МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМ. В. ДАЛЯ
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ЕЛЕКТРОНІКИ
КАФЕДРА КОМП'ЮТЕРНИХ НАУК ТА ІНЖЕНЕРІЇ

До захисту допускається

Завідувач кафедри

_____ Скарга-Бандурова

І.С.

«____»_____ 20__ р.

ДИПЛОМНИЙ ПРОЕКТ (РОБОТА)БАКАЛАВРА

ПОЯСНЮВАЛЬНА ЗАПИСКА

НА ТЕМУ:

Інформаційна система контекстного пошуку в мережі Інтернет

Освітній ступінь "бакалавр"
Спеціальність123 – "комп'ютернаінженерія"

Керівник проекту:	_____ (підпис)	Кривуля Г.Ф. (ініціали, прізвище)
Консультант з охорони праці:	_____ (підпис)	Критська Я.О. (ініціали, прізвище)
Здобувач вищої освіти:	_____ (підпис)	Лєвшин О. В. (ініціали, прізвище)
Група:		КІ-15бд

Сєвєродонецьк 2019

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ

Факультет Інформаційних технологій та електроніки
Кафедра Комп'ютерних наук та інженерії
Освітній ступінь бакалавр
Напрям підготовки _____
(шифр і назва)
Спеціальність 123 – комп'ютерна інженерія
(шифр і назва)

ЗАТВЕРДЖУЮ:

Завідувач кафедри КНІ
_____ І.С. Скарга-Бандурова
«_____» _____20____р.

З А В Д А Н Н Я
НА ДИПЛОМНИЙ ПРОЕКТ (РОБОТУ) БАКАЛАВРА

Лєвшину Олександру Володимировичу
(прізвище, ім'я, по батькові)

1. Тема роботи Інформаційна система контекстного пошуку в мережі Інтернет

керівник проекту (роботи) Кривуля Геннадій Федорович, д.т.н., проф.
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом вищого навчального закладу від "13" 05 2019 р. № 83/15.15
2. Термін подання студентом роботи 16.06.2019
3. Вихідні дані до роботи Тестові файли для завантаження: 1 файл типу doc, 2 файли типу png, 2 файли типу jpg, 1 файл типу log,приклади інформаційно-пошукових систем

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) Поняття інформаційно-пошукових систем, дослідження технології інформаційного пошуку, програмної реалізації пошукової системи, охорона праці

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)
Електронні плакати

6. Консультанти розділів проекту (роботи)

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	ст.викл. Критська Я.О.		

7. Дата видачі завдання 30.04.2019

Керівник _____
(підпис)

Завдання прийняв до виконання _____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломного проекту (роботи)	Строк виконання етапів проекту (роботи)	Примітка
1	Аналіз завдання та робота з	05.05.2019 - 13.05.2019	
2	Аналіз технічних засобів	14.05.2019 - 22.05.2019	
3	Розробка алгоритму	22.05.2019 - 02.06.2019	
4	Програмна реалізація	02.06 .2019- 11.06.2019	
5	Оформлення пояснювальної записки	11.06.2019 - 14.06.2019	
6	Підготовка презентації та доповіді	14.06.2019 - 16.06.2019	

Здобувач вищої освіти _____ Лєвшин О.В.
(підпис) (прізвище та ініціали)

Керівник _____ Кривуля Г.Ф.
(підпис) (прізвище та ініціали)

РЕФЕРАТ

Пояснювальна записка дипломної роботи: 81с., 22 рис., 2 табл., 17 джерел.

Робота присвячена вирішенню проблеми контекстного пошуку. Розроблено методи виявлення ключових слів, які дозволяють проводити пошук інформації, згідно із запитом. Вони полягають в аналізі та обробці пошукового запиту, а також пошуку інформації, на основі цього запиту. В основі методів лежать отримання існуючого запиту у вигляді рядка, його розбір, та пошук інформації, згідно із запитом. Розроблені методи дозволяють знаходити відповідні збіги для подальшого пошуку інформації в Інтернеті. Рішення щодо вибірки ключових слів складається на основі методу, розробленого для контекстного пошуку в Інтернеті.

Розроблений метод реалізації запропонованих методів використано для створення спеціалізованих програмних засобів контекстного пошуку.

Ключові слова: ШТУЧНИЙ ІНТЕЛЕКТ, ПОШУК ІНФОРМАЦІЇ В ІНТЕРНЕТІ, КОНТЕКСТНИЙ ПОШУК, КЛЮЧОВІ СЛОВА, ЗАПИТ, ІНДЕКС, ІНТЕРНЕТ.

Умови одержання дипломного проекту: СНУ ім. В. Даля, пр. Центральний 59-А,. м. Сєвєродонецьк, 93400с.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

ІПС–інформаційно-пошукова система

ПО–програмне забезпечення

WWW - World wide web

БД – база даних

ІПМ – інформаційно-пошукова мова

ВСТУП

Інформаційний пошук — це процес пошуку інформації з певної теми. Основним його завданням є швидке і точне знаходження необхідної інформації. Пошукові системи Інтернету використовують різні методи пошуку інформації. Кожен з них має своїпереваги та недоліки. Виділяють три основні методи для пошуку: простий пошук, розширений пошук та контекстний пошук.

Простий пошук – під час цього пошуку у поле запиту вводять одне або декілька слів, які можуть характеризувати зміст документа. Під час введення одного слова машина видає, як правило, велику кількість посилань, з яких обрати потрібну інформацію буває досить складно. Тому простий пошук використовують для знаходження нескладних, однозначних питань чи теоретичних положень.

Розширений пошук – такий пошук завжди включає запит із групи слів. Під час розширеного пошуку рекомендують зв’язувати ключові слова логічними операторами and (і), or (або), not (ні) тощо. Зазвичай записи ключових слів і логічних операторів у різних пошукових системах або однакові, або досить схожі. Тому, засвоївши один раз прийоми розширеного пошуку, можна ним користуватися де завгодно, переключивши машину в потрібний режим розширеного пошуку.

Контекстний пошук – пошукові машини, що підтримують цей вид пошуку, видають посилання на інформацію, яка точно відповідає ключовим словам у пошуковому вікні.

Мета роботи дати користувачеві можливість користуватися контекстним пошуком, реалізований у сервісі новиннорго порталу. Таким чином скорочується час пошуку необхідних новин, та підвищується зручність використання даної веб-системи.

1 ПОНЯТТЯ ІНФОРМАЦІЙНО-ПОШУКОВИХ СИСТЕМ

Пошук інформації – завдання, яке людство вирішує вже багато століть. У міру зростання обсягу інформаційних ресурсів, потенційно доступних одній людині (наприклад, відвідувачеві бібліотеки), були вироблені дедалі витонченіші і досконалі пошукові засоби і прийоми, що дозволяють знайти необхідний документ.

Автоматизована пошукова система - система, що складається з персоналу і комплексу засобів автоматизації їх діяльності, що реалізує інформаційну технологію виконання встановлених функцій.

Досвід і практика створення систем в різних сферах діяльності дозволяє дати більш широке і універсальне визначення, яке повніше відображає всі аспекти їх сутності.

Інформаційно-пошукова система – це система, що забезпечує пошук і відбір необхідних даних у спеціальній базі з описами джерел інформації (індексами) на основі інформаційно-пошукової мови і відповідних правил пошуку.

Головним завданням будь-якої ІПС є пошук інформації релевантної інформаційним потребам користувача. Дуже важливо в результаті проведеного пошуку нічого не втратити, тобто знайти всі документи, які стосуються запиту, і не знайти нічого зайвого. Тому вводиться якісна характеристика процедури пошуку – релевантність.

Релевантність – це відповідність результатів пошуку сформульованому запиту.

Далі ми будемо, в основному, розглядати ІПС для всесвітньої павутини WWW. Основними показниками ІПС для WWW є просторовий масштаб і спеціалізація. За просторовим масштабом ІПС можна розділити на локальні, глобальні, регіональні та спеціалізовані системи. Локальні пошукові системи можуть бути розроблені для швидкого пошуку сторінок в масштабі окремого сервера. Регіональні ІПС описують інформаційні ресурси певного регіону,

наприклад, російськомовні сторінки в Інтернеті. Глобальні пошукові системи на відміну від локальних прагнуть осягнути неосяжне – по можливості найбільш повно описати ресурси всього інформаційного простору мережі Інтернет.

1.1 Види інформаційно-пошукових систем

Залежно від об'єкта зберігання і типу запиту розрізняють два види інформаційного пошуку: документальний і фактографічний – і, відповідно, два типи ІПС – документальні і фактографічні. Останні також називають інформаційно-довідковими ІПС.

Документальними називаються ІПС, в яких реалізується пошук по тематичним запитам в масиві документів або текстів з подальшим наданням користувачеві підмножини цих документів або їх копій. Поняття документа може змінюватися від системи до системи. У загальному випадку це якийсь інформаційний об'єкт, зафіксований (зазвичай за допомогою деякої знакової системи) на якомусь матеріальному носії (папір, фото- і кіноплівка, магнітна пам'ять і т.п.) і призначений для передачі в просторі та часі в системі соціальних комунікацій.

Фактографічні ІПС реалізують зберігання, пошук і видачу безпосередньо фактичних даних (наукових, технічних, економічних характеристик та властивостей об'єктів, процесів, явищ, адрес, найменувань, кількісних даних і т.п.).

Головна та істотна відмінність між документальним і фактографічним пошуком полягає в підході до семантики документів. У документальних системах описується зміст документів в цілому з точки зору їх тематичного, предметного наповнення. У цьому випадку важливо виявити і назвати (перерахувати) основні теми і об'єкти, яким присвячений документ. У фактографічних системах описуються об'єкти, фіксуються їх ознаки і значення цих ознак. Звідси відмінності в мовах опису і способах зберігання

описів в системі. Відповідно, для кожного виду пошуку існують свої пошукові засоби.

Фактографічні системи припускають накопичення і пошук в масиві документів зі строго регламентованою структурою. Така структура є або результатом попередньої інтелектуальної обробки документів при введенні інформації в систему, або наявністю таких документів в готовому вигляді в конкретних сферах людської діяльності, наприклад, облікові форми, бланки, довідники, розклади і т.п. Існують фактографічні ІПС, які забезпечують накопичення інформації і пошук тільки по одному типу об'єктів і тільки по одному типу запитів. Існують і більш розвинені фактографічні системи, що забезпечують зберігання і пошук даних, різноманітних за змістом і структурою, але це різноманітність завжди звичайна.

Виділяють ще й третій тип систем, які називають інформаційно-логічними. Це системи, що відповідають на запити, на які в інформаційній базі в явному вигляді відповіді немає. Отримати відповідь допомагає екстралінгвістична база знань і інформація, що породжується алгоритмічно з уже наявною (документальною або фактографічною). Ця нова інформація або видається як відповідь на запит, або додатково використовується для пошуку.

Інформаційно-пошукова система документального типу являє собою упорядковану сукупність документів, а також сукупність засобів і методів, призначених для зберігання, пошуку і видачі по запитах документальної інформації. Документальна ІПС видає документи, відповідні запиту по темі та предмету. Документ, центральний предмет або тема якого в цілому відповідає смисловому змісту інформаційного запиту, називається релевантним, а властивість смислової близькості між двома і більше текстами (в даному випадку – між документом і інформаційним запитом) – релевантність. Релевантність – це фундаментальне поняття теорії інформаційного пошуку. Кажуть про два види релевантності: смислова і формальна. Відповідність документа змісту інформаційного запиту називають смисловою релевантністю, а відповідність пошукового образу

цього документа формалізованого пошуковому припису, що виражає даний інформаційний запит, – формальною релевантністю. Також формальну релевантність називають релевантність документа, а смислову релевантність – релевантність інформації (мається на увазі «інформації, що міститься в документі»).

1.2 Архітектура сучасних інформаційно-пошукових систем

Перш ніж описати проблеми побудови інформаційно-пошукових систем WWW та шляхи їх вирішення, розглянемо типову схему такої системи. У різних публікаціях, присвячених конкретним системам, наводяться схеми, які відрізняються один від одного тільки застосуванням конкретних програмних рішень, але не принципом організації різних компонентів системи.

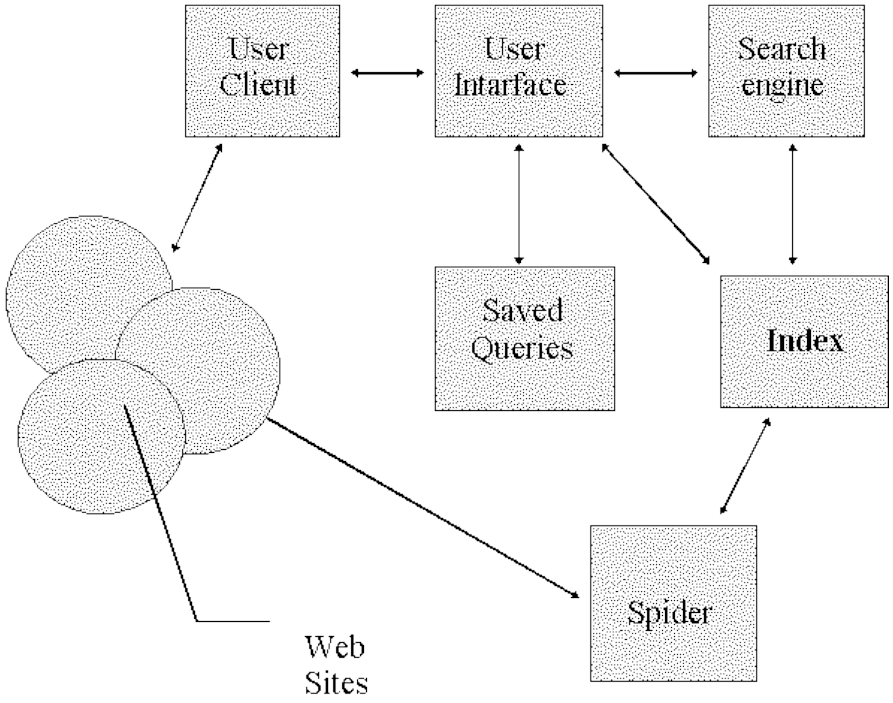


Рисунок 1.1 – Структура ІПС для Internet

На цій схемі позначені:

- UserClient – це програма перегляду конкретного інформаційного ресурсу. В даний час найбільш популярні мультипротокольні програми типу Netscape Navigator. Така програма забезпечує перегляд документів World Wide Web, Gopher, Wais, FTP-архівів, поштових списків розсилки і груп новин Usenet. У свою чергу всі ці інформаційні ресурси є об'єктом пошуку інформаційно-пошукової системи;

- User Interface –це не просто програма перегляду. У разі інформаційно-пошукової системи під цим словосполученням розуміють і спосіб спілкування користувача з пошуковим апаратом системи, тобто з системою формування запитів і переглядів результатів пошуку. Перегляд результатів пошуку та інформаційних ресурсів мережі – це абсолютно різні речі;

- Search Engine – пошукова машина служить для трансляції запиту користувача, який готується ІПМ, в формальний запит системи, пошуку посилань на інформаційні ресурсиInternet та видачі результатів цього пошуку користувачеві;

- Index – це основний масив даних інформаційно-пошукової системи. Він служить для пошуку адреси інформаційного ресурсу. Архітектура індексу влаштована таким чином, щоб пошук відбувався максимально швидко і при цьому можна було б оцінити цінність кожного із знайдених інформаційних ресурсів мережі;

- SavedQueries – запити користувача зберігаються в його особистій базі даних. На налагодження кожного запиту йде досить багато часу, і тому надзвичайно важливо зберігати запити, на які система дає хороші відповіді;

- Spider– робот-індексатор, служить для сканування Internet і підтримки бази даних індексу в актуальному стані. Ця програма є основним джерелом інформації про стан інформаційних ресурсів мережі;

- Web Sites –цевесь Internet.А якщо говорити більш точно, то це ті інформаційні ресурси, перегляд яких забезпечується програмами перегляду.

1.3 Сфери використання сучасних інформаційно-пошукових систем

Сучасні ІПС характерні для так званої інформаційної індустрії – новітньої галузі економіки та соціальної сфери, зайнятої обробкою, систематизацією, накопиченням і поширенням інформації. Бурхливий розвиток ІПС пов'язано з успіхами інформатики. Предметами запиту в ІПС можуть бути бібліографічні дані, управлінська і фактографічна інформація, експертні оцінки, ретроспективний досвід, результати дослідження моделей і т.д. Таке широке коло завдань обумовлює велику різноманітність типів ІПС. Вони розрізняються своїми цілями, обсягом відомостей, видами інформації, способами доведення її до споживача. Поряд з локальними ІПС, що діють в рамках однієї установи (наприклад, поліклініки або лікарні), існують національні і інтернаціональні центри інформаційного обслуговування (наприклад, в області охорони навколишнього середовища). Широке поширення отримали бібліографічні ІПС (наприклад, що містять бібліографію по всіх областях медицини і медико-біологічних наук). Масове виробництво персональних ЕОМ, розвиток засобів комунікацій, можливість об'єднання ЕОМ в інформаційні мережі та звернення зі свого робочого місця до відомостей, що знаходяться в пам'яті інших ЕОМ, істотно розширили діапазон застосування інформації, широту і глибину її пошуку. Якісно новий етап розвитку ІПС пов'язаний з формуванням баз даних на машинозчитуваних носіях. Такі бази даних дозволяють звертатися до них дистанційно, одночасно по багатьом запитам, отримуючи результати пошуку оперативно і в зручному вигляді.

Медицина і охорона здоров'я є надзвичайно специфічною областю впровадження ІПС. Це пов'язано зі складною структурою і різноманіттям форм медико-санітарної інформації, яка містить важкітермінологічні поняття і категорії, а також масиви, які підлягають врахуванню даних. Особливістю медичної інформації є і те, що результати одиничних клінічних або

експериментальних спостережень у міру накопичення і узагальнення стають основою для здійснення великих установ охорони і соціальних заходів. Медико-санітарна інформація є базою прийняття управлінських рішень – від вибору найбільш важливих напрямків науково-дослідної роботи до проведення екстрених санітарно-профілактичних заходів. У масиви інформації, на підставі аналізу якої здійснюється управління охороною здоров'я, входять статистика (демографічна та популяційна, статистика кадрів, дані про захворюваність і смертність та ін.), Узагальнені дані про стан і досягнення медичної та ряду суміжних наукових дисциплін, досвід попередніх років. Саме комплексний характер відомостей послужив причиною розробки єдиної концепції ІПС. Вона включає поетапне створення окремих підсистем, об'єднання яких досягається як на рівні обміну базами даних, так і (або) за допомогою засобів комунікацій.

Процес розробки та інтеграції підсистем в ІПС може здійснюватися по вертикалі і по горизонталі в міру їх створення. Підсистеми, які є допоміжними (наприклад, облік і рух кадрів, планування і фінансування), можуть створюватися незалежно від інших. На нижньому рівні закладу охорони здоров'я (лікарні, клініки, НДІ) користуються ІПС для ведення історій хвороби, контролю ефективності лікувальних заходів, збору і обробки первинних статистичних даних, а також для вирішення управлінських завдань свого рівня компетенції (використання ліжкового фонду та лабораторно-діагностичного обладнання, лікарське забезпечення і ін.).

Здійснюючи оперативні функції, ці ІПС одночасно накопичують, а потім передають необхідну інформацію на більш високий рівень (міський, обласний). Окремо створюються підсистеми довідково-інформаційного обслуговування (в області бібліографії і наукових досліджень, нормативних матеріалів, стандартів). В рамках загальної ІПС можуть розроблятися підсистеми для підтримки і розвитку окремих служб (наприклад, психіатричної, онкологічної) або цільових програм (наприклад, побічна дія лікарських препаратів).

1.4 Постановка задачі

Основною метою атестаційного проекту являється аналіз актуальності інформаційно-пошукової системи та дослідження технологій контекстного пошуку. В процесі виконання даної роботи необхідним є проведення порівняльного аналізу методів реалізації контекстного пошуку, виявлення їх недоліків та переваг для подальшої реалізації різних завдань.

У даній роботі буде детально розглядатися процес розробки програмних продуктів, створених за допомогою методу контекстного пошуку.

Для досягнення мети порівняльного аналізу медодів необхідно вирішити такі завдання:

- вибрати найбільш актуальний з методів;
- возробити та реалізувати алгоритм для вирішення поставленої задачі;
- вослідити якість програмної реалізації, визначити оптимальні параметри його функціонування;
- реалізувати головні задачі, а саме: обробка, аналіз та пошук за запитом користувача;
- розробити інтерфейс веб-системи, що зручний користувачу;
- протестувати програму та викласти на хостинг.

Визначимо основні завдання, які необхідно вирішити в атестаційному проекті:

- аналіз актуальності обраної теми;
- детальний порівняльний аналіз методів контекстного пошуку;
- виявлення переваг та недолік кожного з методів;
- реалізація веб-системи, що дозволить користувачу виконувати пошук у новинному порталі.

Таким чином необхідно розробити програмний продукт для тестування вказаних технологій, а також провести аналіз результатів і зробити висновки щодо доцільності використання даної веб-системи.

2 ДОСЛІДЖЕННЯ ТЕХНОЛОГІЇ ІНФОРМАЦІЙНОГО ПОШУКУ

Інформаційний пошук виконується за певними правилами, які визначають стратегію пошуку, тобто способи досягнення оптимального результату. Стратегія інформаційного пошуку залежить від типу пошукової задачі, критеріїв видачі і характеру діалогу між споживачами інформації і інформаційно-пошуковою системою (ІПС). В загальному вигляді процедура інформаційного пошуку полягає в уточнені інформаційної потреби і формулюванні запиту, визначені сукупності інформаційних масивів, вилучені інформації з інформаційних масивів, ознайомлені користувача з отриманою інформацією і оцінювання результатів пошуку.

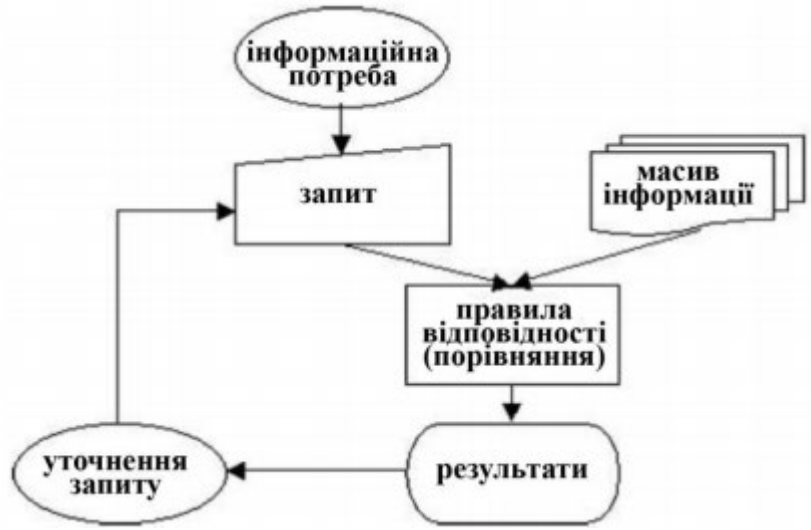


Рисунок 2.1 – Класична модель інформаційного пошуку

2.1 Історія розвитку інформаційно-пошукових систем

Звернемося до історії виникнення мережі Internet, яка була створена в зв'язку з виниклою необхідністю спільного використання інформаційних ресурсів, розподілених між різними комп'ютерними системами. Більшість перших додатків, включаючи FTP і електронну пошту, були розроблені виключно для обміну даними між хост-комп'ютерами Internet.

Інші програми, такі як Telnet, створювалися для того, щоб користувач отримав можливість доступу не тільки до інформації, але і до робочих ресурсів віддаленої системи. У міру розвитку Internet (збільшення користувачів і хост комп'ютерів) колишні методи обміну даними перестали відповідати зрослим потребам користувачів. Виникла необхідність розробки нових способів пошуку мережевих ресурсів і доступу до них, які дозволяли б використовувати інформацією незалежно від її формату і розташування.

Для задоволення таких потреб спочатку були створені пошукова система Archie, вирішальна завдання локалізації ресурсів на FTP-сервері, і система Gopher, що спрощує доступ до різних мережевих ресурсів. Потім були розроблені мережеві інформаційні системи World Wide Web і WAIS, які пропонують абсолютно нові методи отримання інформації. Принципи роботи цих систем дозволяють легко орієнтуватися у величезній кількості інформаційних ресурсів без необхідності надання механізмів роботи самої мережі Internet. Такий підхід дозволяє говорити вже не просто про ресурсах взаємозалежних комп'ютерних систем, а про особливих інформаційних просторах мережі.

Система Archie є комплексом програмних засобів, що працюють зі спеціальними базами даних. У цих базах даних міститься постійно поновлена інформація файлів, до яких можна отримати доступ через сервіс FTP. Користуючись послугами системи Archie, можна здійснити пошук файлу за шаблоном його імені. При цьому користувач отримає список файлів з точним зазначенням місця їх зберігання в мережі, а також з інформацією про тип, часу створення і розмірі файлів. Доступ до інформаційно-пошуковій системи Archie може здійснюватися різними шляхами, починаючи від запитів по електронній пошті і за допомогою сервісу Telnet і закінчуючи використанням графічних Archie-клієнтів. Система Gopher була розроблена для спрощення процесу локалізації FTP-ресурсів Internet і для більш зручного подання відомостей про зміст зберігаються на FTP-серверах файлів.

Система Gopher дає можливість в зручній формі (у вигляді меню) представляти користувачам про наявні файлах та їх зміст. Меню Gopher-серверів можуть містити посилання на інші Gopher- і FTP-сервери. Таким чином, користувач отримує можливість подорожувати по Internet, не звертаючи уваги на місцезнаходження цікавих йому ресурсів, і отримувати доступ до цих ресурсів.

Система Veronica використовується для пошуку інформації в Gopher-просторі по заголовкам пунктів меню. Після введення ключового слова, система Veronica з'ясовує, зустрічається воно в меню на будь-якому Gopher-сервері, і в якості результатів пошуку видає список заголовків пунктів меню, що містять ключове слово. Оскільки система Veronica перестав бути автономної пошукової програмою, а тісно пов'язана з системою Gopher, вона володіє тим же, що і система Gopher, недоліком: далеко не завжди по заголовку можна сказати, що собою представляє той чи інший інформаційний ресурс. Переваги системи полягає в тому, що немає необхідності дізнаватися, де розташована знайдена інформація, досить вибрати потрібну запис зі списку.

2.2 Простий пошук

Під простим пошуком розуміється пошук Web-ресурсів по одному або декільком ключовим словам. Недолік простого пошуку полягає в тому, що зазвичай він видає занадто багато документів, серед яких важко вибрати найбільш підходящі. При простому пошуку в поле запиту вводиться одне або декілька слів, які можуть характеризувати зміст документа. Якщо це слово одне, то, як правило, у відповідь видається велика кількість посилань, з яким незрозуміло що робити. Якщо вводиться кілька слів, то результат залежить від того, як ці слова введені, а це, в свою чергу, залежить від конкретної використовуваної системи.

Наприклад, за запитом «казка» буде видано величезне число різноманітних посилань. Додавання одного або двох ключових слів (наприклад, «українська народна казка») значно звузить область пошуку. При формуванні запиту кількість слів у групі не обмежується. При простому пошуку можливе використання засобів контекстного пошуку. Якщо ключові слова взяти в лапки, то пошукова система знайде документи, у яких дана фраза присутня дослівно. Так можна знайти цитату з художнього твору, наукової праці тощо.

Простий запит дає значну кількість посилань на документи, так як в список потрапляють документи, що містять одне із слів або просте словосполучення, введене при запиті.

Правила оформлення простого запиту:

1. Додавання навпаки. Якщо нам треба, щоб пошукова система знайшла сторінки, на яких одночасно присутні всі використані ключові слова, то перед кожним з них слід поставити знак «+». Використовуючи знак «+», ми звужуємо коло пошуку і зменшуємо кількість можливих посилань. Наприклад, необхідно знайти інформацію про протоколи Internet. Для цього потрібно в рядку пошуку вказати наступне:

+ Інтернет + протокол

Пошукова система видасть список сторінок, на яких зустрічаються обидва ці слова, хоча, звичайно, не виключено, що між ними немає прямого зв'язку.

Ряд систем виконує такий пошук за умовчанням.

2. Вирахування. Конкретизувати коло пошуку інформації можна не тільки знаком «+» а й знаком «-». Наприклад, необхідно знайти інформацію про протоколи Internet, але без урахування тих сторінок, на яких протокол розглядається як нормативний документ. Для цього потрібно в рядку пошуку вказати наступне:

+ Інтернет + протокол-документ

3. Застосування джокера. У запит інформації можна включати спеціальний символ «*», який розширює діапазон пошуку. Символ «*» дозволяє замінити будь-який інший символ або набір символів до кінця слова. Наприклад, необхідно знайти інформацію про протоколи Internet. Для цього в рядку пошуку запишемо:

+ Інтернет * + протокол *

2.3 Розширений пошук

Для спрощення завдання формування складних запитів використовують спеціальні форми, за допомогою яких виконується розширений пошук. Для більш швидкого й успішного пошуку в пошукових системах разом із ключовими словами використовуються різні логічні оператори. Завдяки цьому можна сконструювати запит так, що будуть знайдені не тільки результати на тему, яка вас цікавить, а й конкретні результати і навіть окремі документи. Правила складання складних запитів в одній пошуковій системі можуть відрізнятися від таких в іншій, але в кожному разі будуть використовуватися такі основні логічні оператори й синтаксичні вирази.Наприклад, щоб потрапити на Web-форму розширеного пошуку Google, досить на стартовій сторінці Google знайти посилання «Розширений пошук» і перейти за ним. На стартовій сторінці можна виділити окремі блоки, кожний з яких складається з кількох рядків.

Таблиця 2.1 – Логічні оператори й синтаксичні вирази

Назва	Функція	Приклад використання
1	2	3
І (AND)	Поєднує два або більше слів так, щоб вони всі булиприсутні в шуканому документі	За запитом Червона І Шапочка будуть знайдені документи, що містять і те, іте слово

Продовження таблиці 2.1

1	2	3
АБО (OR)	Забезпечує пошук за кожним із слів групи	За запитом освіта АБО навчання будуть знайдені документи, що містять слово освіта або слово навчання
Логічні дужки	Управляє порядком проходження логічних операторів	За запитом Ломоносов АБО (Михайло І Васильович) будуть знайдені документи, що містять слово Ломоносов або словосполучення Михайло Васильович
НЕ (NOT)	Виключає будь-яке ключове слово з результатів пошуку	За запитом Ссавці НЕ хижаки будуть знайдені документи, що містять слово ссавці і не містять слово хижаки
БЛИЗЬКО (NEAR)	Дозволяє вказати, наякій відстані одне від одного можуть розташовуватися слова в документі	За запитом музика NEAR скачати безкоштовно будуть видані документи, у яких ці слова перебувають недалеко одне від одного, тобто між ними не повинно бути більше десяти слів
Лапки	Забезпечують дослівний пошук виразу або словосполучення	За запитом «відношення долара до євро» будуть видані документи, у яких дослівно міститься ця фраза

Де б не застосовувалися логічні оператори (програмування, логіка і ін.), вони позначають логічний зв'язок між чимось. У пошукових системах, визначаючи логічний зв'язок між пошуковими елементами, оператори дозволяють сформулювати запит.

Наприклад: шукаємо інформацію по темі «Трудове право», тобто з пошуковими елементами «трудове» і «право». Нам потрібні документи, де

обидва слова присутні одночасно: тобто не будь-яке «право», а тільки ті, де є ще й «праця». Втім, можна сказати і так: не будь-який «праця», а тільки в комплекті з «правом».

Коротше кажучи,не важливо в якому порядку, але ОБИДВА СЛОВА в кожному документі.Для цього застосовується оператор «І»:

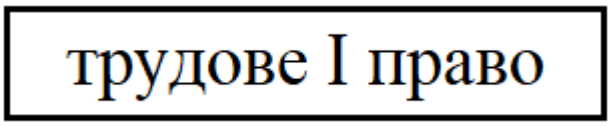


Рисунок 2.2 – Приклад оператору «І»

А якщо запит буде складніше: «Трудове та господарське право»?Якщо написати: «трудове І господарське І право», – будуть видані документи, в яких зустрічаються всі три слова одночасно. Значить окремі «трудове право» і «господарське право» ми у видачі не побачимо.

Якщо потрібно отримати документи з першим словом, а також документи, де зустрічається друге слово (трудове плюс господарське), застосовується логічний оператор АБО:

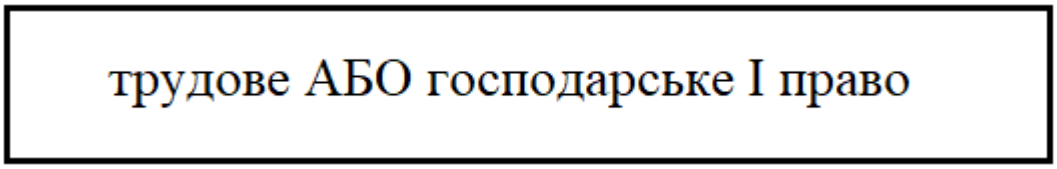


Рисунок 2.3 – Логічний оператор «АБО»

Але в результаті пошуку у видачі буде багато зайвої інформації - більше, ніж після пошуку з первісної формулюванням «трудове право»: про трудовому навчанні, трудових таборах і т.д. Вся справа в операторі «АБО».

Послідовність виконання пошуку в запиті з операторами «АБО» та «І» така ж як в прикладі зі знаками множення і складання.

Оператор І при формулюванні запиту також пріоритетний, а АБО НЕ пріоритетний. Тому в нашій видачі виявилася інформація по «господарського права» плюс до нього різні документи з коренем слова «труд».

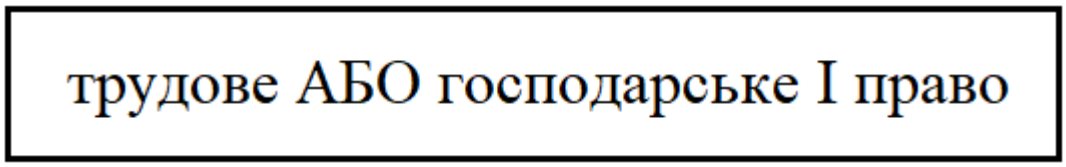


Рисунок 2.4 – Використання операторів «АБО» та «І»

Тепер запит сформульований правильно, і ми отримуємо правильну відповідь – інформацію по трудовому і господарському праву.

Третій оператор НЕ відсікає непотрібну інформацію. Наприклад, якщо ви шукаєте тексти на цю тему, але підручники вам не потрібні, а тільки офіційні документи, монографії, довідники і т. Д. Потрібно продовжити формулювати запит:

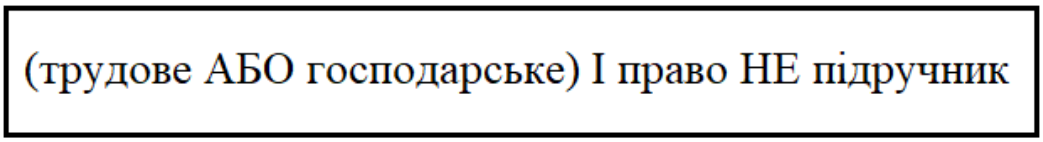


Рисунок 2.5 – Використання операторів «АБО», «І» та «НЕ»

Трудове, трудового, трудових, граматичні закінчення можуть бути різні. Пошукові системи розбираються в них по-різному і про це повідомляють користувачам:

1) шукають «з урахуванням морфології мови», тобто всі можливі граматичні форми кожного з слів;

2) якщо пошукова система не робить цього автоматично, то пропонує не писати граматичне закінчення і усікати його різними значками, щоб не обмежувати пошук;

3) деякі знаходять не тільки різні граматичні форми, але також і синоніми написаних слів;

4) пропонують пошук на точну відповідність того, що написав користувач.

Тобто для тих користувачів, які не бажають безпорадно носитися в океані інформації «без керма і без вітрил» в пасивному режимі лінійного пошуку, пошукові системи пропонують розширений пошук. Він дає можливість сформулювати запит, а значить свідомо направити пошук в потрібну сторону.

2.4 Контекстний пошук

Контекстний пошук буде корисний у тому випадку, якщо не відомо, які ключові слова вибирати для пошуку.

Основною формою контекстного пошуку є процес сканування повнотекстового запиту, щоб зрозуміти, що потрібно користувачеві. Веб-пошукові системи переглядають HTML – сторінки для контенту і повертають рейтинг індексу в залежності від того, наскільки релевантний зміст введеного запиту. HTML-сторінки з більш високим значенням ключових слів запиту в своєму змісті не оцінюються вище. Користувачі мають обмежений контроль над контекстом свого запиту на основі слів, які вони використовують для пошуку. Наприклад, користувачі, які шукають частину меню на веб-сайті, можуть додати «меню» в кінець свого запиту, щоб забезпечити пошукову систему контекстом того, що їм потрібно. Наступним кроком в контекстуалізації пошуку є те, що служба пошуку сама запитує інформацію, яка звужує результати, такі як Google,щоб зменшити діапазон часу для пошуку всередині.

При контекстному пошуку система надає наступні можливості:

– завдання в запиті логічних формул, в тому числі з операторами відстані. Як «слів» можуть виступати цифри, букви і цифри послідовності, слова які повинні бути знайдені в заданому вигляді, а також шаблони –

буквено-цифрові послідовності з символами "*" (будь-яка подстрока, в тому числі порожня) і "?" (Будь-яка буква або цифра);

– вбудований морфологічний аналізатор, що дозволяє автоматично знайти всі існуючі словоформи для більшості слів російської та англійської мов.

Для прискорення контекстного пошуку система створює індекси за текстом. З метою зменшення розміру, ці індекси не містять докладної інформації про становище слова всередині документа, тому в разі використання операторів відстані підсистема пошуку працює в два проходи: на першому відбирає документи, що містять задані слова, на другому зчитує їх тексти в пам'ять і перевіряє виконання умови по відстані.

Здатність знаходити і ранжувати документи, що не містять слів із запиту, часто вважають ознакою штучного інтелекту або пошуку за змістом і відносять апріорі до переваг моделі.

Для прикладу візьмемо одну, мабуть, найпопулярнішу модель, що працює за змістом. У теорії інформаційного пошуку дану модель прийнято називати латентно-семантичним індексуванням (іншими словами, виявленням прихованих смислів). Ця алгебраїчна модель заснована на сингулярному розкладанні прямокутної матриці, що асоціює слова з документами. Елементом матриці є частотна характеристика, яка відображає ступінь зв'язку слова і документа, наприклад, TF * IDF. Замість вихідної мільонноразмерной матриці автори методу Фурнас і Дірвестер запропонували використовувати 50-150 «прихованих смислів», відповідних першим головним компонентам її сингулярного розкладання.

Сингулярним розкладанням дійсної матриці Аразмеров m * n називається всяке її розкладання виду A = USV, гдеU - ортогональна матриця розмірів m * m, V - ортогональна матриця розмірів n * n, S - діагональна матриця розмірів m * n, елементи якої sij = 0, якщо i не дорівнює j, і sii = si> = 0. Величини si називаються сингулярними числами матриці і рівні арифметичним значенням квадратних коренів з відповідних власних значень

матриці AAT. В англомовній літературі сингулярне розкладання прийнято називати SVD-розкладанням.

Давним-давно доведено (Екарт), що якщо залишити в розгляді перші k сингулярних чисел (інші прирівняти до нуля), ми отримаємо найближчу з усіх можливих апроксимацію вихідної матриці рангу k (в певному сенсі її «найближчу семантичну інтерпретацію рангу k»). Зменшуючи ранг, ми відфільтровує нерелевантні деталі; збільшуючи, намагаємося відобразити всі нюанси структури реальних даних.

Операції пошуку або знаходження схожих документів різко спрощуються, оскільки кожному слову і кожній документу зіставляється відносно короткий вектор з k смислів (рядки і стовпці відповідних матриць). Однак через малу осмисленості «смислів» або з якоїсь іншої, але використання LSI в лоб для пошуку так і не набуло поширення. Хоча в допоміжних цілях (автоматична фільтрація, класифікація, поділ колекцій, попереднє зниження розмірності для інших моделей) цей метод, мабуть, знаходить застосування.

Явно наданий контекст ефективно підвищує точність результатів, однак ці служби пошуку, як правило, страждають від поганого користувацького досвіду. Вивчення інтерфейсу таких програм, як Inquirus, може виявитися складним для звичайних користувачів без знання показників пошуку. Аспекти, що поставляється контексту з'являються в основних пошукових системах з найкращим призначеним для користувача взаємодією, таким як Google і Bing. Google дозволяє користувачам фільтрувати по типу: зображення, карти, покупки, новини, відео, книги, авіаквитки і додатки. У Google є великий список операторів пошуку, які дозволяють користувачам явно обмежувати результати відповідно до їх потреб, такими як обмеження певних типів файлів або видалення певних слів. Bing також використовує аналогічний набір операторів пошуку, щоб допомогти користувачам в явному звуженні контексту їх запитів. Bing дозволяє користувачам

виконувати пошук в межах діапазону часу, типу файлу, розташування, мови і т. д.

Існують і інші системи, які працюють над автоматичним обчисленням контексту призначених для користувача запитів на основі вмісту інших документів, які вони переглядають або редагують. Проект Watson від IBM спрямований на створення когнітивної технології, яка динамічно вчиться при обробці запитів користувачів. При поданні запиту Уотсон створює гіпотезу, яка оцінюється в порівнянні з існуючим банком знань на основі попередніх питань. Оскільки відповідні терміни і відповідні документи зіставляються із запитом, гіпотеза Вотсона модифікується, щоб відображати нову інформацію, надану через неструктуровані дані, на основі інформації, отриманої в попередніх ситуаціях. Здатність Уотсона нарощувати колишні знання дозволяє автоматично фільтрувати запити для аналогічних контекстів, щоб забезпечити точні результати.

Основні пошукові служби, такі як Google, Bing і Yahoo, також мають систему автоматичного визначення контексту конкретних запитів користувачів. Google відстежує попередні запити користувача і вибрані результати для подальшої персоналізації результатів для цих осіб. Наприклад, якщо користувач послідовно шукає статті, пов'язані з тваринами, дикими тваринами або тваринами, пошук «ягуара» оцінить статтю про конях ягуарів вище, ніж посилання на автомобілі Jaguar. Подібно Watson, пошукові служби прагнуть вчитися у користувачів на основі попереднього досвіду, щоб автоматично надавати контекст по поточним запитам. Bing також забезпечує автоматичний контекст для конкретних запитів на основі вмісту самого запиту. Пошук «піци» повертає інтерактивний список ресторанів і їх рейтинги на основі приблизного місця розташування комп'ютера користувача. Сервер Bing автоматично повідомляє, що, коли користувач шукає товар, він цікавиться документами в контексті покупки цього продовольчого товару або пошуку ресторанів, які продають цей конкретний товар.

Не всі позатекстові критерії корисні в рівній мірі. Саме посилальна популярність і похідні від неї виявилися вирішальним фактором, поміняв в 1999-2000 рр. світ пошукових систем і повернули їм відданість користувачів. Так як саме з її допомогою пошукові системи навчилися пристойно і самостійно (без підпірок з вручну відредагованих результатів) ранжувати відповіді на короткі частотні запити, що становлять значну частину пошукового потоку.

Найпростіша ідея глобального (тобто статичного) обліку посилальної популярності полягає в підрахунку числа посилань, що вказують на сторінки. Приблизно те, що в традиційному бібліотекознавства називають індексом цитування. Цей критерій використовувався в пошукових системах ще до 1998 року. Однак він легко піддається накрутці, крім того, він не враховує вагу самих джерел.

Природним розвитком цієї ідеї можна вважати запропонований Бріном і Пейджем в 1998 році алгоритмPageRank - ітеративний алгоритм, подібний до того, що використовується в задачі визначення переможця в шаховому турнірі за швейцарською системою. У поєднанні з пошуком по лексиці посилань, що вказують на сторінку (стара, дуже продуктивна ідея, яка використовувалася в гіпертекстових пошукових системах ще в 80-і роки), цей захід дозволив різко підвищити якість пошуку.

Трохи раніше, ніж PageRank, був запропонований локальний (тобто динамічний, заснований на запиті) алгоритм обліку популярності - HITS (Кляйнберг), який не використовується на практиці в основному через обчислювальної дорожнечі. Приблизно з тієї ж причини, що і локальні (тобто динамічні) методи, які оперують словами.

Обидва алгоритми, їх формули, умови збіжності детально описані, в тому числі і в російськомовній літературі. Зазначу тільки, що розрахунок статичної популярності не є самоцінною завданням, він використовується в численних допоміжних цілях: визначення порядку обходу документів, ранжування пошуку по тексту посилань і т.д. Формули розрахунку

популярності постійно покращують, в них вносять облік додаткових чинників - тематичної близькості документів (наприклад, популярна пошукова сістемаwww.teoma.com), їх структури і т.п., що дозволяють знизити вплив непотизму. Цікавою окремою темою є ефективна реалізація відповідних структур даних (Бхарат).

Хоча розмір бази в інтернеті на поверхневий погляд не здається критичним фактором, це не так. Недарма зростання відвідуваності таких машин, як Google і Fast, добре корелює саме з ростом їх баз. Основна причини: «рідкі» запити, тобто ті, за якими знаходиться менше 100 документів, складають в сумі близько 30% від всієї маси пошуків - вельми значну частину. Цей факт робить розмір бази одним з найбільш критичних параметрів системи.

Однак зростання бази крім технічних проблем з дисками і серверами обмежується і логічними: необхідністю адекватно реагувати на сміття, повтори і т.п. Не можу втриматися, щоб не описати дотепний алгоритм, який застосовується в сучасних пошукових системах для того, щоб виключити «дуже схожі документи».

Походження копій документів в інтернеті може бути різним. Один і той же документ на одному і тому ж сервері може відрізнятися з технічних причин: бути представлений в різних кодуваннях і форматах, містити змінні вставки - рекламу або поточну дату.

Широкий клас документів в Інтернеті активно копіюється і редагується - стрічки новинних агентств, документація і юридичні документи, прейскуранти магазинів, відповіді на найбільш поширені питання і т.д. Популярні типи змін: коректура, реорганізація, ревізія, реферування, розкриття теми і т.д. Нарешті, публікації можуть бути скопійовані з порушенням авторських прав і змінені зловмисно з метою утруднити їх виявлення.

Крім того, індексація пошуковими машинами сторінок, що генеруються з баз даних, породжує ще один поширений клас зовні мало

відрізняються документів: анкети, форуми, сторінки товарів в електронних магазинах.

Очевидно, що з повними повторами проблем особливих немає, досить зберігати в індексі контрольну суму тексту і ігнорувати всі інші тексти з такою ж контрольною сумою. Однак цей метод не працює для виявлення хоча б трохи змінених документів.

Для вирішення цього завдання Уди Манбер (автор відомої програми наближеного прямого пошуку agrep) в 1994 році запропонував ідею, а Андрій Бродер в 1997-му придумав назву і довів до розуму алгоритм «шинглів» (від слова shingles - «черепички, лусочки»). Ось його приблизний опис.

Рисунок 2.6 – Опис алгоритму «шинглів»

Для кожного десятисловія тексту розраховується контрольна сума (шингл). Десятісловія йдуть внахлест, з перекриттям, так, щоб жодне не пропало. А потім зі всієї безлічі контрольних сум (очевидно, що їх стільки ж, скільки слів в документі мінус 9) відбираються тільки ті, які діляться на, скажімо, 25. Оскільки значення контрольних сум розподілені рівномірно, критерій вибірки ніяк не прив'язаний до особливостей тексту. Ясно, що повтор навіть одного десятисловія - вагома ознака дублювання, якщо ж їх багато, скажімо, більше половини, то з певною (нескладно оцінити ймовірність) упевненістю можна стверджувати: копія знайдена! Адже один співпав шингл в вибірці відповідає приблизно 25 співпало декалогам в повному тексті.

Очевидно, що так можна визначати відсоток перекриття текстів, виявляти все його джерела і т.п. Цей витончений алгоритм втілив давню мрію доцентів: відтепер болюче питання «у кого студент списував цей

курсовик» можна вважати вирішеним! Легко оцінити частку плагіату в будь-якій статті.

Щоб у читача не склалося враження, що інформаційний пошук - виключно західна наука, згадаю про альтернативний алгоритм визначення майже-дублікатів, придуманий і втілений у нас в Яндексі (Іллінський). У ньому використовується той факт, що більшість пошукових систем вже мають індексом у вигляді інвертованого файлу (або інвертованим індексом), і цей факт зручно використовувати в процедурі знаходження майже-дублікатів.

Хоч би яка була модель, пошукова система потребує «тюнінг» - оцінці якості пошуку і налаштування параметрів. Оцінка якості - ідея, фундаментальна для теорії пошуку. Бо саме завдяки оцінці якості можна говорити про можливість застосування або незастосування тієї чи іншої моделі і навіть обговорювати їх теоретичні аспекти.

Зокрема, одним з природних обмежень якості пошуку служить спостереження, винесене в епіграф: думки двох «асессоров» (фахівців, які виносять вердикт про релевантності) в середньому не збігаються один з одним в дуже великій мірі! Звідси випливає і природна верхня межа якості пошуку, адже якість вимірюється за підсумками зіставлення з думкою асесора.

Зазвичай для оцінки якості пошуку використовують два параметри:

1) точність (precision) - частка релевантного матеріалу у відповіді пошукової системи;

2) повнота (recall) - частка знайдених релевантних документів в загальній кількості релевантних документів колекції.

Саме ці параметри використовувалися і використовуються на регулярній основі для вибору моделей і їх параметрів в рамках створеної Американським інститутом стандартів (NIST) конференції по оцінці систем текстового пошуку (TREC - textretrival evaluation conference) [6]. Розпочата в 1992 році консорціумом з 25 груп, до 12-му році свого існування конференція

накопичила значний матеріал, на якому до цих пір вигострюються пошукові системи. До кожної чергової конференції готується новий матеріал (т.зв. доріжка) по кожному з нас цікавлять напрямків. «Доріжка» включає колекцію документів і запитів.

Трохи осторонь від статистичних моделей і структур даних варто клас алгоритмів, традиційно відносяться до лінгвістичним. Точно кордону між статистичними і лінгвістичними методами провести важко. Умовно можна вважати лінгвістичними методи, що спираються на словники (морфологічні, синтаксичні, семантичні), створені людиною. Хоча вважається доведеним, що для деяких мов лінгвістичні алгоритми не вносять істотного приросту точності і повноти - наприклад, англійської (Стржалковська), - все ж основна маса мов вимагає хоча б мінімального рівня лінгвістичної обробки. Не вдаючись в подробиці, приведу лише список завдань, що вирішуються лінгвістичними або окололінгвістіческімі прийомами:

- автоматичне визначення мови документа;
- токенізація (графематіческій аналіз): виділення слів, кордонів пропозицій;
- виключення неінформативних слів (стоп-слів);
- лематизації (нормалізація, стемінг): приведення словозмінної форм до «словникової» (в тому числі і для слів, що не входять в словник системи);
- поділ складних слів (компаундів) для деяких мов (наприклад, німецької);
- дізамбігуація: повне або часткове зняття омонімії;
- виділення іменних груп.

Ще рідше в дослідженнях і на практиці можна зустріти алгоритми словоосвітнього, синтаксичного і навіть семантичного аналізу. При цьому під семантичним аналізом частіше мають на увазі який-небудь статистичний алгоритм (LSI, нейронні мережі), а якщо толково-комбінаторні або семантичні словники і використовуються, то у вкрай вузьких предметних областях.

2.5 Пристрій повнотекстового індексу

Всі технології повнотекстового пошуку працюють за одним принципом. На основі текстових даних будується індекс, який здатний дуже швидко шукати відповідності ключового слова.

Зазвичай сервіс пошуку складається з двох компонент: пошуковий пристрій і індексатор. Індексатор отримує текст на вхід, робить обробку тексту (вирізання закінчень, незначущих слів і т.п.) і зберігає всі в індексі. Пристрій такого індексу дозволяє проводити по ньому дуже швидкий пошук.

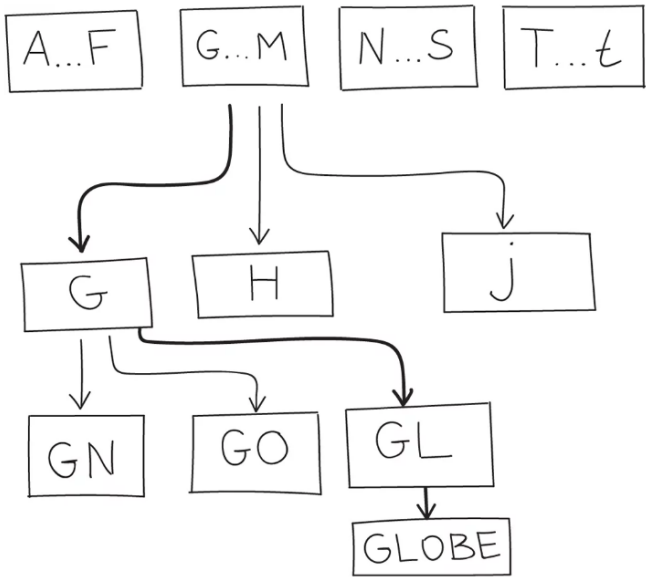


Рисунок 2.7 – Схема індексатора

Пошуковий пристрій– інтерфейс пошуку за індексом – приймає від клієнта запит, обробляє фразу і шукає її в індексі.

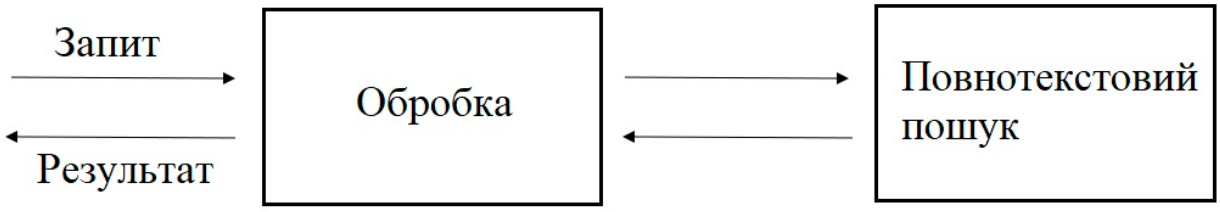


Рисунок 2.8 – Схема повнотекстового пошуку

Пропонований метод можна розділити на наступні підзадачі:

1. Індексування документів.

2. Визначення релевантності документів (пошук),

3. Аналіз статистики з метою визначення якості результатівпошуку.

4. Зміна індексів відповідно до якості результатів пошуку і виділення словосполучень.

2.5.1 Індексування документів

Зміна індексу документа проводиться в наступних випадках

а) якщо в корпус доданий новий документ;

б) якщо існуючий документ змінений;

в) при обробці статистики.

Даний процес реалізований так само, як в більшості пошуковихсистем і тому є тривіальним. Наведемо лише аналіз знаків пунктуації, оскільки цей крок володіє однією особливістю - наявність розділового знака між словами призводить до штучного збільшення позиції наступного слова.

При такому збільшенні конкретне значення залежить від розділового знака. Всі знаки поділяються на чотири класи: роздільники складовихслів (наприклад, «латентно-семантичний» або 1.0 »), роздільники слів(пробіл), роздільники словосполучень (кома, двокрапка і т.п.), роздільники своєї можливості (точка, за якою слідує пробіл і т.п.), Штучне збільшення позиції робиться для того, щоб полегшити розбиття тексту на словосполучення - очевидно, що слова, розділені, наприклад знаком оклику або питальним, знаком словосполученнями не являються.

Вага обчислюється за відомою формулою *tf* (termfrequency), яка описує частоту появи слова в документі:

$$tf_{i,j} = \frac{n_{ij}}{\sum_k n_{k,j}},$$

де $n_{i,j}$ – кількість повторень ключового слова t_i в індексованих документі d_j, а знаменник – загальна кількість слів у документі d_j.

2.5.2 Визначення релевантності

Завдання пошуку зводиться,фактично, до задачі порівняння двох документів $-q$ (query, запиту) і d_j (Документи з корпусу). Результатом кожного такого порівняння є число, яке описує ступінь близькості між документами.Релевантність в пропонованому методі обчислюється поетапно. Спочатку проводиться звичайний логічнийпошук за ключовими словами з урахуванням операторів явно або неявно присутніх в пошуковому рядку.Такими операторамиє:AND, OR, ANDNOT і ExactPhrase (застосовується до слів, укладеним в лапки). Логічний пошук дозволяє зробити первинний відбір документів, що задовольняють запиту,проте він не дає уявлення про релевантність. Наступним етапом є обчислення базової релевантностікожного звідібраних документів за ключовими словами. Для її обчислення пропонується використовувати косинус кута між векторами документів. Також розглядалася і інша популярна формула –OkapiBM25, але експерименти, що проводилися на тестовому безлічі з 7000 документів, показали, що найбільш раціональні результати як за якістю, так і запродуктивності дає саме косинус кута, часто використовується длявизначення смислової близькості при векторної моделі текстів.

$$\cos\left(d_j, q\right) = \frac{d_j \cdot q}{\left\|d_j\right\|\left\|q\right\|} = \frac{\sum_{i=1}^{N} w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2}\sqrt{\sum_{i=1}^{N} w_{i,q}^2}},$$

де $w_{i,j}$ –вага відповідних слів, обчислений за формулою tf.

Після обчислення базової релевантності здійснюється коректування отриманих значень відповідно до близькості документів q та d_j за словосполученнями.Із запиту словосполучення виділяються на основі позицій слів. Близькими вважаються слова, відстань між якими становить1 (слова, розділені пробілом) або 0 (для складних слів, розділених, наприклад, дефісом).Також при виборі словосполучень враховуються дані, накопичені за результатами аналізу статистики попередніх запитів з тими ж словами. Після цього для документа d_j, обчислюється відповідний коригуючий коефіцієнт:

$$K_{phrase}\left(q,d_j\right)=\frac{\sum_i K_p\left(p_i,d_j\right)}{N}.$$

Іншими словами, коефіцієнт обчислюється як середнє арифметичне релевантностсі по кожному з розглянутих словосполучень p_i, виділених із запиту q. Тут K_p - релевантність документа словосполученню p_i, яка обчислюється за такою формулою:

$$K_p\left(p_i,d_j\right)=\sum_k \frac{2^{2n_{t_i}}}{\Delta p_i,k},$$

де n_{t_i} – кількість слів у словосполученні p_i, (в загальному випадку воно можескладатися з двох і більше слів), $\Delta p_i,k$ – сумарне відстань між кожним з цих слів в даному документі d_j, обчислене для кожного входження в документ словосполучення p_i.Для слів, відстань між якими дорівнює нулю (наприклад, слів, розділених дефісом) це значення береться рівним 0,25. Відзначимо, що входження фрази p_i, в документ d_j визначається за наступним правилом: якщо відстань між кожною парою сусідніх слів, складових словосполучення, не перевищує M, вважається, що дані слова в документі d_j є словосполученням p_i. M –константа цілого числа, що позначає величину, на яку збільшується позиція наступного слова після

точки (або іншогорозділового знака з групи знаків, які закінчують речення) прииндексировании.

Таким чином, дана формула виробляє більш точну оцінку близькості текстів по словосполучень за рахунок обліку більшої кількості чинників, ніж, наприклад, що з'явився недавно алгоритм PATeR. Пропонована формула дозволяє використовувати поєднання з трьох і більше слів, причому довжина поєднання дуже впливає на результуючий коефіцієнт. Очевидно, що довгі словосполучення мають більш вузький зміст, ніж короткі, а значить, їх збіг має чинити більший вплив на результуючу смислове близькість. Також враховуються словосполучення, розділені декількома словами (які, в загальному випадку, можуть виявитися вступними фразами), причому кількість таких проміжних слів послаблює підсумковий внесок виділеного в списку в результат.Штучне збільшення позиції ключових слів в залежності від знаків пунктуації та згадане вище правило дозволяє природним чином вирішити проблему проізводітсяьності, що виникає при вимозі знаходження всіх слів в одному реченні (що, по суті, призводить до використання плаваючого вікна).Біліше того, правило, що визначає словосполучення в документі як «слова, що знаходяться в єдиній пропозиції, причому кожна пара розділена не більше ніж M словами», є більш гнучким ніж те, що використовується в алгоритмі PATeR, оскільки враховує можливу відсутність семантичного зв'язку між словами, що знаходяться, наприклад, в різних кінцях досить довгого речення.

Після отримання коефіцієнтів відбувається обчислення підсумкової релевантності для кожного документа:

$$R\left(q,d_j\right)=\cos\left(q,d_j\right)\left(K_{phrase}\left(q,d_j\right)+1\right),$$

де $\cos\left(q,d_j\right)$ – косинус кута між векторами запиту q і документа d_j, що характеризує їх близькість за ключовими словами.

2.5.3 Аналіз статистики

Способи визначення якості результатів пошуку можуть відрізнятися в залежності від типу даних, що зберігаються і реалізації конкретніше системи.Дослідження, проведене під час роботи над запропонованим варіантом алгоритмом, показує, що найочевидніший спосіб - явна оцінка якості користувачами (explicit relevance feedback) є неефективним.В рамках експерименту користувачам опублікованної в Інтернеті довідкової системи було запропоновано оцінити якість статті. В результаті з 346 550 унікальних переглядів, що відбулися за рік, було дано лише 798 оцінок.Іншими словами, експеримент показав, що користувачі схильні явно оцінити переглянуті документи лише в 0,23% випадках, а значить, даний підхід не є достатньо ефективним.

Більш перспективним і універсальним представляється метод не залежить від типу системи і не вимагає від користувачів ніяких додаткових дій. Майже пропонується збирати статистику про пошукові запити та обраних результатах кожного запиту неявно (implicit relevance feedback), після чого використовувати для аналізу кожної сесії роботи користувача з системою наведені нижче евристичні правила. Позначимо за К безліч всіх ключових слів, що зустрічаються як у запиті, так і в уже згадуваному документе (тобто перетин множин ключових слів з документов q і d_j). Якщо документ вважається *хорошим* результатом, вага ключових слів як для цього документа повинен бути збільшений, якщо *поганим* - зменшений.Два запиту пропонується вважати схожими, якщо ступінь їх близькості по ключовим словам досягає певного порогового значення.

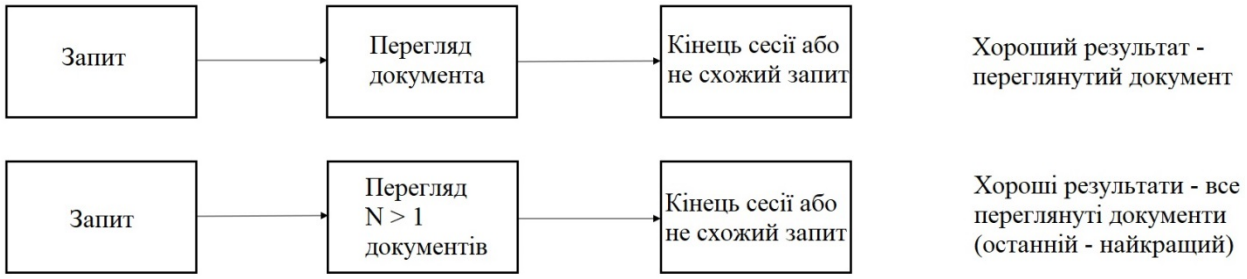


Рисунок 2.9 – Правила визначення хороших результатів

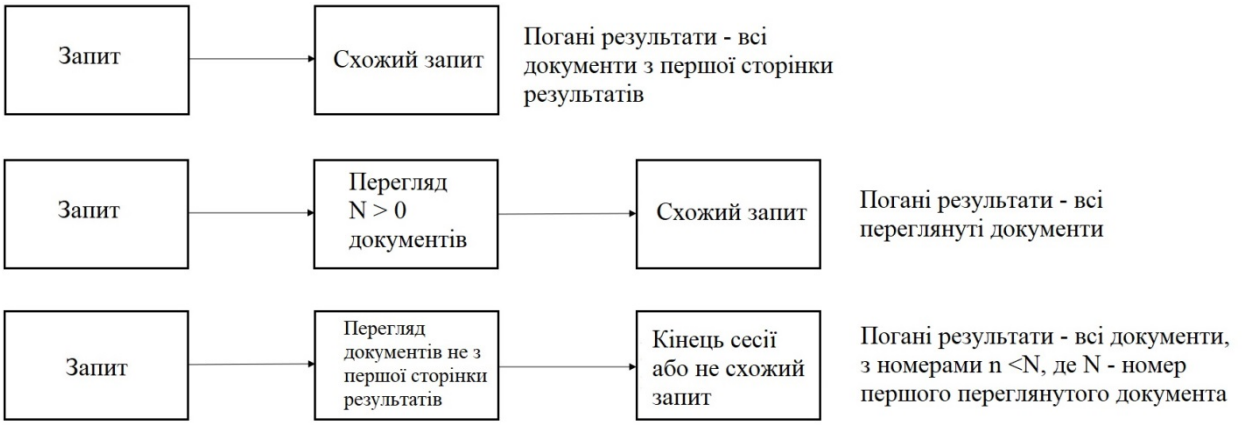


Рисунок 2.10 – Правила визначення поганих результатів

Крім коригування ваг, якість результату пошуку дозволяє зробити висновок про правильність розбиття запиту на словосполучення. Наприклад, із запиту «знайти релевантні документи» за замовчуванням будуть виділені три словосполучення: «знайти релевантні документи» «знайти релевантні» і «релевантні документи».Якщо з результатів користувач вибере документи, в яких останнім зустрічається частіше, ніж передостаннє, це дозволить судити про тіснішу семантичного зв'язку між цими двома словами, і в наступний раз ці дані будуть використані алгоритмом для розбиття на словосполучення аналогічних запитів.

2.6 Порівняння роботи методів пошуку Google

Google.com глобально зайняв звання сайту номер 1 в Alexa Top 500 Global Sites. З огляду на ці цифри, власникам власних веб-сторінок особливо важливо мати хорошу видимість своїх сайтів пошуковою системою.

Але не дивлячись на таку загальну популярність Google, чи знаєте ви, як він дійсно працює і що це за панди, пінгвіни, калибри?

Чим потрібніше стає Google для сучасного маркетингу, тим важливіше розуміти функції пошуку і алгоритми оновлень, які безпосередньо впливають на ранжирування результатів. Moz передбачає, що Google змінює свої алгоритми по 600 разів за рік. Багато з цих змін і пов'язані з ними фактори ранжирування тримаються в секреті. І тільки про великі оновлення оголошують публічно.

Своєю появою пошукові системи геть змінили звичний для нас спосіб збору інформації. Чи цікавить вас оновлення даних фондового ринку або ви хочете знайти кращий ресторан в районі, або пишете академічний звіт про Ернеста Хемінгуея - пошуковик дасть відповідь на всі запити. У 80 роки відповіді на питання зажадали б відвідин місцевої бібліотеки. Тепер же все вирішується протягом мілісекунди з використанням алгоритмічних повноважень пошукача.

В цьому відношенні головна мета пошукової системи полягає в тому, щоб максимально швидко знайти доречну і актуальну інформацію, як відповідь на введені пошукові терміни, також звані ключовими словами. Тому центральним аспектом для будь-якої пошукової системи, яка бажає видати дійсно корисний результат, є поняття мети пошуку, того, як саме люди шукають.

Результат роботи Google можна порівняти з інтернет-каталогом, відібраним за допомогою рейтингової системи на основі алгоритмів. Більш конкретно алгоритм пошуку можна описати як «знаходження елемента із заданими властивостями серед списку елементів».

2.6.1 Сканування

Сканування може бути описано, як автоматизований процес систематичного вивчення загальнодоступних сторінок в Інтернеті. Простіше

кажучи, під час цього процесу Google виявляє нові або оновлені сторінки і додає їх в свою базу. Для полегшення роботи він використовує спеціальну програму. «Googlebots» (можна зустріти альтернативні назви: «боти» або «роботи») відвідують список URL-адрес, отриманих в процесі минулого сканування і доповнених даними карти сайту, яку надають веб-майстри і аналізують їх зміст. При виявленні посилань на інші сторінки під час відвідування сайту, боти також додають їх до свого списку і встановлюють систематичні зв'язку. Процес сканування відбувається на регулярній основі з метою виявлення змін, вилучення «мертвих» посилань і встановлення нових взаємозв'язків. І це при тому, що тільки за даними на вересень 2014 року налічується близько мільярда веб-сайтів. Можете собі уявити складність такого завдання? Тим ні менш, боти не відвідують абсолютно кожен сайт. Щоб потрапити в список перевіряються, веб-ресурс повинен бути розглянутий, як досить важливий.

2.6.2 Індексація

Індексація - процес збереження отриманої інформації в базі даних відповідно до різними факторами для подальшого вилучення інформації. Основна мета процесу індексації: швидко реагувати на пошукової запит користувача. Ключові слова на сторінці, їх розташування, мета-теги і посилання представляють особливий інтерес для індексації Google.

Для того щоб ефективно зберігати інформацію про мільярди сторінок в базі даних пошукової системи, Google використовує великі центри обробки даних в Європі, Азії, Північній і Південній Америці. У цих центрах, як було підраховано, з урахуванням споживання енергії Google в 2010 році, працює близько 900,000 серверів.

2.6.3 Обробка

Коли користувач вводить запит, Google виробляє в базі даних пошук, що відповідає умовам і алгоритмічно визначає актуальність змісту, що виводить до певного рейтингу серед знайдених сайтів. Логічно, що результати, які вважаються більш доречними для користувача пошукової системи, навмисно отримують більш високий ранг, ніж результати, які мають менше шансів забезпечити адекватну відповідь.

Хоча Google і не випустив офіційних даних про це, компанія підтверджує, що використовує понад 200 факторів для визначення релевантності і значущості конкретної сторінки.

Природно, всім веб-розробникам важливо знати, які фактори ранжирування, які впливають на позицію сторінки в пошуковій видачі. Іноді Google дає певні натяки, оголосивши важливі зміни в оновленнях своїх алгоритмів.

Всі вищеописані процеси сканування, індексування та позиціонування можна зобразити за допомогою такої схеми:

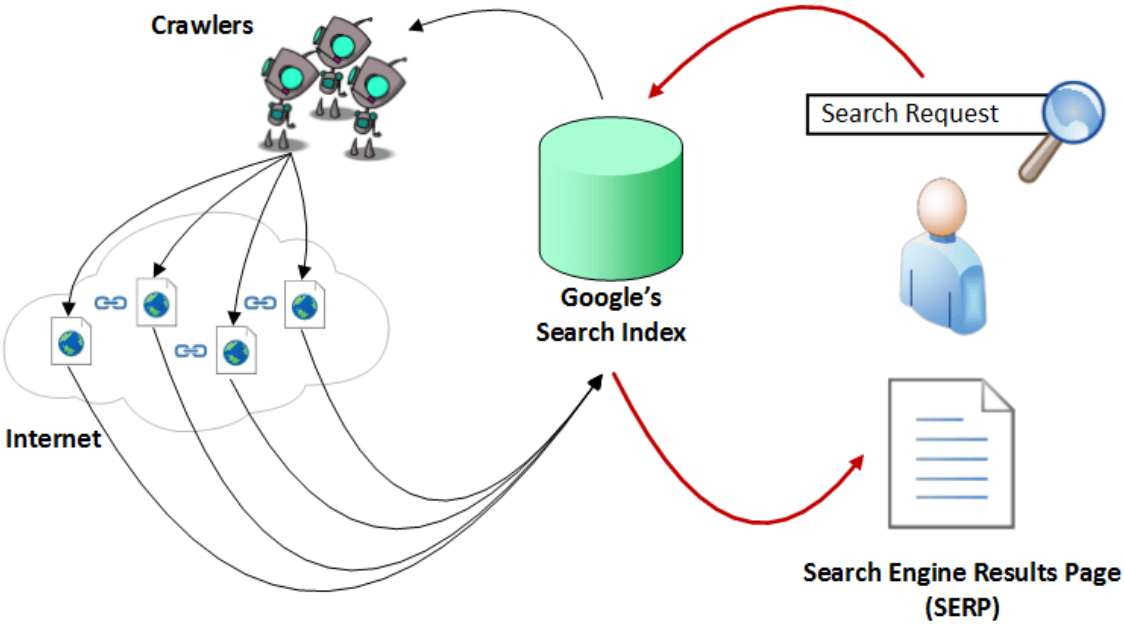


Рисунок 2.11 – Схема пошукувGoogle

Google ніколи не буде публічно розкривати свої алгоритми пошуку і фактори ранжирування результатів. Це було б рівнозначно тому, щоб компанія Coca-Cola викладала рецепти своєї знаменитої газованої води в Інтернет. Тим ні менш, Google хоче покращувати рівень користувацького досвіду і забезпечувати найкращі результати пошуку. Для того, щоб зменшити впроваджуваний контент в результатах пошуку, компанія хоче проінформувати веб-майстрів про те, коли і як змінилися головні стандарти якості відбору. Тому цілком ймовірно, що перед проведенням великого оновлення алгоритму, піде анонс на GoogleWebmasterCentralBlog.

2.6.4 Оновлення методу Panda

Оновлення Panda вперше було представлено в кінці лютого 2011 року. Після було випущено досить багато його апдейтів, на даний момент поточна версія: 4.2. Оновлення можна розглядати, як значне поліпшення алгоритму пошуку, тому що вона спрямована на підвищення якості контенту веб-сайтів. Основна ідея полягає в тому, що оригінальні сайти з авторським контентом в пошуковій системі повинні зайняти місце вище, ніж сторінки з низькою якістю, що повторюють те, що вже і так відомо або ж є копіями інших сайтів. Оновлення Panda встановило новий базовий рівень стандартів якості:

1) Вміст на сторінці повинно мати істотний обсяг. Більший обсяг інформації статистично займає місце вище, ніж містить менше 1500 слів;

2) Інформація, представлена на сайті повинна бути оригінальною. Якщо ви просто копіюєте вміст інших веб-ресурсів, Google покарає за це;

3) Зміст сайту має вносити щось нове до теми. Мало кому буде цікаво в сотий раз перечитувати одне і теж. Для успішного просування контент має бути те, чого немає на інших сайтах;

4) Текст сайту повинен бути орфографічно і граматично правильним і грунтуватися на перевірених фактах;

5) Якщо ви збираєтеся автоматично генерувати контент з бази даних, зміст повинен відповідати описаним стандартам.

2.6.5ОновленняметодуPage Layout (Top Heavy)

Оновлення, вперше випущений в січні 2012 року, передбачає покарання сайтів, які використовують занадто багато реклами у верхній частині сторінки або роблять її надмірно агресивною, відволікає від основного змісту. Це спровоковано великою кількістю скарг від користувачів, яким бувало складно знайти потрібну інформацію і доводилося довго прокручувати сторінку вниз. Даним оновленням Google закликає веб-майстрів розміщувати вміст сайту в центрі уваги. У цьому відношенні велике число реклами заважає зручності засвоєння інформації.

2.6.6 Оновлення методуPenguin

Був випущений в квітні 2012 року. Новий алгоритм, спрямований на боротьбу з пошуковим спамом. Сайти, які використовували спам-методи, були значно знижені в рейтингу або зовсім вилучені з нього.Ще однією особливістю Penguin є здатність аналізувати кількість посилань.

2.6.7 Оновлення методуPirate

З оновленням Pirate, яке було введено в серпні 2012 року, Google знизив рейтинг сайтів, які порушують авторські права та інтелектуальну власність. Для вимірювання цих порушень, Google використовує систему запитів про порушення авторських прав, засновану на Digital Millenium Copyright Act. Правовласники можуть застосовувати інструмент, щоб повідомити і видалити зміст сайту плагіаторів з бази даних Google.

2.6.8ОновленняметодуExact Match Domain(EMD)

Випущено в вересні 2012 року і направлено на боротьбу з доменами, схожими на MFA.

MFA (made-for-adsense) - домен, який створений спеціально для Медійній системи Google. Зазвичай такий домен призначений для якогось одного запиту (або сімейства запитів) і на ньому встановлений Google Adsense. Користувач, який потрапив на цей домен, не бачить нічого, крім реклами і в результаті або закриває сайт, або переходить далі по контекстного оголошенню. Після випуску алгоритму EMD, сайти, що містять в доменному імені запит, були вилучені або дуже істотно знижені в рейтингу.

2.6.9 Оновлення методуPaydayLoan

Випущений в червні 2013 року і спрямований на зменшення сторінок, які містять переспамленние запити. Такі запити часто використовуються веб-майстрами для просування сторінок певної тематики.

Оновлення було запущено в зв'язку з численними скаргами, в яких говорилося, що навіть після впровадження Panda і Penguin чистота видачі залишала бажати кращого.

Розглянемо це оновлення на звичайному прикладі. Припустимо, вам потрібно купити двері. Якщо ввести запит, Google видасть фотографії дверей. З них: 2-3 сторінки, де безпосередньо можна купити двері, 3-4 сайту компаній-виробників дверей і 2-3 сайту про те, як вибрати і поміняти двері. Якби не було оновлення Payday Loan, ви б побачили 15-20 запитів на одну тематику (наприклад, де купити двері).

Критерії, за якими відбувається відбір таких сайтів, Google розкривати не хоче, але даний алгоритм явно спростив життя користувачам пошукової системи.

2.6.10ОновленняметодуHummingbird

З вересня 2013 року Google реалізував заміну алгоритму пошуку, яка була названа Hummingbird. Основні оновлення, як Panda і Penguin, були інтегровані з цим новим алгоритмом. Ім'яHummingbird вибрали в якості синоніма для опису гнучкості, точності і швидкості нового оновлення.

Замість того, щоб повертати точні відповіді на запити, використовуючи введені користувачем ключові слова (як це було раніше), Google інтерпретує наміри і контекст пошуку. Мета полягає в тому, щоб зрозуміти сенс пошукового запиту користувача і повертати відповідні результати. Це означає, що точні співпадіння ключових слів стають менш важливими на користь пошуку наміри. Як приклад: якщо ви вводите запит «погода», то навряд чи очікуєте отримати повне пояснення самого терміна. Швидше в даному випадку маються на увазі погодні умови.

2.6.11ОновленняметодуMobilegeddon

Було випущено в квітні 2015 року. Цеоновленнявпливаєтількинамобільнийпошук, вонодаєперевагусторінкам, дружнімдомобільних пристроїв.

У поточному стані, оновлення не впливає на результати пошуку зі стаціонарних комп'ютерів або планшетів. На відміну від Panda або Penguin, алгоритм працює в режимі реального часу.

Існує спеціальний тест, за допомогою якого веб-майстри можуть перевірити сумісність свого сайту з мобільними пристроями. Також можна використовувати звіти про мобільний юзабіліті в Google Webmaster Tools, тільки вони можуть працювати із затримкою.

3 ПРОГРАМНОЇ РЕАЛІЗАЦІЇ ПОШУКОВОЇ СИСТЕМИ

3.1 Обґрунтування вибору середовища програмної реалізації

У рамках дипломної роботи був розроблений методдля вибору ключових слів контекстного пошуку. Для реалізації було обране середовище WebStorm. Це обумовлено тим, що WebStorm найбільш підходить для розробки на JavaScript, якtй в свою чергу є дуже актуальним у наш час.

JavaScript мультіпарадігменний мову. Він об'єднує об'єктно-орієнтована і функціональний підходи. Так-же є структурне програмування і багато іншого. Це обумовлено високою гнучкістю мови. Функції є об'єктами першого роду, змінні можуть змінювати тип, об'єкти - отримувати нові властивості на льоту. Добре це чи погано? З одного боку, це створює проблеми з розумінням у програмістів, які звикли до визначеності таких мов, як Java або C #. Але це мала плата за ефективність. JavaScript, в більшості випадків, дозволяє створювати додатки набагато швидше, ніж його строго тіпізірованія товариші. Наведу лише один приклад: я працював в компанії, що розробляє біржовий термінал. Необхідно було створити настільний додаток на .Net C # і його точну копію в веб на JavaScript. C # додаток розробляла команда з 12 програмістів протягом двох років, JavaScript - команда з 3 програмістів протягом року. Можна говорити про різницю в кваліфікації, про те, що можливо впливали інші чинники, але так чи інакше різниця в 8 разів показова.

WebStorm – середовище для розробки на JavaScript, яка підходить для client-side-розробки, створення додатків на Node.js і мобільних додатків на React Native.

Головнаперевага WebStorm - це зручний і розумний редактор для JavaScript, HTML і CSS, який також підтримує TypeScript, CoffeeScript, Dart, Less, Sass і Stylus і фреймворки, наприклад, Angular, React і Vue.js.

WebStorm, як і інші IDE на платформі IntelliJ IDEA, робить розробку простіше і зручніше. WebStorm забезпечуючи підсвічування і

автодоповнення коду, перевіряє його на помилки, допомагає швидко навігіроваться за проектом і безпечно вносити зміни за допомогою рефакторингов. У WebStorm є інструменти для налагодження коду і інтеграція з системами управління версіями.

WebStorm по-справжньому розуміє структуру вашого проекту і код, виявляє можливі проблеми ще до того, як ви відкрили проект в браузері, і пропонує їх рішення.

Вбудовані в IDE інструменти для тестування допоможуть в розробці і зроблять її зручніше і продуктивніше.

3.2 Програмна реалізація

При розробці методу контекстного пошуку для новинного порталу були прийняті до відома наступні важливі аспекти:

1) обсяг інформації після обробки пошукового запиту;

2) обсяг самого запиту.

Виходячи з цього, метод був спрощений таким чином, щоб користувач не зміг відправити на обробку форму, у якої понад шістдесяти символів. Суть самого методу полягає в обробці вхідного рядка, і виявленні в ньому ключових слів для пошуку.

Рядок розбивається на слова для спрощення подальшого аналізу. У цьому масиві слів, в першу чергу, видаляються ті слова, які не несуть смислового навантаження. Наприклад, «що», «або», «та», «у» і т. д. Після цього, масив поелементно аналізується на закінчення слів, щоб максимально покращити пошук для всіх відмінків слова. Після аналізу, наш результуючий масив приводиться до типу рядка і запит вже відправляється для пошуку збігів.

Основним методом для пошуку є повнотекстовий пошук в MySQL.Повнотекстові індекси в MySQL позначаються як індекси типу FULLTEXT. Ці індекси можуть бути створені в стовбцях VARCHAR і TEXT

під час створення таблиці командою CREATE TABLE або додані пізніше за допомогою команд ALTER TABLE або CREATE INDEX. Завантаження великих масивів даних в таблицю буде відбуватися набагато швидше, якщо таблиця не містить індекс FULLTEXT, який потім створюється командою ALTER TABLE (або CREATE INDEX). Завантаження даних в таблицю, яка вже має індекс FULLTEXT, буде повільнішою.Повнотекстовий пошук виконується за допомогою функції MATCH ().

Функція MATCH () перевіряє наявність в природній мові, порівнюючи рядок з вмістом тексту (сукупність одного або більше стовпців, включених в індекс FULLTEXT). Рядок пошуку задається як аргумент в вираженні AGAINST (). Пошук виконується без урахування регістру символів. Для кожного рядка стовпця в заданій таблиці команда MATCH () повертає величину релевантності, тобто ступінь подібності між рядком пошуку та текстом, що містяться в цьому рядку зазначеного в списку оператора MATCH () стовпчика.

Коли команда MATCH () використовується у виразі WHERE (див. Приклад вище), повернуті рядки стовпців автоматично сортуються, починаючи з найбільш релевантних. Величина релевантності є невід'ємне число з плаваючою крапкою. Релевантність обчислюється на основі кількості слів в цьому рядку стовпця, кількості унікальних слів в цьому рядку, загальної кількості слів у тексті та числа документів (рядків), що містять окреме слово.

Пошук можливий також в логічному режимі, це пояснюється нижче в даному розділі. Рядки повертаються в порядку зменшення релевантності.У наступному прикладі показано, як витягувати величини релевантності в явному вигляді. У разі відсутності виразів WHERE і ORDER BY повертаються рядки не упорядковуються.

Кожне правильне слово в наборі перевіряються текстів і в даному запиті оцінюється відповідно до його важливістю в цьому запиті або наборі текстів. Таким чином, слово, присутнє в багатьох документах, буде мати

меншу вагу (і навіть, можливо, нульовий), як має більш низьке смислове значення в даному конкретному наборі текстів. З іншого боку, рідко зустрічається слово отримає більш високий вагу. Потім отримані значення ваг слів об'єднуються для обчислення релевантності цього рядка стовпця.

Обмеженнями для повнотекстового пошуку є:

1) всі параметри функції MATCH () повинні бути стовпцями однієї і тієї ж таблиці, тобто частиною одного і того ж індексу FULLTEXT, за винятком роботи MATCH () в режимі IN BOOLEAN MODE;

2) список стовпців в команді MATCH () повинен точно відповідати списку стовпців у визначенні індексу FULLTEXT для цієї таблиці, за винятком роботи даної функції MATCH () в режимі IN BOOLEAN MODE;

3) аргумент у вираженні AGAINST () повинен бути незмінною рядком.

3.3 Інструкція користувача

Новинний портал являє собою веб-сервіс зі зручним розумним пошуком. Всі користувачі можуть з легкістю знайти потрібну їм статтю без якихось ускладнень.

На головній сторінці сайту розміщенні найновіші новини. За допомогою цього, користувач може бачити «свіжі» новини, і не пропустити оновлення статтей.

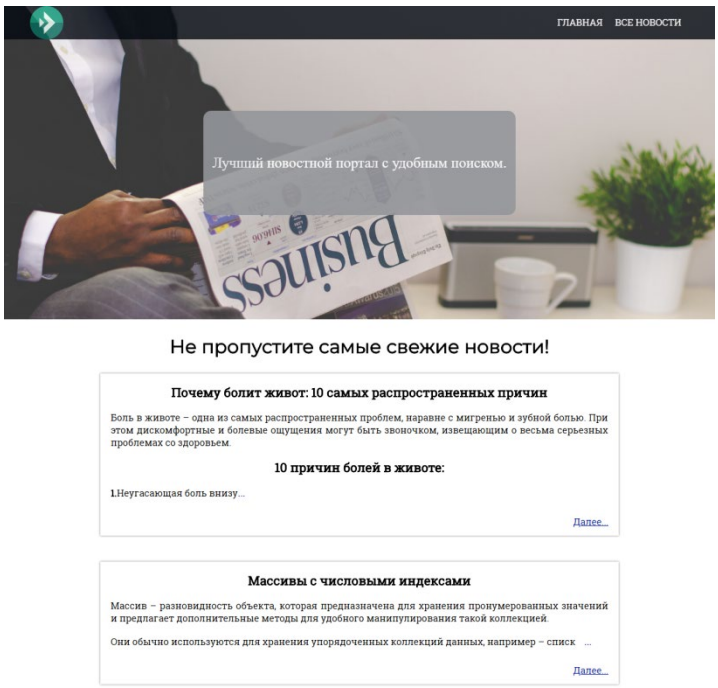


Рисунок 3.1 – Головна сторінка з останніми новими новинами

Щоб скористатися пошуком, треба зайти на сторінку «Усі новини», після чого ввести свій запит на пошук, до якого можуть входити лише символи кирилиці та цифри.

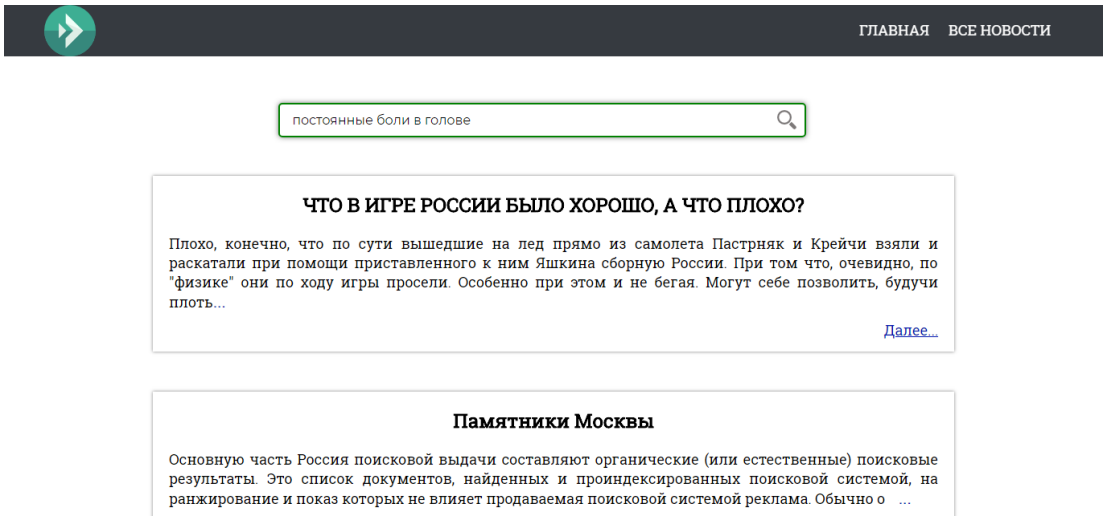


Рисунок 3.2– Пошуковий запит

Вводити коричтувач може не більше 60 символів, інакше запит не буде оброблено.Після цього треба натиснути кнопку пошуку у правій частині форми.

Користувач потрапить на сторінку з результатами, де може вибрати будь-яку с пропонованих новин для читання.

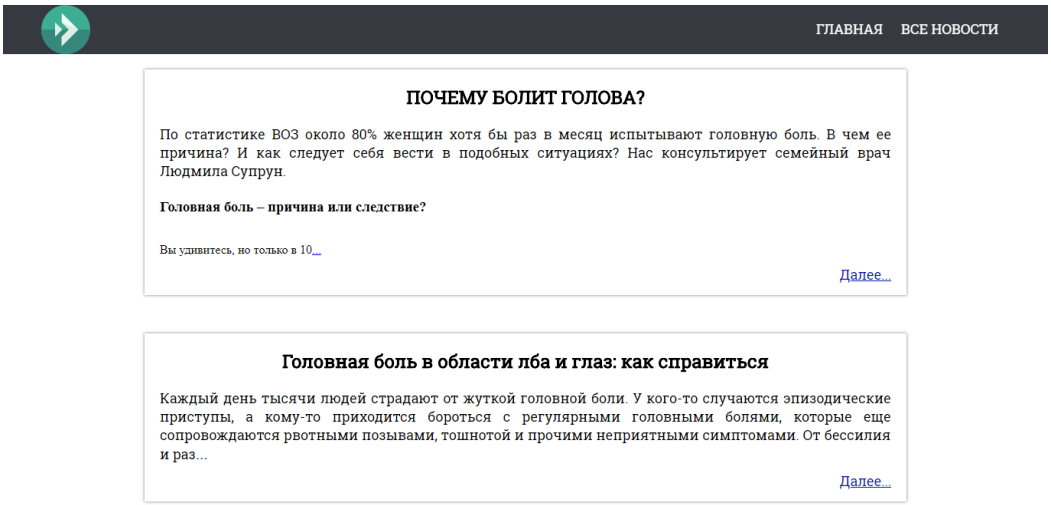


Рисунок 3.3 – Сторінка з результатами пошуку

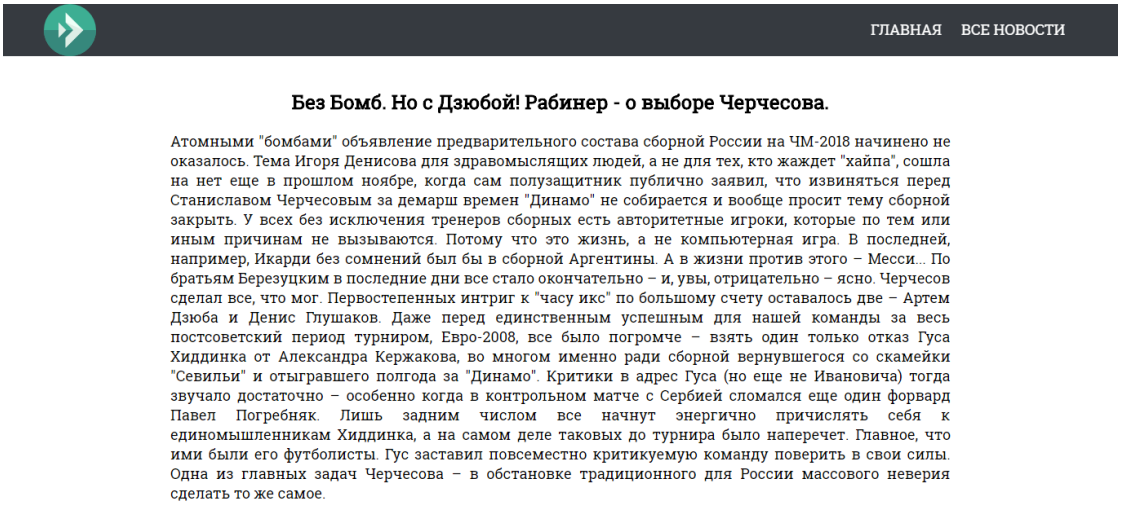


Рисунок 3.4 – Сторінка вибранної новини

3.4 Тестування розробленої моделі

У розділі "Всі новини" можна знайти поле для пошуку потрібної вам статті за ключовими словами.

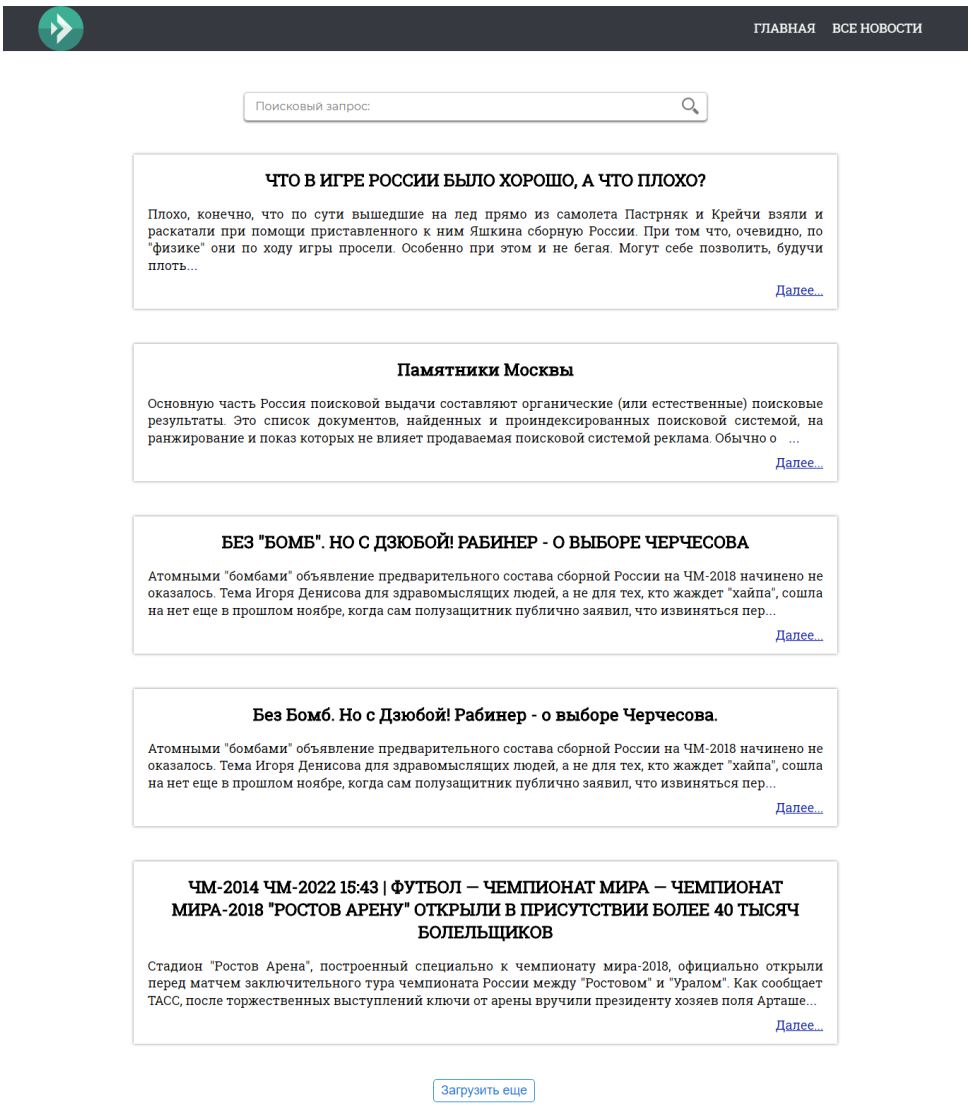


Рисунок 3.5 – Сторінка «Усі новини»

Також на сторінці виводиться по 5 статтей. Натиснувши на кнопку «Завантажити ще» с БД завантажується ще 5 додаткових статтей. Якщо новин в БД більше немає, то користувач побачить відповідне повідомлення.

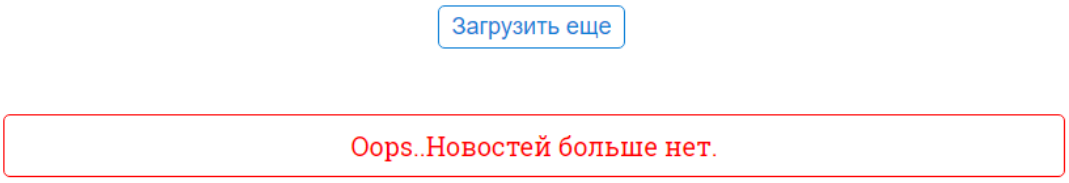


Рисунок 3.6 – Повідомлення користувачу про те, що новин в БД більше немає

При спробі ввести в форму пошуку латинські символи, форма почне підсвічуватися червоним кольором, що говорить користувачеві про валідність даної форми.

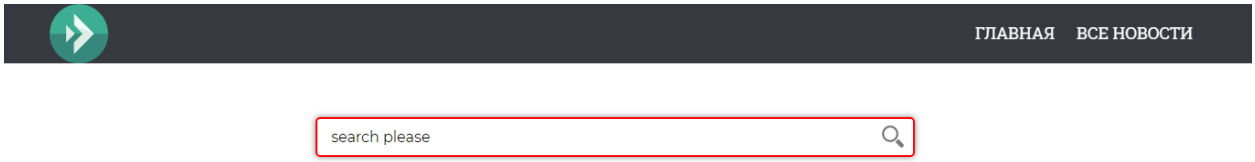


Рисунок 3.7 – Валідність форми пошуку.

Якщо ж користувач спробує відправити такий запит, він отримає відповідне повідомлення.

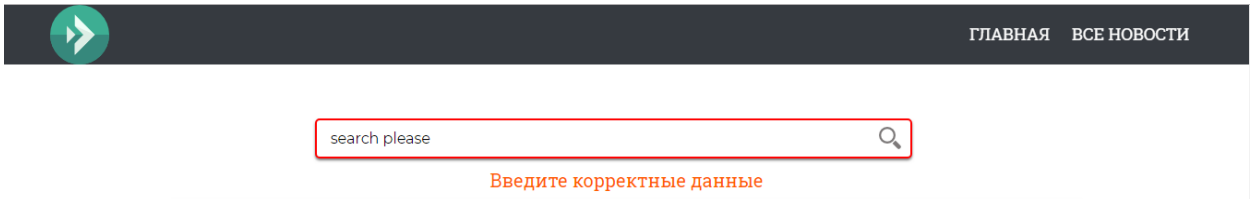


Рисунок 3.8 – Повідомлення про помилку при спробі відправити неправильні дані

Форма підтримує усі символи кирилиці та цифри.

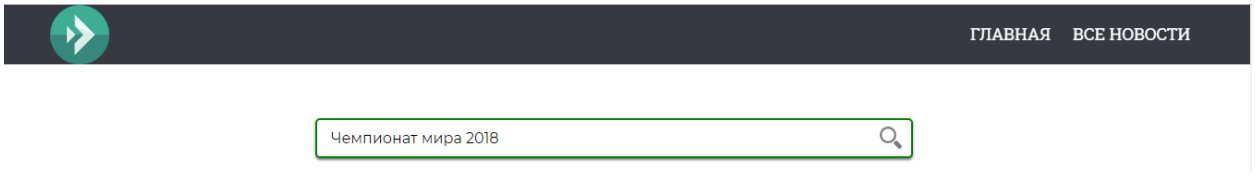


Рисунок 3.9 – Данні введено вірно

Після відправки правильних даних, користувач потрапить на сторінку з результатами, де зможе вибрати найбільш відповідну статтю, відкривши її на новій сторінці.

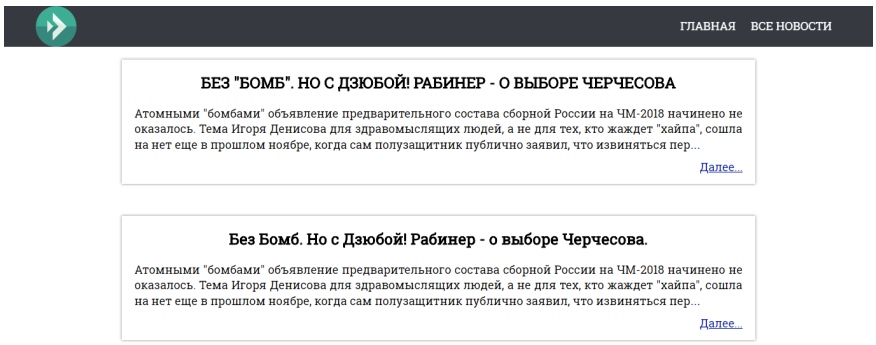


Рисунок 3.10 – Сторінка з результатами пошуку на запит «Чемпионат мира 2018»

4 ОХОРОНА ПРАЦІ

4.1 Аналіз потенційно небезпечних і шкідливих виробничих факторів , що впливають на персонал

Персональні ЕОМ типу IBM PC AT має наступні характеристики:

- споживана потужність 350 Вт;
- робоча напруга 220 В;
- напруга джерел живлення +12 В, -12 В, 5 В;
- робоча частота 50 Гц.

Виходячи з приведених характеристик, очевидно, що для користувача існує небезпека поразки електричним струмом у разі недбалого поводження з комп'ютером і порушення правил експлуатації (невиконання огляду відкритих частин ПЕВМ, що знаходяться під напругою або знятих для ремонту вузлів і т. д.).

Джерелами підвищеної небезпеки можуть служити наступні елементи:

- розподільний щит;
- джерела живлення;

У відповідності з ДСанПіН 3.3.2-007-98 [11] до легкої фізичної роботи відносяться всі види діяльності, вироблювані сидячи і не вимагаючи фізичної напруги. Робота користувача розробленого пакету програм відноситься до категорії Іа.

Згідно з ДСТУ Б А.3.2-13:2011 [17] приміщення для ПЕОМ по ступеню небезпеки поразки людини електричним струмом відноситься до приміщень без підвищеної небезпеки (немає струмопровідної полови, вогкості, підвищеної температури, можливості одночасного дотику до корпусів устаткування з “землею” і до струмонесучих частин).

У відповідності з НПАОП 0.00-7.15-18 [12] при обслуговуванні ПЕВМ мають місце фізичні і психофізичні небезпечні, а також шкідливі виробничі чинники:

– підвищене значення напруги в електричному ланцюзі, замикання якого може відбутися через тіло людини;

– підвищений рівень статичної електрики;

– підвищений рівень електромагнітних випромінювань;

– підвищена або знижена температура повітря робочої зони;

– підвищена або знижена рухливість повітря;

– підвищена або знижена вогкість повітря;

– відсутність або недолік природного світла;

– підвищена пульсація світлового потоку;

– недостатня освітленість робочого місця;

– підвищений рівень шуму на робочому місці;

– розумове перенапруження;

– емоційні навантаження;

– монотонність праці.

Щодо до впливу на довкілля, то програмний засіб, який було розлоблено під час дипломного проекту на довкілля ніяк не впливає.

Діяльність за темою магістерської роботи в процесі її виконання впливає на навколишнє природнє середовище і регламентується нормами діючого законодавства: Законом України «Про охорону навколишнього природного середовища», Законом України «Про забезпечення санітарного та епідемічного благополуччя населення», Законом України «Про відходи», Законом України «Про охорону атмосферного повітря», Законом України «Про захист населення і територій від надзвичайних ситуацій техногенного та природного характеру», Водний кодекс України.

Основним екологічним аспектом в процесі діяльності за даними спеціальностями є процесив пливу на атмосфернеповітря та процеси поводження з відходами, які утворюються, збираються, розміщуються, передаються на відалення (знешкодження), утилізацію, тощо вІТгалузі.

Вплив на атмосферне повітря при нормальних умовах праці не оказує,бо не має в приміщенні сканерів, принтерів та інших джерел викиду забруднюючих речовин в повітря робочої зони.

Відходи в міру їх накопичення збирають у тару, відповідну класу небезпеки, з дотриманням правил безпеки, після чого доставляють до місця тимчасового зберігання відходів відповідно до затвердженої схеми їх розміщення Зазначені для зберігання відходів місця чи об'єкти повинні використовуватися лише для заявлених відходів.

Не допускається зберігання відходів у невстановлених схемою місцях, а також перевищення норм тимчасового зберігання відходів.

Способи тимчасового зберігання відходів визначаються видом, агрегатним станом і класом небезпеки відходів:

– відходи III класу небезпеки зберігаються в тарі, яка забезпечує локалізацію зберігання, дозволяє виконувати вантажно-розвантажувальні і транспортні роботи і виключає поширення в ОС шкідливих речовин;

– відходи IV класу небезпеки можуть зберігатися відкрито на промисловому майданчику у вигляді конусоподібної купи, звідки їх автонавантажувачем перевантажують у самоскид і доставляють на місце утилізації або захоронення;

В разі часового зберігання відходів у стаціонарних складах або промислових приміщеннях повинні бути забезпечені санітарно-гігієнічнимиетичні вимоги до повітря робочої зони згідно з[12].

Не допускається змішування відходів різних видів і класів небезпеки з будівельними і побутовими відходами, відходами дерев'яної, металевої, синтетичної тари, відходами текстильних матеріалів (старийспецодяг, ганчірки) і інш.

Проведення заготовки, здачі, переробки та реалізації метало брухту встановлен іокремо Законом України «Про металобрухт».

Всі відходи, що утворюються в процесі діяльності/роботи, підлягають обліку.

Вимоги безпеки при поводженні з відходами:

Під час роботи з відходами (прибирання виробничих приміщень, збір і сортування, навантаження, транспортування, розвантаження та ін.) працівники та обслуговуючий персонал підприємстваповинні бути забезпечені засобами індивідуального захисту та дотримуватися вимог інструкцій з охорони праці, щодіють на підприємстві.

Наведено перелік деяких відходів, які передаються на утилізацію організаціям, які мають ліцензію на поводження з відходами як вторинної сировини:

- лом і кускові відходи міді, бронзи, латуні, алюмінію, свинцю;
- брухтчорнихметалів;
- макулатура;
- склобій;
- матеріали текстильні вторинні;
- відходи деревини кускові
- відпрацьовані фільтрувальні засоби індивідуальногозахисту
- відпрацьовані вогнегасники
- матеріали пакувальні вторинні

Відвантаження таких відходів здійснюється відповідно до договору (контракту).

Побутові та будівельні відходи вивозяться на полігон твердих побутових відходів міста, також відповідно до договору з комунальним дорожньо-експлуатаційним управлінням.

Особи, винні в порушенні встановленого порядку поводження з відходами (порушення правил обліку відходів, самовільне складування і видалення відходів, передача відходів в інші підприємства/організації з порушенням встановлених правил), згідно законодавства несуть дисциплінарну, адміністративну або кримінальну відповідальність.

4.2 Заходи щодотехнікибезпеки

Основним небезпечним чинником при роботі з ЕОМ є небезпека поразки людини електричним струмом, яка усугубляє тим, що органи чуття людини не можуть на відстані знайти наявності електричної напруги на устаткуванні.

Проходячи через тіло людини, електричний струм надає на нього складну дію, що є сукупністю термічної (нагрів тканин і біологічних середовищ), електролітичної (розкладання крові і плазми) і біологічної (роздратування і збудження нервових волокон і інших органів тканин організму) дій.

Ступінь ураження людини електричним струмом залежить від наступних факторів:

– значення сили струму;

– електричного опору тіла людини і тривалості протікання через нього струму;

– роду і частоти струму;

– індивідуальних властивостей людини і навколишнього середовища.

– Даним проектом передбачаються наступні технічні способи і засоби, застережливі поразки людини електричним струмом:

– заземлення електроустановок;

– занулення;

– захисне відключення;

– електричне розділення сітей;

– використовування малої напруги;

– ізоляція струмоведучих частин;

– огорожа електроустановок.

Проведемо розрахунок заземлюючого пристрою.

Початкові дані для розрахунку заземлюючого пристрою:

– напруга установки, що заземляється, - 220В;

– режим нейтралу мережі - з ізольованою нейтралою;

– питомий опір ґрунту – 100 Ом·м(суглинок);

– гранично допустимий опір заземлюючого пристрою - 4 Ом;

– характеристика кліматичної зони (III):

а) середня багаторічна низька температура, °С- від –14 до -10;

б) тривалість замерзання вод, дні - 150;

в) коефіцієнт сезонності для вертикального електроду завдовжки 3м -1,5.

Визначимо розрахунковий опір ґрунту (Ом·м) по формулі (4.1).

$$\rho_{расч} = \psi \cdot \rho = 1{,}5 \cdot 100 = 150 Ом \cdot м \qquad (4.1)$$

де ρ - питомий опір ґрунту;

ψ_i – кліматичний коефіцієнт, що враховує стан ґрунту під час вимірювань (таблиця 4 [12]).

Розрахуємо опір розтіканню одиночного трубчастого заземлювача по формулі (4.2).

$$R_{з.1} = (\frac{\rho_{расч}}{2 \cdot \pi \cdot l}) \cdot \ln(4 \cdot \frac{l}{d}) \qquad (4.2)$$

де l – довжина заземлювача (l=5м);

d – діаметр труби і стрижня (d=0,05м);

$$R_{з.1} = (\frac{\rho_{расч}}{2 \cdot \pi \cdot l}) \cdot \ln(4 \cdot \frac{l}{d}) = (\frac{150}{2 \cdot 3{,}14 \cdot 5}) \cdot \ln(4 \cdot \frac{5}{0{,}05}) = 28{,}6 \text{ Ом}$$

Розрахуємо кількість паралельно сполучених одиночних заземлювачей по формулі (4.3).

$$n = \frac{R_{з.1}}{R_{доп} \cdot \eta} = \frac{28,6}{4 \cdot 0.47} = 15,2 \qquad (4.3)$$

де Rдоп=4. – самий допустимий опір заземлюючого пристрою;

η - коефіцієнт використовування ґрунтового заземлення (для шістки заземлювачейη=0,47).

Округлятимемо отримане значення у більшу сторону n=[15,2]=16.

Розрахуємо довжину горизонтальної сполучної смуги по формулі (5.4).

$$L = a \cdot (n-1) = 3 \cdot (16-1) = 45м \qquad (4.4)$$

де a – відстань між вертикальними заземлювачами (a=3м);

n – кількість вертикальних заземлювачей (n=16).

Розрахуємо опір сполучної смуги по формулі (4.5).

$$R_n = \frac{\rho_{расч}}{2 \cdot \pi \cdot l} \cdot \ln(\frac{L^2}{d \cdot h}) \qquad (4.5)$$

де d – еквівалентний діаметр смуги шириною l=5 (d=0,05м);

h – глибина заставляння смуги (h=0,8м).

$$R_n = \frac{\rho_{расч}}{2 \cdot \pi \cdot l} \cdot \ln(\frac{L^2}{d \cdot h}) = \frac{150}{2 \cdot 3,14 \cdot 5} \cdot \ln(\frac{45^2}{0,05 \cdot 0,8}) = 51,7 \text{ Ом}$$

Розрахуємо результуючий опір заземлюючого електроду з урахуванням сполучної смуги по формулі (4.6).

$$R_{гр} = \frac{R_{з.1} \cdot R_n}{R_{з.1} \cdot \eta_n + R_n \cdot n \cdot \eta_з} \le R_{доп} \qquad (4.6)$$

де η_n – коефіцієнт використовування сполучної смуги (для 6-і заземлювачей$\eta_{n=0,27}$).

$$R_{гр} = \frac{R_{з.1} \cdot R_n}{R_{з.1} \cdot \eta_n + R_n \cdot n \cdot \eta_з} = \frac{26,6 \cdot 51,7}{26,6 \cdot 0,27 + 51,7 \cdot 16 \cdot 0,47} = 3,47 \text{ Ом}$$

$3,47<4 \Rightarrow$ умова забезпечення електробезпеки персоналу виконується.

Таким чином, остаточна кількість заземлювачей 15 шт.

4.3 Заходи, що забезпечують виробничу санітарію і гігієну праці

Підвищення працездатності людини і збереження його здоров'я забезпечується стабільними метеорологічними умовами.

Мікроклімат виробничих приміщень – це поєднання температури, вогкості і швидкості руху повітря, а також температури навколишніх поверхонь. Значне коливання параметрів мікроклімату приводить до порушення систем кровообігу, нервової і пітовидільної, що може викликати підвищення або пониження температури тіла, слабкість, запаморочення і навіть непритомність.

В приміщенні для виконання робіт операторського типу, пов'язаних з нервово-емоційною напругою, проектом передбачається дотримання наступних нормованих величин параметрів мікроклімату (див. таблицю 4.1).

Таблиця 4.1 - Оптимальні параметри мікроклімату в робочій зоні виробничого приміщення для категорії робіт 1

Період року	Температура, Ос	Відносна вогкість %	Швидкість руху повітря, м/с
Холодний	22.24	40.60	0,1
Теплий	23.25	40.60	0,1

Оскільки в приміщенні немає джерел виділення шкідливих речовин, можна використовувати природну вентиляцію. Площа приміщення складає 18 м^2. Для забезпечення прийнятних параметрів мікроклімату в приміщенні з такою площею можна використовувати 1 кондиціонер типу БК-2000.

Спектр випромінювання монітора комп'ютера включає рентгенівську, ультрафіолетову, інфрачервону області, а також широкий діапазон хвиль інших частот. Небезпека рентгенівського проміння нехтує мала, оскільки цей вид випромінювання поглинається речовиною екрану.

Для зниження дії електромагнітного випромінювання пропонується захист часом і відстанню. Захист часом передбачає обмеження часу перебування людини в зоні дії полів. Тривалість роботи на ПЕОМ повинна складати не більше 3.5–4.5 години.

Також необхідно забезпечити раціональне освітлення в робочому приміщенні. В проекті, що розробляється, передбачається використовувати суміщене освітлення. В світлий час доби приміщення освітлюватиметься через віконні отвори, в решту часу використовуватиметься штучне освітлення.

Штучне освітлення в робочому приміщенні передбачається здійснювати з використанням люмінесцентних джерел світла в світильниках загального освітлення, оскільки люмінесцентні лампи володіють високою світловою віддачею до 75 Лам/Вт і більш, тривалим терміном служби до 10000 годин, спектральним складом випромінюваного світла, близьким до сонячного.

Зорова робота оператора ПЕВМ відповідно до [15] відноситься до розряду Va з світловим потоком Фл=3120 кожна. Нормована освітленість на робочому місці (Ен) при загальному освітленні складає 200 лк.

Проведемо розрахунок кількості світильників в робочому приміщенні завдовжки a=6 м, шириною b=3 м, заввишки c=4 м. Формула розрахунку штучного освітлення при горизонтальній робочій поверхні методом світлового потоку (4.7):

$$Фл = \frac{Ен \cdot S \cdot Z \cdot K}{N \cdot U \cdot M} \quad (4.7)$$

де *Фл*– світловий потік, Лм;

Ен– нормована освітленість;

S – площа підлоги, кв.м;

Z=1.1–1.3 - поправочний коефіцієнт світильника (для стандартних світильників);

K– коефіцієнт запасу, що враховує зниження освітленості в процесі експлуатації світильників;

N – число світильників;

U=0.55–0.6 – коефіцієнт використовування, залежний від типу світильника, показника індексу приміщення і др.;

M – число ламп в світильнику.

З формули (4.7) виразимо N і визначимо кількість світильників для даного приміщення:

$$N = \frac{Ен \cdot S \cdot Z \cdot K}{Фл \cdot U \cdot M}$$

$$N = \frac{200 \cdot 18 \cdot 1{,}2 \cdot 1{,}5}{3120 \cdot 0{,}6 \cdot 2} = 1{,}7$$

Виходячи з цього, рекомендується використовувати 2 світильники. Світильники слід розміщувати рядами, бажано паралельно стіні з вікнами. Схема розташування світильників зображена на рис. (4.1).

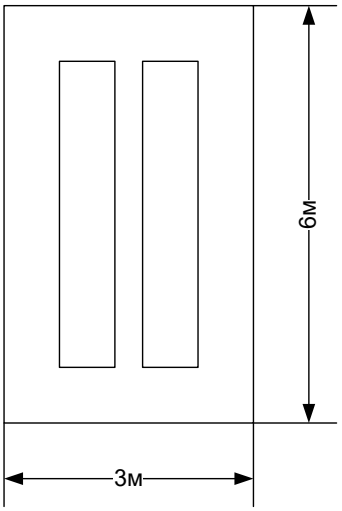


Рисунок 4.1 – Схема розташування світильників

Зниження шуму можна добитися раціонально розпланувавши приміщення, установкою устаткування на спеціальні амортизуючи прокладки. Згідно вимогам [14] рівні звуку не повинні перевищувати 50 дБ.

Для зниження стомлюваності обслуговуючого персоналу в приміщеннях, де розташовані обчислювальні засоби передбачаються використовувати спокійні колірні поєднання і покриття, що не дають відблисків. Від електромагнітного випромінювання, витікаючого від ПЕОМ, використовуються захисні екрани.

Для забезпечення чистоти повітря і відповідних мікрокліматичних умов пропонується застосувати приточування-витяжну вентиляцію. Для зменшення дії шкідливих речовин і загазованості для роботи з розплавленими матеріалами робоче місце забезпечується примусовою витяжною вентиляцією. Цей метод забезпечує притоку потрібної кількості свіжого повітря(30 мм^3 /ч на одного працюючого).

Кількість повітря, яка необхідна подавати в приміщення для забезпечення необхідних параметрів повітряного середовища, визначається на підставі кількість тепла, вологи і шкідливих речовин, що поступають в приміщення, а також враховуючи видалення повітря місцевими відсмоктуваннями від устаткування, загальнообмінною вентиляцією.

4.4 Рекомендації по пожежній профілактиці

Пожежі представляють небезпеку для життя людини і зв'язані як з матеріальними втратами, так і з відмовою засобів обчислювальної техніки, що спричиняє за собою порушення ходу технологічного процесу.

Горючими матеріалами в приміщенні, де розташовані ПЕОМ, є:

– поліамід - матеріал корпусу мікросхеми. Горюча речовина. Температура самозапалення 420 °С, енергія запалення 2мДж;

– полівінілхлорид - ізоляційний матеріал. Горюча речовина. Температура самозаймання 480 єС, енергія запалення 50мДж;

– склостоліт ДЦ - матеріал друкарської платні. Складногорючий матеріал;

– пластикат кабельний No.489 - матеріал ізоляції кабелю. Складногорючий матеріал. Температура самозаймання 1500 єС;

– плита деревостружкова - будівельний і обробний матеріал, матеріал з якого виготовлені меблі. Складнозапалений матеріал. Показник горючості 1.8;

– папір – довідкова і робоча документація, література. Горючий матеріал. Показник горючості більше 2.1.

Відповідно до [16] приміщення відноситься до категорії В (пожежовибухонебезпечної).

Джерелами запалення можуть бути:

– іскри при замиканні і розмиканні ланцюгів;

– іскри і дуги коротких замикань;

– перегріви від тривалого перевантаження і наявності перехідного опору;

– розряди статичної електрики.

Для того, щоб зупинити реакцію горіння, порушують умови її виникнення і підтримки. Звичайно для гасіння використовуються порушення двох основних умов сталого стану – пониження температури і режим руху

газів. Пониження температури може бути досягнутий шляхом введення речовин, які поглинають багато тепла в результаті випаровування і дисоціації (наприклад, вода, порошки).

При повному тому, що згоряє органічних сполук утворюються C, SO, H Про, N,а при тому, що згоряє неорганічних з'єднань – оксиди. Залежно від температури плавлення і тривалості реакції можуть знаходиться або у вигляді розплавів (Al O, Ti O), або підійматися в повітря у вигляді диму (P O, Na Про, MgO).

Склад продуктів неповного згоряє горючих речовин складений і різноманітний. Це можуть бути горючі речовини:

– H, C, CH;

– атомарний водень і кисень;

– різні радикали – BIH, CH .

Продуктами неповного згоряє можуть бути також оксиди азоту, спирти, альдегіди, кетони і високотоксичні з'єднання, наприклад, синильна кислота.

Для захисту персоналу від дій небезпечних і шкідливих чинників пожежі проектом передбачено застосування промислового фільтруючого протигаза з коробкою марки В (жовтий).

До системи запобігання пожежі відносяться: запобігання утворення горючого середовища і освіти в горючому середовищі джерел запалення, забезпечення пожежебеспеки устаткування.

Щоб запобігти пожежі в обчислювальних центрах, проектом пропонується виконання наступних вимог:

– електроживлення ЕОМ має автоматичне блокування відключення електроенергії на випадок перегріву системи, що може бути результатом зупинки системи охолоджування і кондиціонування;

– система вентиляції обчислювальних центрів обладнується блокуючими пристроями, що забезпечують її відключення на випадок пожежі. Система обладнується вогнеперегороджуючими клапанами;

– застосування устаткування, що задовольняє вимогам електростатичної іскробезпеки [17];

– після закінчення роботи, перед закриттям приміщення, всі електроустановки і персональні комп'ютери відключаються від сіті електроживлення;

– в приміщеннях обчислювальних центрів забороняється:

1) влаштовувати електророзетки на основах, що згоряють;

2) використовувати синтетичні доріжки і килими;

3) користуватися побутовими електронагрівальними приладами;

4) захаращувати евакуаційні виходи і проходи;

5) влаштовувати на вікнах глухі грати;

6) залишати без нагляду включену в електромережу апаратуру, що використовується для вимірювань і нагляду.

Для протипожежного захисту проектом пропонується обладнати приміщення площею 18 м2, яке відноситься до категорії В, автоматичною протипожежною сигналізацією із застосуванням датчиків сповіщення РІД-1 (оповіщувач димовий іонізаційний) в кількості 1 штуки і застосовується в первинних засобах пожежегасінні. Площа контрольована оповіщувачем150 м2.

Крім того, необхідно проводити навчання робочого персоналу правилам пожежної безпеки.

Розрахуємо вірогідність виникнення пожежі у виробничому приміщенні у разі запалювання транзистора:

$$Q = л \cdot Т \cdot Ркз/отк \cdot Qвоспл \cdot Рзащ \qquad (5.8)$$

де л – інтенсивність відмов пожежеопасних ЕРІ;

Т – час роботи пожежеопасного ЕРІ за оцінюваний інтервал часу;

Ркз/отк - умовна вірогідність виходу EPI в стан короткого замикання при його відмові;

Qвоспл - вірогідність запалювання EPI, що знаходиться в стані короткого замикання;

Рзащ – вірогідність відмови захисту пожежеопасного EPI. Якщо захист відсутній, Рзащ приймається рівній 1.

Вірогідність виникнення пожежі у разі запалювання транзистора:

$$Q = 1 \cdot 10^{-6} \cdot 1 \cdot 10^{-4} \cdot 0.1 \cdot 1 \cdot 10^{-4} = 1 \cdot 10^{-15}$$

Розрахована вірогідність виникнення пожежі значно менше допустимої, яка складає $1 \cdot 10^{-6}$.

В даному розділі були проаналізовані небезпечні і шкідливі виробничі чинники, що роблять вплив на персонал, розроблені заходи щодо техніки безпеки, заходу, забезпечуючи виробничу санітарію і гігієну праці, а також заходи щодо пожежної профілактики.

ВИСНОВКИ

В дипломній роботі були проведені дослідження щодо актуаль-ності, доцільності та практичності систем інформаційного пошуку в Інтернеті. Був проведений аналіз технологій для реалізації веб-системи.

Зокрема було проведене порівняння з найпопулярнішим постачальником пошукових систем. В результаті були зроблені висновки, щодо можливостей використання вказаних технологій для рішення різних типів задач.

Пошукові системи Інтернету можуть використовувати різні методи для зручного пошуку інформації, і кожен з них має як свої плюси, так і мінуси. Якщо говорити про простий пошук, то він буде актуальним при невеликому об'єму даних. Також при вивченні цього пошуку було виявлено, що він найбыльше підходить під час простих операцій, наприклад пошуку теоритичних відомостей.

Якщо казати про розширений пошук, то тут річ йде вже про більш складний алгоритмічний метод, який включає в себе логічні оператори. Зазвичай пошуковий запит такого методу включає в себе групу слів, серед яких потрібно вибрати та зв'язати ключові слова, задля точних результатів пошуку. Такий пошук є дуже гнучким, а, отже, засвоївши його один раз, ви можете користуватись ним ще дуже довго, переключивши машину у відповідний режим.

Щодо контекстного пошуку, то він є одним з найскладніших методів пошуку інформації в Інтернеті. Зазвичай цей метод краще реалізовувати за допомогої нейронних мереж. Навчаючи такі мережі, пошук можна зробити максимально точним. Пошуковий запит такого методу потрібно обробляти таким чином, щоб машина за сенсом могла зрозуміти, що саме має на увазі користувач. Такий пошук реалізують найвідоміші компанії у цій галузі, наприклад, Google. Їхні методи використовують нейронні мережі, як і було сказано вище.

У ході виконання дипломної роботи була розроблена веб-система новинного порталу, яка дозволяє переглядати та робити пошук статтей, які зацікавлять користувача. Був проведений аналіз результатів та зроблені висновки щодо доцільності та актуальності теми.

В результаті проведеної роботи було зроблено аналіз умов праці, шкідливи та небезпечних чинників, з якими стикається робітник. Було визначено параметри і певні характеристики приміщення для роботи над запропонованим проектом написаному в кваліфікаційній роботі, описано, які заходи потрібно зробити для того, щоб дане приміщення відповідало необхідним нормам і було комфортним і безпечним для робітника. Приведені рекомендації щодо організації робочого місця, а також важливу інформацію щодо пожежної та електробезпеки. Була наведена схема, розміри приміщення та наведено значення температури, вологості й рухливості повітря, необхідна кількість і потужність ламп та інші параметри, значення яких впливає на умови праці робітника, а також – наведені інструкції з охорони праці, техніки безпеки при роботі на комп’ютері

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1) Семакин, И.Г., Хеннер, Е.К. Информационные системы и модели [Текст]//И.Г. Семакин, Е.К. Хеннер – Москва : БИНОМ – 2005.

2) Колисниченко, Д. Н. Поисковые системы и продвижение сайтов в Интернете[Текст]//Д. Н. Колисниченко – Москва : Диалектика – 2007.

3) Маннинг, К. Введение в информационный поиск [Текст]//К. Маннинг– Москва : ВИЛЬЯМС – 2011.

4) Ананько,М. С., Москович, В.А., Негуляев, Г.А. Информационно-поисковые системы[Текст]//М. С. Ананько, В.А. Москович, Г.А. Негуляев – Москва : ИСЕРЕПАТ – 2006.

5) Черный, А.И., Горькова, В.И. Зарубежные автоматизированные справочно-информационные системы интегрального типа [Текст]//А.И. Черный, В.И. Горькова – Москва : ИНТ – 1980.

6) Краюшкин, Д.В. Анализ технологий предварительной обработки документальной информации [Текст]//Д.В. Краюшкин – Москва : РАНИ.А. – 2005.

7) Кузнецов, И.П.Системы и средства информатики[Текст]//И.П. Кузнецов – Москва : РАН И.А. – 2003.

8) Сомин, Н.В., Соловьева, Н.С., Кузнецова, Е.В., Шарнин М.М. Система морфологического анализа: опыт эксплуатации и модификации. [Текст]//Н.В. Сомин, Н.С. Соловьева, Е.В. Кузнецова, М.М. Шарнин. – Москва : РАН И.А. – 2005.

9) Батура Т.В., Еркаева О.Н., Мурзин Ф.А. К вопросу об анализе текстов на естественном языке. // Новые информационные технологии в науке и образовании[Текст]//Т.В. Батура, О.Н. Еркаева, Ф.А. Мурзин – Новосибирск :Под ред. проф. В.Н Касьянова – 2003.

10) Батура Т.В., Мурзин Ф.А. Логические методы представления смысла текста на естественном языке. // Новые информационные технологии

в науке и образовании[Текст]// Т.В. Батура, Ф.А. Мурзин – Новосибирск : Под ред. проф. В.Н Касьянова. – 2003.

11) ДСанПіН 3.3.2-007-98 Державні санітарні правила і норми. Гігієнічні вимоги до організації роботи з візуальними дисплейними терміналами електронно-обчислювальних машин Режим Доступу- :https://zakon.rada.gov.ua/rada/show/v0007282-98

12) НПАОП 0.00-7.15-18 Вимоги щодо безпеки та захисту здоров`я працівників під час роботи з екранними пристроями Міністерство доходів і зборів України Наказ від 05.09.2013 р. № 443 "Про затвердження Примірної інструкції з охорони праці під час експлуатації електронно-обчислювальних машин") Режим доступу- : https://zakon.rada.gov.ua/laws/show/%D0%BD%D0%BF%D0%B0%D0%BE%D0%BF_0.00-7.15-18

13) ДСТУ 7237:2011. Система стандартів безпеки праці. Електробезпека. Загальні вимоги та номенклатура видів захисту. Режим доступу- : https://zakon.rada.gov.ua/laws/show/ru/z0052-13

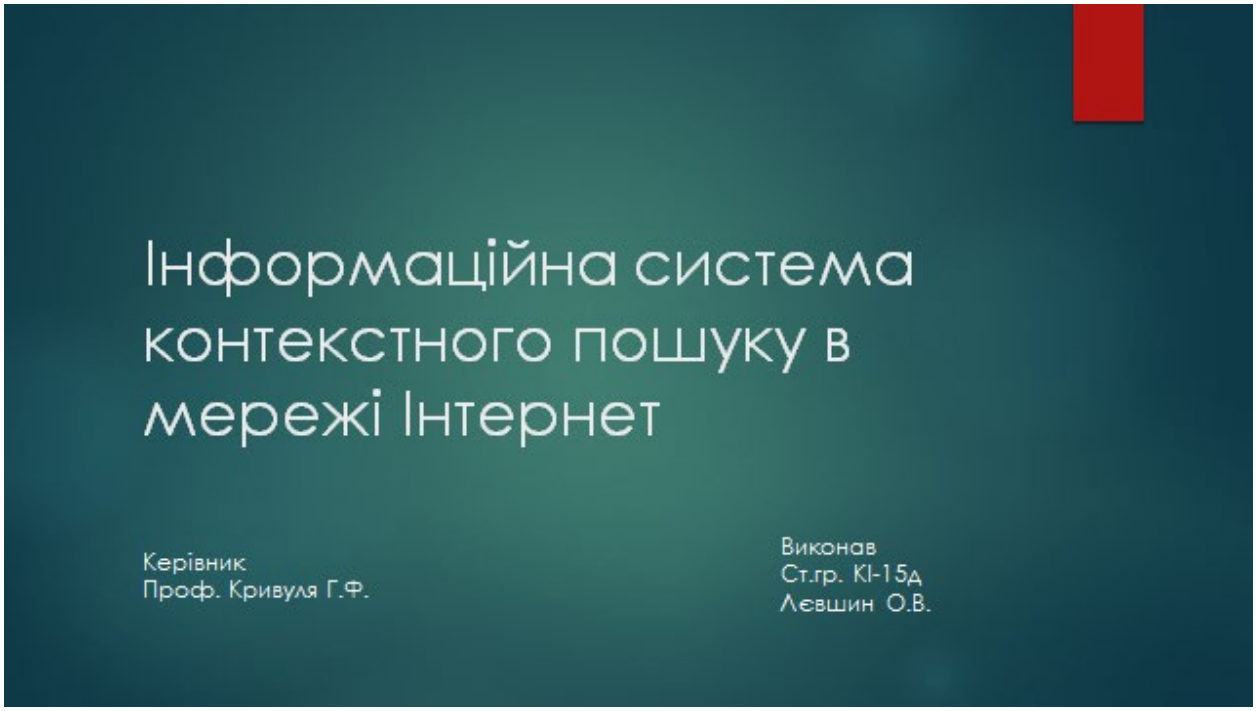
14) ДСН 3.3.6.037-99 Державні санітарні норми виробничого шуму, ультразвуку та інфразвуку. РежимДоступу-: https://zakon.rada.gov.ua/rada/show/va037282-99

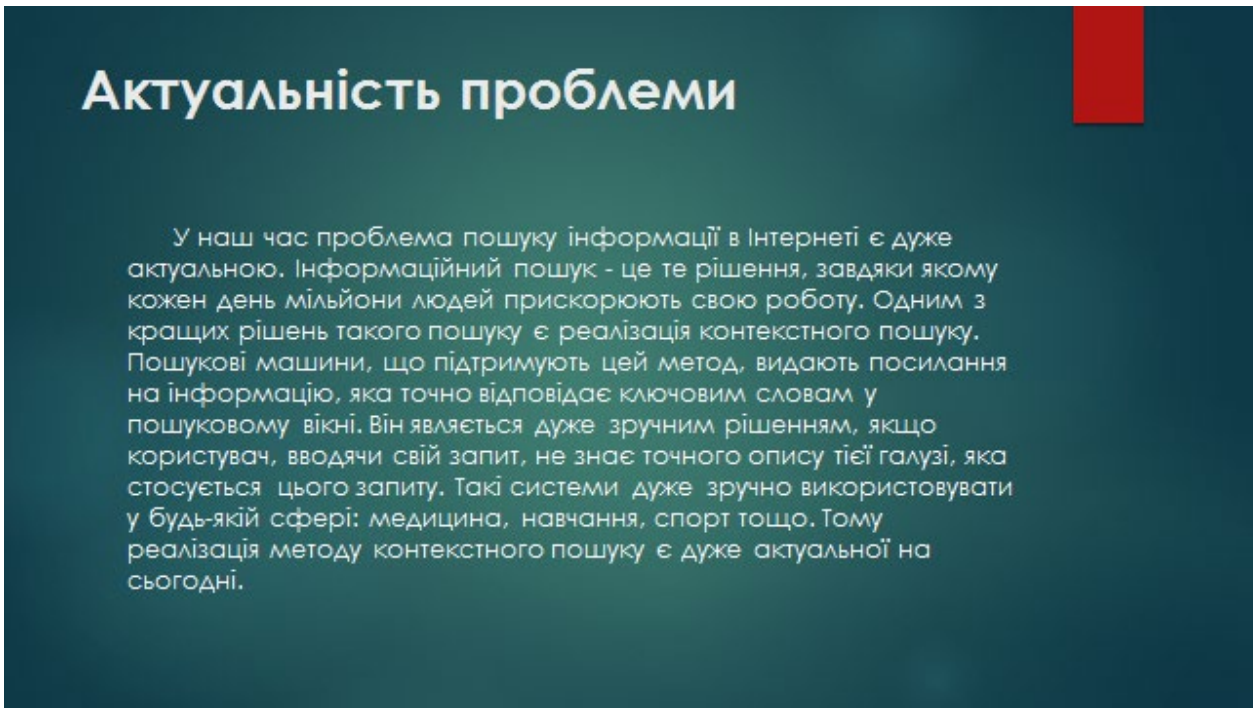
15) ДБН В.2.5-28-2006. Природне і штучнеосвітлення. Режим доступу - : https://zakon.rada.gov.ua/rada/show/v0168667-06

16) ДБН В.1.2-12-2008 Система забезпечення надійності та безпеки будівельних об`єктів. Будівництво в умовах ущільненої забудови. Вимоги безпеки. Режим доступу- : https://zakon.rada.gov.ua/rada/show/v0385661-08

17) ДСТУ Б А.3.2-13:2011Система стандартів безпеки праці. Будівництво. Електро безпечність. Загальні вимоги. Режим доступу- : https://zakon.rada.gov.ua/laws/show/z0633-12

ДОДАТОК А.
Електронні плакати





Мета роботи

Метою атестаційної роботи є розробка Web-системи для новинного порталу з пошуком, реалізованим контекстним методом. Контекстний пошук буде корисний у тому випадку, якщо не відомо, які ключові слова вибирати для пошуку.

Таким чином скорочується час пошуку необхідних новин, та підвищується зручність використання даної веб-системи.

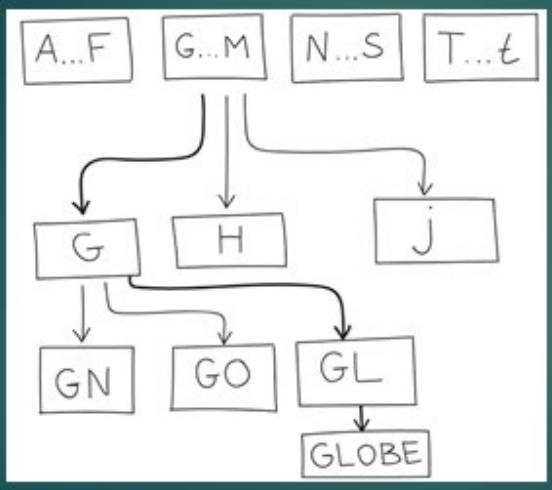
Веб-система повинна реалізувати наступні функції: пошук потрібних новин та їх перегляд.

Метод контекстного пошуку

Основною формою контекстного пошуку є процес сканування повнотекстового запиту, щоб зрозуміти, що потрібно користувачеві.

Всі технології повнотекстового запиту працюють за одним принципом. На основі текстових даних будується індекс, який здатний дуже швидко шукати відповідності ключового слова. Така система, зазвичай, складажться з двох компонент: пристрій та індексатор.

Метод контекстного пошуку

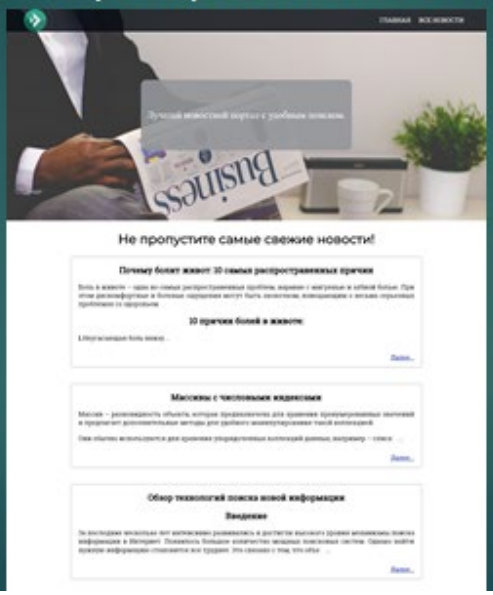


Метод контекстного пошуку

Пропонований метод можна розділити на наступні підзадачі:

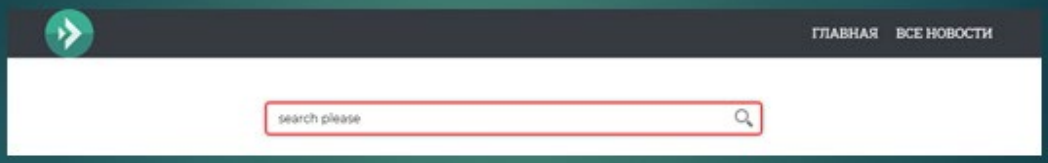
1. Індексування документів.
2. Визначення релевантності документів (пошук).
3. Аналіз статистики з метою визначення якості результатів пошуку.
4. Зміна індексів відповідно до якості результатів пошуку і виділення словосполучень.

Інтерфейс розробленої системи

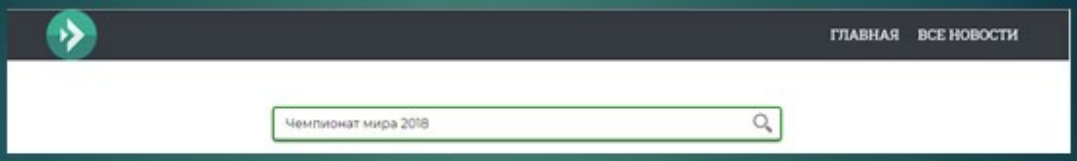


Інтерфейс розробленої системи

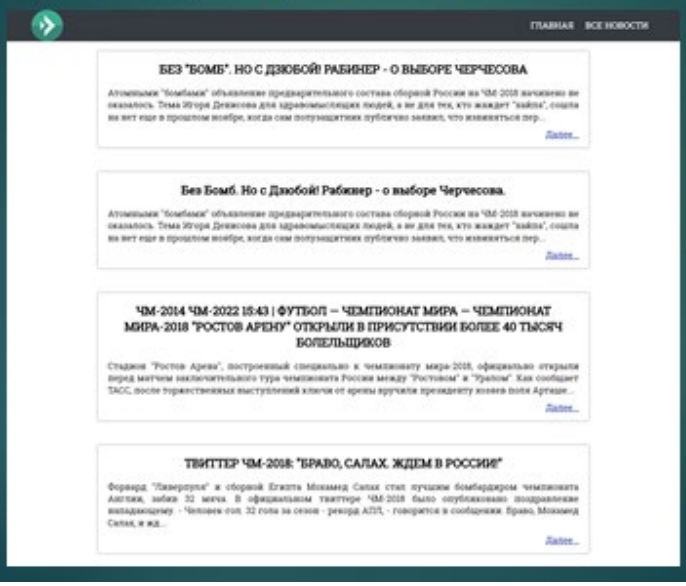
Інтерфейс розробленої системи



Інтерфейс розробленої системи



Інтерфейс розробленої системи



Висновки

В атестаційній роботі були проведені дослідження щодо актуальності, доцільності та практичності інформаційно-пошукової системи з контекстним методом. Була розроблена веб-система новинного порталу, яка дозволяє переглядати та робити пошук статтей, які зацікавлять користувача. Зокрема було проведене порівняння з найпопулярнішим постачальником пошукових систем.. В результаті були зроблені висновки, щодо можливостей використання вказаних технологій для рішення різних типів задач.