Рамський Ю.С., Рєзіна О.В.
НПУ ім. М.П. Драгоманова
Кіровоградський державний педагогічний університет ім. В. Винниченка

Формування інформаційно-пошукових та дослідницьких умінь майбутніх учителів інформатики та математики

Важливою складовою інформатичної компетентності особистості є інформаційно-пошукові та дослідницькі уміння.

Науково-дослідницька діяльність студентів є одним із найважливіших засобів підвищення якості підготовки фахівців з вищою освітою. З умінням виконувати дослідницьку діяльність пов'язується: 1) здатність до постановки різноманітних дослідницьких цілей; 2) спроможність до виконання розумових і практичних дій, які підпорядковуються логіці наукового дослідження; 3) здатність до пізнавального інформаційного пошуку і аналітико-синтетичного опрацювання одержаних результатів; 4) готовність до одержання різноманітних, у тому числі несподіваних, непрогнозованих результатів дослідження, з'ясування їх сутності, узагальнення і використання для подальшого пізнання.

Успішність науково-дослідницької діяльності значною мірою залежить від здатності особистості здійснювати інформаційний пошук. Інформаційно-пошукове уміння можна визначити як складний комплекс розумових і практичних дій, який передбачає: 1) усвідомлення інформаційної потреби і формулювання її в інформаційному запиті; 2) визначення сукупності інформаційних масивів, у яких відбуватиметься пошук; 3) планування і добір засобів виконання інформаційно-пошукової діяльності; 4) аналіз результатів пошуку.

Сьогодні інформаційні ресурси набули *глобального* масштабу і актуальним стало уміння знаходити релевантні дані *швидко*. Тому формування інформаційно-пошукових та дослідницьких умінь студентів може відбуватися на основі використання інформаційних ресурсів і пошукових засобів мережі Інтернет [1].

У вересні 2011 року в Кіровоградському державному педагогічному університеті імені В.Винниченка проведено опитування, завданнями якого були:

- визначити рівень доступності і популярності Інтернету серед студентів у порівнянні з іншими джерелами даних та відомостей;
- виявити значущість інформаційно-пошукової діяльності у порівнянні з іншими видами активності студентів в Інтернеті;
- з'ясувати рівень ефективності інформаційно-пошукової діяльності;
- виявити найбільш використовувані пошукові системи.

В опитуванні взяли участь 101 студент II-IV курсів фізико-математичного факультету спеціальностей «Математика та основи інформатики», «Фізика та основи інформатики».

За даними опитування переважна більшість (63%) студентів мають постійний і зручний доступ до Інтернету. На запитання «Яку з двох технологій – телебачення чи Інтернет Ви б обрали, якщо б мали можливість користуватися тільки однією з них?» 96% опитаних відповіли «Інтернет». При написанні реферату (курсової або кваліфікаційної роботи) тільки 6% респондентів постійно звертаються до послуг традиційної бібліотеки, а до ресурсів Інтернету постійно звертається 75% опитаних (рис.1).

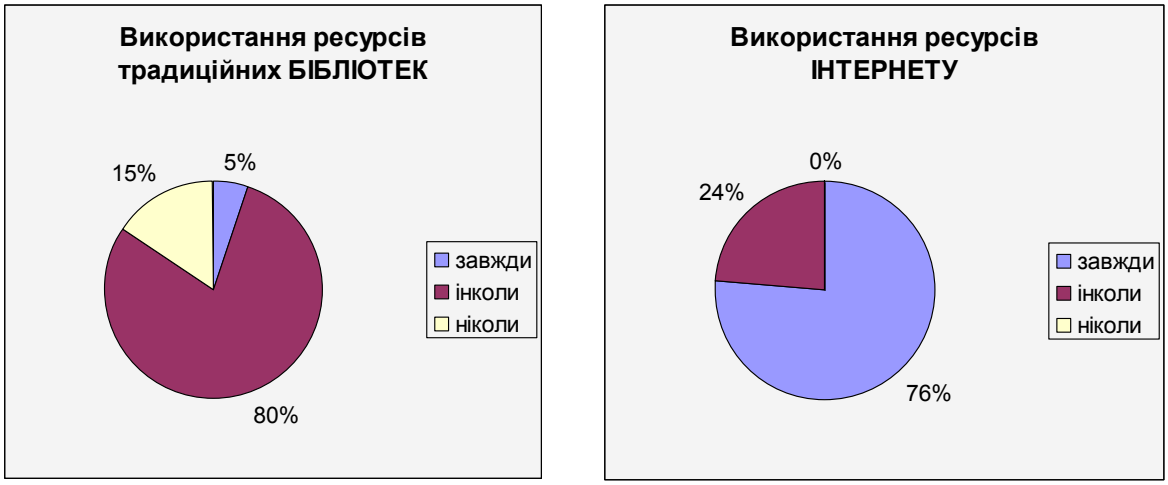

Рис. 1.

Досить активно студенти використовують пошукові системи. Значний відсоток (84%) опитаних під час кожного сеансу роботи в Інтернеті здійснює інформаційно-пошукову діяльність. Результативність пошуку в Інтернеті студенти оцінили «50 на 50»: половина опитаних завжди знаходять необхідні відомості, а інша половина – іноді стикаються з труднощами і не досягають бажаного результату. Якщо пошук не дав результату, то переважна більшість опитаних намагається переформулювати запит (67%) або перейти до іншої пошукової системи (25%), і тільки 8% респондентів використовують мову запитів.

За популярністю серед пошукових систем на першому місці – Google (96% респондентів), на другому – Яндекс (78%), на третьому – <META> (24%) (була надана можливість назвати три найчастіше використовувані пошукові системи).

Жоден з опитуваних не має уявлення про архітектуру пошукової системи і не знає, на основі яких алгоритмів працюють ці системи. Хоча володіння цими знаннями дає змогу кваліфікованіше формулювати запити до пошукової системи й аналізувати результати пошуку.

Близькі результати такого самого опитування одержано в НПУ імені М.П. Драгоманова (тільки два студенти IV МІ курсу мали початкові знання про архітектуру пошукової системи і алгоритми, на основі яких вони працюють.

Опитування показало, що:

- інформаційні ресурси і пошукові системи Інтернету є популярними серед студентів;
- ресурси Інтернету активно використовуються студентами в інформаційно-пошуковій та дослідницькій діяльності;
- необхідно посилити теоретичну складову дисциплін, що передбачають вивчення принципів будови та функціонування мережі Інтернет (це може бути окрема дисципліна, наприклад «Основи Інтернету» чи певний спецкурс).

Для вирішення цього питання до змісту навчання доцільно включити такі питання:

1. Поняття індексної пошукової системи та її характеристики.
2. Складові індексних пошукових систем.
3. Призначення та принципи побудови інвертованого файлу.
4. Архітектура індексних пошукових систем.

Індексна пошукова система (ІПС) – це програмно-апаратний комплекс, призначений для здійснення пошуку в мережі Інтернет. ІПС «реагує» на запит користувача, що задається у вигляді текстової фрази, наданням списку результатів пошуку (списком посилань на web-документи).

Для того, щоб передати сутність пошукової системи, автор роботи [2] перефразовує назву відомої книги Вірта і стверджує, що

Алгоритми + структури даних = пошукова система.

Для оцінювання якості пошуку використовуються два параметри

– Повнота (recall) – частина знайдених релевантних документів у загальному числі релевантних документів колекції.

– Точність (precition) – частина релевантних документів у відповіді пошукової системи (рис. 2).

Повноту та точність можна виразити формулами:

$$\text{Повнота} = \frac{\text{Кількість знайдених релевантних документів}}{\text{Загальна кількість релевантних документів}}.$$

$$\text{Точність} = \frac{\text{Кількість знайдених релевантних документів}}{\text{Кількість знайдених документів}}.$$

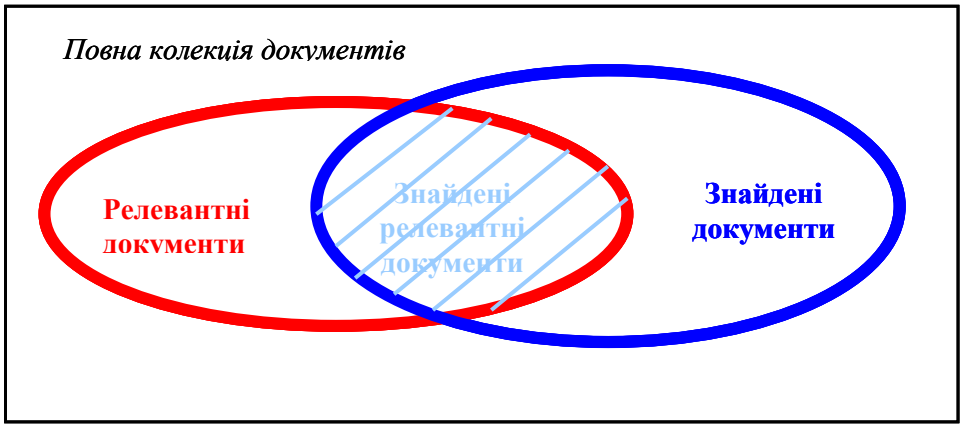

Рис. 2

Крім повноти та точності, характеристиками індексних пошукових систем є:

- розмір та якість (відсутність дублікатів документів) індексної бази даних;
- швидкість пошуку;
- актуальність (відсутність посилань на документи, що не існують);
- наочність подання результатів (інформативність сніпетів; сніпет – коротке текстове повідомлення про сайт, яке з'являється в результаті пошуку відразу під адресою).

Етапи роботи індексних пошукових систем

- Сканування WWW-простору
- Індексація просканованих документів
- Опрацювання запиту користувача

Можна виділити дві головні складові індексних пошукових систем [3]:

- Індексна база даних/індекс (Indexing).
- Система опрацювання запитів і виведення результатів (Query Processing).

Перед тим, як розглянути зі студентами об'єкти індексної бази даних, необхідно зазначити, що запити до пошукових систем формулюються природними мовами, тому створення і функціонування цих систем вимагає вирішення низки лінгвістичних проблем:

- автоматичне визначення мови документа;
- парсінг (parsing) – синтаксичний аналіз тексту, який автоматично проводиться за спеціальною програмою (парсером);
- токенізація (tokenization) – графематичний аналіз тексту: визначення слів, меж речень;
- стемінг (stemming) – процес визначення основи слова;
- лематизація (lemmatization) – зведення словоформ до словарної форми, тобто леми;
- виключення неінформативних, так званих, стоп-слів;
- позбавлення омонімії [2].

З термінами парсінг і токенізація студенти зустрічалися при вивченні мов програмування. Далі можна перейти до розгляду об'єктів індексної бази даних.

Об'єктами індексної бази даних є:

- Індексатор (IndexBuilder) – призначений для управління процесом індексації.
- Синтаксичний аналізатор (Parser) – призначений для опрацювання вхідних документів, створення списку об'єктів документа.
- Лексичний аналізатор (Tokenizer) – призначений для розпізнавання та виокремлення лексем (токенів) із вхідної послідовності символів.
- Пошуковий образ документа – містить лексичні одиниці тексту документа, що відображають його зміст, разом із частотою їх входження.
- Інвертований файл/індекс (inverted index) – призначений для ефективного зберігання кожного терміну колекції документів: 1) список документів, що містять термін; 2) частоту його згадування в документі tf (term frequency).

Використання інвертованого файлу збільшує швидкість та ефективність пошуку. Відсутність такої структури передбачає здійснення прямого пошуку, тобто послідовного перегляду документів, що містять слова запиту.

Можна поставити перед студентами проблемне запитання: «Чому пошук файлу на персональному комп'ютері, жорсткий диск якого містить кілька тисяч файлів, триває кілька секунд (іноді і більше хвилини), а пошук у вмісті мільярдів документів в Інтернеті – менше секунди?» Відповідь на це запитання є такою. Для виконання операції «Пошук» у Windows використовується алгоритм прямого пошуку. Чи можна використати такий алгоритм для пошуку в Інтернеті? Наявність комп'ютера, за допомогою якого опрацьовується один мільйон документів за секунду, надасть змогу здійснювати пошук у колекції із одного мільярда документів протягом 1000 секунд, тобто 16 хвилин. Обсяг Web-ресурсів Інтернету давно сягнув за межі 1 мільярда документів. Жоден користувач не буде чекати результатів пошуку 16 хвилин. Таким чином, прямий пошук є практично неприйнятним.

Швидким та ефективним пошук стає завдяки тому, що у процесі опрацювання запиту за допомогою пошукової системи здійснюється звертання до завчасно підготовлених (у режимі off-line) інвертованих/індексних файлів.

Прикладами завчасно підготовлених для пошуку даних є:

- телефонний довідник;
- бібліотечний каталог;
- предметний покажчик у книзі.

Інвертований файл – це впорядкований за алфавітом список слів, для кожного з яких вказані всі документи, в яких це слово зустрічається. За алгоритмами опрацювання інвертованих файлів здійснюється відшукання слів запиту та виведення списку відповідних документів.

Інвертований файл можна порівняти із конкордансом – переліком усіх слів деякого тексту, розташованих в алфавітному порядку, з мінімальним контекстом (онлайн-конкорданс роману Івана Франка «Перехресні стежки»).

Доцільно розглянути процес побудови інвертованого файлу на прикладі двох документів.

Документ 1

```
<html>
<head>
<title>Електронні ресурси сучасної регіональної бібліотеки</title>
</head>
<body>
<p>
```

Інформаційно-пошуковий тезаурус (далі – Тезаурус)- це спеціально створений словник, що базується на штучній мові, призначеній для відображення змісту документів і запитів користувачів з метою їх подальшого пошуку в автоматизованих інформаційних системах (АІС) у тому числі – автоматизованих інформаційно-бібліотечних системах (АІБС). Тезаурус – це інструмент, завдяки якому можна сформувати уніфікований пошуковий образ документу (ПОД) і пошуковий образ запиту (ПОЗ).</p>

```
</body>
</html>
```

Документ 2

```
<html>
<head>
<title>Основи інформаційних систем</title>
</head>
<body>
<p>
```

Для виконання якісного пошуку відомостей недостатньо провести лексикографічний контроль та побудувати список дескрипторів і ключових слів. Необхідно створити спеціальний нормативний словник. Крім внутрішніх текстових взаємозв'язків такий словник — його називають тезаурусом — має містити позатекстові зв'язки.

Тезаурус у перекладі з грецької означає скарб, багатство, запас. Отже, множина дескрипторів і ключових слів з їх відношеннями утворюють тезаурус.</p>

```
</body>
</html>
```

Пошуковий образ документа містить перелік слів (за винятком стоп-слів) і частоту їх згадування в документі.

Пошуковий образ Документа 1

інформаційний	3	зміст	1	бібліотечний	1
пошуковий	3	документ	2	аібс	1
тезаурус	3	запит	2	це	2
далі	1	користувач	1	інструмент	1
спеціально	1	мета	1	завдяки	1
створений	1	їхній	1	який	1
словник	1	подальший	1	можна	1
що	1	пошук	1	сформувати	1
базується	1	автоматизований	2	уніфікований	1
штучний	1	система	2	образ	2
мова	1	аіс	1	запит	1
призначений	1	той	1	поз	1
відображення	1	число	1		

Пошуковий образ Документа 2

виконання	1	створити	1	позатекстовий	1
якісний	1	спеціальний	1	зв’язок	1
пошук	1	нормативний	1	переклад	1
відомості	1	словник	2	грецький	1
недостатньо	1	крім	1	означати	1
провести	1	внутрішній	1	скарб	1
лексикографічний	1	текстовий	1	багатство	1
контроль	1	взаємозв’язок	1	запас	1
побудувати	1	такий	1	отже	1
список	1	він	1	множина	1
дескриптор	2	називається	1	їхній	1
ключовий	2	тезаурус	3	відношення	1
слово	2	має	1	утворювати	1
необхідно	1	містить	1		

Виключені зі списку стоп-слова зберігаються в окремому файлі.

На основі пошукових образів документів можна створити інвертований файл. Його складовими є:

- список термінів;
- список документів, в яких ці терміни зустрічається з указанням частоти їх згадування (рис. 3).

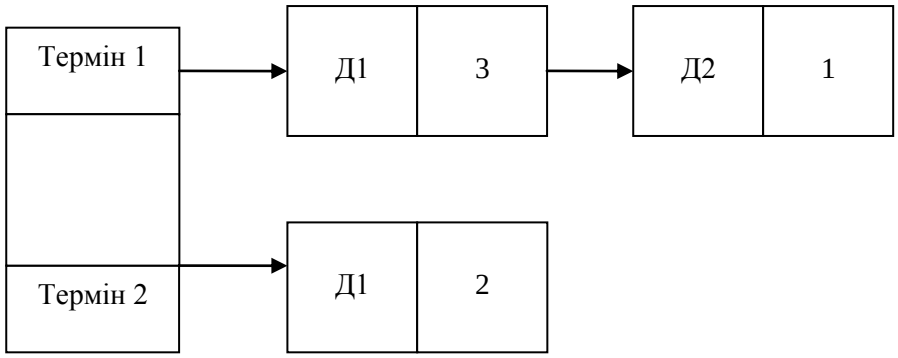

Рис. 3.

Для наведеного прикладу інвертований файл набуде вигляду:

автоматизований:	(Д1,2)
аібс:	(Д1,1)
аіс:	(Д1,1)
багатство:	(Д2,1)
базується:	(Д1,1)
бібліотечний:	(Д1,1)
…	
пошук:	(Д1,1) (Д2,1)
…	
словник:	(Д1,1) (Д2,2)
…	
тезаурус:	(Д1,3) (Д2,3)
…	
якісний:	(Д2,1)

Для великих різноманітних колекцій документів неможливо гарантувати збереження індексної бази даних в пам’яті комп’ютера. У таких випадках, якщо індекс перевищує виділену пам’ять, використовують бінарні дерева (B-Trees, B+ Tree, B* Tree). Для великих баз даних бінарне дерево – це метод доступу первинного ключа у вторинній пам’яті. Для пошуку ключа відбувається спускання деревом з вибором відповідної гілки на кожному кроці. Число дискових доступів дорівнює висоті дерева. Бінарні дерева визначають лексикографічний порядок даних, значення ключа (слова запиту) використовується для визначення напряму пошуку (рис. 4).

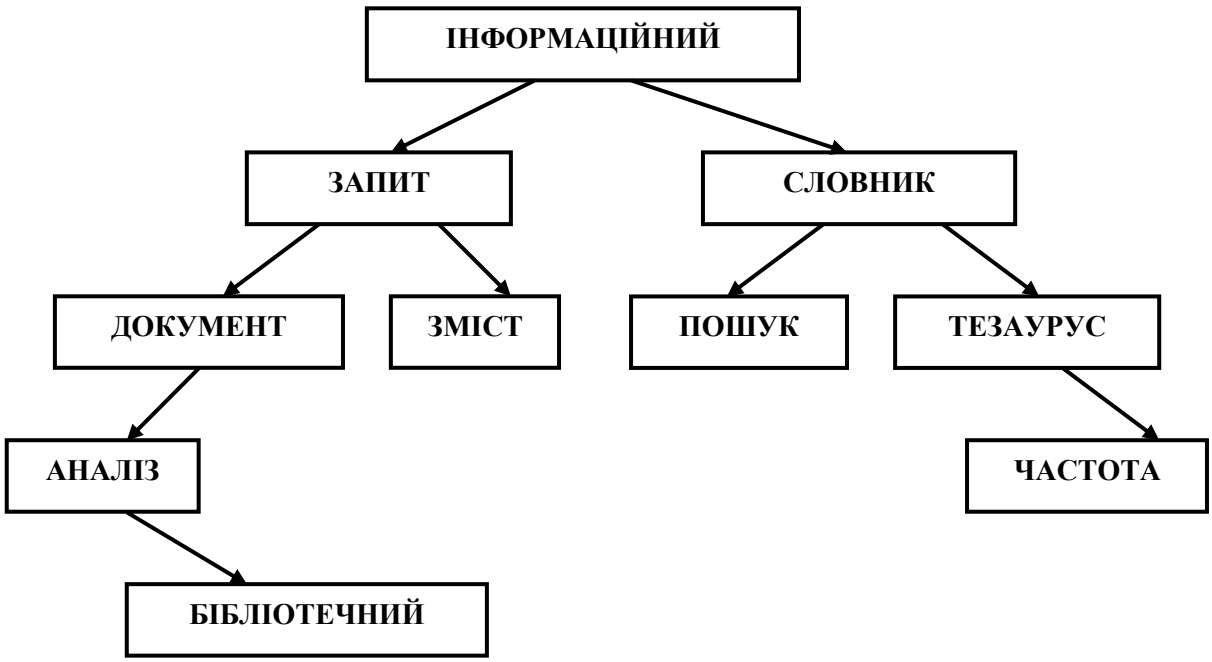

Рис. 4.

Переваги використання бінарних дерев:

- придатність до розширення (збільшення обсягів);
- підтримка збалансованої структури при здійсненні оновлень;
- O(log n) звернень до списку, в якому відбувається пошук, де n – число елементів дерева (множина термінів колекції документів) [4].

Розглянута пошукова структура вважається основною в класичній теорії інформаційного пошуку.

Але інвертований файл можна побудувати й за іншим принципом. Для кожного слова повинні бути зазначені всі позиції, в яких це слово зустрічається в документі. За пошуковим алгоритмом здійснюється відшукання необхідного слова і завантаження у пам’ять комп’ютера розгорнутого списку позицій [2].

При такому підході проблема в тому, щоб найкомпактніше зберігати цей розгорнутий список позицій (адже мова іде про мільярди документів). Для розв’язування цієї проблеми позиції кожного слова впорядковуються за зростанням адрес і для кожної позиції зберігається не повна адреса, а різниця від попереднього. Для терміну «тезаурус» Документа 1 такий список набуде вигляду:

тезаурус: [Д1],[+2],[+1],[+29]

Додатково список позицій може бути стиснутий за допомогою одного з алгоритмів стиснення даних.

Для підвищення ефективності роботи пошукових систем розв’язуються такі задачі:

- класифікація, маршрутизація, фільтрація, анотування документів;
- кластеризація результатів, організація зворотного зв’язку з користувачем, створення мови запитів та пошукового інтерфейсу [2].

Єдина видима для користувача частина пошукової системи – інтерфейс. Основні критерії, за якими можна оцінювати інтерфейс пошукової системи:

- наявність динамічної підказки при введенні запиту;
- наявність підказки альтернативних формулювань запиту;
- швидкість пошуку;
- інформативність сніпетів;
- можливість сортування результатів пошуку за різними критеріями;
- наявність посилань на файли форматів .pdf, .doc, .ppt тощо;
- перевірка орфографії;
- непереобтяженість інтерфейсу елементами, що відволікають увагу (реклама, новини тощо) [5].

Серед завдань до лабораторних робіт з цієї теми можна запропонувати студентам порівняти інтерфейси трьох пошукових систем за зазначеними критеріями.

Важливим є розгляд питання архітектури пошукових систем. Архітектурно пошукові системи – це складні багатокомп’ютерні комплекси. Для того, щоб за секунду опрацювати сотню запитів, індекс ділять на частини і розміщають на тисячі комп’ютерів. Зараз використовується термін «пошуковий

кластер». Наприклад, на сервері компанії Яндекс приймається запит і розсилається тисячам інших комп'ютерів. На кожному з них проводиться пошук у своєму сегменті. Далі знайдені дані передаються на сервер, об'єднуються і оформлюються у список результатів пошуку. Сервер, через який розсилається запит, називається «Метапошук». Сервери, які містять складові індекса, називаються «базовим пошуком». Тобто, пошуковий кластер – це тисячі комп'ютерів, що розташовані в декількох дата-центрах і за допомогою яких забезпечується опрацювання мільйонів запитів на добу [6].

Цікавим і світоглядним є той факт, що «один відсоток продуктивності системи (наприклад, невдало написаний оператор у циклі) для десятитисячної комп'ютерної системи вартий приблизно ста комп'ютерів. Тому програмний код, що призначений для пошуку та ранжирування результатів, ретельно перевіряється, використання ресурсів (кожного байта пам'яті, кожного звернення до диску) максимально оптимізується» [2].

Включення розглянутих питань до змісту навчання сприяє фундаменталізації навчання дисциплін інформатичного циклу.

Важливим є добір інформаційно-пошукових завдань до лабораторних робіт. Досвід показує, що завдання слід добирати таким чином, щоб правильні відповіді на них змінювалися із плином часу. Наприклад:

1. Коли буде найближче сонячне затемнення?
2. Коли буде найближче місячне затемнення?
3. О котрій годині сьогодні був схід сонця у Вашому місті?
4. Як називається перша стаття в останньому номері Наукового часопису НПУ імені М.П. Драгоманова «Комп'ютерно-орієнтовані системи навчання»?

Завдання можуть носити міжпредметний характер:

1. У звіт до лабораторної роботи запишіть назву однієї з тем розділу «Основи алгоритмізації та програмування» шкільного курсу інформатики. Знайдіть методичні рекомендації до навчання цієї теми. Збережіть окремим файлом.

Або просвітницький характер:

5. Скільки людей у світі щорічно помирають від паління?

Індивідуалізувати завдання можна таким чином:

6. У звіт до лабораторної роботи запишіть назву футбольного клубу, за який Ви вболіваєте. Коли, де і з ким відбудеться найближча зустріч цього клубу.
7. У звіт до лабораторної роботи запишіть назву столиці країни, яку Ви хотіли б відвідати. На який найближчий час є авіаквитки до цього міста з аеропорту Бориспіль? Якої авіакомпанії? Номер рейсу? Як називається аеропорт призначення? Яка погода буде в цьому аеропорту в день прильоту шуканого рейсу?

Існують методи наближеного оцінювання відносних розмірів індексних баз даних пошукових систем, в яких використовуються як інструменти вимірювання запити з рідкісних слів [7]. Можна запропонувати студентам завдання дослідницького характеру: придумати запит (одне слово), за яким буде знайдено менше 500 документів. Результати були, наприклад, такими: фенилбензотеллуразол (рос.), довціпнії (гострі дотепні відповіді, жарти), рожнії (різні), ростепель (рос., неохайна людина), нестелепний (неспритний).

Проблема вдосконалення змісту навчальних дисциплін, в яких вивчаються проблеми інформаційного пошуку, є актуальною і потребує подальшого розвитку. Розгляду потребують алгоритми ранжирування результатів пошуку. Цікавими є перспективи розвитку індексних пошукових систем: персоналізований пошук з урахуванням демографічних, психологічних та поведінкових характеристик користувача.

Література

1. Ramsky, Y. Study of Information Search Systems of the Internet /Ramsky, Y., Rezina, O. // From Computer Literacy to Informatics Fundamentals / Ed. Roland T. Mittermeir. – Vol.3422. – Klagenfurt (Austria): Springer, 2005. – P. 84-91.

2. Сегалович И. Как работают поисковые системы [Электронный ресурс] / Сегалович И. // Мир Интернет. – 2002. – №2 – Режим доступа: http://download.yandex.ru/company/iworld-3.pdf

3. Grossman, D. Information Retrieval. Algorithms and Heuristics [Electronic resource] / Grossman, David A., Frieder, Ophir // Springer The Information Retrieval Series. – 2004. – Vol. 15. – Mode of access : http://ir.iit.edu/~dagr/cs529/files/ir_book/

4. Monz C. Inverted Index Construction. Introduction to Information Retrieval [Electronic resource] / Monz Christof, Maarten de Rijke // Spring. – 2002. – Mode of access : http://www.n3labs.com/pdf/clean-w-05.pdf

5. Hearst, M.A. Search User Interfaces [Electronic resource] / Marti A. Hearst. – Cambridge University Press. – 2009. – Mode of access : http://searchuserinterfaces.com/

6. Орлов А. Как устроены поисковые системы [Электронный ресурс]: доклад / Орлов А. – 2011. – Режим доступа: http://www.searchengines.ru/articles/how_are_search_.html

7. Сегалович И.В. Методы сравнительного анализа современных поисковых систем и определения объема Рунета [Электронный ресурс] / Сегалович И.В, Зеленков Ю.Г., Нагорнов Д.О. // RCDL – 2006. – №2 – Режим доступа: http://download.yandex.ru/company/ paper_76_v1.pdf