DOI:
УДК 004.912

В. В. Костенко, старший викладач
кафедри комп'ютерних наук та інженерії
програмного забезпечення Університету
митної справи та фінансів
Д. Є. Костенко, старший викладач
кафедри комп'ютерних наук та інженерії
програмного забезпечення Університету
митної справи та фінансів
Є. Д. Замотаєв, студент Університету
митної справи та фінансів
В. О. Широченко, студент Університету
митної справи та фінансів

ВИЯВЛЕННЯ ПРОБЛЕМ СТРУКТУРИ ІНФОРМАЦІЙНИХ РЕСУРСІВ ПІД ЧАС ОБРОБКИ ТА ПОШУКУ ІНФОРМАЦІЇ

Досліджено одну з важливих проблем процесу пошуку інформації. Виявлено основні відмінності між пошуком (як автоматизованою процедурою) і виділенням потрібної інформації. Визначено питання, які доцільно аналізувати під час розв'язання проблеми пошуку або виділення необхідної інформації. Створено тестовий програмний модуль, який розв'язує частину описаних у статті проблем. Розкрито основні моменти реалізації механізму виділення потрібної інформації. Надано рекомендації та пропозиції щодо розв'язання проблеми отримання необхідної для користувача інформації серед великої кількості "інформаційного шуму".

Ключові слова: *інформація; дані; інформаційний пошук; фільтрація інформації; сортування інформації.*

Исследована одна из важных проблем процесса поиска информации. Выявлены основные различия между поиском (как автоматизированной процедурой) и выделением нужной информации. Определены вопросы, которые целесообразно анализировать при решении проблемы поиска или выделения необходимой информации. Создан тестовый программный модуль, который решает часть описанных в работе проблем. Раскрыты основные моменты реализации механизма выделения нужной информации. Даны реко-

мендации и предложения по решению проблемы нахождения необходимой для пользователя информации среди большого количества "информационного шума".

Ключевые слова: *информация; данные, информационный поиск; фильтрация информации; сортировка информации.*

There are investigated important problems of the information search process. Long-term accumulation of information has its negative sides. The main problem is the glut of so-called "digital noise". "Digital noise" is information which doesn't assist the user in information searching process. "Digital noise" makes it difficult to find the right answer. At the same time there is another problem – the number of materials that are freely available is constantly growing, doubling on average once every 2 years. Sometimes – more often. In this growth there are hidden one of the main problems. Finding the right information often turns into a difficult problem that takes time and effort. There are shown the main differences between the search (as an automatical procedure) and the necessary information allocation. Identified some questions which appropriate to analyze when solving a search problem or highlighting the necessary information (site content, data storage and processing methods, site values for information support of the relevant type of activity). The main idea of this article is in optimize the filtering and retrieving information process for users. It is necessary to hide from the user information which doesn't solve his problems.

There are create a test-software module which solve some problems described in this article. Revealed the main points of implementation the selection mechanism of the necessary information. The idea is in allocate absolutely or most useful answers. It wouldn't be necessary to re-read or scroll through the whole discussion board to find out what is the ultimate correct answer or the right decision. Recommendations and suggestions are given to solve the problem of finding the necessary information for the user among a large amount of "information noise". It is necessary to create the more specialized and thematic resources with high-quality content. Using this portals, the information searching will be faster and better. It's necessary to develop comparison services. If it possible, there are necessary to use geolocation services and base searching process on user's location. It is necessary to emphasize the quality of information and divide it according to the tasks.

Key words: *information; data; information search; information filtering, information sorting*.

Постановка проблеми. Нині в глобальній мережі Інтернет існує інформація щодо будь-якого аспекту, будь-якої галузі.

Також існують розвинені (не завжди повною мірою) пошукові системи, які не тільки знаходять інформацію, пов'язану із пошуковим запитом, але й мають досить потужні фільтри та ключові слова, завдяки яким пошук проходить, наприклад, з точною відповідністю запиту і лише серед документів зазначеного формату. Але це працює, якщо проводиться пошук інформації загального спрямування.

Багаторічне накопичення інформації має і також свої негативні аспекти, головним із яких є перенасичення так званим "цифровим шумом".

Під "цифровим шумом" розуміють інформацію, яка під час пошуку жодним чином не допомагає користувачу, а дуже часто навіть заважає знайти потрібну відповідь серед усієї маси цього самого "шуму".

Розглянемо проблему на прикладі двох абстрактних сервісів для розробників програмного забезпечення: Сервіс № 1 та Сервіс № 2.

Нехай обидва сервіси мають численну аудиторію, яка безперервно ставить питання щодо виконання різноманітних завдань та зазвичай отримує робочі рішення від інших користувачів.

Таким чином, коли у нового користувача виникає питання, яке вже було обговорено, він з великою вірогідністю натрапить на відповідь. Сервіс № 2 має структуру звичайного "діалогу", всі коментарі додаються у нижню частину сайту лише з розподілом за сторінками. Простіше кажучи, довге "полотно" повідомлень (розташованих за датою додавання).

Але в усіх сервісів, які мають таку ж структуру, як і Сервіс № 2, є велика проблема: після того, як користувач знайшов своє питання, яке хтось уже ставив, йому доводиться прокручувати сотні відповідей, у яких інші користувачі пишуть: "Маю таку ж проблему …" або дають неправильні відповіді. А дочитавши до кінця цієї гілки відповідей, майже завжди бачить повідомлення: "Робоче рішення знаходиться за таким посиланням ..." з переходом на іншу гілку відповідей. Не виключено, що з десяток наступних гілок будуть з такими ж посиланнями в кінці. Отже, процес стає нескінченним.

При цьому зазвичай у разі переходу в іншу гілку дещо змінюється і проблема, розв'язання якої шукають. Таким чином, виявляється, що розробник, якому необхідно отримати відповідь, втрачає великий обсяг робочого часу, даремно витрачає свої сили на перечитування сотень непотрібних відповідей. Це і є "інформаційний шум".

На відміну від Сервісу № 2, Сервіс № 1 менш популярний, але він частково позбавлений цієї проблеми. Реалізовано це завдяки винесенню робочого рішення на самий верх сторінки одразу ж під питанням. Правильною вважається та відповідь, яку або позначив такою автор самого питання, або яка набрала максимальну кількість позначок "правильності" від усіх користувачів. Виникає питання щодо "професійної адекватності" тих, хто вважає відповідь правильною.

Та факт залишається фактом: відкривши сторінку з питанням, користувач одразу ж бачить відповідь, яка є правильною на 100 %, або ту, яка допомогла найбільшій кількості користувачів.

Роблячи висновок з вищезазначеної проблеми, нескладно дійти до розв'язання проблеми для всіх сервісів, що працюють за типом Сервіс № 2. Беручи до уваги надлишкову кількість інформації, вважаємо обов'язковим додавання алгоритму "рейтингів" відповідей до кожного подібного сервісу.

Аналіз останніх досліджень і публікацій. Інформаційний пошук – процес пошуку неструктурованої документальної інформації, що задовольняє інформаційні потреби [1].

Головне завдання інформаційного пошуку – допомогти користувачеві задовольнити його інформаційну потребу.

Здавалося б, в епоху інформаційних технологій, Інтернету і пошукових систем стало набагато простіше знайти потрібну інформацію. Але чи дійсно це так?

Чим більший обсяг інформації, тим більша можливість використання корисної її частини для прийняття рішень [2].

Сучасні технічні засоби (комп'ютери, мережа Інтернет) значно спрощують доступ до величезної кількості матеріалів, що перебувають у вільному доступі. Причому ця величезна кількість постійно зростає, подвоюючись у середньому раз на 2 роки. Саме в такому зростанні й приховано "підводне каміння": пошук потрібної інформації часто перетворюється на досить непросту проблему [2].

У роботі з інформаційними потоками діє закон Парето. Згідно зі статистичними дослідженнями, якщо інформація збільшується вдвічі, її корисність становить не більше 20 %, а 80 %, які залишилися, не мають корисного характеру [2; 3].

Не зайво нагадати, що однією з ознак інформаційного суспільства є розвинута інфраструктура, що забезпечує створення достатньої кількості інформаційних ресурсів [4].

Інформація стає предметом масового використання. Інформаційне суспільство забезпечує індивіду доступ до будь-якого джерела інформації, що гарантується законом і технічними можливостями [4].

Закладені в працях К. Муерса і Дж. Солтона фундаментальні основи пошуку інформації актуальні й донині. Однак тут є невеликий нюанс у використанні термінології цих праць [5].

Слід підкреслити відмінність між пошуком як автоматизованою процедурою і виділенням потрібної інформації в знайдених документах [5].

Суть відмінностей полягає в такому [5]:

1) виділення інформації – це діяльність людини, яка використовує пошукову машину. Вона є інтерактивною, ітераційною і пов'язана з іншими видами інтелектуальної діяльності людини;

2) користувач шукає не документи як такі, а інформацію, що міститься в них для яких-небудь власних цілей (навчання, прийняття рішень тощо.);

3) користувач потребує доступу до різних джерел даних, щоб отримати всеосяжне уявлення про об'єкт пошуку;

4) якими б досконалими не були апаратне і програмне забезпечення, що використовуються людиною, вони залишаються інструментами, а інтелект – це атрибут користувача.

Однак інтернет-сайт являє собою не просто набір документів, а досить складну систему, вивчаючи яку доцільно аналізувати такі питання [5]:

а) інформаційне наповнення сайту;

б) методи зберігання й обробки даних (що розглядаються разом з програмними засобами);

в) значення сайту для інформаційного забезпечення відповідного виду діяльності.

Ці питання тісно взаємопов'язані.

Згадаємо, як проходить процес пошуку. Вводиться запит у пошуковий рядок і в результаті ми отримуємо занадто багато відповідей (інколи – тисячі або десятки тисяч, а інколи більше). Тут можна поставити такі досить актуальні питання: що зі знайденого було корисним? як швидко була знайдена відповідь на поставлене питання?

Формування результатів пошуку (наприклад, в Інтернеті) проходить у кілька етапів.

1. Спочатку потрібно визначити, які є сторінки. Оскільки їх офіційного реєстру не існує, доводиться постійно шукати нові сторінки й додавати їх до списку вже відомих. Цей процес називається скануванням.

2. Після виявлення сторінки потрібно визначити, які темі присвячено її зміст. Цей процес називається індексацією. Він полягає в тому, що відбувається аналіз контенту сторінки і проводиться систематизація знайдених на ній зображень і вбудованих відео. Отримана інформація зберігається у величезній базі даних, яка розміщена на багатьох комп'ютерах.

3. Коли користувач уводить запит, відбувається пошук найбільш відповідних результатів за низкою факторів. До таких факторів належать розташування, мова, тип пристрою користувача тощо.

Ніби то ідеальний механізм. Але тут є свої проблеми, що стосуються пошуку інформації.

1. Накопичення "порожньої" і застарілої інформації.

2. Ресурси та сервіси з невеликою кількістю інформації.

3. "Чорний" копірайтинг.

Мета статті – вдосконалення процесу фільтрації та пошуку інформації для користувачів. Тобто надання максимально корисної для користувача інформації.

Виклад основного матеріалу. Зазвичай на форумах, де обговорюється та чи інша проблема, доводиться читати всі повідомлення, щоб знайти хоча б невеличку частину потрібної інформації. А коли потрібне посилання знайдено, виявляється, що ресурс уже закритий або не працює. Тут потрібно було б ввести більш чітку модерацію або адміністраторами, або користувачами. Або зробити її автоматичною. Адміністратори відбиратимуть теми, які протягом тривалого часу перебувають без відповіді і можуть їх закривати. На великих порталах це буде неефективно і витратно, тому що за це треба платити, а ентузіазм зазвичай не оплачується. Користувачі за бажанням могли б відправити свою тему у розділ актуальних або видалити її, якщо на цьому порталі вже була дана відповідь на подібне питання, але тут можна зіштовхнутися із проблемою під назвою "війна за місце" [6].

До подібних сайтів можна зарахувати сайти для дизайнерів, ілюстраторів, фотографів або, наприклад, сервіси, що містять корисні функції, та не мають на своїх сторінках хороших текстів, щоб їх знайшли пошукові роботи. У більшості випадків, такі сайти знаходиш випадково й заносиш їх у закладки, щоб "потім, коли-небудь" ними скористатися. Із розв'язанням цієї проблеми нам має допомогти або пошукова система, або якийсь ресурс, який збиратиме такі сайти в один список. В обох випадках доведеться вводити теги для більш зручного пошуку. Найцікавіше починається з цього моменту: кількість і рівень SEO-оптимізаторів збільшується щороку, разом з ними зростає кількість сайтів для пошукових роботів з метою заробітку на рекламі.

Завдання полягає в тому, щоб звертаючись, скажімо, до теми з заголовком "Яка мінімальна версія підтримки N-го софту", користувач бачив напис "Версія 7 і вище" замість п'ятисторінкового чату, в самому кінці якого розміщена потрібна відповідь.

При цьому немає необхідності знищувати всю інформацію з коментарями. Серед усіх коментарів для кого-небудь іншого знайдеться і своя потрібна інформація.

Ідея полягає у так званому виділенні абсолютно або максимально корисної відповіді, щоб тисячам нових відвідувачів не треба було б перечитувати-перегортати всю стрічку обговорення, щоб дізнатися, що ж у підсумку є правильною відповіддю/правильним рішенням?.

Реалізація цього механізму така.

Кожному з коментарів в обговоренні присвоюється особливе число – рейтинг (рис. 1, табл. 1). Спочатку воно дорівнює нулю.

id	message	rating	owner
1.	Як цьому запобігти?	0	11
2.	В мене не виходить	−5	43
3.	Треба оновити ядро	7	76
4.	Мені це допомогло!	−2	54

Рис. 1. Фрагмент таблиці з бази коментарів

Таблиця 1

Властивості коментарів, які зберігаються в базі для можливості розподілу їх за рейтингом

Стовпець	Пояснення
id	Унікальний ключ ідентифікації коментаря
message	Текст коментаря
rating	За цим полем відбувається сортування
owner	Ключ власника коментаря: він необхідний для того, щоб коментарі могли видаляти ті, хто їх створив

Будь-який користувач має право один раз або підвищити рейтинг будь-якого з коментарів на 1 одиницю значення, або знизити (за принципом “лайків” у соціальних мережах). При цьому обов’язково проводиться перевірка коментаря на “порожність”.

$$message = \begin{cases} null, \textit{ то вивести помилку} \\ not\ null, \textit{ то додати коментар.} \end{cases}$$

Через деякий час у кожного коментаря обговорення сформується певний рейтинг. В одних коментарів він буде негативним, що свідчить про його непотрібність, оскільки максимальна кількість користувачів забажала відняти бал рейтингу, натиснувши кнопку “непотрібний коментар”. Процедура обробки зниження рейтингу аналогічна до процедури збільшення.

В інших записів, навпаки, рейтинг буде високим. На цьому етапі стає зрозуміло, що коментарі з найвищим рейтингом найбажаніші. Отже, необхідно якомога швидше показати цю інформацію всім наступним користувачам. Таким чином, згодом під час виведення діалогу обговорення повідомлення з найвищим рейтингом просто виводяться у верхню частину обговорення.

Тобто змінна сортування дорівнює SQL-запиту, в якому зазначено: “вибрати все з таблиці, відсортувати рейтинг за зменшенням”.

На рис. 2 наведено фрагмент програмного коду тестового модуля.

```
// insert a quote if submit button is clicked
if (isset($_POST['submit']))                                    // якщо натиснута кнопка submit
{
        if (empty($_POST['task']))                              // якщо поле порожнє
        {
                $errors = "Порожній коментар  .";               // якщо порожнє, вивести помилку
        }
        else                                                    // в іншому випадку
        {
                $task = $_POST['task'];
                $query = "INSERT INTO `data` (`message`) VALUES ('$task')";// змінна query = цьому SQL-запиту
                mysqli_query($db, $query);                      // звертаємося до бази db за допомогою змінної query
                header('location: index.php');                  // відкрити index.php
        }
}
```

Рис. 2. Фрагмент програмного коду тестового модуля

На рис. 3 зображено діаграму варіантів використання, яка відображає дії користувачів

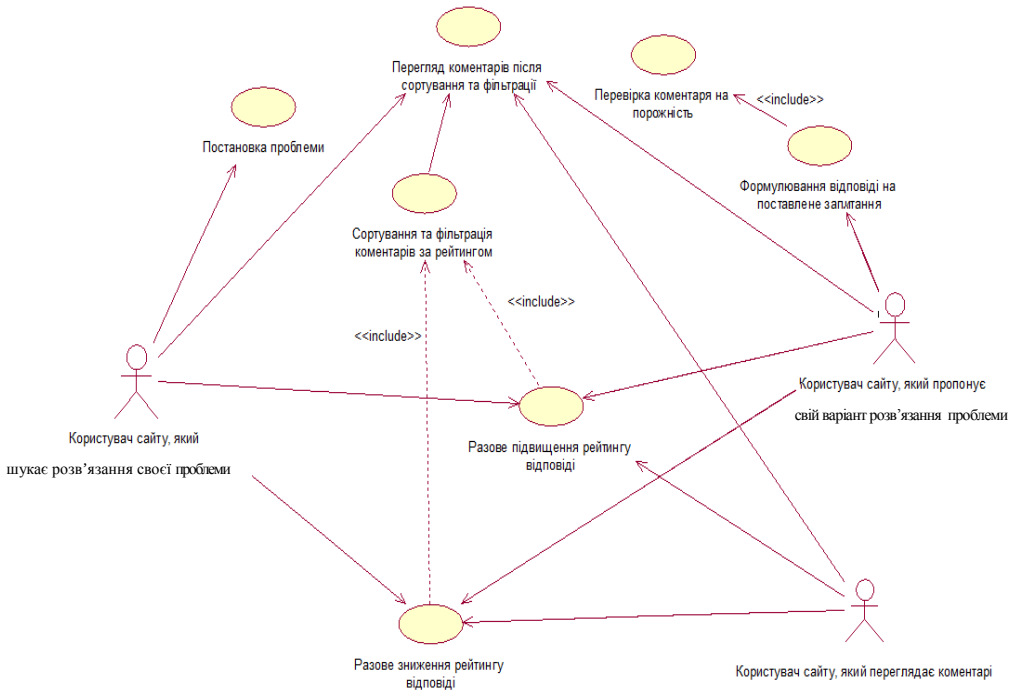

Рис. 3. Діаграма варіантів використання

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі. Інтернет – це “звалище даних”, у якому легко знайти розважальний матеріал, творчість різних людей тощо. Знайти це інколи легше, ніж потрібну робочу інформацію. Особливо, якщо вона вузькоспрямованого характеру. Ідеальним було б проводити повну модерацію у всесвітній мережі, але ресурсів зараз стільки, що таке завдання буде занадто проблематичним.

У статті досліджено важливу проблему пошуку інформації. Створено тестовий програмний модуль, який розв’язує проблему фільтрації й пошуку інформації для користувачів. Пропонується продовжити проект і вдосконалити цей модуль для простішого інтегрування в будь-які системи серверів.

Не зайве нагадати, що потрібно створювати більше спеціалізованих і тематичних ресурсів з якісним контентом. Чим більше таких порталів, тим швидше і якісніше триватиме пошук інформації.

Необхідно розвивати так звані сервіси порівняння. Це сервіси, що надають зведену таблицю за певною проблемою, ґрунтуючись на результатах, отриманих з багатьох сайтів і сервісів. Як наприклад, конкретні сервіси порівняння цін на товари. Користувач пише назву товару і такий сервіс видає порівняльну таблицю цін і відгуків з усіх можливих магазинів. ІТ-сфера значною мірою потребує подібного сервісу. Вводячи проблему, користувач отримував би звіт готових рішень за всіма ІТ-сайтами без необхідності заходити на кожен із сайтів і перечитувати масу інформації. Але тут з’являється ще одна необхідність: за можливості потрібно використовувати сервіси геолокації і засновувати пошукові видачі на основі місця розташування користувача. Це робиться, на жаль, рідко.

Зараз необхідно робити акцен на якості інформації і розподіляти її залежно від завдань. Якщо користувач налаштований на роботу і йому необхідно максимально швидко знайти потрібний контент, то слід прибирати з пошуку розважальні ресурси. А коли користувач хоче відпочити, то навпаки.

Список використаних джерел:

1. *Маннинг К. Д., Рагхаван П., Шютце Х.* Введение в информационный поиск / пер. с англ. Д. А. Клюшин. М.: Вильямс, 2011. 528 с.

2. *Поташова А. В.* Проблеми пошуку інформації в глобальній мережі Інтернет // Науковий огляд. № 5 (48). 2018. С. 130–139.

3. *Лук’янчикова Ю. В., Попова Ю. М.* Інформаційні потоки в системі управління організацією. URL: https://conferdsum.ucoz.ua/_fr/0/7120405.pdf

4. *Пожуєв В. І.* Глобальне інформаційне суспільство як новий соціальний та економічний феномен XXI століття // Гуманітарний вісник Запорізької державної інженерної академії. 2013. Вип. 52. С. 5–14.

5. *Шокин Ю. И., Федотов А. М., Барахнин В. Б.* Проблемы поиска информации. Новосибирск: Наука, 2010. – 220 с.

6. *Карпенко О.* Эпическая битва за пиксели: как участники Reddit воевали за место на графическом полотне. URL: https://ain.ua/2017/04/10/bitva-za-pikseli

References:

1. Manning C.D., Raghavan P., Schütze H. (2011), *Vvedenie v informazionnyi poisk* [Introduction to Information Retrieval], translated from english D.A.Klushin, Williams, Moscow, Russia, 528p. [Rus.]

2. Potashova A.V. (2018), *"Problemy poshuku informacii v globalniy merezhi Internet"* [Problems of searching information in the Internet], journal *Naukovyi Oglyad* [The Scientific Screening], vol. 5(48), pp. 130-139. [Ukr.]

3. Lukyanchikova Y.V., Popova Y.M., *Informaciyni potoki v sisteme upravlinnya organizaciey* [The information flows in the management system organization], available at: https://conferdsum.ucoz.ua/_fr/0/7120405.pdf. [Ukr.]

4. Poghuev V.I. (2013), *"Globalne informaciyne suspilstvo yak novyi socialnyi ta ekonomichnyi fenomen 21 stolittya"* [Global information society as new social and economic phenomenon of the 21st century], journal *Gumanitarnyi visnyk Zaporizhskoi derghavnoi ingenernoi akademii* [The Humanities Bulletin of Zaporizhzhe State Engineering Academy], vol. 52, pp. 5-14. [Ukr.]

5. Shokin Y.I., Fedotov A.M., Barahnin V.B. (2010), *Problemy poiska informacii* [The problems of information searching], Nauka, Novosibirsk, Russia, 220 p. [Rus.]

6. Karpenko O. (2017), *Epicheskaya bitva za pikseli: kak uchastniki Reddit voevali za mesto na graficheskom poligone* [Epic battle for Pixels: how Reddit participants fought for a place on a graphic canvas], available at: https://ain.ua/2017/04/10/bitva-za-pikseli/ [Rus.]