УДК: 004.738.52 DOI: 10.30929/1995-0519.2019.5.116-122

ПЕРСПЕКТИВИ ВДОСКОНАЛЕННЯ МЕТОДІВ ІНФОРМАЦІЙНОГО ПОШУКУ

В. В. Терещенко
Кременчуцький національний університет імені Михайла Остроградського
вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E-mail: darkwolfthehunter@gmail.com

В умовах розвитку інформаційного суспільства однією з найважливіших задач залишається вирішення проблеми ефективного пошуку і збору інформації. Це, насамперед, пов'язано з зростаючим різноманіттям інформаційних джерел спрямованих на розвиток різних сфер людської діяльності. У роботі проаналізовано принципи функціонування систем інформаційного пошуку та, спираючись на вимоги сьогодення, виокремлено найважливіші тенденції розвитку. Відповідно, проаналізовано ряд наукових досліджень у сфері інформаційного пошуку. В ході дослідження встановлено перспективність використання методики прецендентів у рамках вдосконалення пошукових методів та, зокрема, при побудові систем інформаційного пошуку. Зокрема, наголошено що організація пошуку на основі прецедентів дозволяє об'єднати в собі різні підходи до вирішення завдання інтелектуалізації та персоналізації пошуку і знизити навантаження на індекс пошукового інструменту, а також спростити вирішення проблеми, пов'язаної із забезпеченням конфіденційності даних. Результати, що отримані при проведенні даного дослідження можуть бути використанні при подальшому опрацюванні пошукових методик, розвитку засобів забезпечення пошукових систем, вдосконаленні пошукових алгоритмів.

Ключові слова: пошукова оптимізація, пошукова система, пошукова видача, інформаційний пошук.

ПЕРСПЕКТИВЫ УСОВЕРШЕНСТВОВАНИЯ МЕТОДОВ ИНФОРМАЦИОННОГО ПОИСКА

В. В. Терещенко
Кременчугский национальный университет имени Михаила Остроградского
ул. Первомайская, 20, г. Кременчуг, 39600, Украина. E-mail: darkwolfthehunter@gmail.com

В условиях развития информационного общества одной из важнейших задач остается решение проблемы эффективного поиска и сбора информации. Это прежде всего связано с растущим многообразием информационных источников направленных на развитие различных сфер человеческой деятельности. В работе проанализированы принципы функционирования систем информационного поиска и, опираясь на требования современности, выделены важнейшие тенденции развития. Соответственно, проанализирован ряд научных исследований в области информационного поиска. В ходе исследования установлено перспективность использования методики прецедентов в рамках усовершенствования поисковых методов и, в частности, при построении систем информационного поиска. В частности, отмечено что организация поиска на основе прецедентов позволяет объединить в себе различные подходы к решению задачи интеллектуализации и персонализации поиска и снизить нагрузку на индекс поискового инструмента, а также упростить решение проблемы, связанной с обеспечением конфиденциальности данных. Результаты, полученные при проведении данного исследования могут быть использованы при дальнейшей обработке поисковых методик, развития средств обеспечения поисковых систем, совершенствовании поисковых алгоритмов.

Ключевые слова: поисковая оптимизация, поисковая система, поисковая выдача, информационный поиск.

АКТУАЛЬНІСТЬ РОБОТИ. У сучасних умовах розвитку інформаційних технологій та пошукових машин виникає потреба у нових методах забезпечення ефективного інформаційного пошуку. Проблема досконалого пошуку і збору інформації, яка може бути використаною при вирішенні важливих завдань в ході науково-дослідної діяльності, для підтримки прийняття рішень в науково-технічній, соціальній та інших сферах залишається відкритою впродовж десятиліть. Це зумовлено феноменом стрімкого перенасичення інформаційного простору [1]. До числа причин, які зумовили таке накопичення значних об'ємів інформації та зростання вимог до достовірності даних, що надаються користувачеві слід віднести популяризацію так званих віртуальних університетів та електронних форм навчання, збільшення важливості Internet-простору у сферах дозвілля та розвитку суспільства.

Відомо що сучасна пошукова машина здатна вдосконалюватися в різних напрямках: з'являються нові чинники ранжування або змінюється їх пріоритет, змінюється формат взаємодії інструментарію пошуку з користувачем, посилюються вимоги до якості побудови сайтів, а також з'являються нові сервіси, що спрощують пошук інформації. Відповідно вимоги до швидкості пошуку, актуальності інформації з кожним днем зростають, що в свою чергу впливає на розробку методів та алгоритмів пошуку і подання даних.

В свою чергу, розвиток комп'ютерної техніки також тягне за собою суттєве зростання обсягу інформації, що подається в електронному вигляді. Вплив цього процесу на розвиток сучасних інформаційних технологій, включаючи пошук, відзначається в більшості наукових публікацій в періодичних виданнях [2]. Хоча на сьогоднішній день й існує значна кількість методів та алгоритмів інформаційного пошуку, проте неперервний розвиток цієї галузі та вищезазначені проблеми вказують на необхідність постійного покращення існуючих методів та розробку якісно нових підходів. Тож, відповідно, проблема вдосконалення методів інформаційного пошуку є актуальною. Наукова новизна дослідження полягає в тому, що вперше було розглянуто можливість використання методики прецендентів з точки зору інформаційного пошуку та встановлено перспективність даної методики у рамках вдосконалення пошукових алгоритмів.

Головною метою даної роботи є суттєве поліпшення результатів інформаційного пошуку за рахунок застосування методики прецедентів.

МАТЕРІАЛ І РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ. У загальному випадку під поняттям «інформаційний пошук» розуміють процес відшукання серед деякої множини текстів (документів) таких, які присвячені саме зазначеній в пошуковому запиті темі, або містять потрібні користувачеві факти чи відомості. Пошук може здійснюватися як вручну, так і за допомогою інформаційно-пошукової системи з використанням засобів автоматизації. Залежно від характеру інформації, яка міститься в документах, які містить пошукова видача, – пошук може бути документальним, в тому числі й бібліографічним чи фактографічним.

Проблемам організації інформаційного пошуку було присвячено багато наукових праць. Зокрема, деякі теоретичні аспекти було розглянуто в публікаціях як вітчизняних, так і закордонних дослідників: Урвачова В. А [1], Шокін Ю. І. [2], Климчук С. О. [3], Маннінг К. Д. [4] та ін.

Варту уваги інформацію, з точки зору дослідника, містить стаття В. А. Урвачової [1]. У роботі вченої наводиться короткий огляд сучасних методів та алгоритмів інформаційного пошуку. В огляд також включені класичні алгоритми, які покладені в основу сучасних пошукових методів. Окремої уваги заслуговують зроблені нею висновки щодо перспективності застосування технологій інтелектуальних агентів для пошуку інформації. З урахуванням давності викладеної інформації, слід зазначити, що в монографії Ю. І. Шокіна [2] детально розглянуто загальні аспекти розробки та створення інформаційно-пошукових систем. Зокрема, наводиться докладний виклад моделей, структур і алгоритмів, що описують окремі різновиди інформаційно-пошукових систем. Заслуговують уваги розкриті автором перспективи напряму інформаційного моделювання при організації пошуку.

У своїй роботі [3] С. О. Климчуком пояснюються важливі, з точки зору дослідника, принципи організації прецедентної системи (Case-Based Reasoning System). Зокрема, проаналізовано переваги методики прецедентів у рамках створення інтелектуальних засобів підтримки прийняття рішень. Публікація заслуговує уваги, з урахуванням можливості застосування відповідної методики для побудови системи інформаційного пошуку. Незважаючи на те що роботу К. Д. Маннінга [4] задумано як вступний курс з інформаційного пошуку та написано з точки зору інформатики; в ньому поряд з класичним пошуком розглядаються веб-пошук, принципи роботи пошукових механізмів а також класифікація та кластеризація текстів. Книга містить сучасний виклад всіх аспектів проектування та реалізації систем збору, індексування та пошуку документів, методів оцінки таких систем, а також введення в методи машинного навчання.

Очевидно, що проблема широко обговорюється науковим співтовариством. Однак, попри значну кількість публікацій дослідників, проблема вдосконалення методів інформаційного пошуку не розв'язана повністю та залишається актуальною.

Проблема пошуку, збору та оптимізації інформації з'явилася ще в період розвитку пошукових систем [1]. У той час пошукові системи надавали велике значення аспектам, якими власники сайтів могли легко маніпулювати. Це призвело до того, що у видачі багатьох пошукових систем перші кілька сторінок займали сайти, наповнення яких було нерелевантним.

У загальному випадку під терміном «релевантність» розуміють міру відповідності отримуваного результату бажаному. В термінах пошуку – це міра відповідності результатів пошуку завданню, поставленому в пошуковому запиті [5]. Відповідно, нерелевантним називається такий, що був відібраний у результаті інформаційного пошуку, але зміст якого не відповідає запиту користувача.

Пошуковими системами, при вирішенні завдань збору, зберігання, обробки і видачі інформації, виконуються такі операції [2]:

1) пошук документів;

2) аналіз вмісту документів;

3) побудова пошукових образів документів (отримання з документів інформації, яка використовується системою як відомості про документ);

4) зберігання пошукових образів документів (відомостей про документи);

5) аналіз запитів користувачів (споживачів інформації);

6) пошук релевантних (відповідних) запиту документів;

7) організація пошукової видачі користувачам.

Ефективність інформаційного пошуку характеризується двома відносними показниками: коефіцієнтом точності (відношенням числа документів, що відповідають критеріям інформаційного запиту до загальної кількості документів, отриманих в пошуковій видачі) і коефіцієнтом повноти (відношенням числа документів, що відповідають критеріям інформаційного запиту до загальної кількості таких документів, що містяться в просторі, оброблюваному інформаційно-пошуковою системою) [1]. Необхідні значення цих показників залежать від специфіки інформаційних потреб. Наприклад, якщо відбувається пошук патентних описів з метою проведення експертизи патентної заявки на новизну необхідна 100% повнота результату пошукової видачі, а при пошуку, орієнтованому на звичайного дослідника, або випадкового користувача, то прийнятною вважається точність (релевантність) результату близько 80%, а повнота - близько 50% [4].

Однак, при визначенні релевантності пошукові системи в першу чергу звертають увагу на те, скільки разів на сторінці зустрічається фраза, тотожна запиту користувача. Цей параметр називається частотою ключового слова. Чим він вищий, тим релевантнішим вважається сайт.

Більшість пошукових машин знаходять величезну кількість «релевантних» сторінок за запитом користувача. Кожен знайдений документ ранжується за ступенем його відповідності до запиту. Релевантність кожного документа оцінюється за допомогою різних технологій: обліку частоти появи на сторінці шуканих слів, «відстані» між шуканими словами,

вмісту META-тегів, просторово-часового контексту документа, популярності ресурсу в рейтингах, використання індексу цитування.

Типову організацію машин пошуку можна розглянути на прикладі машини WebCrawler, розробленої в університеті Вашингтон (Сіетл, США). WebCrawler починає процес пошуку нових сайтів з відомих йому документів і переходить за посиланнями на інші сторінки. Він розглядає мережевий простір як орієнтований граф і використовує алгоритм обходу графа, працюючи в такому циклі [1]:

1) знайти новий документ;

2) зазначити документ як вилучений;

3) розшифрувати посилання з цього документу;

4) проіндексувати зміст документу.

Пошуковий механізм працює в двох режимах: пошук документів в реальному часі та режимі індексування документів. У режимі індексування система будує індекс інформації зі знайдених документів, в режимі пошуку знаходить документи, які максимально відповідають запиту користувача. Агенти в системі WebCrawler відповідають за вилучення документів з мережі.

Для виконання цієї роботи пошуковий механізм знаходить вільного агента і передає йому завдання на пошук. Агент приступає до роботи і повертає або зміст документа, або пояснення, чому документ не може бути доставлений. Агенти запускаються як окремі процеси, що дозволяє ізолювати основний процес роботи системи від помилок і проблем з пам'яттю.

Одночасно можуть бути використані до 15 агентів [1]. У базі даних зберігаються метадані документів, зв'язки між документами і повнотекстовий індекс. База оновлюється кожного разу, коли надходить новий документ. Для відсікання семантично незначущих слів існує стоп-словник. Словам з документу приписується вага, рівна частоті їх появи в даному тексті, поділеній на частоту появи слова в посиланнях на інші документи. Такий індекс дозволяє швидко знаходити по заданому слову посилання на документи, що містять його.

Аналогічним чином влаштовані і інші машини пошуку. Вони не можуть налаштовуватися на переваги користувача і не мають достатніх ресурсів для аналізу інформації, а мережевим роботам стає все важче справлятися з постійним зростанням кількості Інтернет-ресурсів [5]. Головним завданням машин пошуку є індексація ресурсів глобальної мережі. Фактично в базах даних машин пошуку зберігається інформація про те, де і що лежить в мережі. Тому можна вважати, що існуючі машини пошуку забезпечують низькорівневий сервіс для клієнтських пошукових програм більш високого рівня.

У загальному випадку вважається, що будь-яка пошукова машина повинна відповідати наступним вимогам [2]:

1) простота у використанні;

2) чітко організований і оновлюваний індекс;

3) швидкий пошук за індексом і швидке реагування;

4) надійність і точність результатів пошуку.

В результаті проведеного дослідження було зазначено, що всі сучасні пошукові системи мають деякі серйозні недоліки:

1) стандартний механізм пошуку за ключовими словами в сучасних інформаційно-пошукових системах видає результати з великим показником інформаційного шуму;

2) велика кількість пошукових машин з різними призначеними для користувача інтерфейсами породжує проблему когнітивного перевантаження;

3) методи індексування баз даних, як правило, не пов'язані з інформаційним змістом;

4) часто видаються посилання на інформацію, якої в Інтернеті вже давно немає, а також немає можливості в реальному часі враховувати динаміку зміни змісту Інтернет-ресурсів;

5) в пошукових машинах немає розвинених засобів розуміння природних мовних конструкцій.

Також серйозним недоліком сучасних пошукових систем є їх централізація, що вимагає колосальних ресурсів (величезні обсяги бази даних, безліч серверів, апаратури і т.д.) [5]. Ще однією з основних проблем при створенні сучасних інформаційно-пошукових систем є недостатнє врахування думок і бажань користувачів. Існуючі механізми персоналізації володіють певними обмеженнями і не можуть враховувати в повному обсязі потреби користувачів. Явні методи персоналізації вимагають активного залучення користувача в процес накопичення персональних даних (потрібна наявність зворотного зв'язку), що не зовсім зручно для користувача пошукової системи, а неявні методи персоналізації не завжди адекватно можуть реагувати на зміни в перевагах користувачів.

Нині досить важливою проблемою в галузі інформаційного пошуку є проблема конструювання інтелектуальних систем, орієнтованих на відкриті і динамічні бази даних [4]. В основі таких систем лежить інтеграція здатних до адаптації, модифікації і навчання моделей пошуку, виявлення та оперування знаннями, орієнтованих на специфіку шуканої (предметної) області та відповідний тип невизначеності, що відображає їх здатність до розвитку і зміни свого стану. Організація пошуку на основі прецедентів дозволяє об'єднати в собі різні підходи до вирішення завдання інтелектуалізації та персоналізації пошуку.

У більшості енциклопедичних джерел термін «прецедент» визначається як випадок, що стався раніше і слугував прикладом або виправданням для наступних випадків подібного роду [3]. Міркування на основі прецедентів (Case-Based Reasoning) являється методикою здатною вирішити нове або невідоме завдання, використовуючи або адаптуючи рішення вже відомої задачі, тобто використовуючи вже накопичений досвід вирішення подібних завдань. Підхід на основі прецедентів виник в процесі розвитку досліджень в області створення експертних систем (систем, заснованих на знаннях).

Як правило, процес виведення на основі прецедентів підлягає декомпозиції на чотири основні етапи, що утворюють так званий цикл міркування на основі прецедентів [3]. Основними етапами такого циклу є:

1) вилучення найбільш відповідного (подібного) прецеденту (або прецедентів) до ситуації, що склалася з бібліотеки прецедентів;

2) повторне використання вилученого прецеденту для спроби вирішення поточної проблеми;

3) перегляд і адаптація в разі необхідності отриманого рішення відповідно до проблеми;

4) збереження нового прийняття рішення як частини нового прецеденту.

До переваг міркувань на основі прецедентів можна віднести наступні аспекти [3]:

1) можливість безпосередньо використовувати досвід, накопичений системою без інтенсивного залучення експерта в тій чи іншій предметної області;

2) можливість скорочення часу пошуку рішення поставленої задачі за рахунок використання вже наявного рішення для такого завдання;

3) можливість виключити повторне отримання помилкового рішення;

4) відсутня необхідність повного та поглибленого розгляду знань відносно конкретної предметної області.

До недоліків міркувань на основі прецедентів можна віднести наступне [3]:

1) при описі прецедентів зазвичай обмежуються поверхневими знаннями про предметну область;

2) велика кількість прецедентів може привести до зниження продуктивності системи;

3) проблематичним є визначення критеріїв для індексації і порівняння прецедентів;

4) проблеми з налагодженням алгоритмів визначення подібних (аналогічних) прецедентів;

5) неможливість отримання рішення задач, для яких немає прецедентів або ступінь їх схожості (подібності) менше заданого порогового значення.

Основна мета використання апарату прецедентів в інформаційно-пошуковій системі полягатиме в видачі відповіді на запит користувача на основі прецедентів, які вже мали місце в минулому при виконанні подібних запитів. Інформація про новий запит використовуватиметься для вилучення з бібліотеки прецедентів найбільш підходящого прецеденту (прецедентів). Витягнутий прецедент використовується повторно для отримання рішення нової проблеми (завдання) [3]. Потім запропоноване рішення в разі необхідності може бути адаптоване до особливостей нової ситуації і застосовано на практиці. У разі успішного застосування, перевірене рішення спільно з описом запиту утворює новий прецедент, який зберігається в базі прецедентів.

Вибір методу отримання прецедентів безпосередньо пов'язаний зі способом уявлення прецедентів і відповідно зі способом організації бібліотек прецендентів [3]. Останні є важливою складовою бази знань інтелектуальної системи, але можуть виступати як і окремий компонент пошукової системи. Таким чином, їхня структура істотно впливає на різні показники роботи системи і, зокрема, на час пошуку та вилучення прецедентів.

Існують різні способи подання та зберігання прецедентів: від простих (лінійних) до складних ієрархічних. Варто зазначити, що прості способи зберігання та подання прецедентів, що базуються на технології реляційних баз даних, вимагають значно менше витрат на реалізацію, а також підтримку і супровід бібліотек прецедентів системи на відміну від більш складних, але може знадобитися значно більше часу для здійснення пошуку рішення при простому поданні прецедентів порівняно з іншими способами представлення і збереження прецедентів.

Прецеденти можуть бути представлені у вигляді списку параметрів, концептуальних графів, семантичної мережі, деревовидних структур, предикатів, фреймів, малюнків і мультимедійної інформації [3]. Прецедент може включати наступні компоненти:

1) опис завдання (проблеми або проблемної ситуації);

2) рішення задачі;

3) результат застосування рішення.

Опис результату може включати список виконаних дій, додаткові коментарі та посилання на інші прецеденти. прецедент може мати як позитивний, так і негативний результат застосування рішення, а також в деяких випадках може приводитися обґрунтування вибору даного рішення і можливі альтернативи.

Слід зазначити, що у простих задачах класифікації деякі відмінності просто ігноруються і клас рішення витягнутого прецеденту переноситися на клас рішення нового прецеденту [3]. Однак, багато систем, враховують відмінності між знайденим і наявним прецедентом, і тому рішення витягнутого прецеденту не може бути безпосередньо перенесено на нову ситуацію.

Для успішної реалізації пошуку за допомогою міркувань на основі прецедентів, відповідно, необхідно забезпечити коректне вилучення прецедентів з бібліотеки прецедентів системи.

Існують різні способи такого отримання прецедентів, наприклад [3]:

1) метод найближчого сусіда і його модифікації;

2) метод пошуку на основі дерев рішень;

3) метод вилучення на основі знань;

4) метод вилучення з урахуванням застосування прецедентів.

З точки зору інформаційного пошуку було прийнято рішення скористатися методом найближчого сусіда. Одним з ключових моментів даного вибору стало те, що даний метод є найпоширенішим з методів порівняння і вилучення прецедентів [3]. Він дозволить досить легко обчислити ступінь подібності поточного запиту і прецедентів з бібліотеки системи. Принцип роботи полягає у тому, що з метою визначення ступеня подібності на множині параметрів, використовуваних для опису прецедентів і поточного запиту, вводиться певна метрика. Далі відповідно до обраної метрики визначається відстань від цільової точки, відповідної поточному запиту ї, до точок, що представляють прецеденти з бібліотеки прецедентів і вибирається найближча до цільової точка.

Безумовно, ефективність методу найближчого сусіда багато в чому залежить від вибору метрики (міри схожості).

Основні метрики, які можуть бути використані в методі найближчого сусіда [3]:

1) Евклідова відстань. Евклідова відстань є геометричною відстанню в багатовимірному просторі; відстань між точками С і Т в n-вимірному просторі визначається за наступною формулою (1):

$$d(C,T)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}, \quad (1)$$

де d (С, Т) – відстань між С і Т, х(1,2,3 ... n) та у(1,2,3 ... n) – значення ознак для прецедента та поточної ситуації, n – кількість змінних, якими описуються прецеденти та поточна ситуація.

2) Відстань Хемінга. Найпростішою мірою схожості, а точніше кажучи, відмінності між закодованими представленнями є відстань Хемінга. Хоча першочергово вона була введена для двійкового коду, вона цілком може примінятися для порівняння будь-яких упорядкованих наборів, які складаються з елементів, здатних набувати дискретних значень.

Розглянемо для прикладу два упорядковані набори х та у які складаються з дискретних, нечислових символів (наприклад логічних 0 та 1). Порівняння на несхожість полягатиме у кількості неспівпадаючих символів у цих наборах. Таку величину d(С, Т) називають відстанню Хемінга; визначається вона лише для послідовностей однакової довжини. Тобто якщо х = (1,0,1,1,1,0), у = (1,1,0,1,0,1), то d(С, Т) = 4.

3) Міра схожості Танімото. В загальному випадку може примінятися для оцінки релевантності (ступеня відповідності) документів при інформаційному пошуку. Дескрипторам в цих документах можуть бути присвоєні індивідуальні ваги.

Якщо a_{ik} – вага, відносна до k-го дескриптору і-го документу, то ступінь схожості двох документів визначених через С і Т можна визначити за формулою (2):

$$(C,T)=\sum_k a_{Ck}a_{Tk}=a_{CT} \quad (2)$$

при цьому вираз набуде наступного вигляду (3):

$$d(C,T)=\frac{a_{CT}}{a_{CT}+a_{TT}-a_{CT}}. \quad (3)$$

Як вже наголошувалося, вибір відповідної метрики творче і досить трудомістке завдання, від успішного вирішення якої безпосередньо залежить результативність пошуку та вилучення прецедентів.

Для вирішення поставленої задачі інформаційного пошуку [5] було обрано у якості метрики використовувати евклідову відстань. Відповідно, спираючись на дану метрику було створено та запропоновано відповідний алгоритм вилучення прецендентів для організації інформаційного пошуку.

Нехай заданий прецедент (С) і поточна ситуація (пошуковий запит) (Т) в n-вимірному просторі ознак (властивостей), тоді ступінь подібності або близькості можна визначити, використовуючи евклідову метрику для визначення відстані між С і Т (4).

$$d(C,T)=\sqrt{\sum_{i=1}^{n}(x-y)^2}. \quad (4)$$

Вхідні дані: поточна ситуація (Т), бібліотека прецедентів – непорожня множина прецедентів CL, m – кількість прецедентів в бібліотеці, порогове значення ступеня схожості К.

Вихідні дані: тимчасовий контейнер зберігання прецедентів SC порогового значення К.

Алгоритм:

Крок 1. Для визначення значення ступеня подібності S(C,T) необхідно знайти максимальне значення відстані d_{max} в евклідовій метриці, використовуючи границі діапазонів параметрів для описання прецедентів ($x_{поч}$ та $x_{кін}$, i = 1, ..., n).

Крок 2. Поки j ≤ m обираємо прецедент C_j з множини CL та переходимо до наступного кроку, інакше вважаємо що всі прецеденти бібліотеки розглянуті й переходимо до кроку 6.

Крок 3. Розраховуємо відстань в евклідовій метриці між обраним прецедентом C_j та поточною ситуацією Т (5) та переходимо до наступного кроку.

$$d(C_j,T)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}. \quad (5)$$

Крок 4. Обчислюємо ступінь подібності (6) S(Cj,T); переходимо до наступного кроку.

$$S(C_j,T)=1-\frac{d(C_j,T)}{d_{\max}}. \quad (6)$$

Крок 5. Якщо S(Cj,T) ≥ К , то id-номер даного прецеденту розміщуємо в SC (тимчасовому контейнері для зберігання прецедентів, які задовольняють умову) та переходимо до кроку 2.

Крок 6. Якщо після всіх ітерацій контейнер SC залишається пустим, значить прецеденти для поточної ситуації не знайдені і потрібно перейти до кроку 7 та передати користувачеві повідомлення про необхідність зниження порогового значення К; в іншому випадку вважається що прецеденти для поточної ситуації успішно вилучені із бібліотеки за id-номерами, що зберігаються в тимчасовому контейнері SC та переходимо до наступного кроку.

Крок 7. Кінець (завершення алгоритму).

Слід також зазначити, що при повторному використанні знайденого прецедента в контексті нової проблемної ситуації важливо звернути увагу на наступні особливості: різниця між вилученим та новим прецедентом а також те яку частину вилученого прецеденту можна застосувати до поточної ситуації (пошукового запиту).

Таким чином, можна зробити допущення щодо перспективності використання даної методики у рамках вдосконалення системи інформаційного пошуку. Відповідно, між побудованими таким чином інформаційно-пошуковими системами (з метою їх навчання) та користувачами повинен встановлюватися ефективно працюючий зворотний зв'язок (абонент повідомляє, якою мірою цей документ відповідає запиту і чи потрібно продовжувати пошук, вказує на ступінь відповідності цього документа його інформаційним потребам), який дозволяє уточнювати потреби абоне-

нтів, своєчасно реагувати на зміни цих потреб і оптимізувати роботу системи. Таким чином, можна підсумувати, що використовуючи методику прецедентів [3] при ретроспективному пошуку, пошуковою системою знаходитимуться в першу чергу документи, які містять необхідну інформацію.

ВИСНОВКИ. В результаті проведеного дослідження було зазначено, що всі сучасні пошукові системи мають деякі серйозні недоліки:

1) стандартний механізм пошуку за ключовими словами в сучасних інформаційно-пошукових системах видає результати з великим показником інформаційного шуму;

2) велика кількість пошукових машин з різними призначеними для користувача інтерфейсами породжує проблему когнітивного перевантаження;

3) методи індексування баз даних, як правило, не пов'язані з інформаційним змістом;

4) часто видаються посилання на інформацію, якої в Інтернеті вже давно немає, а також немає можливості в реальному часі враховувати динаміку зміни змісту Інтернет-ресурсів;

5) в пошукових машинах немає розвинених засобів розуміння природних мовних конструкцій.

Виходячи з проведеного огляду сучасного стану досліджень встановлено що все очевиднішими стають потреби в розробці розвинених засобів інтелектуалізації та персоналізації пошуку. Окрім того, встановлено, що хоча проблема широко обговорюється науковим співтовариством; попри значну кількість публікацій дослідників, проблема вдосконалення методів інформаційного пошуку не розв'язана повністю та залишається актуальною

На сьогоднішній день практично всі сучасні інформаційно-пошукові системи Інтернету активно працюють у сфері розробки інструментів інтелектуалізації та персоналізації пошуку [6–13], але більшість серйозних проблем в цій галузі поки не вирішені. Відповідно, перспективним напрямом залишається дослідження і розробка методів та програмних засобів інтелектуалізації та персоналізації в інформаційно-пошукових системах Інтернету.

В ході дослідження встановлено перспективність використання методики прецедентів у рамках вдосконалення пошукових методів та, зокрема, при побудові орієнтованих на розподілену структуру інформаційно-пошукових систем. Окрім того, автором виділено основні напрями для розробки алгоритму вилучення прецедентів в рамках організації інформаційного пошуку. Наголошено, що організація пошуку на основі прецедентів дозволяє об'єднати в собі різні підходи до вирішення завдання інтелектуалізації та персоналізації пошуку і знизити навантаження на індекс пошукового інструменту, а також спростити вирішення проблеми, пов'язаної із забезпеченням конфіденційності даних.

Висновки та пропозиції в рамках даного дослідження можуть бути використані в науково-дослідній та викладацькій діяльності. Зокрема, результати, отримані при проведенні даного дослідження можуть бути використанні при подальшому аналізуванні та вдосконаленні методів інформаційного пошуку.

ЛІТЕРАТУРА

1. Урвачева В. А. Обзор методов информационного поиска. *Вестник Таганрогского института имени А.П. Чехова.* 2016. № 1. С. 457–463

2. Шокин Ю. И. Проблемы поиска информации. Новосибирск: Наука, 2010. 220 с.

3. Климчук С. О. Розроблення прецедентної системи підтримки прийняття рішень. *Вісник Національного університету «Львівська Політехніка»*. 2010. № 689. С. 169–176.

4. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2017. 640 с.

5. Терещенко В. В., Терещенко В. Л. Перспективність вдосконалення систем інформаційного пошуку. *Четверта Всеукраїнська науково-практична конференція «ІТ-Перспектива»*. Кременчук: КрНУ, 2017. С. 26–28.

6. Цивільський Ф. М., Дроздова Є. А., Григорова А. А. Використання сучасних internet-технологій для історичних досліджень. *Вісник Кременчуцького національного університету імені Михайла Остроградського.* Кременчук: КрНУ, 2018. Вип. 5 (112). С. 59–64.

7. Слабченко О. О., Сидоренко В. Н. Покращення якості первинних даних в задачах моделювання інтернет-співтовариств на основі комплексного застосування моделей сегментації, імпутації і збагачення даних. *Вісник Кременчуцького національного університету імені Михайла Остроградського.* Кременчук: КрНУ, 2013. Вип. 6 (83). С. 50–58.

8. Заїка А. В., Філенко М. І. , Остапченко А. С., Григорова Т.А. Моделювання архітектурних рішень підтримки мультисайтовості для організації інформаційних систем. *Вісник Кременчуцького національного університету імені Михайла Остроградського.* Кременчук: КрНУ, 2015. Вип. 3 (92) ч.1. С. 54–59.

9. Костенко П. П., Левченко І. В. Веб-сервіс уточнення релевантності веб-документів пошукової видачі Google на основі поведінки користувача. *Інженерні та освітні технології. Щоквартальний науково-практичний журнал*. Кременчук: КрНУ, 2014. Вип. 4 (8). С. 49–62.

10. Терещенко В. В. Аналіз сучасних методів інформаційного пошуку. *Вісник Кременчуцького національного університету імені Михайла Остроградського.* Кременчук: КрНУ, 2018. Вип. 3 (110). С. 26–32.

11. Alexandros N., Mark M. Detecting Spam Web Pages through Content Analysis. *Microsoft Research*, 2012. PP. 1–6.

12. Brin S., Page L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 2004. PP. 107–117.

13. Ganz A., Sieh L., Behavioral factors and SEO. *Proceedings of 24th International Conference on Computer Communications and Networks (ICCCN 2015)*, Las Vegas, Nevada, USA August 3 - August 6, 2015, Scottsdale, Arizona, USA. PP. 218–223.

PERSPECTIVES IN IMPROVING METHODS OF INFORMATION SEARCH

V. Tereschenko
Kremenchuk Mykhaylo Ostrohradskiy National university
vul. Pershotravneva, 20, 39600, Kremenchuk, Ukraine. E-mail: darkwolfthehunter@gmail.com

Purpose. To analyze current methods of information search. In the conditions of the development of the information society, one of the most important tasks remains the solution of the problem of effective search and collection of information. This is primarily due to the growing diversity of information sources aimed at developing different areas of human activity. Accordingly, there is a need for new methods to ensure effective information search. **Methodology.** In this paper the principles of information search systems functioning were investigated, a number of scientific works in the field of information search was analyzed, the prospect of using the precedent methodology in the improvement of search methods was established, the feasibility of using that method to improve the accuracy of document evaluation was emphasized. **Results.** Based on the conducted review of the current state of research in the field of optimization of methods and algorithms for information retrieval, the following problems have been identified: the accumulation of duplicate content; lack of thematic breakdown of web search results; the prevalence of information spam when viewing documents, which significantly affects the time of searching and viewing documents. Was been found that Case-Based Reasoning method able to solve a new or unknown problem by using or adapting a solution to a known problem, that is, by using experience gained in solving such problems. Therefore, it is possible to make assumptions about the prospect of using this technique as part of improving the information retrieval system. From the point of view of increasing the reliability of the assessment of relevance of the document to the request, it will be expedient to use that method. **Originality.** For the first time, considered the prospect of using the precedent method in terms of improving the information retrieval system rather than organizing decision support systems; based on the requirements of the search organization, the most important aspects are outlined. Within the framework of the analysis of publications was identified the most adapted methods that are acceptable for refining the search engine. **Practical value.** The results obtained during this study can be used for further analysis and improvement of methods and algorithms of information search. References 13.

Key words: search engine optimization, search engine, search engine results, information search.

REFERENCES

1. Urvacheva, V. A. (2016), *Obzor metodov informacionnogo poiska* [Overview of information retrieval methods], *Vestnik Taganrogskogo instituta imeni A.P. Chekhova*, №1, pp. 457-463, Taganrog, Russia

2. Shokin, Y. I. (2010), *Problemy poiska informacii*. [Problems of information search], Nauka, Novosibirsk, Russia

3. Klymchuk, S. O. (2010), *Rozroblennya pretsedentnoyi systemy pidtrymky pryynyattya rishen*. [Develop a precedent-based decision support system], *Visnik Natsionalnoho universitetu «Lvivska Politekhnika»*, № 689, pp. 169-176.

4. Manning, K., Raghavan, P., Shutce, H. (2017), *Vvedenie v infirmacionnyi poisk* [Introduction into information search], Willyams, Moscow, Russia

5. Tereschenko, V. V., Tereshchenko, V. L. (2017), *Perspektivnist vdoskonalennya sistem informatsiynogo poshuku* [The prospectivity of improving information search systems], *IV Vseukrainska naukovo-praktychna konferentsiya «IT-Perspektiva»*, pp. 26-28, Ukraine.

6. Tsivilsky, F. M., Drozdova, E. A., Hryhorova, A. A. (2018), *Vykorystannya suchasnykh internet-tekhnolohiy dlya istorychnykh doslidzhen* [Use of modern internet technologies for historical research], *Transactions of Kremenchuk Mykhailo Ostrohradskyi National University*, Vip. 5 (112), pp. 59-64, Ukraine.

7. Slabchenko, O. O., Sidorenko, V. N. (2013), *Pokrashhennya yakosti pervinnih danih v zadachah modelyuvannya internet-spivtovaristv na osnovi kompleksnogo zastosuvannya modelej segmentacii, imputacii i zbagachennya danih* [The improvement of initial data quality in modeling problems of online communities on the base of combined implementation of segmentation, imputation and data enrichment models], *Transactions of Kremenchuk Mykhailo Ostrohradskyi National University*, Vip. 6 (83), pp. 50-58, Ukraine.

8. Zaika, A. V., Filenko, M. I., Ostapchenko, A. S., Hryhorova, T. A. (2015), *Modelyuvannya arhitekturnih rishen pidtrimki multisajtovosti dlya organizacii informaciynih sistem* [Design of architecture for support multi-site in information systems], *Transactions of Kremenchuk Mykhailo Ostrohradskyi National University*, Vip. 3 (92) part 1, pp. 54-59, Ukraine

9. Kostenko, P. P., Levchenko, I. V. (2014), *Web-servis utochnennya relevantnosti web-dokumentiv poshukovoi vidachi Google na osnovi povedinki koristuvacha* [Web-service for clarification relevant web-documents of search results of Google based on user behavior], *Inzhenerni ta osvitni tehnologii*, Vip. 4 (8), pp. 49-62, Ukraine

10. Tereschenko, V. V. (2018), *Analiz suchasnyh metodiv informatsiynogo poshuku* [Analysis of current methods of information search], *Transactions of Kremenchuk Mykhailo Ostrohradskyi National University*, Vip. 3 (110), PP. 26-32, Ukraine

11. Alexandros, N., Mark, M.(2012), Detecting Spam Web Pages through Content Analysis, *Microsoft Research*, pp. 1-6.

12. Brin, S., Page, L. (2004), The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, pp. 107-117.

13. Ganz, A., Sieh, L., (2015), Behavioral factors and SEO, *Proceedings of 24th International Conference on Computer Communications and Networks (ICCCN 2015)*, Las Vegas, Nevada, USA August 3 - August 6, 2015, Scottsdale, Arizona, USA. pp. 218-223.

Стаття надійшла 03.10.2019.