УДК 681.3.06.003.1:004.369.6

Р. М. Камінський, Л. Я. Нич, Н. Б. Шаховська
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

ФОРМАЛЬНЕ ПОДАННЯ ДІЯЛЬНОСТІ КОРИСТУВАЧА З ВИЯВЛЕННЯ ІНФОРМАЦІЙНО-ЗНАЧУЩИХ ОБ'ЄКТІВ

Розглянуто діяльність користувача щодо пошуку інфораційно-значущих об'єктів розв'язання поставленої перед ним задачі. Для формалізації цієї діяльності використано апарат теорії множин. Наведено математичні моделі формулювання запиту, аналізу документів у видачі та відбір інформаційних об'єктів для розв'язання поставленої перед користувачем задачі. Крім того, розроблено моделі в аспекті інтелектуальної діяльності користувача та когнітивних процесів. Наведено результати експериментальних досліджень з такими пошуковими системами, як Google, Yandex, META, Rambler, Yahoo. Для оцінювання ефективності інформаційного пошуку знайдені документи поділені на пертинентні, релевантні та нерелевантні. Ефективність пошуку визначено відношенням кількості пертинентних та релевантних документів до кількості всіх документів у видачі. Інформаційні характеристики пошукових систем запропоновано враховувати відповідним коефіцієнтом.

Ключові слова: інформаційна система, інформаційний пошук, пошукова діяльність, пертинентність, релевантність, модель користувача, когнітивний процес, інтелектуальна діяльність.

The user activity for finding informative meaningful objects needed to solve the problem is considered. To formalize this activity, set theory is used. The mathematical models of formulation of request, analysis of the issued documents and selection of information objects for the solution to the user problem are presented. In addition, mathematical models of intellectual activitiesof user and cognitive processes are developed. The experimental results of the search engines such as Google, Yandex, META, Rambler, Yahoo are given. To evaluate the effectiveness of information retrieval the found documents were divided by pertinent, relevant and irrelevant. Search Performance Ratio is determined with the ratio of pertinent and relevant documents to the number of all the documents in issue. It is advised to take into account information characteristicsof the offered search engines with appropriate coefficient.

Key words: information system, information retrieval, search activity, pertynent, relevance, user model, cognitive process, intellectual activity.

Вступ. Загальна постановка проблеми

Розвиток комп'ютерної техніки та інформаційних технологій значною мірою стимулював створення і наповнення різноманітною інформацією як загальні, так і спеціалізовані бази даних, забезпечуючи управління ними. Крім того, з'явилась можливість опрацювання не лише текстових матеріалів, але і різноманітних зображень практично у всіх областях діяльності людини. Проте, з іншого боку, величезні обсяги даних практично унеможливлюють безпосередню роботу користувача з ними, що, своєю чергою, стимулювало розвиток відповідних пошукових систем, основною метою яких є своєчасне і повне забезпечення користувача необхідними йому даними.

Особливість пошуку інформаційно-значущих об'єктів полягає в тому, що якщо пошук текстових документів здійснюється за кількома десятками символів об'єднаних в групи – слова, для яких існують спеціалізовані словники, то пошук інформаційно-значущих об'єктів для розв'язання конкретних задач є значно складнішим, оскільки створення запиту на їх пошук потребує чималих

зусиль на його формулювання. Тому найактуальнішою проблемою, з якою стикаються користувачі, є забезпечення надійного, постійного та повнофункціонального доступу до актуальних даних в сенсі пошуку інформаційно-значущих об'єктів. Вирішення цієї проблеми, на нашу думку, варто розпочати з побудови моделі користувача в системі людина-комп'ютер.

Загальна постановка проблеми

У системному аспекті пара людина – комп'ютер є поєднанням двох підсистем: специфічних психофізичних та функціональних особливостей людини та можливостей сучасної обчислювальної техніки в сенсі пошуку потрібної інформації в розподілених базах даних та в мережі Internet. Перша з них, тобто людина, є переважно не програмуючим користувачем, але може бути як висококваліфікованим фахівцем, добре обізнаним з класом розв'язуваних задач, методами їх розв'язання та підходами і принципами інтерпретації отриманих результатів, так і звичайним користувачем, який принаймні вміє включити комп'ютер і вийти в мережу. Друга – це сучасні високопродуктивні комп'ютери з високою швидкістю обробки інформації, величезними обсягами пам'яті, об'єднані у мережі та побудовані на їх основі інформаційно-пошукові системи.

Проблема організації та забезпечення високої функціональної ефективності інформаційного пошуку в базах, сховищах та просторах даних полягає в тому, що шукана інформація зберігається в різних формах її кодування, створених в різний час і з різною метою; вона є складно структурованою, для різних задач має різну інформаційну цінність, і різними користувачами сприймається по-різному. Натомість за високої надійності і стабільності апаратного та програмного забезпечення вся відповідальність за результати пошуку покладена на людський фактор в сенсі укладання пошукового запиту та відбору знайденого матеріалу. В цьому плані об'єктивно оцінити ефективність пошуку можна лише на підставі виданих документів.

Історично, а в певному сенсі і політично (з метою захисту інформації) різні джерела інформації (електронні бібліотеки, загальні та локальні бази, сховища, простори даних) мають свої особливості стосовно організації форм збереження, пошуку, виявлення, видачі потрібної інформації, які в основному полягають у видах і тонкощах мов запитів та способів кодування збережуваної інформації. Сьогодні такий пошук здійснюють спеціальні пошукові системи, наприклад, Google, Yandex, META, Rambler, Yahoo. Робота з однією чи навіть декількома базами даних практично полягає у правильному формулюванні запиту і власне тут існуюча пошукова система допомагає знайти необхідну інформацію. Наприклад, локальні бази даних навіть великих підприємств доволі швидко дають інформацію про виготовлені вироби, товари, зарплату працівників тощо. Проте, пошук даних в "чужих" базах даних може стати складною проблемою. Тут найкращим прикладом є пошукова система Google та аналогічні, які видають десятки тисяч документів, з яких користувач вибирає лише декілька, витрачаючи величезну кількість часу на пошук потрібних серед наданих пошуковою системою.

Задача пошуку полягає в тому, щоб отримати з одного або декількох інформативних джерел системно інтегровані набори з максимальною кількістю релевантних і пертитентних інформаційно-значущих об'єктів, які в сукупності наділені ознаками повноти, цілісності та несуперечності. Вони фактично подаються у формі адекватної інформаційної моделі проблемної області для її аналізу, опрацювання та ефективного використання в процесах підтримки прийняття рішень. Як правило, різноманітні джерела інформації було створено в різний час і за різними принципами та мовами запитів, а головне, за різними фаховими ознаками та онтологіями.

Фактично в задачах пошуку документів з інформаційно-значущими об'єктами, що використовують пошукові системи основна роль належить саме користувачу – укладання правильного запиту та відбору релевантних і аналізу пертинентних документів, отриманих у видачі.

У загальному вигляді процедура пошуку є ітеративною процедурою, тобто за етапом видачі результатів пошуку переважно здійснюється корекція запиту та проводиться новий пошук вже за виправленим запитом і т.д. Схематично таку процедуру показано на рис. 1. Корегує запит користувач за результатами аналізу лише виданих документів та їх релевантності.

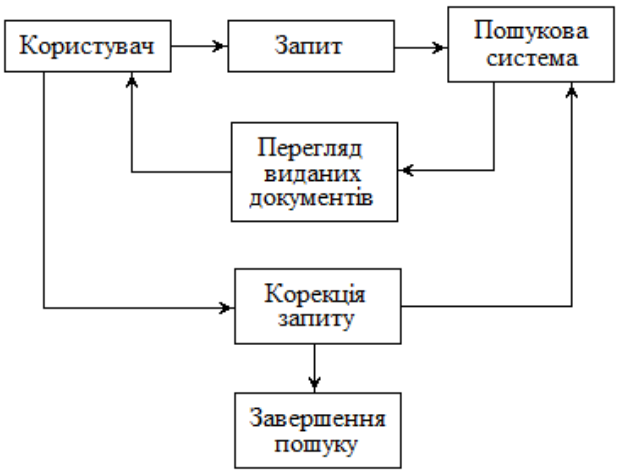

Рис. 1. Загальна схема процедури пошуку

Залежно від співвідношення повноти і точності знайдених документів користувач може звузити або розширити область пошуку, перейшовши до загальніших або, навпаки, більш специфічних термінів, а також використовуючи споріднені поняття. У разі пошуку за декількома термінами таке корегування області пошуку може відбуватися за одним або за декількома термінами, що дозволяє змінювати цю область доволі плавно. Якщо їх немає у списку виданих документів, область пошуку має бути розширена. Крім того, виявляється надзвичайно корисною апріорна інформація, збережена в пам'яті користувача як про завідомо релевантні документи з потрібними інформаційно-значущими об'єктами, так і про методи розв'язання аналогічних задач.

Аналіз останніх досліджень та публікацій

Поняття ефективності пошуку має доволі широке тлумачення і переважно має запозичений економічний аспект. У роботі [1] для оцінювання якості роботи пошукової системи використовуються такі оцінки: точність, повнота, акуратність, помилка, F-міра, які визначаються як метрики на множинах документів і фактично дають кількісну характеристику самого пошуку. За результатами аналізу існуючих пошукових систем у [2] роблять висновок, що для пошуку документів гіпертекстових базах даних існуючі загальновизнані оцінки мають певні обмеження. Запропоновано використовувати додаткові характеристики, зокрема M-різновид вибірки та U-впорядкованість вибірки. На цій підставі наводять коефіцієнт впорядкованості та коефіцієнт пошукового шуму. Виділено низку факторів, що впливають на успішність пошуку. Оцінці ефективності інформаційних систем як одній з проблем інформаційного суспільства присвячена стаття [3], в якій на основі аналізу практичного застосування інформаційних систем показано, що в оцінці ефективності інформаційних систем можна виділити три типи ефектів: врахування додаткової інформації, нормування та врахування організаційних процесів та планування, оптимізації, управління процесами та ресурсами. Підкреслено роль врахування витрат, які ділять на дві складові: капітальні (бюджетні) або прямі витрати і позабюджетні, пов'язані з користувачами. Для оцінювання трудовитрат наведено модифіковану формулу, яка враховує модель оцінювання вартості розроблення програмного забезпечення. Кількісні показники оцінювання функціональної ефективності інформаційно-пошукових систем наведені в [4]. До них належать такі: повнота, точність, акуратність, помилки. Для оцінювання функціональної ефективності інформаційно-пошукових систем запропоновано використовувати методи теорії статистичних рішень. Значну увагу звернено на модифікацію відомого критерію зваженої комбінації та показано його ефективність на прикладі експериментального пошуку в масиві патентів США. У роботі [5] розглянуто проблему пошуку інформації в Інтернеті, її зв'язок з традиційною проблемою пошуку інформації. Описано нові завдання, відмінності пошуку в Інтернеті від традиційного пошуку інформації, відомі методи пошуку інформації в Інтернеті. Модель розв'язання задачі інформаційного пошуку, яка передбачає математичний опис послідовного та бінарного пошуків, наведено в [6]. Зміст послідовного пошуку полягає в порівнянні записів. Для бінарного пошуку

використовується бінарне дерево. Показано, що ефективність пошуку визначається принаймні двома основними – точністю і повнотою, та чотирма додатковими – специфічністю, вибірковістю, коефіцієнтом втрати інформації та коефіцієнтом пошукового шуму – показниками. Зазначено, що для оцінювання роботи пошукової системи потрібна репрезентативна кількість запитів. В [7] сформульовано принципи оцінювання ефективності функціонування сучасних інформаційно-пошукових систем Інтернету. Наведено результати тестування шести інформаційно-пошукових систем на основі методу визначення глибини користувацького пошуку. Різні моделі діяльності користувача в системі "людина–комп'ютер" наведено в [8]. Тут визначено особливості цих моделей, сфери їх використання, дано деякі характеристики та наведено їхні аналітичні вирази. Важливим класом задач, які розв'язуються в інформаційних системах, як зазначено в [9], є задачі прийняття рішення. Наведено дворівневу системну модель опису технічного об'єкта, в якій не прямо враховано інтелектуальну складову в сенсі ознак, властивостей, зв'язків, відповідностей, характерних для інтелектуальної діяльності людини. В останніх двох роботах для дослідження і побудови моделей використовують апарат теорії множин.

Формулювання мети статті

Очевидно, що процес пошуку вимагає не лише коректної постановки задачі пошуку "що треба знайти", але насамперед коректної постановки питання "де і як треба шукати". Зрештою виникає логічне запитання: "чи знайдено те, що треба". На підставі аналізу існуючих підходів до оцінювання результатів інформаційного пошуку можна зробити такі висновки.

1. У теоретичному плані оцінюють результат на підставі математичних моделей інформаційного пошуку. Для цього використовують переважно теоретико-множинний апарат, рідше ймовірнісний, і розглядають відношення множин релевантних та нерелевантних документів у видачі їх пошуковою системою.

2. На практиці використовують критерії точності і повноти, рідше включають і частку нерелевантних документів у видачі.

3. Відсутність інтегрального критерію оцінювання результатів інформаційного пошуку.

4. Відсутність математичної моделі користувача, оскільки саме він, в сенсі поставленої задачі, визначає якість і відповідність отриманих у видачі документів.

Метою цього дослідження є моделювання та оцінювання результатів інформаційного пошуку в сенсі математичної моделі інтелектуальної діяльності користувача і прийняття ним рішення на підставі документів, отриманих у видачі.

Такий підхід повинен враховувати отриманий результат пошуку як вибірку документів у видачі і має здійснюватись виключно на підставі видачі першого запиту, а вже потім пошук можна уточнювати додатковими змінами в запиті, наприклад, доповненнями до ключових слів, термінів, використанням окремих фрагментів тексту тощо.

Виклад основного матеріалу

Процес діяльності користувача як складової системи "людина–комп'ютер" можна формально подати у вигляді когнітивної моделі інтелектуальної роботи з матеріалами у видачі. Формалізація пошукової діяльності людини певною мірою стосується моделювання її трудового процесу [8]. Структура інтелектуальної діяльності людини є багаторівневою, адаптивною, суттєво залежить від психофізіології організму, а тому погано піддається формалізації. На загальному рівні діяльність користувача можна подати у вигляді послідовності процесів інформаційного пошуку, сприйняття інформації, аналізу інформації та прийняття рішення. В цьому аспекті діяльність користувача є інформаційно-аналітичною діяльністю, скерованою на розв'язання довідкових та інформаційно-аналітичних задач.

Доволі часто перед користувачем стоїть задача Z, розв'язання якої вимагає конкретних даних, які відсутні у нього. В таких випадках користувач звертається до пошукових систем з відповідним запитом для виявлення і отримання документів (матеріалів), які містять дані, потрібні для розв'язання його задачі. Очевидно, організація такого пошуку, як і вибір потрібних

інформаційно-значущих об’єктів, отриманих в документах видачі, та прийняття рішення, а точніше, формулювання розв’язку, фактично може бути віднесено до інтелектуальної діяльності користувача. Структуру такої діяльності можна подати схемою, зображеною на рис. 2, яка містить такі три складові: формулювання запиту, аналіз виданих матеріалів та прийняття рішення.

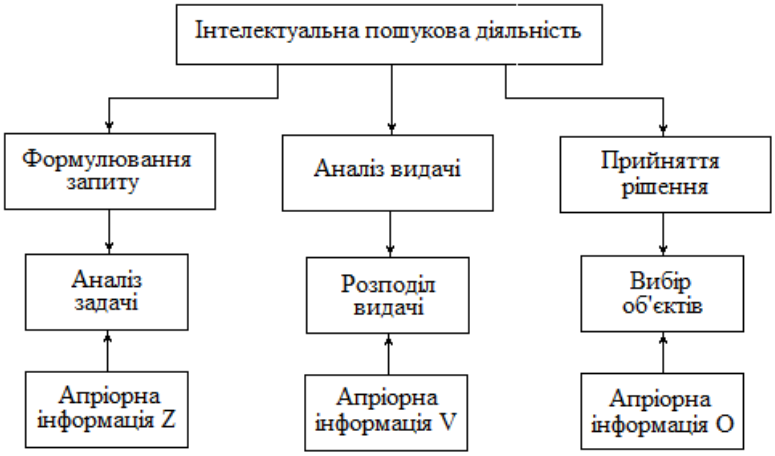

Рис. 2. Схема інтелектуальної пошукової діяльності користувача

Формулювання запиту, аналіз видачі та прийняття рішення щодо відібраних інформаційно-значущих об’єктів для розв’язання задачі фактично є когнітивними процесами, значною мірою пов’язаними з апріорними даними, збереженими в пам’яті користувача. Апріорні інформації $A(Z)$, $A(V)$, $A(O)$ відрізняться між собою, оскільки відповідають різним процесам. Ці процеси можна формально подати в такий спосіб. Нехай для розв’язання задачі $Z = \{ z_i : i = 1, 2, \mathbf{K}, q \}$, яка декомпонується на z_q підзадач, потрібно використати $m \geq q$ інформаційно-значущих об’єктів (констант, конкретних величин, формул, моделей, методів, принципів, законів тощо). Користувач уявляє собі спосіб розв’язання, проте йому також потрібні об’єкти, що знаходяться у відповідних документах в бібліотеках та базах даних. Назви таких документів доволі часто не відповідають назвам потрібних об’єктів, а пошукові системи переважно видають саме назви документів. Тому перед користувачем постає задача: так сформулювати запит, щоб отримати у видачі якомога більше документів з інформаційними об’єктами. Отже, якщо у видачі присутні релевантні P, пертинентні Π та нерелевантні H об’єкти, то запит має забезпечити умову $(P + \Pi) >> H$.

У процесі формулювання запиту користувач використовує апріорну інформацію $A(Z)$ у вигляді збереженої в його пам’яті еталонної моделі $E(D_z)$, де D_z – документи з інформаційними об’єктами стосовно цієї задачі. У результаті когнітивного аналізу x він формує множину ключових слів K_z, тобто існує відображення

$$x : E(D_z) \mathbf{I} A(Z) \rightarrow K_z .$$

Результатом реалізації запиту є видача $V(D)$ – вибірка довільних документів, які містять релевантні P, пертинентні Π та нерелевантні H об’єкти. Користувач, аналізуючи видані документи, використовує також апріорну інформацію $A(V)$, яку асоціює з об’єктами у видачі і поділяє їх на класи P, Π, H, тобто виникає відображення:

$$c : V(D) \mathbf{I} A(V) \rightarrow \Omega, \quad \Omega = \{ \Omega_P, \Omega_\Pi, \Omega_H \mid \Omega_P \mathbf{U} \Omega_\Pi \mathbf{U} \Omega_H = \Omega, \ \Omega_P \mathbf{U} \Omega_\Pi \mathbf{U} \Omega_H = \varnothing \}.$$

Використовуючи апріорну інформацію $A(O)$ щодо необхідних інформаційних об’єктів, користувач формує їх концептуальну модель у вигляді інформаційно-значущих об’єктів

$w_q \in (\Omega_P \cup \Omega_П)$, тобто вибір інформаційно-значущих об'єктів для розв'язання задачі можна подати відображенням

$$J : (\Omega_P \cup \Omega_П) \cap A(O) \to R_Z(w_q).$$

Формальне подання видачі документів. Діяльність користувача з пошуку потрібних матеріалів – інформаційно-значущих об'єктів – розпочинається з аналізу поставленої перед ним задачі. Очевидно, розв'язок такої задачі перш за все вимагає даних про способи розв'язання аналогічних задач, з'ясування, які дані, критерії, умови, вимоги були для них використані, що дали отримані розв'язки. Крім того, користувач, опираючись на власні знання та досвід, укладає у своїй уяві шляхи її розв'язання і, головне, намагається виявити, якої інформації йому бракує і як її компактно сформулювати, щоб збільшити ймовірність отримання релевантної інформації з доступних джерел. Власне результат такої розумової діяльності використовується ним для побудови запитів для різних баз даних. Подамо множину документів w, виданих користувачу в результаті запиту $z_m \in Z$, де Z – множина з m зроблених запитів, деякою множиною Ω, тобто $w \in \Omega$. Серед документів w містяться, можливо, потрібні користувачеві. Розіб'ємо цю множину на декілька неперетинних класів $k = 1, 2, \ldots K$, де K – множина класів, тоді:

$$\Omega = \left\{ \Omega_k : \bigcup_{k=1}^{K} = \Omega, \quad \bigcap_{k=1}^{K} \Omega_k = \varnothing, \; k = \overline{1, K} \right\}.$$

Позначимо класи в такий спосіб: k_1 – пертинентні; k_2 – релевантні; k_3 – нерелевантні. Тоді будь-який з виданих об'єктів w_j^k належатиме до своєї підмножини – класу, тобто $w_j^{k_i} \in \Omega_{k_i}$ де $j = \overline{1, J_k}$, J_k – кількість об'єктів в класі Ω_k. Об'єкти w_j^k представлені ключовими словами та певними вимогами у запиті.

Фактично кожен об'єкт у цьому класі документів можна описати певною підмножиною $A_k^* \subseteq A_k$ ознак – ключових слів

$$A_k = \left\{ a_i^k : \bigcup_i a_i^k = A_k, \; i = 1,2,\ldots,n \right\},$$

де A_k – множина всіх ключових слів у цьому класі, а в деяких ситуаціях виданий об'єкт може бути віднесений до свого класу і за відсутності в нього деяких з них, тобто за набором ознак, що є булеаном $B(A_k)$ – множини A_k. Ознаки та їх комбінації, які входять у цей набір, визначають рішення про віднесення виданого об'єкта до його класу розпізнавання та забезпечують своєю інформативністю ефективність пошуку.

За невеликої кількості чітко виражених інформативних ознак визначається клас виданого об'єкта практично миттєво, оскільки такий об'єкт сприймається цілісно і збігається зі своїм образом в уяві. Проте в більшості випадків, навіть коли належність об'єкта даному класу очевидна, можуть виникати певні сумніви щодо його релевантності чи пертинентності. Наприклад, невідомі автор, час і місце публікації, авторитет видання тощо.

Користувач як динамічна система

Діяльність користувача в таких випадках вимагає значного розумового напруження, посиленої роботи пам'яті і зорового аналізатора, причому за умови мінімальної рухливості, монотонного режиму роботи. В результаті прогресує психічне напруження, знижується концентрація уваги, зростає втома, а відтак знижується якість роботи внаслідок помилок, повторного опрацювання матеріалів, збільшується тривалість пошуку.

Тому в організації пошукової діяльності користувача основною складовою перегляду виданих пошуковою системою матеріалів є когнітивний процес.

Позначимо через $\vec{g}(V)$ вектор назв документів у видачі, які відповідають ключовим словам у запиті, а через $\vec{g}(Z)$ вектор ключових слів у запиті.

Тоді побудувати когнітивну модель пошукової діяльності користувача можна, розглядаючи користувача як складну динамічну систему S, в сенсі теоретико-множинного підходу, наведеного в [10] і подану відношенням на множинах: X – об'єктів у видачі і Y – відібраних користувачем.

Система S є функціональною, якщо кожному елементу множини Y однозначно відповідає єдиний елемент множини X, тобто відношення S є функцією

$$S : X \to Y .$$

Множина X отриманих документів є областю її визначення

$$D(S) = \{x : (\exists y)((x,\ y) \in S)\} = X ,$$

а множина прийнятих рішень Y є областю значень цієї системи

$$R(S) = \{y : (\exists x)((x,\ y) \in S)\} = Y .$$

У результаті попереднього перегляду виданих матеріалів функцію S буде позитивно визначено лише для деяких документів множини X, тобто релевантних, а при повторних переглядах до них можуть бути долучені і пертинентні. Очевидно, функцію буде негативно визначено для всіх інших – не релевантних, відкинутих документів.

Якщо система S є динамічною системою, то для неї існує довільна множина C така, що функція R реалізує деяке відображення $R : (C \times X) \to Y$, для якого існує умова

$$(x,\ y) \in S \Leftrightarrow (\exists c)[R(c,\ x) = y].$$

Якщо ця умова виконується, тоді множина C є множиною станів системи – рівнями функціонального стану користувача. Власне рівні функціонального стану користувача визначаються змінами його нормального робочого стану, які можуть бути зумовлені зниженням психомоторних функцій та концентрації уваги, дискомфортом робочого середовища та зовнішніми впливами (шумом, звуком, відволіканням тощо).

Оскільки процес аналізу і відбору виданих документів реалізується в часі, то, в принципі, в роботі користувача можна виділити такі два моменти:

по-перше, функціональний стан C користувача на момент часу t відповідає деякому рівню C_t, і стосовно документа X_t він приймає відповідне рішення Y_t, що можна подати такою сім'єю відображень

$$\bar{r} = \{\, r_t : C_t \times X_t \to Y_t \quad \& \quad t \in T \,\};$$

по-друге, сенс, зміст, форма, належність документа X_t можуть змінити рівень його функціонального стану C_t і навіть значно збільшити термін прийняття рішення Y_t, що відповідатиме такому відображенню

$$\bar{j} = \{\, j_{tt'} : C_t \times X_t \to C_{t'} \quad \& \quad t,\ t' \in T \quad \& \quad t < t' \,\}.$$

Отже, формально діяльність користувача протягом деякого часу T можна подати моделлю, яка містить останні два відображення $\bar{r}$ і $\bar{j}$.

Когнітивна модель опрацювання отриманих документів користувачем

Нехай V – множина виданих пошуковою системою документів за зробленим користувачем запитом Z. Користувач в результаті перегляду цих документів поділяє їх на релевантні V_r, нерелевантні V_n та пертинентні V_p. У результаті отримані з видачі документи поділено на три класи, тобто множина виданих документів $V = V_r \cup V_n \cup V_p$ складається з трьох підмножин.

На “вхід” користувача з використаної інформаційно-пошукової системи надходить видача V – множина виданих документів, відповідно до зробленого ним запиту Z. У результаті побіжного перегляду частину $L \subseteq V$ документів відразу можна відкинути як не релевантну для цієї задачі. Відкидання завідомо непотрібних документів зумовлене існуючою в пам’яті користувача еталонною моделлю стосовно розв’язання задач такого типу, яка містить апріорні знання про ті чи інші документи, які мають допомогти розв’язати поставлену задачу. Отже, користувач працюватиме з множиною документів $X = V \setminus L$. До цієї множини входять всі документи множини $X = \{x_i : x_i = x(t_i),\ i = 1, 2, \mathbf{K}, m \ \&\ m = |X|\}$, причому зв’язок з часом вказує на послідовність роботи з документами, чим підкреслюється розгляд користувача в сенсі теорії динамічних систем, а m – реальна потужність множини. Множина X є фактично інформаційною моделлю $M(X(x(t_i)))$ на “вході” користувача. Зазначимо, що в цьому дослідженні подання документа як x_i означає конкретний документ у множині X, а подання документа як $x(t_i)$ означає роботу з документом в момент часу t_i, тобто можливе повторне звертання до вже переглянутого документа, що й визначає реальну потужність m множини X. Інакше кажучи, інформаційна модель допускає повторний перегляд документів, тобто повторення документа у множині X, адже кількість елементів в моделі визначається кількістю моментів часу роботи з документом.

Еталонна модель містить: назви документів, терміни, ключові слова, на підставі яких розроблено запит. Вона в інформативному плані містить більше даних і знань, ніж їх містить запит, а тому є значно ширшою, ніж отримана інформаційна модель. Еталонна модель $E(G(g(t_i)))$ містить назви відомих користувачу документів, зміст багатьох документів, термінів, способів формулювання ключових слів, можливі шляхи та способи розв’язання поставленої задачі.

Маючи на “вході” інформаційну модель $M(X(x(t_i)))$ і використовуючи свою еталонну модель $E(G(g(t_i)))$, користувач для вироблення рішення стосовно кожного об’єкта видачі будує у своїй уяві когнітивну модель для вибору відповідного рішення з множини альтернативних рішень $Y = \{y_j : y(t_j),\ j = 1, 2, \mathbf{K}, r \ \&\ r = |Y| \ \&\ r \le m\}$.

Множина альтернативних рішень містить такі:

- документ є релевантним, відповідає зробленому запиту;
- документ є пертинентним, хоча і не відповідає ключовим словам, сформульованим у запиті;
- проте його побіжний перегляд вказує на те, що він містить потрібні для розв’язання поставленої задачі дані;
- документ є не релевантний, оскільки його перегляд не дав нічого нового, що може бути використано для розв’язання задачі;
- документ вже з першого погляду відкидається як такий, що не відповідає розв’язуваній задачі і не вартий того, щоб його докладніше опрацьовувати;
- необхідно модифікувати запит і повторити пошуковий запит;
- змінити пошукову систему.

Когнітивна модель за таких альтернативних рішень відповідає перетину моделей $M(X(x(t_i)))$ і $E(G(g(t_i)))$, тобто маємо $K(M(X(x(t_i))) \mathbf{I}\ E(G(g(t_i))))$.

Ці три моделі фактично відображають інтелектуальну діяльність користувача в процесі перегляду, аналізу і відбору документів, отриманих у видачі за конкретним запитом.

З формального погляду таку діяльність користувача можна подати системою відображень:

$$\begin{cases} a: Z(A_k)\times B\times \Pi\times T\to V; \\ h: V\setminus\Pi\times T\to R\cup P=X; \\ m: X\to M(X(x(t_i))); \\ k: M(X(x(t_i)))\times E(G(g(t_i)))\times C(d_q,t_q)\times T\to K(M(X(x(t_i)))\cap E(G(g(t_i)))); \\ r: K\times Y\times C\times T\to R, \end{cases} \quad (1)$$

де a – відображення виявлення пошуковою системою в бібліотеках, базах даних та інших джерелах документів за зробленим користувачем запитом; h – відображення візуалізації назв і фрагментів анотацій документів, знайдених пошуковою системою, які в результаті поверхневого перегляду вже на цьому етапі можуть бути прийняті як релевантні або нерелевантні. Останні можуть бути віднесені до пошукового "шуму" Π, оскільки ймовірно вони потрапили у видачу через недосконаліть "розуміння" запиту пошуковою системою. Такі документи виключають з множини X внаслідок різниці $V\setminus\Pi$; m – відображення сприйняття отриманої видачі користувачем, тобто формування інформаційної моделі на "вході" користувача; k – відображення побудови когнітивної моделі інтелектуальної діяльності користувача з опрацювання отриманих у видачі документів шляхом взаємодії інформаційної моделі $M(X(x(t_i)))$ і еталонної моделі $E(G(g(t_i)))$, яка стосовно розв'язаної задачі сформована в його пам'яті. Ця модель створюється в результаті докладного і аналітичного перегляду та відбирання документів, отриманих у видачі; r – відображення вибору та прийняття стосовно кожного елемента розв'язання R. Отже, пошукову діяльність користувача, тобто його взаємодію з пошуковою системою, результатом якої є створення вибірки релевантних документів, можна подати моделлю – системою відображень. Результат видачі може задовольнити користувача, але може поставити перед ним нові завдання: зміна або уточнення ключових слів, розширення запиту, зміна пошукової системи тощо. Як правило, користувач здійснює декілька пошукових ітерацій, часто з залученням різних інформаційних джерел. В останньому випадку пошук проводять на конкретних базах даних, спеціалізованих електронних бібліотеках та на відповідних сайтах.

Відповідність видачі запиту. Найскладнішим моментом в оцінюванні ефективності будь-якого інформаційного пошуку є встановлення відповідності між знайденими і виданими документами і документами, а точніше, пошуковими ознаками документів, поданих у запиті. Річ у тім, що рішення про відповідність (тобто чи є релевантними видані документи, чи ні) є вельми суб'єктивним. Крім того, якщо можна точно відповісти, чи документ є релевантним, чи нерелевантним, то чітко вказати, чи документ є пертинентним, не можна, оскільки він може бути пертинентним різною мірою. Зі змісту понять релевантності та пертинентності випливає, що оцінювання ефективності пошуку має принципові дві складові. Нагадаємо, що поняття релевантності означає відповідність інформаційного пошуку зробленому користувачем запиту, а пертинентність – відповідність інформаційній потребі користувача.

Перша з них – це оцінювання, а точніше, розуміння пошуковою системою складеного користувачем запиту. В цьому аспекті інформаційно-пошукова система відбирає ті документи, ознаки яких вказано у запиті. Очевидно, що в такому разі семантичний аналіз виявлених документів в базі або сховищі даних, у файлах чи бібліотеках не проводиться, а лише зіставляються ознаки виявлених документів, і за умови повного чи часткового збігу документи подаються у видачу.

Друга складова – це оцінювання документів у видачі, отриманих користувачем, у результаті інформаційного пошуку. Тут користувач поділяє документи на три групи: релевантні (Р), пертинентні (П) та нерелевантні (Н). Документи у видачі, як правило, сортуються інформаційно-пошуковою системою за певними критеріями: за датою (власна дата документа або остання дата звертання до нього), за рейтингом користування (скільки разів цей документ фігурував у запитах різних користувачів загалом чи за певний період). Можливі й інші критерії, наприклад, за обсягом чи датою останнього звертання тощо. Отримавши видачу, тобто перелік знайдених документів,

користувач послідовно або вибірково ознайомлюється з документами, відбираючи релевантні та пертинентні і відкидаючи нерелевантні. Послідовність релевантних, пертинентних та нерелевантних документів у кожній конкретній видачі практично завжди є випадковою. Перевірено цей факт експериментально в такий спосіб.

Організація експериментального дослідження

Під час пошуку необхідної інформації для проведення наукових досліджень крім відбору пертинентних документів фіксувалися релевантні та нерелевантні. Зміст експериментального дослідження такий.

Відбір ключових слів. Для цього було сформульовано такі ключові слова, а точніше, словосполучення: інформаційний пошук; моделі інформаційного пошуку; інформаційно-пошукова система; ефективність інформаційно-пошукових систем; оцінювання ефективності інформаційного пошуку.

Уточнення понять. *Пертинентні документи* – це документи, які за змістом максимально відповідають потребі користувача, хоча їх назви та анотації не містять зазначених у запиті ключових слів і мають усі реквізити для посилання на них, тобто документи, що є електронними копіями паперових: монографій, статей у наукових журналах та збірниках праць, тезисах та працях наукових форумів та статті, подані в енциклопедіях та довідниках.

Релевантні документи – це ті, які за змістом і ключовими словами цілком відповідають потребі користувача і мають реквізити своїх паперових оригіналів. Ними можуть бути також і фрагменти різних матеріалів, але тоді для посилання на них треба використовувати їхню електронну адресу, яка в деяких випадках є або громіздкою, або неточною, і для виявлення цього документа необхідно провести додатково ще й окремий спеціальний пошук, причому результат не гарантується. Всі інші документи визнаються нерелевантними.

Зміст експерименту. Для експериментальних досліджень використано інформаційно-пошукові системи Google, Яndex, МЕТА, Rambler, Yahoo, які за ключовими словами видають веб-сторінки знайдених документів. Налаштування пошуку забезпечило оптимальний варіант видачі результату – 10 електронних документів на кожній сторінці. На основі попередніх результатів пошуку і з власного досвіду відомо, що потрібна інформація стосовно цього питання знаходиться переважно на перших п’яти сторінках. Тому для експериментів вибрано обмеження 5 повних сторінок, тобто обсяг видачі за кожним ключовим словом становив 50 документів.

Для кожної сторінки в результаті перегляду кожному з десяти наведених документів присвоювалися індекси Р, П, Н.

Завдання користувача полягає в тому, щоб серед цієї множини вибрати саме ті, які відповідають або сприяють розв’язанню його задачі. Очевидно, за будь-якого пошуку перегляд отриманих документів буде аналогічним. Оскільки надана вибірка є скінченна, можемо оцінити ефективність пошукової системи відношенням сприятливих подій до всіх можливих, тобто відношенням, наприклад, кількості релевантних документів до кількості всіх наданих документів, отриманих за цим запитом. Якщо документи класифікувати як в цьому прикладі, то можна отримати три частоти появи документів кожного класу

$$f_P = \frac{1}{N}\sum_{i=1}^{n} P_i, \qquad f_П = \frac{1}{N}\sum_{k=1}^{l} П_k, \qquad f_H = \frac{1}{N}\sum_{j=1}^{m} H_j, \tag{1}$$

де N – кількість документів, виданих конкретною пошуковою системою, а величини n, m і l вказують на кількість документів P, $П$ і H відповідно.

На практиці, як правило, інформаційний пошук здійснюється за різними запитами залежно від поставлених задач. Своєю чергою, задачі можуть стосуватися різних предметних областей, обсягу їх онтологій, специфіки конкретних об’єктів, що потребують розв’язання. З іншого боку, не можна бути впевненим у тому, що інформаційні джерела мають усю необхідну інформацію з будь-якої області знань та діяльності людини. А тому кількості наданих користувачам документів є різними. Зазвичай пошук у джерелах інформації здійснюється пошуковою системою, яка працює за

певним алгоритмом і визначеними формальними критеріями відповідності, а тому можна припустити, що результати різних пошуків в одному і тому самому джерелі інформації будуть статистично однорідні, тобто матимуть певні статистичні закономірності, які можуть відбитися принаймні на співвідношенні частостей розглянутих вище класів.

Результати експериментального дослідження. Послідовність документів у видачі можна зобразити графічно у вигляді діаграми, наведеної на рис. 3.

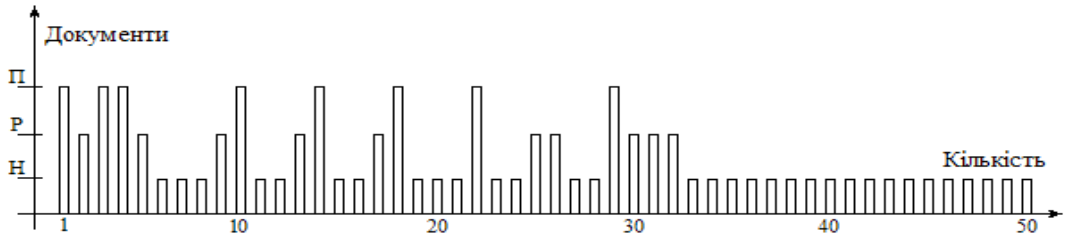

Рис. 3. Розподіл документів у видачі: П – пертинентні, Р – релевантні, Н – не релевантні

Як кількісну оцінку використано відносну частоту появи того чи іншого виду документів. Для кожної пошукової системи результат визначаємо за співвідношеннями (1): релевантних f_P, пертинентних $f_П$, нерелевантних f_H.

Очевидно, що усі ці значення значною мірою залежать від обсягу документів в інформаційній системі (базі, сховищі даних, папках з файлами, бібліотеці), можливостей інформаційно-пошукової системи, форми запиту, а також від інформаційної потреби користувача – наскільки глибоко він розуміє завдання, для вирішення якого він здійснює цей пошук.

Оцінювання ефективності інформаційного пошуку

Сформовані практично відповідними пошуковими мовами, властивими тому чи іншому інформаційному фонду, запити мають доволі обмежену кількість пошукових ознак – ключових слів, певного типу розширень та пояснень чи обмежень. Алгоритми інформаційно-пошукових систем, використовуючи ці дані в процесі сканування-пошуку існуючого каталогу, переважно використовують як дані: ім’я, назву та анотацію документів, хоча можливим є і сканування самого документа. Оскільки ключові слова залежно від контексту можуть мати декілька значень у видачу потрапляють абсолютно нерелевантні документи.

Загалом оцінювання ефективності ґрунтується на визначенні, як було сказано вище, на оцінках точності і повноти. Спроба використати додаткові показники пошуку вимагає врахування не лише обсягу самого інформаційного фонду, але і обсягу релевантних та нерелевантних стосовно даного запиту документів. Отримати такі дані практично неможливо, оскільки для однієї задачі документи можуть бути релевантними, а для другої вже ні. З іншого боку, якщо знати всі релевантні документи у фонді, то можна здійснити пошук лише для них, і тоді у видачі будуть лише релевантні документи, а це здійснити практично не можливо, принаймні з двох причин: ніхто не буде з багатотисячного інформаційного фонду відбирати релевантні для даної задачі окремого користувача документи, присутність конфіденціальної інформації та відсутність інформації про сам фонд, за винятком лише загальних його характеристик. Тому найбільш правомірним є оцінювання ефективності пошуку за його результатами, тобто на основі документів, які є у видачі.

За наявності трьох типів документів оцінити ефективність інформаційного пошуку можна так. Очевидно, що пертинентні документи мають найбільшу цінність для користувача, оскільки можуть містити потрібні нові, невідомі або забуті інформаційні об’єкти.

Для своєї задачі користувач, як правило, використовує лише релевантні та відібрані пертинентні документи, тобто ефективною видачею вважається сума $E_{пош} = П + P$, а тому в межах обсягу видачі ефективність пошуку можна визначати як відношення

$$E_{пош} = \frac{П + Р}{П + Р + Н}. \quad (2)$$

Враховуючи особливості форми запиту, яка тісно пов'язана з конкретною інформаційною системою, тобто з її інформаційним фондом та його системою індексування, необхідно ввести деякий корегувальний множник – коефіцієнт пропорціональності b, в результаті чого отримаємо

$$E_{пош} = b \cdot \frac{П + Р}{П + Р + Н}. \quad (3)$$

Оцінка ефективності пошуку у вигляді (3) характеризує здійснений інформаційний пошук у конкретній системі, для конкретного ключового слова та за результатами отриманої видачі, обмеженої $N = Р + П + Н$ документами. Інакше кажучи, показник (3) характеризує надання переваги джерелу (системі пошуку) стосовно розв'язку задачі користувача.

Визначити показник b можна лише на підставі отриманих у видачі даних двома і більше пошуковими системами. Для цього необхідно.

1. Чітко сформувати множину ключових слів.
2. Визначити інформаційно-пошукові системи, якими буде здійснюватись пошук.
3. Знайти для кожної пошукової системи середнє значення $E_{пош}$ за всіма ключовими словами.
4. Обчислити суму значень $E_{пош}$ всіх використаних пошукових систем.
5. Поділити усереднені показники ефективності для кожної системи на цю суму.

Значення цього показника для п'яти пошукових систем наведено в таблиці.

Очевидним є факт: що більший обсяг інформаційного фонду, то більше релевантних документів буде знайдено. Однак, тут треба мати на увазі і популярність чи розвиненість цієї тематики, оскільки саме її популярність і затребуваність визначають обсяг документів у фонді. Тому значення показника b з часом змінюватимуться. Вирази (2) або (3) дають об'єктивну оцінку ефективності інформаційного пошуку, але лише за умови, що у видачі будуть присутніми також і нерелевантні документи – принаймні хоча б один.

Оцінка ефективності інформаційно-пошукових систем за кількістю релевантних та пертинентних документів

№ з/п	Ключові слова	Google	Yandex	META	Rambler	Yahoo
1	Оцінка ефективності інформаційного пошуку	0,42	0,46	0,22	0,42	0,34
2	Інформаційний пошук	0,50	0,32	0,36	0,42	0,36
3	Модель інформаційного пошуку	0,34	0,44	0,22	0,76	0,46
4	Інформаційно-пошукова система	0,46	0,44	0,10	0,40	0,26
5	Ефективність інформаційно-пошукових систем	0,36	0,52	0,34	0,36	0,54
6	Усереднений показник ефективності	0,41	0,42	0,24	0,47	0,39
7	Показник відношення до системи *b*	0,212	0,218	0,124	0,244	0,202

Вирази (2) і (3) дають об'єктивну оцінку ефективності інформаційного пошуку, але в перших (лівих) варіантах, лише за умови, що у видачі будуть присутніми і нерелевантні документи – принаймні, хоча б один. Невиконання цієї умови означає ділення на нуль. Тобто, буде неправильний результат оцінювання. Така ситуація може виникнути тоді, коли кількість релевантних документів в інформаційному фонді перевищує обсяг видачі.

У цьому випадку оцінюється ефективність за видачею для одного чи декількох запитів. Якщо такий показник використати для кожного з декількох запитів, але таких, що стосуються конкретної

теми, можна оцінити якість і самого запиту, точніше встановити, який з запитів чи які ключові слова є найбільш ефективними і вже за ними модифікувати наступні запити.

Висновки та перспективи подальших наукових розвідок

Ефективність інформаційного пошуку в різних системах зберігання інформації в сенсі побудови інтегрального показника практично не можна визначити, оскільки крім двох показників – повноти і точності – усі інші вимагають знання кількості релевантних та нерелевантних документів у цьому інформаційному фонді стосовно цієї задачі. Отримати такі дані для великих за обсягами фондів неможливо, оскільки: по-перше, здійснити такий підрахунок означає перегляд кожного документа, по-друге, у великих базах даних перехід від нерелевантних документів до релевантних практично за будь-яким запитом є нечітким і розмитим, по-третє – для різних задач поняття релевантності документів різняться. Найпростішим способом побудови оцінки ефективності пошуку є використання логічного підходу, який полягає у поданні відношенням – кількості потрібних замовлених документів до кількості документів у видачі, які не відповідають потребам користувача. На ефективність пошуку впливає не лише наявність в інформаційному фонді потрібних документів, але й правильність побудови самого запиту згідно з вимогами пошукової системи. Наведений приклад оцінювання ефективності інформаційного пошуку демонструє правомірність використання знайдених і виданих документів на пертинентні, релевантні та нерелевантні. В результаті такого поділу оцінку ефективності можна подати як усереднену, або сумарну, за результатами проведення інформаційного пошуку в одному або в кількох інформаційних фондах і на різних пошукових системах за одного набору ключових слів.

Розроблений підхід до побудови оцінки інформаційного пошуку має практичне значення, оскільки отримані кількісні значення локальних оцінок дають підстави для оптимізації набору ключових слів та визначення найбільш відповідних інформаційних фондів і пошукових систем.

1. Агеев М. Официальные метрики РОМИП 2010/ М. Агеев, И. Кураленок, И. Некрестьянов // Российский семинар по оценке методов информационного поиска: Труды РОМИП, 2010. (Казань, 15 октября 2010 г.) Казань, 2010. С. 172–187. 2. Целых А. Н. Оценка эффективности информационного поиска / А. Н. Целых, Э. М. Котов // Известия ТРТУ. Тематический выпуск "Управление в математических системах". – Таганрог: Изд-во ТРТУ. – 2006. – № 10 (65). – С. 43–45. 3. Яхина Е. П. Методы оценки информационных систем / Е. П. Яхина // В мире научных открытий. – 2010. – № 3 (09). – Ч. 1. – С. 63–66. 4. Попов С. В. Оценка функциональной эффективности систем текстового поиска на примере поиска патентных документов / С. В. Попов // Патентная информация сегодня. – 2010. – № 1. – С. 22–25. 5. Козлов Д. Д. Информационно-поисковые системы в Internet: текущее состояние и пути развития [Електронний ресурс] / Д. Д. Козлов // Информационно-поисковые системы в Internet: текущее состояние и пути развития. Технологический обзор. – М. 2000. – [28 с.]. – Режим доступу: http://lvk.cs.msu.su/~ddk/ir_and_ia_review.pdf. 6. Тявкин И. В. Математическая модель информационного поиска и оценка єффективности поисковой системы / И. В. Тявкин, В. М. Тютюнник // Вестник ТГТУ. – 2008. – Т. 14. – № 3. – С. 478–481. 7. Козлов М. В. Метод оценки эффективности функционирования современных информационно-поисковых систем Интернета [Електронний ресурс] // М. В. Козлов, В. А. Яцко. – Режим доступу: – http://www.dialog-21.ru/dialog2006/materials/html/Kozlov.htm. 8. Кирхар Н. В. Модели деятельности пользователя компьютеризованной системы / Н. В. Кирхар, Д. В. Ходаков // Вестник ХНТУ № 4(27), 2007. – С. 370–378. 9. Багаев Д. В. Разработка системной модели технического объекта [Електронний ресурс]. – Режим доступу: http://systech.miem.edu.ru/2.doc. 10. Месарович М. Общая теория систем: математические основы / М. Месарович, Я. Такахара; под ред. С. В. Емельянова. – М.: Мир, 1978. – 312 с.