УДК 681.3.06.003.1:004.369.6

Нич Л. Я.[1], Шаховська Н. Б.[2], Камінський Р. М.[3]

[1]Асистент, кафедра інформаційних систем та мереж, Національний університет «Львівська політехніка», Львів, Україна
[2]Д-р техн. наук, професор, кафедра інформаційних систем та мереж, Національний університет «Львівська політехніка», Львів, Україна
[3]Д-р техн. наук, доцент, кафедра інформаційних систем та мереж, Національний університет «Львівська політехніка», Львів, Україна

ОЦІНЮВАННЯ ЕФЕКТИВНОСТІ ІНФОРМАЦІЙНОГО ПОШУКУ В СИСТЕМАХ КОНСОЛІДОВАНОЇ ІНФОРМАЦІЇ

Для оцінювання ефективності інформаційного пошуку запропоновано поділяти знайдені документи на пертинентні, релевантні та нерелевантні. Ефективність пошуку визначати відношенням кількості пертинентних та релевантних документів до кількості нерелевантних документів, а властивості пошукової системи повинні бути подані відповідним коефіцієнтом. Основною метою цього дослідження є розробка інтегрального критерію оцінювання ефективності інформаційного пошуку за результатами видачі в системах консолідованої інформації. Введено поняття консолідованої інформації.

Розроблено метод оцінювання ефективності інформаційного пошуку. Він демонструє використання поділу знайдених і виданих документів на пертинентні, релевантні та нерелевантні. Введено інтегральний показник релевантності документу пошуковому запиту, який враховує негативну та позитивну оцінку. Оцінку ефективності подано як сумарне значення різних компонентів. Експеримент виконано на підставі 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.

Ключові слова: інформаційна система, інформаційний пошук, ефективність, пертинентність, релевантність.

НОМЕНКЛАТУРА

P – клас релевантних документів;

$П$ – клас пертинентних документів;

H – клас нерелевантних документів;

D – множина інформаційних джерел;

K – користувач, який формує запит для інформаційного пошуку в джерелах;

N – кількість документів у видачі;

$f_{П}$ – відносна частоту появи пертинентних документів;

$f_{Р}$ – відносна частоту появи релевантних документів;

$f_{Н}$ – відносна частоту появи нерелевантних документів;

$f_{Р+П}$ – відносна частоту появи релевантних і пертинентних документів;

γ – відношення загальної кількості релевантних і пертинентних документів до кількості нерелевантних документів.

ВСТУП

Розвиток комп'ютерної техніки та інформаційних технологій значною мірою стимулював створювання і наповнення різноманітною інформацією як загальні, так і спеціалізовані бази даних, забезпечуючи управління ними. Проте, з іншої сторони, величезні обсяги даних практично унеможливлюють безпосередню роботу користувача з ними, що у свою чергу стимулювало розвиток відповідних пошукових систем, основною метою яких є своєчасне і повне забезпечення користувача необхідними йому даними. Тому найкритичнішою проблемою, з якою зустрічаються користувачі, – це забезпечення надійного, постійного та повнофункціонального доступу до актуальних даних.

1 ПОСТАНОВКА ЗАДАЧІ

Проблема побудови критеріїв оцінювання функціональної ефективності інформаційного пошуку в системах консолідованої інформації полягає в тому, що шукана інформація зберігається в різних джерелах, 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. В цьому плані, об'єктивна оцінка ефективності пошуку, а в даному випадку ще й консолідації знайдених і виданих документів може бути зроблена саме 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 кодування збережуваної інформації. Очевидно, що вихід з такої ситуації для користувача є і йому немає потреби вивчати премудрості мов різних запитів, оскільки потрібний пошук здійснюють спеціальні пошукові системи. Робота з однією чи навіть декількома базами даних практично полягає у правильному формуванні запиту і тут існуюча пошукова система допомагає знайти необхідну інформацію. Наприклад, локальні бази даних навіть великих підприємств досить швидко дають інформацію про виготовлені вироби, товари, зарплату працівників тощо. Проте, пошук даних в «чужих» базах

DOI 10.15588/1607-3274-2016-2-13

даних може стати складною проблемою [11]. Тут найкращим прикладом є пошукова система Google та аналогічні з нею, які видають десятки тисяч документів, з яких вибирають лише декілька, витрачаючи величезну кількість часу на пошук потрібних серед наданих пошуковою системою.

Не меншою є проблема інформаційного пошуку в системах консолідованої інформації. Термін консолідована інформація означає одержані з декількох інформативних джерел системно інтегровані різнотипні інформаційні ресурси 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, такі системи є результатом інтеграції різноманітних джерел інформації, які були створені в різний час і за різними принципами та мовами запитів, а головне за різними фаховими ознаками та онтологіями. Досить часто основні техніко-економічні дані зосереджені в системах, які реалізують численні офісні, адміністративні і технологічні процеси, а в результаті такі дані не можуть спільно використовуватись в масштабах всього підприємства.

2 ЛІТЕРАТУРНИЙ ОГЛЯД

Поняття ефективності має широке тлумачення і переважно в економічному аспекті. В роботі [1] для оцінки 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. З результатів аналізу існуючих пошукових систем в [2] робиться висновок, що для пошуку документів гіпертекстових базах даних існуючі загальновизнані оцінки мають певні обмеження. Запропоновано використовувати додаткові характеристики, до яких відносять M-різновид вибірки та U-впорядкованість вибірки. На цій підставі наводиться коефіцієнт впорядкованості та коефіцієнт пошукового шуму. Виділена низка 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 оцінки трудовитрат приведена модифікована формула, яка враховує модель оцінки вартості розробки програмного забезпечення. Кількісні показники – оцінки функціональної ефективності інформаційно-пошукових систем приведені в [4]. До них віднесено такі: повноту, точність, акуратність, помилки. Для оцінювання функціональної ефективності інформаційно-пошукових систем запропоновано використовувати методи теорії статистичних рішень. Значна увага приділена модифікації відомого критерію зваженої комбінації, та показано його ефективність на прикладі експериментального пошуку в масиві патентів США. У роботі [5] розглянута проблема пошуку інформації в Інтернет, її зв'язок з традиційною проблемою пошуку інформації. Описано нові завдання, відрізняють проблему пошуку в Інтернет від традиційної проблеми пошуку інформації, даний огляд існуючих методів пошуку інформації в Інтернет. Модель розв'язку задачі інформаційного пошуку, яка включає математичний опис послідовного та бінарного пошуків приведена в [6]. Зміст послідовного пошуку полягає в проведенні порівнянь 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 оцінки роботи пошукової системи потрібна репрезентативна кількість запитів. У [7] формулюються принципи оцінки ефективності функціонування сучасних інформаційно-пошукових систем Інтернету. Наводяться результати тестування шести інформаційно-пошукових систем на основі методу визначення глибини користувацького пошуку.

На підставі аналізу існуючих підходів до оцінювання ефективності інформаційного пошуку можна зробити такі висновки.

1. В теоретичному плані оцінювання ефективності проводиться 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 та інколи у інформаційному фонді.

2. В практичному використанні використовують критерії точності і повноти, рідше включають і частку нерелевантних документів у видачі.

3. Відсутність інтегрального критерію ефективності інформаційного пошуку.

Основною метою дослідження є розробка інтегрального критерію оцінювання ефективності інформаційного пошуку за результатами видачі в системах консолідованої інформації.

Такий інтегральний показник повинен враховувати не лише позитивний результат пошуку, але і негативний – частку нерелевантних документів та частку релевантних, але не виданих документів. Релевантні невидані документи за одним запитом можуть бути знайдені і включені у видачу або за рахунок іншого (нового) запиту або за рахунок модифікації даного запиту. Проте в першу чергу базове оцінювання ефективності пошуку має здійснюватись виключно на підставі видачі першого запиту, а вже далі такий інтегральний показник можна уточнювати додатковими оцінками.

3 МАТЕРІАЛИ І МЕТОДИ

Пошук в системі консолідованої інформації. Розглянемо роботу системи консолідованої інформації як діяльність користувача, пов'язану з відбором відповідної інформації стосовно поставленої задачі. В результаті зроб-

леного запиту інформаційно-пошукова система здійснює видачу знайдених документів. Як правило, не всі видані документи відповідають зробленому запиту і потребам користувача. З точки зору його задачі видані документи можуть бути поділені принаймні на три класи: релевантні P, пертинентні Π та не релевантні H.

Позначимо, через D множину різноманітних інформаційних джерел $D = \{d_1, d_2, \dots, d_n,\ d_i \in D, d_i \cap d_j \neq \neq \varnothing,\ i, j = 1, 2, \dots, n\}$, які можуть мати спільні фонди інформаційного ресурсу; K – користувач, який формує запит для інформаційного пошуку в різнотипних джерелах.

Тоді, процес інформаційного пошуку в системах консолідованої інформації можна подати у вигляді схеми цілеорієнтованої роботи трьох блоків – «Опрацювання запитів», «Консолідації даних» та «Опрацювання даних» зображеної на рис. 1. Перший з цих блоків функціонально забезпечує переклад мови запиту користувача на мову запитів кожного з інформаційних джерел $d_i \in D$. В результаті, кожне таке джерело розуміє отриманий запит і процес пошуку може здійснюватися переважно його власною пошуковою системою.

Функціонально другий блок здійснює консолідацію знайдених даних, тобто приводить дані різних форматів у формат користувача, тобто вирішує обернену задачу – приведення різнотипних даних до типу запиту, сформованого користувачем. Консолідовані дані передаються в блок «Опрацювання даних», робота якого полягає у ранжуванні даних за частотою використання, часовими характеристиками, важливістю, доступністю, терміном використання тощо. Іншими словами, знайдена в результаті пошуку інформація має бути подана користувачеві у тій самій формі, у якій він сформував свій запит або в іншій, зрозумілій для нього формі. У такій ситуації користувач може отримати надзвичайно велику кількість, випадково перемішаних, як релевантних так і не релевантних документів. Для зменшення кількості нерелевантних документів у цьому блоці здійснюється відповідний логічний аналіз наявності збігів виданих документів з визначеними в запиті. Тут, фактично здійснюється фільтрація виданих документів, шляхом використання відповідних критеріїв, попередньо заданих користувачем. У цьому плані, інформаційний пошук в системах консолідованої інформації суттєво відрізняється від пошуку в звичайних базах чи сховищах даних. Тому, оцінювання ефективності інформаційного пошуку в системах консолідованої інформації має враховувати і особливості нормалізації різнотипних даних, тобто приведення їх до форми запиту користувача.

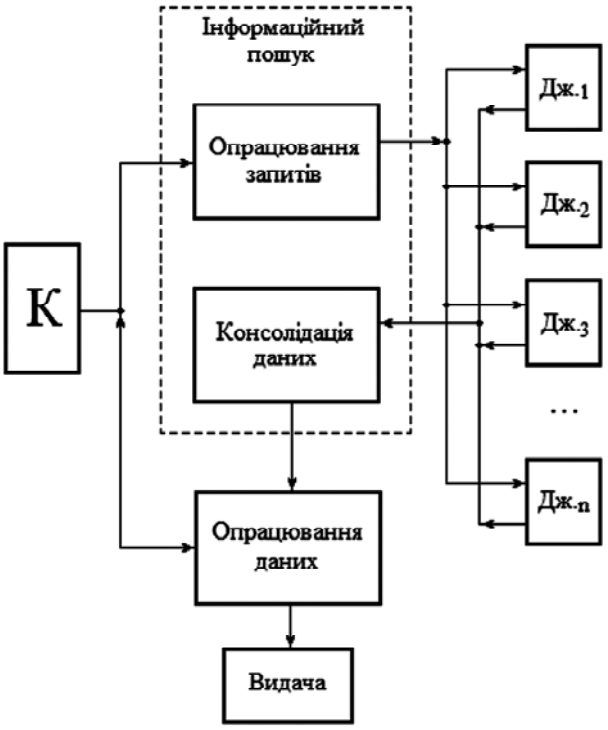

Рисунок 1 – Інформаційний пошук в різнотипних системах

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. Справа в тому, що ступінь відповідності, тобто чи є релевантними видані документи чи ні, є вельми суб'єктивним. Крім того, якщо можна точно відповісти даний документ є релевантний або нерелевантний, то чітко вказати, чи даний документ є пертинентним чи ні, 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 плані інформаційно-пошукова система відбирає ті документи, ознаки яких вказані у запиті. Очевидно, що в такому разі семантичний аналіз виявлених документів в базі або сховищі даних, у файлах чи бібліотеках не проводиться, а лише здійснюється зіставлення ознак 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 розділяє документи на три групи: релевантні (Р). пертинентні (П) та нерелевантні (Н).

Документи у видачі як правило сортуються інформаційною-пошуковою системою за певними критеріями: за датою (власна дата документа або остання дата звертання до нього), за рейтингом користування (скільки разів даний документ фігурував у запитах різних користувачів загалом чи за певний період). Можливі і інші критерії, наприклад за обсягом. 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. Перевірка цього факту здійснена експериментально в такий спосіб.

Організація експериментів. В процесі пошуку необхідної інформації для проведення наукових досліджень поряд з відбором пертинентних документів також фіксувалися релевантні та нерелевантні. Зміст експерименту поданий планом дослідження.

4 ЕКСПЕРИМЕНТИ

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

Уточнення понять. Пертинентність – документи, які за змістом максимально відповідають потребі користувача і мають усі реквізити для посилання на них у магістерській кваліфікаційній роботі (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).

Релевантність – документи, які за змістом цілком відповідають потребі користувача, але не мають реквізитів своїх паперових оригіналів і для посилання на них треба використовувати їхню електронну адресу, яка в деяких випадках є або громіздкою або неточною і для виявлення цього документу необхідно провести додатково ще й окремий спеціальний пошук, причому результат не гарантується.

Усі інші документи визнаються як нерелевантні.

Хід експерименту. Для експериментальних досліджень використано інформаційно-пошукові системи Google, Яndex, Мета, Rambler таYahoo, 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 даного питання буде знаходитись на перших п'яти сторінках. Тому для експериментів вибрано обмеження 5 повних сторінок, тобто обсяг видачі становив 50 документів.

Для кожної сторінки, за результатами перегляду, кожному з десяти виданих документів присвоювалися індекси P, $П$ та H.

У табл. 1 приведені результати одноразової видачі знайдених документів для вказаних пошукових систем за ключовим словом «Оцінювання ефективності інформаційного пошуку».

Попередні результати. Задача користувача полягає в тому, щоб серед цієї множини вибрати саме ті, які йому потрібні. Очевидно за будь-якого пошуку перегляд отриманих документів буде аналогічним. Оскільки надана вибірка є скінчена, можемо оцінити ефективність пошукової системи відношенням сприятливих подій до всіх можливих, тобто відношенням, наприклад, кількості релевантних документів до кількості всіх наданих документів, отриманих за даним запитом. Якщо документи класифікувати як в даному прикладі, то можна отримати три частоти появи документів кожного класу:

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

де N – кількість документів у видачі.

Таблиця 1 – Оцінювання ефективності інформаційного пошуку

Ключове слово «Оцінювання ефективності інформаційного пошуку»					
	Google	**Яndex**	**Мета**	**Rambler**	**Yahoo**
1	Р Р Р П Н	Р Р Н Р Н	Н Н Н Н Н	Н Р Н Р Н	Н Р Р Н Н
2	Р Р П Н Н	Р Р Н Н Н	Н Р Н Р Н	Н Р Р Н Р	Н Н Р Н Н
3	Р Н П Р Н	Р П Н Р Н	Н Р Р Р Н	Р Р Р Р Н	Р Р Р Р Н
4	Н Н Н Р Н	Н Р Н Р Н	Р Н Н Н Н	Р Р П Р Н	Р Р П Р Н
5	Н Н Р Н Н	Р Н Р Н Н	Р Н Н Н Н	Р Н Н Н Р	Р Н Н Н Н
6	Н Н Н Р Н	Р Р Р Н Н	Р Н Р Н Н	Р Н Н Н Н	Р Н Р Н Н
7	Н Н Н Н Н	Р Р Н Н Н	Н Н Р Н Н	Н Н Н Н Н	Н Н Н Р Н
8	Н Р Н Н Н	Р Н Н Н Н	Н Н Н Н Н	Н Н Н Н Н	Н Н Н Н Н
9	Н П Н Н Н	Р Н Н Н Н	Н Н Н Н Н	Н Н Р Р Н	Н Н Н Н Н
10	Н Р Н Р Н	Р Р Р Н Н	Р Н Н Н Н	Р Р Н Н Н	Р Н Н Н Н

Таблиця 2 – Зведена таблиця експериментів

Ключові слова	Google	Яndex	Мета	Rambler	Yahoo
Оцінка ефективності інформаційного пошуку	П – 4 Р – 13 Н – 33	П –1 Р –21 Н – 28	П –0 Р –11 Н – 39	П –1 Р –19 Н – 30	П –1 Р –15 Н – 34
Інформаційний пошук	П – 7 Р –11 Н – 32	П –2 Р –12 Н – 36	П –2 Р –14 Н – 34	П –3 Р –15 Н – 32	П –4 Р –10 Н – 36
Модель інформаційного пошуку	П –3 Р –11 Н – 36	П –2 Р –18 Н – 30	П –1 Р –9 Н – 40	П –9 Р –20 Н – 21	П –1 Р –21 Н – 28
Інформаційно-пошукова система	П – 1 Р –21 Н – 28	П –0 Р –22 Н – 28	П –0 Р –5 Н – 45	П –1 Р –18 Н – 31	П –1 Р –11 Н – 38
Ефективність інформаційно-пошукових систем	П –1 Р –16 Н – 33	П –2 Р –22 Н – 26	П –0 Р –17 Н – 33	П –1 Р –16 Н – 33	П –3 Р –21 Н – 26

На практиці як правило інформаційний пошук здійснюється за різними запитами, в залежності від поставлених задач.

У свою чергу, задачі можуть стосуватися різних предметних областей, обсягу їх онтологій, специфіки конкретних об'єктів, що потребують їхнього розв'язку. З другої сторони, не можна бути впевненому в 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 в 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 ж джерелі інформації будуть статистично однорідні, тобто матимуть певні статистичні закономірності, які можуть відбитися, принаймні, на співвідношенні частоти появи розглянутих вище класів.

5 РЕЗУЛЬТАТИ

Послідовність документів у видачі можна зобразити графічно, у вигляді діаграми приведеної на рис. 2.

В якості кількісної оцінки використано відносну частоту появи того чи іншого виду документів. Для поданого результату пошуку маємо такі співвідношення:

пертинентних $f_{\text{П}} = \dfrac{П}{50}$, релевантних $f_{\text{Р}} = \dfrac{P}{50}$, нерелевантних $f_{\text{Н}} = \dfrac{H}{50}$, релевантних і пертинентних (корисних) $f_{\text{Р+П}} = \dfrac{P+П}{50}$, а також відношення загальної кількості релевантних і пертинентних документів до кількості нерелевантних документів

$$\gamma = \frac{P+П}{H}.$$

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 даний пошук.

Оцінювання ефективності інформаційного пошуку. Сформовані, практично на відповідних пошукових мовах, властивих тому чи іншому інформаційному фонду, запити мають досить 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 в якості даних автора, назву та анотацію документів, хоча можливим є і сканування самого документа. Оскільки ключові слова в залежності від контексту можуть мати декілька значень у видачу потрапляють абсолютно нерелевантні документи.

У загальному оцінювання ефективності базується 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 одної задачі документи можуть бути релевантними, а для другої вже ні. З другої сторони, 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 в такий спосіб логічного виведення. Очевидно, що пертинентні документи мають найбільшу цінність для користувача.

Очевидним є те, що для своєї задачі користувач використовує лише релевантні та пертинентні документи, тому ефективність пошуку у загальному випадку є пропорційна кількості релевантних і пертинентних документів, що є сприятливою подією для користувача, тобто

$$E_{\text{пош}} = \frac{P+П}{П+P+H}. \qquad (1)$$

Наявність у видачі не релевантних документів є обернено-пропорційною подією до кількості пертинентних і релевантних документів у видачі, а тому, ефективність відносно не релевантних документів можна подати як

$$E_{\text{пош}} = \frac{P+П}{H}. \qquad (2)$$

Враховуючи особливості форми запиту, яка тісно пов'язана з даною конкретною інформаційною системою, тобто з її інформаційним фондом та його системою індексування необхідно ввести деякий коректую-

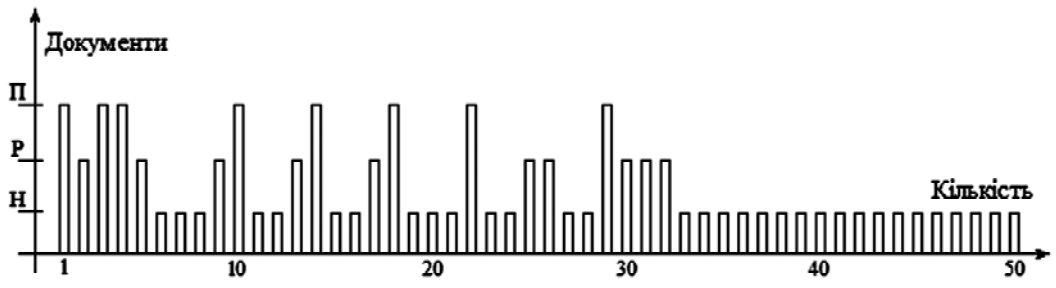

Рисунок 2 – Розподіл документів у видачі: П – пертинентні, Р – релевантні, Н – нерелевантні

чий множник – коефіцієнт пропорціональності β, в результаті чого отримаємо

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

Оцінка ефективності пошуку у вигляді (3) дає характеристику здійсненого інформаційного пошуку для даного конкретного ключового слова та конкретної пошукової системи за результатами отриманої видачі, обмеженої, наприклад, 50 документами.

За кількістю релевантних документів стосовно вибраних ключових слів та інформаційно-пошукових систем результати пошуку наведені в табл. 3.

У поданій оцінці ефективності, залишається невідомим показник β. Оцінити цей показник можна лише на підставі даних про пошуковий алгоритм і саму пошукову систему, яка використовується даним інформаційним фондом.

Якщо припустити, що пошуковий алгоритм інформаційної системи переглядає всі документи інформаційного фонду або, принаймні, усі документи конкретної рубрики згідно з інформацією поданою у запиті, тоді можна прийняти значення показника $\beta = 1$. Проте докладну інформацію про характеристики цього алгоритму отримати практично неможливо.

Очевидним є той факт, що чим більший обсяг інформаційного фонду, тим більше релевантних документів буде знайдено. Однак, тут треба мати на увазі і популярність чи розвиненість даної тематики, оскільки саме вона, тобто її популярність і затребуваність визначають обсяг документів у фонді [9–10].

6 ОБГОВОРЕННЯ

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 не релевантні документи – принаймні, хоча б один. Невиконання цієї умови означає ділення на нуль. Тобто, буде невірний 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

Ефективність інформаційного пошуку в системах консолідованої інформації в сенсі побудови інтегрального показника практично не може бути визначена, оскільки крім двох показників – повноти і точності, усі інші вимагають знання кількості релевантних та нерелевантних документів у даному інформаційному фонді стосовно даної задачі. Отримати такі дані для великих за обсягами фондів є 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 подається відношенням – кількості отриманих потрібних і замовлених документів до кількості документів у даній видачі. В цьому плані, на ефективність пошуку впливає не лише наявність в інформаційному фонді потрібних документів, але й правильність побудови самого запиту згідно з вимогами даної пошукової системи.

Наведений приклад оцінювання ефективності інформаційного пошуку демонструє використання поділу знайдених і виданих документів на пертинентні, релевантні та нерелевантні. В результаті якого, оцінку ефективності можна подати як усереднену, або сумарну, на підставі 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.

ПОДЯКИ

Роботу виконано в рамках держбюджетної науково-дослідної теми «Методи та засоби консолідації баз даних в інформаційних системах електронного урядування», тематика кафедри інформаційних систем та мереж Національного університету «Львівська політехніка», 2010/2012, № держреєстрації 0110U005022.

СПИСОК ЛІТЕРАТУРИ

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). – Часть 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 / Д. Д. Козлов // Техноло-

Таблиця 3 – Оцінка ефективності інформаційно-пошукових систем за кількістю релевантних документів

Ключові слова	Google	Яndex	Мета	Rambler	Yahoo
Оцінка ефективності інформаційного пошуку	0,34	0,44	0,22	0,40	0,32
Інформаційний пошук	0,36	0,28	0,32	0,36	0,28
Модель інформаційного пошуку	0,28	0,40	0,20	0,58	0,44
Інформаційно-пошукова система	0,44	0,44	0,10	0,38	0,24
Ефективність інформаційно-пошукових систем	0,34	0,48	0,34	0,34	0,48
Усереднений показник ефективності	0,35	0,40	0,23	0,41	0,35

гический обзор [Электронный ресурс]. – Режим доступа: 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. – Том 14. – № 3. – С. 478–481.
7. Козлов М. В. Метод оценки эффективности функционирования современных информационно-поисковых систем Интернета / М. В. Козлов, В. А. Яцко [Электронный ресурс]. – Режим доступа: http://www.dialog-21.ru/dialog2006/materials/html/Kozlov.htm.
8. Лекции по введению в информатику и информационные системы. – Лекция 13. Эффективность информационных систем. [Электронный ресурс]. – Режим доступа: http://informling.narod.ru/lectures.html
9. Kirchgassner G. Introduction to Modern Time Series Analysis / G.·Kirchgassner, J. Wolters. – Berlin : Springer, 2007. – 274 p.
10. Hegger R. Practical implementation of nonlinear time series methods: The TISEAN package / R. Hegger, H. Kantz, T. Schreiber // CHAOS 9. – 1999. –P. 413–435.
11. Kuhlthau, C. C. Seeking Meaning: A Process Approach to Library and Information Services. / C. C Kuhlthauю –. Westport, CT: Libraries Unlimited. – 2004. – 342 p.

Стаття надійшла до редакції 10.02.2016.
Після доробки 15.02.2016.

Ныч Л. Я.[1], Шаховска Н. Б.[2], Каминский Р. М.[3]
[1]Ассистент, кафедра информационных систем и сетей, Национальный университет «Львовская политехника», Львов, Украина
[2]Д-р техн. наук, профессор, кафедра информационных систем и сетей, Национальный университет «Львовская политехника», Львов, Украина
[3]Д-р техн. наук, доцент, кафедра информационных систем и сетей, Национальный университет «Львовская политехника», Львов, Украина

ОЦЕНКА ЭФФЕКТИВНОСТИ ИНФОРМАЦИОННОГО ПОИСКА В СИСТЕМАХ КОНСОЛИДИРОВАННОЙ ИНФОРМАЦИИ

Для оценки эффективности информационного поиска предложено разделять найденные документы на пертинентни, релевантные и нерелевантные. Эффективность поиска определяется отношением количества пертинентних и релевантных документов в количестве нежелательных документов, а свойством поисковой системы должно быть возможность учесть соответствующие коэффициенты. Основной целью данного исследования является разработка интегрального критерия оценки эффективности информационного поиска по результатам выдачи в системах консолидированной информации. Введено понятие консолидированной информации.

Разработан метод оценки эффективности информационного поиска. Он демонстрирует использование разделения найденных и выданных документов на пертинентные, релевантные и нерелевантные. Введено интегральный показатель релевантности документа поисковому запросу, который учитывает негативную и положительную оценку. Оценку эффективности определено как суммарное значение различных компонентов. Експеримент выполнен на основании проведения информационного поиска в одном или в нескольких информационных фондах и на разных поисковых системах с одним набором ключевых слов.

Разработанный подход к построению оценки информационного поиска имеет практическое значение, поскольку полученные количественные значения локальных оценок дают основания для оптимизации набора ключевых слов, и определение наиболее подходящих информационных фондов и поисковых систем.

Ключевые слова: информационная система, информационный поиск, эффективность, пертинентность, релевантность.

Nych L. Ya.[1], Kaminskyj R. M.[2], Shakhovska N. B.[3]
[1]Assistant professor, department of information systems and networks, Lviv Polytechnic National University, Lviv, Ukraine
[2]Dr.Sc., Professor, department of information systems and networks, Lviv Polytechnic National University, Lviv, Ukraine
[3]Dr.Sc., Professor, department of information systems and networks, Lviv Polytechnic National University, Lviv, Ukraine

EFFECTIVENESS EVALUATION OF SEARCH IN INFORMATION SYSTEMS WITH CONSOLIDATED INFORMATION

To evaluate the effectiveness of information retrieval there is proposed to share the found documents on pertinent, relevant and irrelevant. Search Performance is ratio to determine the number of pertinent and relevant documents to the number of irrelevant documents and search engine properties have been submitted by the coefficient. The goal of this paper is to develop integrated criterion of evaluating the effectiveness of information retrieval on the results of the issuance of consolidated information systems. The concept of consolidated informationis given.

The method of evaluating the effectiveness of information retrieval is built. It demonstrates the usage of the division found and published documents on pertinent, relevant and irrelevant. There is given integral indicator of the relevance of the document search query that takes into account the negative and positive features. Evaluation of effectiveness presented as the total value of the different components. The experiment was performed on the basis of information search in one or several search machines and information on the various search engines for one set of keywords.

The approach to building assessment information retrieval is of practical importance because quantitative values obtained local assessments give grounds to optimize the set of keywords and determine the most appropriate information collection and search engines.

Keywords: information system, information search, efficiency, pertinence, relevance, irrelevance.

REFERENCES

1. Aheev M., Kuralenok Y., Nekrestianov Y.Ofytsyalnye metryky ROMYP 2010, *Rossyiskyi semynar po Otsenke Metodov Informatsyonnoho Poiska. Trudy ROMYP 2010. (Kazan, 15 october 2010.)* Kazan, 2010, pp. 172–187.
2. Tselykh A. N., Kotov E. M. Otsenka эffektyvnosty ynformatsyonnoho poyska, *Yzvestyia TRTU. Tematycheskyi vypusk «Upravlenye v matematycheskykh systemakh»*. Tahanroh, Yzd-vo TRTU, 2006, No. 10 (65), pp. 43–45.
3. Yakhyna E.P. Metody otsenky ynformatsyonnykh system, *V myre nauchnykh otkrytyi,* 2010, No. 3 (09), Chast 1, pp. 63–66.
4. Popov S. V. Otsenka funktsyonalnoi effektyvnosty system tekstovoho poyska na prymere poyska patentnykh dokumentov, *Patentnaia informatsyia sehodnia,* 2010, No. 1, pp. 22–25.
5. Kozlov D. D. Ynformatsyonno-poyskovye systemy v Internet: tekushchee sostoianye i puty razvytyia, *Tekhnolohycheskyi obzor.* Access mode: lvk.cs.msu.su/~ddk/ir_and_ia_review.pdf
6. Tiavkyn Y. V., Tiutiunnyk V. M. Matematycheskaia model informatsyonnoho poiska i otsenka yeffektyvnosti poyskovoi sistemy, *Vestnyk THTU,* 2008, Vol 14, № 3, pp. 478–481.
7. Kozlov M. V., Yatsko V. A. Metod otsenky effektyvnosti funktsyonyrovanyia sovremennykh informatsyonno-poyskovykh system Interneta, Access mode: http://www.dialog-21.ru/dialog2006/materials/html/Kozlov.htm.
8. Lektsyy po vvedenyiu v informatyku i informatsyonnыe systemy. Lektsyia 13. Yeffektyvnost ynformatsyonnыkh system. Access mode: http://informling.narod.ru/lectures.html
9. Kirchgassner G., Wolters J. Introduction to Modern Time Series Analysis. Springer Berlin Heidelberg, New York , 2007, 274 p.
10. Hegger R., Kantz H., Schreiber T. Practical implementation of nonlinear time series methods: The TISEAN package. *CHAOS 9*, 1999, pp. 413–435.
11. Kuhlthau, C. C. Seeking Meaning: A Process Approach to Library and Information Services, *2nd. ed. Westport*, CT, Libraries Unlimited, 2004, 342 p.