізувати структуру полів команд (кодів операції, адрес, посилань на операнди, даних);

– оптимізувати пропускні здібності каналів зв’язку та ін.

Для звичайного користувача веб-сервісу це занадто трудомістка робота, яка потребує додаткових знань і вмінь. Основною ж потребою користувача постає отримання в мінімально короткий проміжок часу релевантної інформації за його запитом (пошуковий або довідковий сервіс). Відповідно, задачею розробника довідкових та пошукових сервісів на веб-платформах постає забезпечення живучості інформації з дотриманням усіх властивостей при її перетвореннях у формалізовані дані.

3. Живучість інформації на платформах у мережі Інтернет

Користувач Інтернету задає запитання через веб-портал для отримання довідки за допомогою спеціально розробленого інтерфейсу для зв’язку машини і людини. Користувачу не обов’язково знати, як працює цей зв’язок, але для цілеспрямованого пошуку слід логічно розуміти причинно-наслідкові зв’язки. Тобто, для чого потрібна вишукувана інформація і чи знайдене відноситься до області пошуку, які наслідки виникнуть із моменту реалізації дії до отримання результату. Для користувача довідкового сервісу через веб-інтерфейс є найважливішою відповідність програмно-апаратних ресурсів, бо у разі, якщо сайт понад 6-8 с. (останнім часом – 2-3 с.) [11] завантажує необхідну інформацію, користувач перейде на інший ресурс [12], навіть, якщо там джерело інформації важко ідентифікувати за критеріями повноти, несуперечності і своєчасності. Там же відзначається, що перехід користувача на інший ресурс можливий як за умов отримання невідповідної інформації, так і у разі, якщо користувача не задовольняє інтерфейс веб-порталу. Саме для того модель інтерфейсу сервісу виступатиме складною інформаційною системою, яка поєднує в собі блоки термінів у предметній області, задач користувача, структури та властивостей подання, сценарію діалогу та зв’язків з іншими додатками [13]. Якщо портал одночасно є і довідковою, і пошуковою системою, то виникає вимога створення блока вибірки даних і класифікації

задач [14]. Якщо для сервісу пошуку інформації достатньо на запит отримати перелік структурованих визначень та характеристик об'єкта, щодо якого задано пошук, то для сервісу отримання довідки треба представити конкретне визначення та характеристики об'єкта запиту, тобто представлення знань та даних про об'єкт як систему [15]. Математично у випадку пошукового сервісу це можна описати розпливчастими множинами, зокрема, тим, що розпливчаста множина A, не зважаючи на нечіткість її кордонів, може бути точно визначеною шляхом співставлення кожному її об'єкту X числа, яке розташовано між 0 та 1 і яке являє собі ступінь його приналежності до A. А у випадку сервісу довідки виникає вимога конкретизації, що може бути представлена, наприклад, онтологічною залежністю у вигляді упорядкованої множини довільних графів, де ребра графа визначають властивості між концептами. Так реалізований, наприклад, портал Т.Г. Шевченка www.kobzar.ua, який представляє собою унікальну тематичну пошукову систему та довідкову систему щодо культурної спадщини Т.Г. Шевченка. Щоб подібна система була живучою, до неї пред'являється ряд вимог, які можна описати за прикладом вимог до сайту наукового проєкту [16]:

– реалізація внутрішніх задач, знання про які не важливі для кінцевого споживача інформації, але їх виконання необхідно для отримання релевантної інформації (моделі, прогнози, вибірки та ін.);

– обробка та структуризація первинних даних із вичленуванням того, що може бути використане як довідкові дані, а що може бути представлене у вигляді запиту на пошук;

– подрібнення і розподіл виконання окремих завдань порталу, як-то: актуалізація термінів або гіперпосилань, збільшення окремих масивів інформації за рахунок оновлення даних, розширення функціонала пошукового запиту.

Перелічене націлене на відокремлення мір достовірності та невизначеності інформації, отриманої з системи моніторингу якогось об'єкта. З математичної точки зору, отриманий результат пошуку відповіді на питання або довідки тут є похідною від запиту, який зробив користувач.

Але якщо припустити, що ситуація змінилася, користувач задає новий пошук, але до інформаційних масивів пошукової і довідкової системи не були внесені необхідні зміни і доповнення. Такий розвиток взаємодії користувача та машини із приводу пошуку релевантної інформації та забезпечення живучості інформаційної системи можна представити за Д.О. Поспєловим [17], а саме, використати його погляди на ситуаційне управління, коли інформаційний контекст побудовано двома системоутворюючими компонентами – ситуаційною моделлю об'єкта та алгоритму виділення і порівняння ознак розвитку ситуації. А в залежності від того, яка це система – конкретно орієнтована на пошук чи на отримання довідкової інформації слід відповідно використовувати «лабіринтну» гіпотезу мислення у вигляді прямого перебору варіантів ситуації, щоб знайти відповідь на запитання з максимально можливим забезпеченням повноти, несуперечності і своєчасності, або «модельну» гіпотезу з перебором комбінацій ознак для вибору ознак, які найбільше підходять за запитом користувача і можуть бути визначені як релевантні дані. У цьому випадку можна визначити виконання певного алгоритму, який повинен вибрати з певних масивів інформації ті дані, які максимально відповідають запиту користувача. Обрання такого алгоритму утримує в собі дослідження множин щодо ознак, властивостей та критеріїв. При цьому формування алгоритмічного базису під вирішення конкретної задачі залежить від різноманітності алгоритмів, представлених у інформаційній системі, масовості набору алгоритмів для виконання різноманітних задач та можливості їх адаптації до умов використання. Як зазначається у Г.С. Теслера [18], останні три таксони є основою класифікації алгоритмічного базису як фактора процесу обчислень і можуть виступати критеріями оцінки живучості таких інформаційних систем.

4. Формалізація та збереження структури інформації при її перетвореннях

Викладене вище можна продемонструвати на прикладах Вікіпедії, яка є інформаційно-довідковою системою, або бібліографічної бази даних Google Scholar, яка є інформаційно-пошуковою системою.

Користувачі он-лайн енциклопедії Вікіпедії зустрічаються з фактом того, що не мають можливості перевірити знайдену інформацію. Наприклад, джерело є невизначене або джерело визначене, але посилання на це першоджерело вже не працює (рис. 1).

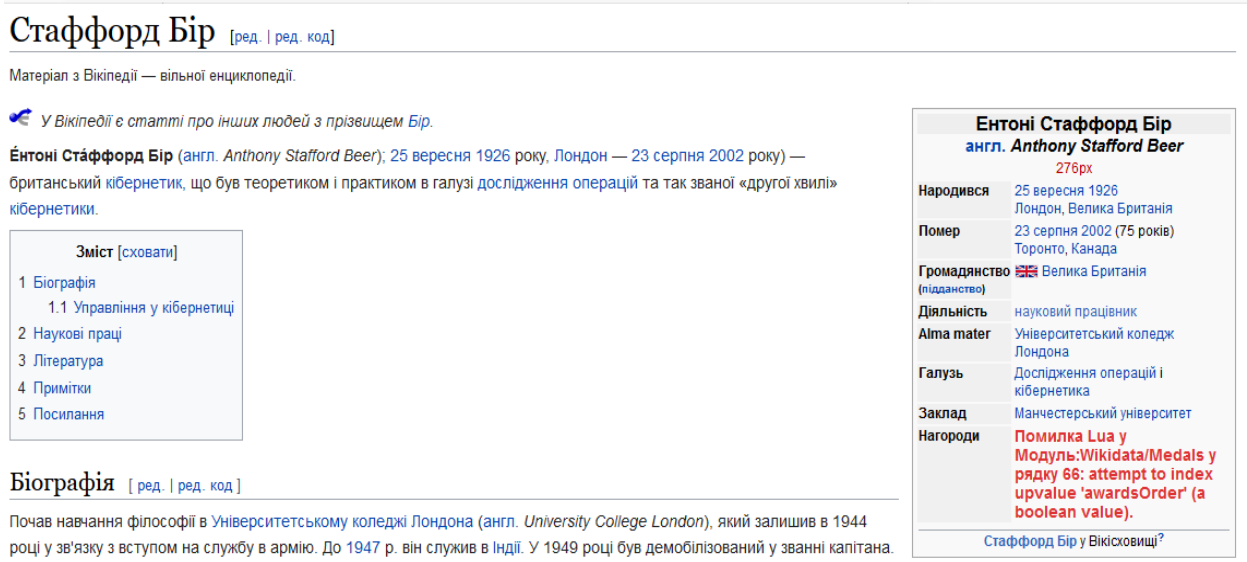

Рисунок 1 – Приклад із Вікіпедії з помилкою посилання на сторінку з шаблоном

Існують деякі механізми відслідковування недійсних посилань, проте остаточно виправити це можна за допомогою користувача, який має змогу/повноваження щодо внесення змін. Такий користувач повинен знайти новий ресурс, перевірити, чи відповідає цей ресурс вимогам Вікіпедії, чи утримує в собі релевантну інформацію стосовно встановленого раніше посилання і відредагувати обрану статтю.

Тобто, дана довідкова система без своєчасного втручання адміністратора не є живучою, а інформація, що надається за допомогою системи в певний момент часу, стає неповною. Для забезпечення повноти такої системи потрібен масив із вибіркою даних. У даному випадку перелік інформаційних ресурсів, на яких представлена підтверджуюча інформація або механізм автоматичного пошуку необхідного посилання, наприклад, при трьох хибних переходах за наведеним першоджерелом.

Бібліографічна база даних Google Scholar у своєму алгоритмі має багато помилок, які призводять до суперечності та несвоєчасності інформації, що надається кінцевому користувачу. Наприклад, пошук нових публікацій або додавання посилань на вже існуючі публікації проходить за якимось лінійним алгоритмом, у зв’язку з чим до профілю додаються статті іншого автора або втрачаються посилання, які за параметрами підпадають під умови невизначеності, хоча за контекстом відносяться до робіт певного автора. Крім того, з’являються помилки – проблеми великих даних, коли неструктурована інформація переробляється в окремі дані, які потім треба поєднувати в режимі адміністрування, як то: деякі відмінності у написанні заголовку чи прізвища автора, похибки у HTML-коді цитувань, помилки при трансформації кириличної інформації у структуровані дані. Все це призводить до накопичування таких помилок, втрачання даних і часу для оптимізації окремих профілів Google Scholar.

5. Формалізація та збереження структури інформації при її перетвореннях

Основу пошукових (ПС) та довідкових сервісів (ДС) складає статистична і аналітична інформація, яка формує масив інформації для вибору оптимальної відповіді на запит користувача. Це можна представити через показник ефективності ($\Delta\Omega$), як це сприймає користувач інформації:

$$\Delta\Omega = \Omega^{ПС/ДС} - \Omega, \tag{1}$$

де Ω – показник ефективності наявної інформації у користувача, на основі якої він має змогу самостійно, без допомоги пошукових та довідкових сервісів дати відповідь на запит;

$\Omega^{ПС/ДС}$ – показник ефективності інформації, яку користувач отримує за допомогою пошукових та довідкових сервісів для більш точної відповіді на запит.

Але слід враховувати той факт [19], що як тільки користувач отримав інформацію від пошукових та довідкових сервісів і використав її, ефективність (користь) цієї інформації починає зменшуватися:

$$\Delta\Omega \to 0 \text{ при } \Omega \to \Omega^{ПС/ДС}. \tag{2}$$

Проблема взаємодії користувача та пошукових і довідкових сервісів постає в узгодженні розуміння необхідного/достатнього [20] для вирішення задач користувача, що виражається через імплікацію

$$A \to B. \tag{3}$$

Тобто, якщо інформація A є повною, то вона достатня для задоволення запиту B. Ця імплікація повинна бути незалежною від впливу розробників пошукових і довідкових сервісів та користувачів і визначати функціональні основи системи, де протиріччя відсутні. Протиріччя виникають при забезпеченні дотримання певної структури інформації при її формалізації з виконанням (3).

Таку формалізацію можна розглянути на прикладі онтологій за допомогою мережевого графа [15]. Саме там прослідковуються чіткі зв'язки відношення між об'єктами, зберігається ієрархія між об'єктами, що забезпечує дотримання структури інформації при її обробці і формалізації, у тому числі великої кількості різнорідної інформації. Цей підхід апробований як у режимі прикладних програм, так і на веб-порталах, зокрема, і для рішення задач щодо мінімізації непрацюючих посилань [13]. В основі онтологічного підходу лежить механізм динамічного формування та використання ієрархій у вигляді таксономій [21]. Відбуваються організація інформації, її класифікація та відображення упорядкованої множини інформаційних ресурсів.

Аналізуючи онтологічний граф [22], можна побачити вершини та терми-об'єкти відповідної онтології, пов'язані з цією вершиною [23]. Щоб отримати інформацію, слід пройти за ієрархічними відношеннями між різними класами об'єктів, роблячи переходи за заданими зв'язками.

Але припустимо, що непуста множина об'єктів не задовольняє вимогам [15], зокрема, немає визначеної ієрархічної структури скінченної множини понять щодо предмета дослідження, існує деяка вільна інтерпретація понять і відношень, функції інтерпретації не формалізовані, аксіоми не визначені. Тобто, масив інформації є необробленим і до початку формалізації слід вирішити задачу щодо забезпечення структури інформації таким чином, щоб вона повністю відповідала вимогам інформаційної системи і могла бути представлена у вигляді певних залежностей і відображалася графами [24].

У такому випадку обробку інформації та формалізацію можна провести за допомогою адаптивних алгоритмів, які дозволяють подавати і структурувати інформацію за певними правилами. Для цього спочатку виконується вибірка інформації, яка явно чи неявно відноситься до теми запиту, потім проводяться функціональні перетворення й застосову-

ються методи породжуваних алгоритмів у системі генерування алгоритмів із використанням нечіткої логіки та побудовою рішень у вигляді лінгвістичних правил-продукцій, подальшої формалізації з застосуванням одного з базових методів наближення для отримання результату за максимумом чи мінімумом відповідності.

За таким підходом, у підсумку, пошуковий та довідковий сервіси будуть утримувати не просто масиви інформації, структуровані за тематикою, а:

– сформовані масиви інформації на засадах символьних перетворень;

– використання як загальних, так і окремих схем виведення інформації на запит;

– переформулювання задач і запитів для виведення максимально повної та різної за структурою інформації за запитом.

Сукупність взаємопов'язаних засобів формального визначення інформації і засобів маніпулювання цими визначеннями являє собою те, що складає термін «база знань» [25–26]. Від того, наскільки повно визначено знання про об'єкт, процес, предметну галузь із позиції отримання максимально повної відповіді на запит, залежать функціональні можливості пошукових та довідкових сервісів.

Якщо визначити C_r як реальні (перевірені, формалізовані, структуровані) дані об'єкт, процес, предметну галузь із позиції функції мети роботи ПС чи ДС, а через C_p виразити поточну інформацію про об'єкт, процес, предметну галузь, яка отримана в результаті моніторингу на якийсь момент часу t, то можна отримати залежність

$$C_r - C_p = \Delta m_t, \tag{4}$$

що можна пояснити, як залежність повноти бази знань від інформації про стан щодо об'єкта, процесу, предметної галузі в конкретний момент часу. Звичайно, що пошуковий (довідковий) сервіс буде живучим, у разі $\Delta m_t \to 0$.

Проте, як зазначено вище, первинна інформація поступатиме неформалізованою і неструктурованою (або слабо систематизованою). Неформалізована інформація N, яка поступатиме до ПС/ДС, повинна бути перетворена на формалізовані дані F, але з виникненням деякої неадекватності:

$$C_r - (N + F) = \Delta m_t. \tag{5}$$

Знову виконуватиметься вимога живучості, коли $(N + F) \to C_r$, а показник $\Delta m_t \to 0$.

Поєднання формалізованих знань із неформалізованими знаннями у ПС/ДС з урахуванням концепту живучості інформаційної системи дозволяє отримати рішення проблемної задачі – забезпечення повної, несуперечної і своєчасної інформації про стан об'єкта чи системи на конкретний момент часу. Такий сервіс можна розглянути за продукційними правилами IF-THEN структур [27], тобто, в залежності від наявних фактів. Тоді залежність (4) можна навести так:

$$C_r - (\sum_{j=1}^{n} N_j + F) = \Delta m_t, \tag{6}$$

де n – кількість вибірок з масиву інформації, який характеризує подію j.

Але формула (6) може бути представлена і так:

$$C_r - (\sum_{j=1}^{n} N_j) = \Delta m_t, \tag{7}$$

що означає неможливість формалізації інформації на момент часу t, наприклад, через відсутність такої інформації.

Слід урахувати ще один випадок формалізації інформації на основі (6):

$$\sum_{k=1}^{L}(C_k - N_k) = \Delta m_t, \tag{8}$$

який має місце бути при запиті і відповідь на який може представляти собою ієрархію (наприклад, з представленням гіперпосилань на інші варіанти відповіді або додаткові інформаційні ресурси). L – кількість рівнів ієрархії відносно деякої події k. Тоді на вищому рівні ієрархії буде представлена інформація більш релевантною, ніж на нижчому рівні, де вона буде обмеженою. Наведену залежність (8) можна назвати частковою формалізацією, а виправляється вона шляхом відсіювання нерелевантної інформації за кожним рівнем ієрархії та встановленням додаткових залежностей, наприклад, між вершинами та термами-об'єктами онтології.

Відповідно до (4), повна формалізація інформації може бути представлена так:

$$C_r - F = 0. \tag{9}$$

Наведені залежності (4)–(9) дозволяють також зазначити, що для їх досягнення можна в кожному окремому випадку використовувати різноманітні функції (прямі і зворотні тригонометричні, гіперболічні, експонента, логарифм), які в підсумку дозволять побудувати зв'язки між концептами декількох онтологій.

6. Виявлення істини та хиби при отриманні відповідей на запит від сервісу

Додавання нових даних до наявної структури інформації пошукових та довідкових сервісів вимагатиме збереження сформованої структури для забезпечення механізму надання відповідей на запити користувача. Це може бути представлено за правилами нечіткої логіки операцією Заде, коли розуміння множини можна замінити розумінням характеристичної функції. Тоді кожен елемент M (концепт) буде розумітися як такий, що належить до деякої нечіткої множини A. А завдяки тому, що операції над нечіткими множинами задаються поелементно, можна через узагальнення булевих функцій отримати невідомі раніше нові дані $x \in M$ із дотриманням тотожностей:

– ідемпотентності, коли відповіді на ідентичні запити будуть представлені однаковими концептами чи однаковими даними з визначеного масиву інформації:

$$A \cap A = A, \; A \cup A = A;$$

– комутативності, коли один концепт чи дані можна замінити іншими:

$$A \cap B = B \cap A, \;\; A \cup B = B \cup A;$$

– асоціативності, коли не виникає черговості у явних пріоритетах при подачі інформації, що особливо важливо при ієрархічності запитів:

$$A \cap (B \cap C) = (A \cap B) \cap C, \;\; A \cup (B \cup C) = (A \cup B) \cup C;$$

– поглинання, коли додаткові дані додаються до існуючого масиву:

$$A \cap (A \cup B) = A, \; A \cup (A \cap B) = A,$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C);$$

– дистрибутивності, що дозволяє оперувати кон'юнкцією і диз'юнкцією в логічних доказах при представленні запитів за умов невизначеності:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C);$$

– інволютивності, яка дозволяє деяку суперечливу інформацію представити як зворотну самій собі (що може бути представлене, наприклад, матричною згорткою):

$$|(|A) = A;$$

– за допомогою правил Де Моргана, коли диз'юнкцію «і» або кон'юнкцію «або» при визначенні концепту можна замінити через іншу операцію, операцію доповнення або заперечення для уточнення інформації за умов невизначеності:

$$|(A \cap B) = |A \cup |B, \ |(A \cup B) = |A \cap |B;$$

– за допомогою граничних умов, коли для формування інформаційного масиву задаються межі внутрішньої та зовнішньої відповідності, а інформація формалізується за встановленими обмеженнями:

$$A \cap \phi = \phi, \ A \cup \phi = \phi, \ A \cup U = U, \ A \cap U = A,$$

що у підсумку забезпечує збереження структури інформації ПС та ДС.

Викладене можна проілюструвати таким прикладом.

Поставлено завдання: знайти всі посилання x на монографію (підмножина A) за умовною назвою «ABC» у пошуковому сервісі Google Scholar (універсальна множина Google – χ). Це можна записати за допомогою характеристичної функції:

$$\chi_a(x) = \begin{cases} 1, \ якщо \ x \in A \\ 0, \ у \ всіх \ інших \ випадках \end{cases}.$$

Маємо $\chi_a(x) = 0$, що не відповідає значенню «істина», бо маємо посилання у друкованих наукових виданнях, які мають інтернет-версії, тобто, можуть бути знайдені пошуковими інструментами Google.

Тоді приймаємо, що також $x \in B$, а підмножини перетинаються $A \cap B$. Тобто відбувається часткове упорядкування $\subseteq$ елементів за решіткою Де Моргана:

$A \subseteq B$ лише і тільки лише у випадку $A \cap B \equiv A$, $A \cap B \equiv B$.

У цьому випадку пошук істини або хиби проходитиме за системою аксіом для операцій Заде:

– кон'юнкції $x \wedge y = \min(x, y)$;

– диз'юнкції $x \vee y = \max(x, y)$,

де x, y – деякі набори операторів із різних множин.

Використовуючи тотожності, зазначені вище, можна вийти на умову ідентичності диз'юнкції в рамках поставленої задачі:

$$x = x \wedge 1 = x \wedge (1 \vee 1) = (x \wedge 1) \vee (x \wedge 1) = x \vee x,$$

де 1 є константою і приймається як «істина».

Відповідно можна вивести й умову хиби для зазначеного випадку, але в рамках вирішення задачі цікавим є лише істинне значення.

Кон'юнкції Заде мають властивості булевих операцій і можуть бути виражені математично, що дозволяє їх використовувати в моделях обробки даних [28], у задачах оптимізації з нечіткою моделлю, що і має місце за умов невизначеності.

Використовуючи алгоритм [29], можна представити формалізацію задачі в термінах нечіткої логіки:

1. Початковою точкою пошуку обираємо назву монографії «ABC».

2. Обираємо терми лінгвістичних змінних, які відповідають підмножині A.

3. Для кожного терму обираємо значення, яке його найкраще характеризує.

4. Робимо крок до підмножини B та відшукуємо там ідентичні терми, чітко дотримуючись правила комутативності щодо термів із множин A і B.

5. Здійснюємо пошук відповідних значень з обраних на етапі 4 характеристик у підмножині B із присвоєнням «1» або «0» кожному з отриманих значень.

6. Після отримання екстремальних значень обираємо проміжні значення, кроки до яких можуть відбуватися не по прямій, а описуватися різноманітними функціями.

7. Присвоюємо «1» або «0» кожному з отриманих проміжних значень.

8. Усім значенням присвоюємо відповідні функції стандартних приналежностей.

9. Формуємо інформаційний масив для пошуку.

10. Визначаємо продукційні правила пошуку.

При роботі з ІПС Google Scholar продукційні правила пошуку можуть бути задані таким чином:

– якщо монографія «ABC»=ISBN 978-966-0000-00-0;

– якщо монографія «ABC»=ББК (номер);

– Якщо монографія «ABC»=УДК (номер);

– якщо автор Z=Наукова установа Q;

– якщо наукова установа Q=Рішення вченої ради (номер, дата) та інші, які виконуються послідовно, у тому числі з заміщенням лівої та правої частин.

Результат пошуку буде обумовлений відповідністю хоча б за одним продукційним правилом або відповідністю частин продукційних правил пошуку з виконанням (5).

Фактично, вирішуючи цю задачу, проводиться доповнення нечіткої підмножини A у множині X підмножиною $\neg B$ з функцією приналежності:

$$\mu_{\neg B}(x) = 1 - \mu_{\neg A}(x), \quad \forall x \in X .$$

Функція $\mu_{(A;B)} : X \to [0;n]$ є функцією приналежності [30] підмножини $A(B)$ до базової множини χ (Google).

Існують декілька підходів [31] та комерційних рішень для роботи з неструктурованими даними, у тому числі з вичленуванням структурованих даних із неструктурованих джерел [32]. Запропонований підхід дозволяє знаходити відповідні за структурою фрагменти з масиву неструктурованої інформації, де в одному записі одночасно представлена інформація різної структури, що значно підвищує живучість пошукових та довідкових сервісів за будь-яких змін зовнішнього/внутрішнього середовища об’єкта запиту.

5. Висновки

1. У процесі аналізу особливостей живучості інформації в мережі Інтернет зазначена одна з важливих особливостей – забезпечення довідкового/пошукового сервісу масивом із вибіркою даних для підтвердження інформації з першоджерел або механізмом переходу на посилання, яке підтвердить інформацію або забезпечить відповідь за запитом у разі збою сервісу.

2. Процес формалізації та збереження структури інформації при обробці інформації на вебсервісах повинен орієнтуватися на забезпечення властивостей інформації про стан об’єкта запиту на конкретний момент часу. Тоді сервіс пошуку чи довідки можна розглядати за продукційними правилами IF-THEN структур у залежності від наявних фактів. Виведено показник ефективності формалізованої інформації через визначник рівня погрішності інформації при її обробці за умов невизначеності, який висвітлює ті дані, яких не вистачає для опису складної системи і серед яких можливо є вирішення системних проблем. Наведено його представлення через оціночні значення.

3. Запропоновано механізм виявлення істини та хиби при отриманні відповідей на запит від сервісу за правилами нечіткої логіки, коли розуміння множини можна замінити розу-

мінням характеристичної функції. Наведено приклад алгоритмічної реалізації шляхом побудови рішень у вигляді лінгвістичних правил, що дозволяє формалізувати інформацію з застосуванням одного з базових методів наближення для отримання результату за максимумом чи мінімумом відповідності та збереженням вже існуючої структури інформації.

Усе наведене вище може бути використане для оптимізації довідкових та пошукових сервісів на веб-платформах, вдосконалення роботи пошукових систем, різноманітних сервісів розподіленої обробки інформації.

СПИСОК ДЖЕРЕЛ

1. Ашманов И., Иванов А. Оптимизация и продвижение сайтов в поисковых системах. СпБ.: Питер, 2011. 464 с.
2. Characterization of Specifications. Characterization of Proposed Standards. IETF. 2014. January, sec. 3. doi:10.17487/RFC7127. RFC 7127. Retrieved 11 March 2016. 5 p.
3. Беллман Р., Заде Л. Принятие решений в расплывчатых условиях / Вопросы анализа и процедуры принятия решений. М.: Мир, 1976. С. 172–215.
4. Новые слова и значения. Словарь-справочник по материалам прессы и литературы 90-х годов XX века. СПб.: Дмитрий Буланин, 2014. 1360 с.
5. Lawrence S., Lee G.C. Accessibility of information on the web. *Nature*. 1999. N 400 (6740): 107–9. Bibcode: 1999Natur.400..107L. doi:10.1038/21987.
6. Технічна енциклопедія TechTrend. URL: http://techtrend.com.ua.
7. Воройский Ф.С. Информатика. Энциклопедический систематизированный словарь-справочник. М.: Физматлит, 2006. С. 432–945.
8. Ромашев В. CMS Drupal: Система управления содержимым сайта. Питер, 2010. 255 p.
9. Carston R., Uchida S. Relevance Theory: Applications and Implications. Amsterdam: John Benjamins Publishing, 1998. 312 p.
10. Фон Нейман Дж. Теория самовоспроизводящихся автоматов. М.: Мир, 1971. 384 с.
11. Федотова Ю. Блог GetGoodRank. URL: http://blog.getgoodrank.ru/vremya-zagruzki-sajta-vliyanie-na-nastroenie-polzovatelej.
12. Быстров И. «Справочник оптимизатора» блога GetGoodRank. URL: http://blog.getgoodrank.ru/stepen-otkazov-bounce-rate.
13. Стрижак О.Є., Попова М.А., Ляшук К.В. Методика створення онтологічного інтерфейсу у середовищі WEB-порталу. *Радіоелектронні і комп'ютерні системи*. 2014. № 2. С. 78–84.
14. Коваленко О.В. Концептуальні основи створення бази даних наукового експерименту та спостереження. *Математичні машини і системи*. 2016. № 2. С. 91–101.
15. Стрижак О.Є. Засоби онтологічної інтеграції і супроводу розподілених просторових та семантичних інформаційних ресурсів. *Екологічна безпека та природокористування*. 2013. № 12. С. 166–177.
16. Коваленко О.В. Сайт наукового проекту: особливості реалізації. *Математичні машини і системи*. 2017. № 3. С. 120–129.
17. Поспелов Д.А. Ситуационное управление: теория и практика. М.: Наука, 1986. 288 с.
18. Теслер Г.С. Новая кибернетика. К.: Логос, 2004. 404 с.
19. Itzhak G. Theory of Decision under Uncertainty. Cambridge: Cambridge University Press, 2009. 232 p.
20. Эдельман С.Л. Математическая логика. М.: Высшая школа, 1975. 176 с.
21. Шаталкин А.И. Таксономия. Основания, принципы и правила. М.: Товарищество научных изданий КМК, 2012. 600 с.
22. Вишняков В.Ю., Стрижак О.Є., Трофимчук О.М. Застосування онтологічного підходу при створенні інструментів геоінформаційних систем на прикладі визначення температурних процесів на території України за даними космічної зйомки. *Екологічна безпека та природокористування*. 2013. Вип. 13. С. 96–113.
23. Довгий С.О., Величко В.Ю., Глоба Л.С., Стрижак О.Є. Комп'ютерні онтології та їх використання у навчальному процесі. Теорія і практика: монографія / Нац. акад. пед. наук України, Інститут обдарованої дитини. Київ: Інститут обдарованої дитини, 2013. 308 с.

24. Стрижак О.Є. Онтологічні інформаційно-аналітичні системи. *Радіоелектронні і комп'ютерні системи.* 2014. № 3. С. 71–76.
25. Субботін С.О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень: Навчальний посібник. Запоріжжя: ЗНТУ, 2008. 341 с.
26. Cordell G., Luckham D., Balzer R., Cheatham T., Rich C. Report on a knowledge-based software assistant. *Readings in artificial intelligence and software engineering*. Morgan Kaufmann, 1986. P. 377–428.
27. Триус Ю.В., Манько М.О. Web-орієнтована консультаційна експертна система з методів оптимізації. *Вісник Черкаського університету. Прикладна математика. Інформатика*. 2014. № 18. С. 99–114.
28. Крянев А.В., Лукин Г.В. Математические методы обработки неопределенных данных. М.: Физматлит, 2003. 216 с.
29. Гриняев С. Нечёткая логика в системах управления. Компьютерра. 2001. N 38. URL: http://masters.donntu.org/2011/fknt/godetskiy/library/docs/4.htm.
30. Рыжов А.П. Элементы теории нечетких множеств и ее приложений. М.: Диалог-МГУ, 1998. 81 с.
31. Гришковский А. Интегрированная обработка неструктурированных данных. *Открытые системы. СУБД*. 2013. № 6. URL: https://www.osp.ru/os/2013/06/13036849.
32. Shilakes C.C., Tylman Ju. Enterprise Information Portals. New York: Merrill Lynch, 1998. 64 p.

Стаття надійшла до редакції 31.01.2021