МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Тернопільський національний економічний університет
Навчально-науковий інститут інноваційних освітніх технологій
Кафедра комп'ютерної інженерії

Новосад Христина Василівна

Алгоритми агрегації новин на основі попередніх запитів користувачів / News aggregation algorithms based on previous user requests

спеціальність: 123 - Комп’ютерна інженерія
магістерська програма - Комп’ютерна інженерія

Магістерська робота

Виконав студент групи КІзм-21
Х. В. Новосад

Науковий керівник:
к.т.н., Г. Мельник

ТЕРНОПІЛЬ -2019

РЕЗЮМЕ

Випускна кваліфікаційна робота 87 сторінки пояснюючої записки, 27 рисунків, 8 таблиць, 6 додатків.

Метою роботи є розробка алгоритму агрегації новин на основі попередніх запитів користувачів.

В роботі на основі аналізу існуючих агрегаторів обгрунтовано вибір обов'язкових модулів, розроблено алгоритм формування новин на основі попередніх запитів користувачів, розроблено структуру бази даних.

Враховуючи сучасні тенденції до розробки веб – сайтів та недоліки існуючих сайтів пошуку новин, використано фреймворк Semantic UI для клієнтської частини та Laravel 5.3 для серверної, що дозволило забезпечити адаптивність сайту до дисплеїв із різними розширеннями, наприлад, мобільних телефонів, планшетів, широкоформатних моніторів. Розроблений сайт складається з таких модулів: виводу та фільтрації новин,модуль виводу попередніх запитів. Така інформації дозволить підвищити інформативність сайту, а наявність адаптивного графічного інтерфейсу забезпечить зручність отримання інформації на різних пристроях та покращить індексацію сайту пошуковими системами.

КЛЮЧОВІ СЛОВА: ВЕБ - САЙТ, ГРАФІЧНИЙ ІНТЕРФЕЙС, АГРЕГАТОР, НОВИНИ,ЗАПИТИ,КОРИСТУВАЧІ.

RESUME

Graduation qualififcation work contains 87 pages of explanatory note, 27 drawings, 8 tables, 6 appendices.

The aim of the diploma project is to develop a news aggregation algorithm based on previous user requests.

The project based on the analysis of existing aggregators justified the choice of required modules, developed an algorithm for generating news based on previous user queries, developed a database structure.

Taking into account current web development trends and shortcomings of existing news search sites, Semantic UI framework for the client side and Laravel 5.3 for the server side were used, which allowed to adapt the site to displays with different extensions, for example, mobile phones, tablets, widescreen monitors. The developed site consists of the following modules: news output and filtering, output module of previous queries. Such information will increase the informative content of the site, and the presence of an adaptive graphical interface will provide convenience of obtaining information on different devices and will improve the indexation of the site by search engines.

KEYWORDS: WEB SITE, GRAPHIC INTERFACE, AGGREGATE, NEWS, REQUESTS, USERS.

ЗМІСТ

ВСТУП

Актуальність теми. Актуальність даної задачі полягає у тому, що кількість інтернет ЗМІ досить велика, а для того, щоб швидко отримати всю нову інформацію, необхідно існування одного ресурсу, який би міг збирати новини та агрегувати їх за змістом.

Мета і завдання дослідження. Дослідити методи оцінювання тексту за подібністю та розробити програмний продукт з автоматизованою агрегацією новин та застосувати його до конкретної задачі.

Об’єкт дослідження - задача агрегації тексту новин.

Предмет дослідження - аналіз методів комп’ютерної обробки природньої мови.

Методи досліджень. При розробці алгоритму агрегації новин використовувались методи LSA та W- shingle.

Наукова новизна одержаних результатів. Розроблений алгоритм зберігає у базі даних ключові слова, які використав користувач у попередніх сесіях та на їх основі підбирає новини, групуючи їх по категоріях. Також користувачу відображається контент, з яким найбільше взаємодіють інші користувачі тієї ж категорії новин. Якщо користувач переглядає певну новину, у базі даних збільшується лічильник переглядів даної новини і зростає імовірність її відображення для інших користувачів.

Практичне значення отриманих результатів. Матеріали проведеного дослідження стануть у нагоді для подальшого вивчення специфіки алгоритму агрегації і основою для майбутнього дослідження з використанням нейронних мереж для класифікації новин за тематикою та регіоном.

Публікації та апробація ВКР. Результати наукового дослідження опубліковано в матеріалах науково – практичних конференцій молодих вчених та студентів «Інтелектуальні комп’ютерні системи та мережі» (Тернопіль,2019) [51,52].

У першому розділі проведено аналіз підходів до агрегації контенту,розглянуто основні принципи формування новин їхні тенденції застосування у світі та розглянуто існуючі агрегатори контенту види та класифікацію та зроблено відповідні висновки.

У другому розділі описано алгоритм роботи веб-сайту агрегації новин,розроблено алгоритм формування новин на основі попередніх запитів користувачів описано структуру бази даних та досліджено методи алгоритмів LSA та W – shingle та зроблено відповідні висновки.

У третьому розділі здійснена програмна реалізація алгоритму агрегації пошуку новин,зроблена структура програмного модуля яка включає в себе серверну та клієнтську частини також розроблено модуль роботи з базою даних та проведено порівняльний аналіз розробленого сайту з аналогами та зроблено відповідні висновки.

У додатках наведено код розробленої системи та зроблено графічний опис структури бази даних та графічно зображено алгоритм формування новин на основі попередніх запитів користувачів.

1 АНАЛІЗ ПІДХОДІВ ДО АГРЕГАЦІЇ КОНТЕНТУ

1.1 Основні принципи формування новин

Інтернет − мережа яка складається з мільйонів локальних мереж, які зв'язані з використанням різноманітних технологій [3]. Інтернет становить фізичну основу для розміщення величезної кількості інформаційних ресурсів і послуг, таких як гіпертекстові документи загальновідомої мережі та електронна пошта [4].

Сучасний Інтернет дуже різноманітний - це середовище для спілкування, розваг та навчання. Зі інтернетом можна робити покупки та оплачувати послуги. Інтернет − це спосіб отримання грошей, а в цілому − це відображення суспільства та світосприйняття [3]. Мережа пропонує широкі можливості для спілкування тут легко знайти людей зі спільними інтересамиО крім того, в мережі простіше спілкуватись, ніж при особистій зустрічі такі умови створюють розвиток інтернет − спільнот людей зі спільними інтересами, які спілкуються переважно через Інтернет. Таке середовище поступово відіграє важливу роль у житті сучасного суспільства [4].

Європа у розпалі інформаційної війни, одними із пристроїв застосовування яких є фейки (розповсюдження неправдивої інформації). Їх поширюють через засоби масової інформації і соціальні мережі. Неправдивого є багато не тільки в інтернеті, а й в традиційній масовій інформації [5]. Зараз складно відрізнити неправдиву інформацію від правдивої. Фейки − це фотографії зроблені у фотошопі, спеціально створені відео, написані брехливі новини. Фейками є створені акаунти вигаданих людей у соцмережах, через які розноситься брехлива інформація.Завданням фейкових повідомлень є рознести сумніви і вплинути на аудиторію у правдивості інформації. А мета: дезінформувати аудиторію, висувати власну позицію, викликати агресію змусити користувача засумніватися, змінити свою думку в товаристві, змусити до певної дії, активувати увагу і зацікавити аудиторію, зацікавити аудиторію за

допомогою вигаданих фактів. Фейк − це спеціально створена новина, подія або особа,доставляє брехливу інформацію і заставляє повірити людину та змінити свою думку [6].

Класифікацію фейків можна переглянути на рисунку 1.1

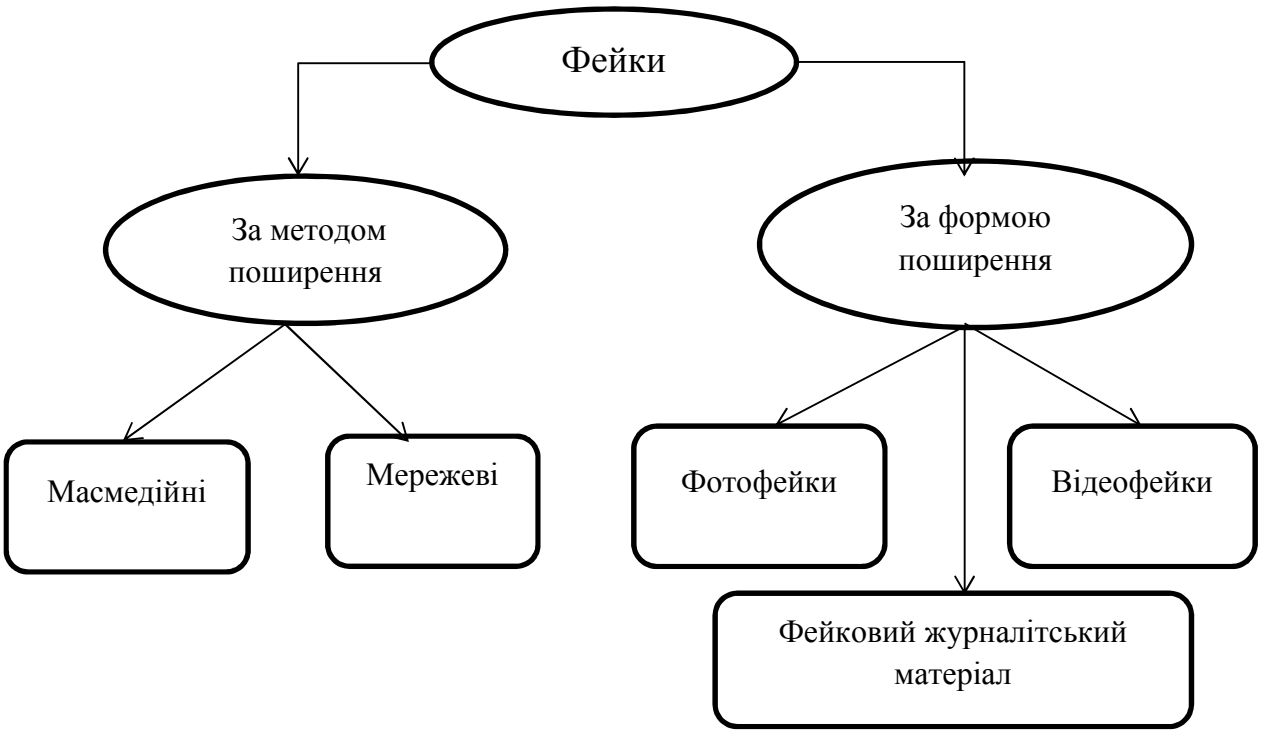


Рисунок 1.1 − Класифікація фейків

Фотофейк поширений як найбільш легкий тип фейків. Щоб розпізнати фотофейк потрібно зайти в Google Chrome достатньо лише клікнути по підозрілому зображенню правою кнопкою миші і вибрати пункт «Знайти це зображення в Google» (рисунок 1.2) [7]. Якщо ми користуємося іншим браузером, в якому за замовчуванням немає функції пошуку по зображеннях, можна встановити для нього інший додаток.

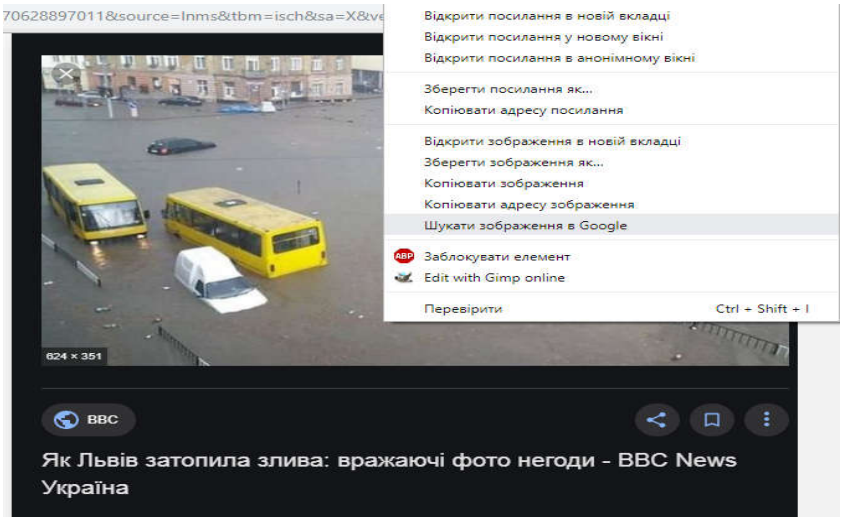


Рисунок 1.2 – Розпізнавання фотофейку

Відеофейки − цей фейк складніший, ніж картинка, оскільки простого способу пошуку по відео немає. Якщо відео неправдиве то потрібно перейти на YouTube, щоб мати доступ до відео та більше інформації . Явних фактів відеофейка немає, то потрібно звернути увагу на назву відео де вказана нова дата і якщо цей відео декілька разів заливається на YouTube декілька разів, то можна сказати, що це фейк. Відео з найбільшим переглядом та з коментарями це перший показник, що відео подивилися люди яких воно зацікавило. Потрібно добре придивитись на деталі на відео де є назви предметів, номери, таблички [8].

Фейкові матеріали журналістів. Часто у таких матеріалах посилаються засоби масової інформації, перекручують повідомлення та коментарі. Це доволі часто використовується для посилення переконливості повідомлення. У деяких повідомленнях є безліч вигадок незначних сайтів, які потрібно перевіряти. Щоб перевірити чи правдива інформація потрібно знайти новину у масовій інформації, на яку посилаються. Журналістські матеріали спеціально створюють неправдиву новину і дають до неї підґрунтя [9].

Виділяємо кілька способів як відрізнити фейки:

− новини, факти які опускаються на повідомлення із соціальних мереж;

− сенсаційні новини, які починаються із запитання;

− новина повинна бути лояльною. Емоції вказують на те що це є показником присутності пропаганди.

Витримати напади провокаторів потрібно особисто. Враховуючи те, що українська сторона має перевагу в інформаційному конфлікті то потрібно кожного дня підтверджувати інформацію, а якщо цього не буде то фейків буде достатньо щоб пошири брехню у країні.

Дезінформування − спосіб психологічного впливу інформацією яка обманює. Є різні методи дезінформування які мають позитивні та негативні риси. Вибір методу залежить від ситуації, яка складається на вибраній ділянці [10].

Виділяють три методи дезінформування:

− тенденційне викладення фактів − вид дезінформування, який полягає в необ'єктивному висвітленні фактів за допомогою спеціально підібраних істинних даних;

− дезінформування від зворотного − відбувається шляхом надання справедливих відомостей у спотвореному вигляді, коли сприймаються об'єктом як брехливі;

− термінологічне «мінування» − полягає у викривленні правильної суті термінів і тлумачень світосприймання.

Основне завдання емоційно упередженої новини − показати обурення та звернути увагу на винного. Застосовуються емоційні слова. Повідомляється про щось несправедливе. Мета цієї новини − дати настільки емоційний посил,що читач вимкне логіку і діятиме керуючись обуренням [11].

1.2 Тенденції застосування онлайн – новин у світі

Онлайн новини стають популярнішим в європейських країнах, однак ситуація в різних країнах має свої особливості. Актуальність онлайн новин у тому що це є другий найпоширеніший спосіб проведення часу в Інтернеті. Багатьом людям новини в Інтернеті мають перше місце, куди вони заходять, щоб дізнатись найсвіжіші новини [12]. Особливо це стосується споживачів у США та Великобританії, де онлайн новини повноцінно конкурують з телебаченням. Американські дослідники виявили наступну закономірність: онлайн новини є найпоширенішою формою споживання новин серед молоді, тоді як телебачення залишається найпопулярнішим для старших вікових груп. Особи які росли з Інтернетом, демонструють зовсім інші звички інформаційного споживання онлайн. Користувачі знаходять більше новин в соцмережах діляться ними, а також демонструють меншу нейтральність традиційним медіа платформам.

У сучасному світі онлайн новини лишають позаду телевізійні. Спільнота віком до 45 років у всіх країнах говорить що онлайн-новини важливіші за телевізійні. 51 % користувачів використовує соцмережі щотижня, так як 12 % опитаних кажуть, що паперові носії є для них основним джерелом новин. Серед 18–24-річних соцмережі усунули телевізор як перший доступ до новин.

Особи які мають смартфони читають новини дуже часто, ніж люди, які використовують комп'ютери, планшети. Перехід на смартфони йде в ногу із рухом до класифікованого наповнення. Із відповідей на запитання як до користувачів доходять новини відомо, що усі використовують свої телефони для доступу до соцмереж. На рисунку 1.5 показано значні відмінності залежно від пристрою користування. Наприклад, жінки частіше використовують соціальні мережі для отримання новин і менше використовують веб-сайти або додатки. Фейсбук − пошукова система, яка вабить усіх своїм контентом та доступністю до нього [13].

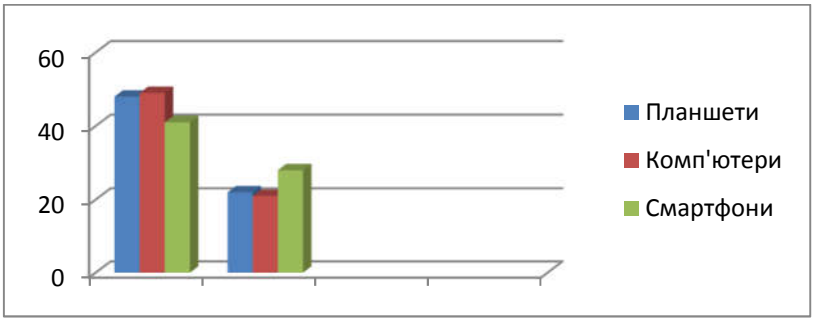


Рисунок 1.3 – Головні шляхи пошуку новин з пристроїв

Тимчасом розділений контент зосереджений здебільшого на Фейсбуці, агрегатори певних веб-платформ сьогодні панує у багатьох країнах − особливо у частинах Азії та у Європі. Останнім часом в усіх країнах (див. Рисунок 1.7), для перегляду новин все більше використовуються смартфони. Та все ж популярність ноутбуків зменшується, а зростання використання планшетів сповільнюється [15].

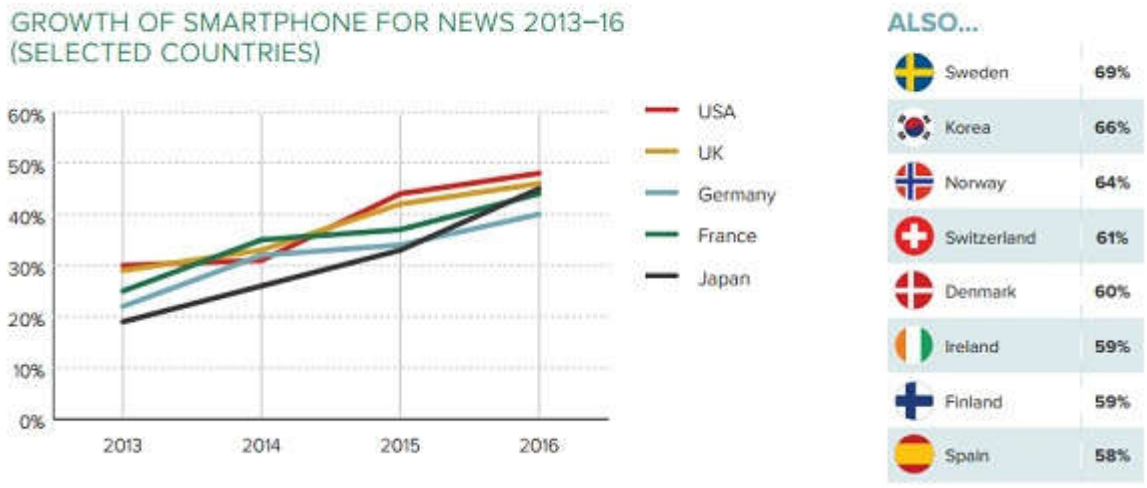


Рисунок 1.4 – Використання смартфонів протягом 2013-2016рр

З рисунку помітно, що активні користувачі смартфонів частіше дивляться новини, ніж користувачі комп'ютерів або планшетів. Дослідження показало, що люди, які використовують багато пристроїв частіше дізнаються про новини через те що особа зацікавлена в свіжих новинах, звідси і виходить практика

користування кількома пристроями. Зараз формується залежність від цифрових носіїв, і це також впливає на те, як ці люди починають свій день. Мобільні телефони займають дуже багато вільного часу, вони все частіше боряться за першість з радіо, телебаченням та пресою за право надати перші новини дня. Відеоновини онлайн зростають значно повільніше, ніж очікується [16]. Активні користувачі соцмереж більше отримують доступ до відеоновин онлайн, ніж інші групи населення. Агрегатори й соцмережі є джерелами новин, більше контенту споживачі отримують із друкованих видань, сайтів телерадіомовників. У всіх 26 країнах більш ніж дві третини вибірки (69 %) заходять на сайти популярних газет в інтернеті щотижня.

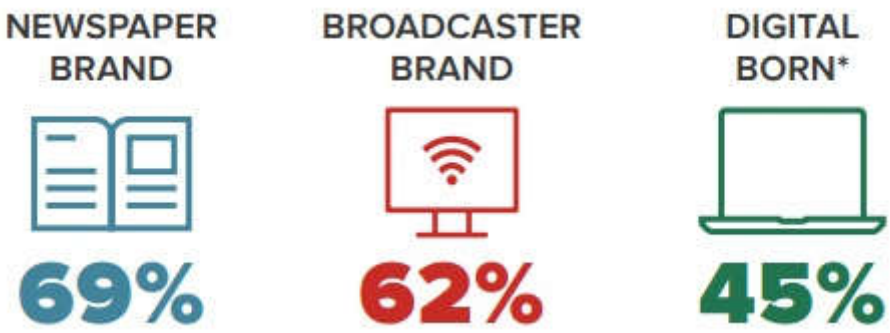


Рисунок 1.5 – Джерела якими користуються люди для перегляду новин протягом останнього тижня

Також на рисунку 1.8 зображено якими саме платформами користуються особи для перегляду онлайн – новин у Великій Британії,Японії,Італії,Новергії та США [17].

Рисунок 1.6 – Статистика онлайн платформ для перегляду новин у світі

Отримані дані вказують, що ЗМІ є джерелом інформації. Формат онлайн−новин додає кореспондентам можливості для самореалізації, а в деяких випадках і для заснування власного бізнесу. Журналісти мають низьку репутацію, бо завдяки розповсюдженим пліткам їхня самооцінка падає. Соцмережі й агрегатори осмислюють потребу в високоякісному контенті для залучення й утримання аудиторії. Видавництва хочуть отримати доступ до великих аудиторій і компенсації за вкладений обсяг інформації у контент.

Складність полягає в тому як знайти шлях до отримання доходу від онлайн - контенту. Рух до контенту, відбитий у цілому світі, складна економіка мобільних платформ і зростання блокування реклами більше не задовольняє стійкість традиційних бізнес-моделей [19].

У попередніх роках, помітно зміни цифрової доби як між різними поколіннями, так і в межах країн. Помітно, що деякі країни захищені своїм рівнем культури, інші відчувають прихід нової цифрової доби. Зараз здебільшого молодь різними шляхами отримує доступ до новин, це відбувається через мобільні платформи та соціальні мережі. Далеко дійде до розділення медіа, але це вже відчувається як початок нового етапу руйнування колишньої інформації. Інформаційні компанії повинні готуватись до адаптації та до змін знаючи, що репутація кореспондентів, нейтральність та довіра залишатимуться необхідною та головною складовою успіху.

1.3 Агрегатори контенту види та їх класифікація

Агрегатор контенту − це сайт, який збирає контент з різних джерел. Матеріали можуть бути різними від простих текстових до відео. Контент агрегатора збирає всю інформацію і спрощує користування. Колись дуже цінувалася унікальність і об'ємність матеріалів. Кореспонденти кожної носі сиділи перед екранами моніторів, збираючи і обробляючи інформацію. Вони

писали статті, замітки, нариси із цього виходила велика стаття. У ній широко розкривалася тема, наводилися думки експертів. Люди, які можуть обробляти і цікаво подавати такі вагомі матеріали, цінувалися видавництвами, редакціями інтернет−ресурсів. Але на підготовку таких статей йшло багато часу [20]. В інтернеті присутня незліченна кількість способів розплатитися за покупки. Але чи звертали ви увагу, що оплата відбувається за одним і тим же принципом, незалежно від того якою карткою і на якому сайті ви розраховуєтесь. Справа в платіжних агрегаторах, які покликані зробити наше життя легше при здійсненні оплат в інтернеті. Платіжний агрегатор – це система електронної комерції, яка об'єднує всі можливі способи оплати в один для полегшення проведення платежів в інтернеті. Компанія-платіжний агрегатор встановлює домовленості з окремими платіжними системами і операторами та спрощує процес здійснення покупок в інтернеті. Адже самостійно налаштувати функціонал навіть мінімального набору платіжних сервісів − завдання непросте й трудомістке. Робота безпосередньо без агрегатора − це багаторазова реєстрація, ознайомлення, встановлення контакту, настройка API та укладення договорів з кожним платіжним сервісом окремо. В подальшому регулярний підпис актів, відстеження змін API, підтримка користувачів, самостійна робота з помилковими платежами. І так десятки і сотні разів. Навіть уявити складно, скільки часу піде на підключення навіть найпопулярніших платіжних систем. В результаті − це просто втрата часу, сил, нескінченний потік невирішених питань і складності на кожному з етапів підключення. Агрегатори встановлюють фіксований відсоток від транзакції. Сума комісії встановлюється з огляду на такі фактори. Щомісячний грошовий обіг сайту. Чим вище грошовий обіг − тим вигідніше умови для власника інтернет-магазину, аж до розробки індивідуального тарифного плану. Тип бізнесу. Для оплати комунальних послуг, сайтів-пошуковиків авіаквитків та бірж благодійності розробляється окрема тарифна сітка. Способи оплати. На кожен спосіб оплати встановлюється окремий відсоток. Наприклад, при оплаті банківською картою комісія становитиме 2% від суми платежу, а при оплаті в терміналі вже 10%.

Іноді можливе укладання договору без комісії. В такому випадку встановлюється фіксована щомісячна абонентська плата. Найвідоміші світові платіжні агрегатори − 2checkout, Ccnow, Paysera і Paymentwall. Сервіс EasyPay також надає можливість підключення системи EasyPay до інтернет-сайту для повноцінного прийому і здійснення інтернет-платежів. EasyPay надає максимальну в Україні функціональність та фінансово-технологічні рішення: гнучкі тарифи, антифрод-систему та моніторинг платежів, рекурентні платежі, холд, маршрутизацію платежів та багато іншого. Інформація завжди була однією з основних цінностей у розвитку суспільства, і саме бібліотеки були тим місцем, де надавався доступ до виявленої та документованої інформації. Проте за останні десятиліття ситуація змінилася – від суспільства дефіциту інформації ми перейшли до суспільства інформаційного надлишку. На сьогодні інформаційний потік формується сотнями інтернет-ресурсів, серед яких: новинні та соціально-публіцистичні портали, стрічки новин державних установ і громадських організацій, а також персональні сайти, блоги та сторінки в соціальних мережах публічних осіб тощо.

Стрімке зростання обсягів інформаційних ресурсів привело до потреби використання як спеціальних програм, так і фахівців, які займатимуться опрацюванням цих ресурсів. Сьогодні одним з найбільш актуальних завдань інформаційно-аналітичних центрів різного спрямування – бізнесових, економічних, політологічних, правових, соціальних та інших – є обробка й аналіз великих обсягів структурованих і неструктурованих даних. Очевидно, що зростання обсягу інформації в епоху цифрових технологій актуалізує проблему оптимальних методів роботи з нею. Дедалі більше користувачів глобальної мережі мають потребу в отриманні актуальної інформації, оптимізуючи час для її пошуку. Пошуки шляхів реалізації цього завдання сприяє розвитку нових інструментів, серед яких, зокрема, спеціальні сервіси – агрегатори, які покликані акумулювати інформацію з різних інформаційних джерел.

У цьому контексті актуальним видається дослідження використання можливостей новинних агрегаторів у практиці створення інформаційно-аналітичних продуктів, чому і присвячено цю статтю.

Питання використання сучасних веб-сервісів та переваг їх використання бібліотеками в процесі трансформації у сучасні соціокомунікаційні центри з науково-консультативними функціями є актуальним серед наукової спільноти. Так, В. Горовий у статті «Бібліотека в інформаційному суспільстві» детально досліджує нові виклики, що постають перед бібліотечною системою України у зв'язку з розвитком у системі інформаційних обмінів електронних інформаційних ресурсів. Фахівець наголошує на необхідності розроблення інноваційних методик управління в бібліотечних установах цими ресурсами, організації ефективного використання їх користувачами.

Зараз змінився темп життя і інтернет-портали вступили в вічну гонку один з одним за першість публікації гарячих новин звідси і виходить що відвідувач рідко буде читати одну і те ж саме. Редактори поспішають першими написати той матеріал, який буде цікавий аудиторії. Найчастіше, на тривалу підготовку публікації просто не залишається часу щоб завжди бути першим почали користуватися агрегаторами контенту. Це дозволяє з одного місця збирати матеріал, який буде цікавий користувачеві.

Контент – агрегатори дозволяють:

- економити час на пошук матеріалів;
- економити гроші;
- знаходити цікаву інформацію для себе;
- архівувати контент,розбивати його на категорії;
- переглядати персональні дані в будь – якому місці та пристрої.

Сайти – агрегатори контенту збирають різні матеріали з джерел. Процес збору відбувається автоматично. Вибираються джерела, які підходять по тематиці порталу. Потім ці ресурси моніторяться з певною частотою, а матеріали копіюються зі збереженням авторства на сайт-агрегатор. Агрегатори

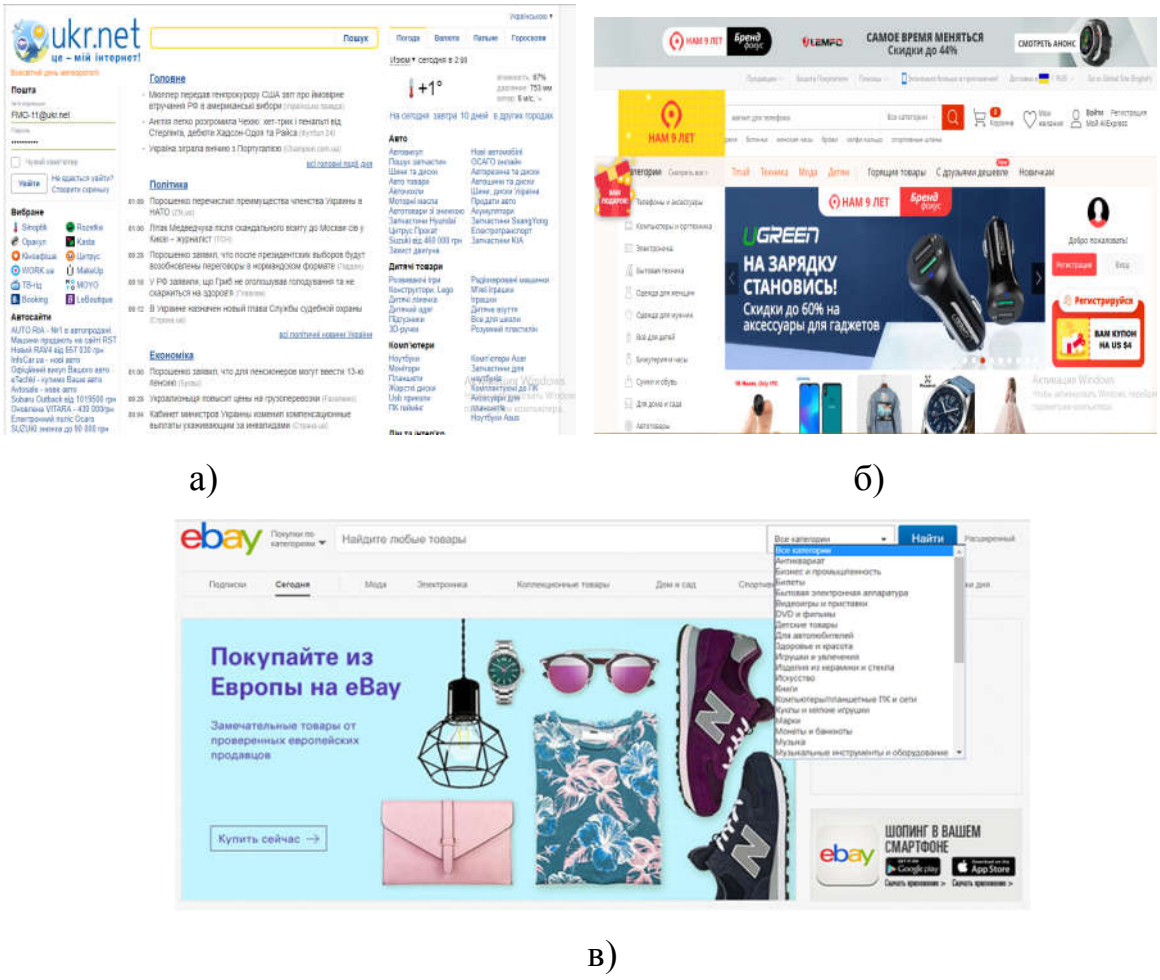
можуть збирати різну інформацію. На прикладі розглянемо що ж таке товарний агрегатом [21].

У таблиці 1.1 наведемо плюси та мінуси товарних агрегаторів.

Таблиця 1.1 – Плюси та мінуси товарних агрегграторів

Плюси	Мінуси
− зручний пошук по фільтрам і параметрам. Якщо покупець шукає, наприклад, жіноче пальто - йому пропонується вибрати фасон, колір, ціну і матеріал; − можливість сортувати товари за популярністю або ціною. Це дуже зручно: спочатку показуються найдешевші або самі ходові; − агрегатори отримують величезні прибутки: можемо уявити, який трафік щодня йде через кожен подібний сайт; − агрегатори потрапляють в топ: на відміну від простих інтернет-магазинів, вони мають більшу кількість посилань і корисного контенту, а значить, піднімаються в топ пошукової видачі.	− агрегатори за розміщення на своєму ресурсі отримуть гроші. Великі маркетплейси знову ж перевіряють інформацію - так вони стежать за своєю репутацією і якістю виставлених товарів; − навіть якщо з якістю все добре, на деяких агрегаторах можна заблукати: настільки великий перелік категорій і товарів в них. Для власників торговельних майданчиків це добре: більше переходів, більше трафіку, більше конверсії. Покупцям не дуже: якщо часу мало, а вибір великий, процес прийняття рішення може затягнутися надовго; − різні дошки оголошень, в яких відсутня виразна перевірка і модерація.

Зараз розглянемо найбільш відомі в Україні агрегат ори (рисунок 1.9)



а) б)

в)

Рисунок 1.7 – Найвідоміші агрегатори: а) «Ukr.net» – найвідоміший український сайт; б) «Aliexpress» – найпопулярніший у світі інтернет магазин; в) «eBay» – інтернет аукціон

«Ukr.net» – це ще один вітчизняний агрегатор, що набирає більше 10 мільйонів переглядів в день. Тут можна вибирати товари, переглядати описи, порівнювати ціни. Інтернет-магазини можуть розміщувати товарні пропозиції з прайс-листом, пропозиції торгового майданчика. «Aliexpress» – тут можна знайти все і навіть більше,включаючи шурупи і закінчуючи весільним одягом. Доречі на Aliexpress багато інтернет-магазинів, кожен з них продає свої товари і встановлює власні ціни. «eBay» – інтернет-аукціон, в принципі те ж саме, що і "Алі" – велика торгова площадка, де можна купити все що душа забажає. І на

відміну від Aliexpress, на eBay продаються і старі товари [22].

Для того щоб краще дізнатись чим же відрізняється агрегатор і інтернет − магазин переглянемо переваги та недоліки [23].

1. Агрегатор сам не продає товари, на відміну від інтернет-магазинів, він збирає їх усіх під крильцем і бере за це відсоток.

2. Агрегатор зробити складніше, ніж інтернет-магазин.

3. Трафік −у агрегатора високий. Якщо саме він знаходиться в топі видачі то користувач не буде перегортати сторінки, а перейде по першим посиланнях.

4. Асортимент товарів − у агрегатора більший за рахунок того, що в ньому накопичуються сотні і тисячі різних інтернет-магазинів і постачальників.

5. Контент різноманітніший у агрегатора: багато постачальників, багато описів, багато фото. Інтернет-магазини можуть виправити цю ситуацію за допомогою блогу, ведення сторінок у соцмережах.

6. Відгуки − і тут агрегатор рулить. Чим більше товарів − тим більше покупців і більше відгуків.

7. Зворотній зв'язок − тут перше місце займає інтернет-магазин. Агрегатор може працювати з клієнтами за допомогою розсилок або СМС. Магазин більш інтерактивний: заводите групу в соціальних мережах, проводите опитування, особисто питайте кожного клієнта про його побажання і пропозиції.

8. Особистий бренд. У агрегатора можливість оприлюдненння відсутня. Як власник інтернет-магазину можете робити ставку на особистий бренд і всіляко співвідносити себе зі своїм орендарями. Копіюючий фільтр – це агрегатом якого фільтри в основному відносяться до новинних агрегаторів і до мас – медіа. Будь-який інший агрегатор контенту, що копіює фільтр-агрегатор збирає інформацію по всій мережі, а потім фільтрує контент згідно з визначеними характеристиками. Але звичайний агрегатор контенту видає дані у вигляді певної інформаційної стрічки і зберігає посилання на оригінали, а копіює − агрегує прямі копії матеріалів. Фільтр − агрегатор маскує запозичений

контент настільки сильно, що іноді не можна сказати, що перед вами - авторське видання з унікальними матеріалами.

Останнім часом копіюючих фільтрів − агрегаторів стає все більше. Це призводить до серйозної проблеми дублювання інформації. Контент збирається агрегатором по всьому інтернеті, а потім матеріал дуже часто позиціонується як унікальний і авторський. В результаті однакові матеріали з'являються в різних копіюючих фільтрах-агрегаторах [24]. Важливим також видається висновок дослідження про домінування у багатьох країнах, особливо в деяких частинах Азії і континентальної Європи, агрегаторів на базі певних веб-платформ. Так, найпопулярнішим джерелом новин у Кореї є Нейвер (Naver) (66 % охоплення щотижня), агрегатор і портал, який надає повний спектр послуг, включаючи відео, ігри та електронну пошту. Подібна ситуація в Японії, де Yahoo (59 % охоплення щотижня) поширює новини від багатьох видавців через Інтернет або мобільні телефони. Ці портали є основним джерелом новин для значної частини населення цих країн .

Згідно з результатами опитувань, перевагами соціальних мереж і агрегаторів у пошуках онлайн-новин користувачі називають швидкість оновлення та зручність розташування багатьох джерел інформації в одному місці. Респонденти вважають, що агрегатори виконують свою роботу на «відмінно», забезпечуючи швидкий і легкий доступ до різних джерел інформації, але віддають перевагу соціальним мережам – з огляду на більшу інтерактивність .

Отже, оскільки використання агрегаторів новин впевнено стає однією з визначальних тенденцій споживання новинної інформації, а також враховуючи наявний досвід використання цього інструмента для аналітичної обробки інформації, актуальним видається осмислення його в контексті науково-практичних досліджень сучасних інформаційних сервісів. Агрегатор новин (іноді RSS-агрегатор) – клієнтська програма або веб-застосунок для автоматичного збору повідомлень із джерел, що експортуються у формати RSS

(англ. Really Simple Syndication – дійсно простий збір новин) або Atom. Так, кожен новий матеріал на сайті, який застосовує технологію RSS, автоматично перетворюється на коротке повідомлення, яке містить лише мінімум інформації (наприклад, назву матеріалу, прізвище автора, дату публікації, повний або частковий текст, зображення) . Агрегатори бувають двох типів – веб-агрегатори і програмні, але завдання у них однакові: робота з RSS та отримання оновлень. Програмний агрегатор – це програма, що встановлюється на комп'ютер для роботи з RSS та може бути вбудована в браузер чи поштовий клієнт, в операційну систему або може бути окремою програмою. Більшість сучасних браузерів (включно з Opera, MozillaFirefox, Maxthon та Internet Explorer) мають вбудовані інструменти збору RSS. Серед інших популярних програм з підтримкою стрічок новин можна виділити програвачі мультимедіа iTunes (агрегація підкастів) та Miro (можливість отримання оновлень відеоблогів). Прикладом окремо існуючого агрегатора може служити RSS Bandit. Веб-агрегатор – це агрегатор, що розміщений в Інтернеті та до якого є доступ з будь-якого комп'ютера, що під'єднаний до глобальної мережі. Прикладами таких агрегаторів є Google Reader, Bloglines та ін. Зважаючи на кількість, швидкість поширення та зростання популярності агрегаторів новин, на сьогодні можна говорити про два напрями їх використання: як користувацький інструмент (спрямований на задоволення особистих інформаційних потреб) і як професійний (спрямований на автоматизацію, покращення та полегшення збору інформації). У цьому дослідженні пропонується детально розглянути питання саме професійного використання агрегаторів новин, оскільки у зв'язку з постійно зростаючими обсягами інформації в процесі підготовки інформаційно-аналітичних продуктів додаткової уваги потребує відбір інформації. Можна виокремити різні підходи до відбору й структурування інформації, що залежатимуть від завдань, поставлених перед дослідником. Так, дослідження, результатом яких є створення інформаційно-аналітичного продукту, здійснювані співробітниками аналітичних підрозділів Національної бібліотеки

України ім. В. І. Вернадського, спрямовані на моніторинг та опрацювання зростаючих інформаційних потоків, аналіз і виявлення тенденцій новинного інформаційного поля, відстеження особливостей та динаміки найактуальніших повідомлень у ЗМІ. У цьому контексті використання новинних агрегаторів на етапі первинного збору та попередньої систематизації інформації видається доцільним. Враховуючи зростаючий вплив соціальних мереж, блогів і твітів на інформаційні процеси, на сьогодні їх можна розглядати як важливе альтернативне джерело інформації, що надає доступ до фактографічної та оцінювальної інформації, необхідної в процесі аналізу суспільно-політичних процесів, громадської думки з актуальних питань суспільного життя, визначення і прогнозування тенденцій подальшого розвитку подій. Отже, актуалізується необхідність використання інформаційних ресурсів соціальних мереж, блогів і твітів у процесі підготовки інформаційно-аналітичного продукту. Так, новинний сервіс Zmiya.com.ua (див. рис. 1) дає можливість у режимі реального часу отримувати інформацію, яка публікується в соціальних мережах і блогах. Щоб налаштувати сервіс індивідуально, необхідно пройти швидкий процес авторизації. Головною особливістю цього проекту є відображення всіх новин на одній сторінці, структурованих за хронологією, та можливість пошуку за ключовими словами. Головна сторінка проекту представлена в кількох блоках, у яких візуалізовано аналітику по тих чи інших зонах: хмара тегів, рейтинги користувачів у соцмережах, рейтинг найпопулярніших статей у блогах, ЗМІ, рейтинг твітів. Відображається також список найпопулярніших новин. Ранжування інформації відбувається не тільки за результатами пошуку, а й за відгуками на новину в соцмережах. Для інформаційного аналітика використання цього агрегатора дасть можливість в одному місці переглянути інформацію як соціальних мереж, так і блогів та твітів, а оперативне її оновлення – не втратити інформацію про важливі події і новини в українському сегменті Інтернету. Корисною функцією агрегатора в професійному використанні видається функція ранжування – рейтинги ЗМІ,

користувачів Facebook і Twitter ранжуються не тільки за кількістю переглядів/передплатників, а й за внутрішніми алгоритмами сайту, які враховують «живу» активність, коментарі, лайки та поширення в соцмережах. Треба зазначити, що на сьогодні одним з головних трендів розвитку Інтернету є персоналізація – користувачу пропонується тільки та інформація, що відповідає його потребам. Розглядаючи персоналізацію агрегаторів новин з позиції використання їх інформаційними аналітиками як інструмента в процесі роботи, варто відмітити таку перевагу, як можливість використання не тільки дайджесту, пропонованого сервісом, а і використання функції «ключових слів», які полегшують пошук потрібних новин для формування джерельної бази досліджень. Наприклад, працюючи з повідомленнями, аналітик може виділити слово «Конституція», додати його у свої ключові слова і в подальшому отримувати пов’язану з ним інформацію. Таким чином, чим більше ключових слів додасть аналітик, тим більше сервіс персоналізуватиметься під його інформаційні потреби і пропонуватиме кращі рекомендації. Якщо ж аналітик хоче систематично отримувати інформацію за відповідною тематикою, тоді є можливість підписатися на отримання новин за заданим запитом. Так, уже сьогодні такі інформаційні агентства, як Інтерфакс або Reuters ,пропонують користувачеві вказати, які теми йому цікаві та згодом транслюють контент з відповідних рубрик у першу чергу. Ще однією альтернативою може стати створення «власного агрегатора новин». Для цього існують зручні сервіси, які мають функції індивідуальної персоналізації, наприклад Feedly.com. По-суті, це RSS-стрічка, у яку можна додати потрібні інтернет-ресурси й отримувати інформацію виключно з них; мобільний застосунок Flipboard також дає можливість сформувати список обраних інформаційних ресурсів; LinkedIn Today отримує інформацію про інтереси користувача з його особистого профілю в соціальній мережі LinkedIn, його друзів і активності та, проаналізувавши ці дані, формує пропоновану стрічку новин. Розглянемо використання агрегаторів на прикладі роботи над проектами інформаційно-

аналітичних підрозділів НБУВ. Ці проекти спрямовані на корпоративних користувачів, серед яких органи державної влади, наукові установи, громадські, бізнесові, соціальні організації, які потребують якісної, тематичної, достовірної електронної інформації. Специфіка забезпечення інформаційних запитів такої категорії користувачів полягає в необхідності дотримання певних принципів, серед яких: оперативність виконання інформаційних запитів, якість продукції, дотримання усіх вимог і стандартів представлення інформації (достовірність, повнота, релевантність викладених матеріалів з додаванням експертних висновків авторитетних фахівців); дотримання законодавства України про інформацію та ЗМІ тощо. Задовольнити інформаційні потреби таких категорій користувачів здатні бібліотечні інформаційно-аналітичні служби, які на сьогодні мають репутацію надійних інформаційних партнерів. Робота таких служб «полягає у відборі, обробці й аналізі інтернет-інформації за тематичними запитами, у створенні нових знань за формулою “інформація на базі інформації”, у подальшій розробці аналітичного дослідження, що пояснює події, які відбуваються або відбулися, і містить прогноз на майбутнє, а також у представленні його результатів у вигляді інформаційно-аналітичних продуктів» Треба зазначити, що використання агрегаторів новин – це один з етапів відбору інформаційних ресурсів у процесі підготовки інформаційно-аналітичних продуктів. Для оптимізації процесу пошуку та відбору інформації в інформаційно-аналітичних підрозділах НБУВ уже напрацьовано достатній досвід, визначено стратегії пошуку, розроблено поетапні рекомендації щодо створення інформаційної бази, яке починається «з моніторингу інформаційного простору за заданими параметрами (з використанням пошукових систем Інтернету, вузькотематичних сайтів, бібліотечних каталогів), пошуку інформації з поглибленням в архів ресурсу (сайти, журнали, газети, монографічна література), перевірки її якісних характеристик, обробки з метою створення інформаційного середовища для правильної оцінки досліджуваних фактів, подій і явищ» .

1.4 Висновки до першого розділу

Сучасний Інтернет різноманітний насоціальні та культурні грані − це універсальне середовище для спілкування, розваг та навчання.За допомогою Інтернету можна робити покупки та оплачувати послуги. Інтернет − спосіб заробітку. У всесвітній павутині є багато агрегаторів спрощують життя користувачів, оскільки збирають цікаву для них інформацію з різних джерел на одному сайті. Для того щоб запустити свій агрегатор контенту, потрібно розібратись над тим, яку користь він принесе звичайному користувачеві. Якщо просто збирати контент, навряд чи сайт буде потужним. Потрібно дати можливість порівнювати, якщо запускати агрегатор товарів то дозволити налаштовувати фільтри по інтересам, якщо ж запускати агрегатор новин то він повинен спрощувати життя користувача.

Отже,у першому розділі наведено інформацію про основні принципи формування новин у світі, про сучасні тенденції формування новин та здійснено порівняльний аналіз існуючих найпопулярніших агрегаторів у світі.

2 АЛГОРИТМИ РОБОТИ ВЕБ – САЙТУ АГРЕГАЦІЇ НОВИН

2.1 Критерії формування рейтингу новин

Проблему відбору новин усвідомлюють журналісти, оскільки у них існують свої критерії, які не завжди збігаються з уподобаннями аудиторії. Ми зараз не говоримо про передвиборчу ситуацію, в якій відносини преси з політичними громадськими організаціями регламентуються законодавчо або тимчасовими регламентами. У цей період увага преси зосереджується залежно від реальної постановки сил у діючих законодавчих зборах, інститутах сьогоднішньої представницької влади. Це також обумовлено розкладом громадської думки але по відношенню до певних політичних платформ. Мова йде про повсякденну практику, коли журналісти вибирають, вирішують, чому віддати перевагу при висвітленні суспільного життя.На відбір інформації впливають характеристики події (масштаб того, що відбувається, його екстремальність - катастрофа, скандал), так і риси інформаційного органу: його політичні симпатії, природа самого засобу (аудіо, відео або друк) [25].

Новини повинні задовольняти велику кількість вимог. Але при цьому не будь-яка новина, написана за всіма правилами, може потрапити в друк. Пов'язано це з тим, що на відбір і подачу новин в ЗМІ впливають і зовнішні фактори. Ось основні з них: запити аудиторії, конкуренція, загальна лінія видання, тиск з боку видавця або рекламодавця.

Є такі основні характеристики новин:

− оперативність − інформація про певну подію має бути максимально швидко оприлюднена через відповідний канал передачі інформації, інакше вона перестане бути новиною;

− актуальність − інформаційне повідомлення показує важливі для суспільства питання, привертає суспільну увагу;

− інтерес − новина має висвітлювати питання, які є цікавими суспільству, великим соціальним групам;

− об'єктивність − незмінна подача інформації, відсутність її викривлення чи перекручування, висвітлення різних точок зору;

− достовірність − подача правдивої інформації з перевірених джерел;

специфічність побудови інформаційного повідомлення − повідомлення містить лише найважливішу інформацію з теми без заглиблення у деталі [26].

За статистикою, навіть ті, хто клікнули на заголовок і відкрили новину, присвячують їй від п’яти до двадцяти секунд. До кінця не дочитують навіть важливі й цікаві новини. За увагу користувачів інтернету змагається купа джерел і подразників, які не дають сфокусуватись надовго. А ще, зрозумівши суть новини, вони поспішають нею поділитись. Тому новина має завжди починатися з головного, продовжуватись важливим, а найменш важливе має бути в кінці. Таку структуру називають перевернутою пірамідою. Її вигадали в дев’ятнадцятому столітті, коли новини передавали телеграфом. Зв’язок коштував дорого, проходив повільно й був нестабільним, тому журналісти навчились обходитись без передмов. Зараз проблема та сама − зв’язок онлайн-видання з аудиторією нестабільний і нетривалий. Структуру перевернутої піраміди розглянемо на рисунку 2.1

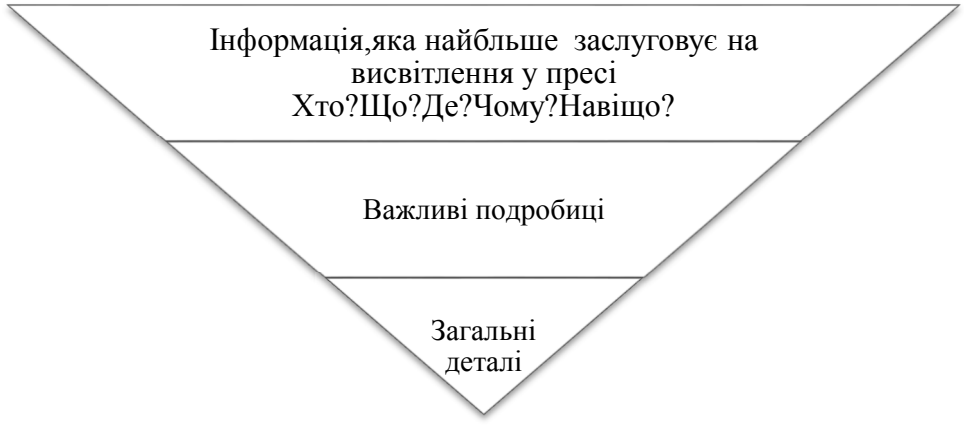


Рисунок 2.1 – Перевернута піраміда

Інформованість залежить від джерел, яким ми довіряємо. Чим більше цих джерел, чим більш вони компетентні та різноманітні, тим краще ми

орієнтуємося у ситуації. Відповідно, читаючи газету чи сайт, повинні звертати увагу на те, звідки вони взяли ту або іншу інформацію.Здавалося б, за часів інтернету, проблем із інформацією взагалі не має бути. Адже її так багато. До того ж завдяки соціальним мережам дуже легко знайти інформацію про кожного. Однак, на жаль, інтернет тільки створює ілюзію різноманітності. Насправді саме через появу мережі журналістам усе складніше надавати аудиторії ексклюзивні новини. Розглянувши структурні особливості існуть основні типи новин [27]:

надзвичайні ситуації;

− злочини;

− місцеві і центральні органи влади;

− плани і проекти;

− конфлікти і суперечки;

− господарство;

− здоров'я;

− визначні особи;

− спорт;

− погода;

− ситуації на шляхах.

Отже, сучасні новини під впливом інтернету змінюються. З одного боку, новин стає багато, з’являються вони швидше. Але з іншого – втрачається унікальність контенту, і всі видання стають схожими одне на одного. Оскільки заснувати сайт сьогодні дуже просто, місцеві медіаринки стають популярними, багато сайтів можуть просто дублювати новини.За таких умов дуже важливо знайти власний стиль роботи з інформацією, зосереджуватися не тільки на новинах, але й шукати нові формати. Адже якщо місцеве видання зосереджується тільки на новинах, воно не займатиме на медіа ринку своєї ніші й не матиме свого постійного читача [28].

2.2 Існуючі алгоритми агрегації

2.2.1 Метод W- shingle з алгоритмом MinHash

Інформаційне суспільство продовжує свій розвиток. Ключовою технологією, яка допомагає задовольняти інформаційні потреби людини, залишаються пошукові системи. Вони реалізують механізми інформаційного пошуку, шукають різні об'єкти інформації, дуже часто це текстові документи. Одним із завдань щодо поліпшення якості інформаційного пошуку текстової інформації є вирішення проблеми дублювання інформації [29].

Метод W − shingle або шингл − це певні фрагменти тексту, за якими проводиться перевірка оригінальності документа, сайтами перевіряючими текст на антиплагіат. Поняття шингл можна зрозуміти, розглянувши поняття як крок шинглів. При перевірці тексту на унікальність, системи антиплагіату використовують певний алгоритм перевірки. Цей алгоритм заснований на кроці шингли. Розглянемо на прикладі сайту антиплагіат , який протягом довгого часу використовує крок шингли . Це значить, що при перевірці тексту на унікальність, антиплагіат аналізує кожне третє слово в перевіряється тексті, і якщо знаходить збіги, робить висновок про те, що текст не унікальний, і суміжні права [30].

Більш детальніше продемонструємо метод шингли на малюнку 2.2

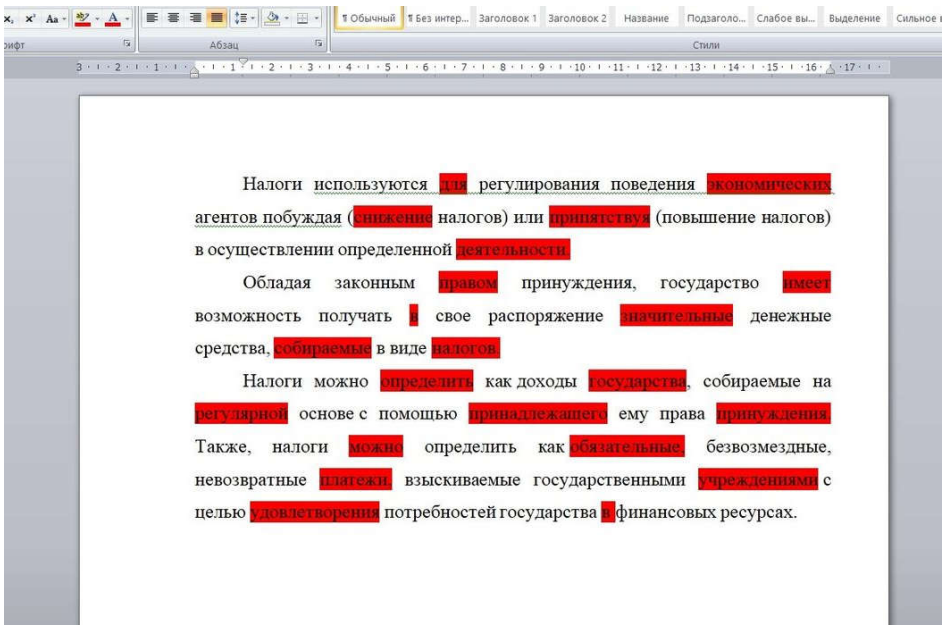


Рисунок 2.2 – Метод W - shingle

Цей текст має унікальність 0%. Він повністю завантажений з інтернету. Якби антиплагіат, перевіряв на унікальність ось цей текст, то його алгоритм шингли при аналізі тексту був би таким (червоні слова). Тобто антіплагаіт буде аналізувати кожне третє слово в тексті . Відповідно, якщо всі ці слова залишити на своєму місці і нічого в тексті не поміняти, то унікальність тексту покаже унікальність 0%. Багато задач, які важко вирішити, досить просто вирішуються за допомогою існуючих алгоритмів. У своїй повсякденній роботі програміст постійно має справу з ймовірними алгоритмами і структурами даних [31].

Алгоритм MinHash це продовження алгоритму Shingles для зменшення часу витрачається на порівняння образів різних текстів і зменшення кількості займаної пам'яті. На рисунку 2.3 зображено поділ тексту на послідовності шинглів:

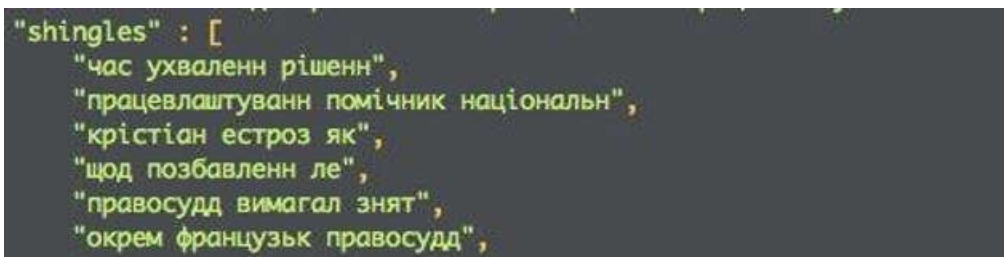


Рисунок 2.3 – Поділ тексту на послідовності шинглів

Після обробки тексту алгоритмом шинглування алгоритм MinHash здійснює обчислення хеш-функцій і відбір їх мінімальних значень.Алгоритм MinHash використовує h хеш − функцій. Кожній хеш-функції ставить у відповідність число − найменше її значення для всіх шинглів досліджуваного рядка. Отриманий в результаті вектор довжини h утворює сигнатуру або короткий числовий образ документа. На вхід алгоритму MinHash подається масив шинглів довжини N. На виході алгоритм видає одновимірний масив (вектор) хеш-кодів довжини h. Сигнатуру або короткий числовий образ документа [32].

Отже, можна зробити висновок що Метод W-shingles з алгоритмом MinHash ділить текст на послідовності шинглів, MinHash хешує усі шингли, для зменшення обчислювальної складності та використовує коефіцієнт Джакарда для оцінювання подібності множин.

2.2.2 Латентно – семантичний аналіз + TF-IDF

Латентно-семантичний аналіз (LSA, Latent Semantic Analysis) − це теорія і метод для витягання контекстно-залежних значень слів за допомогою статистичної обробки великих наборів текстових даних. LSA був запатентований в 1988 році. Об'єднання слів у теми і уявлення текстових одиниць в просторі тим здійснюється шляхом застосування до даної матриці одного з матричних розкладань. Найбільш популярними є: сингулярне розкладання і факторизация невід'ємних матриць.Згідно теоремі про сингулярний розклад, будь-яка матеріальна прямокутна матриця може бути розкладена на три матриці: $A = USV^T$, де $A \in Rn \times m$, матриці $U \in Rn \times k$ і $V \in Rm \times k$ − ортогональні, а $S \in Rk \times k$ − діагональна матриця, значення на діагоналі якої називаються сингулярними значеннями матриці A, V^T − транспортна матриця. Істотним недоліком методу LSA є значне зниження швидкості обчислення при збільшенні обсягу вхідних даних (наприклад, при

SVD перетворенні. Легко помітити що переважна кількість осередків частотної матриці індексованих слів, створеної на першому кроці, містять нулі. Матриця сильно розріджена і ця властивість може бути використано для поліпшення продуктивності і споживання пам'яті при створенні більш складної реалізації. Щоб тексти були приблизно однієї і тієї ж довжини, в реальних ситуаціях частотну матрицю слід нормалізувати. Стандартний спосіб нормалізації матриці TF-IDF [33].

TF – IDF – це алгоритм який використовується для розрахунку важливості слова в документі, ваги слова. Це простий і зручний спосіб оцінити важливість терміну для будь-якого документа щодо всіх інших документів. Принцип такий - якщо слово зустрічається в будь-якому документі часто, при цьому зустрічаючись рідко у всіх інших документах − це слово має велику значимість для того самого документа. Принцип роботи TF − це частотність терміна, який вимірює, наскільки часто термін зустрічається в документі. Логічно припустити, що в довгих документах термін може зустрітися в великих кількостях, ніж в коротких, тому абсолютні числа тут не проходять. Застосовують відносні − ділять кількість разів, коли потрібний термін зустрівся в тексті, на загальну кількість слів у тексті.

IDF − це зворотна частотність документів. Вона вимірює безпосередньо важливість терміну. Тобто, коли ми вважали TF, всі терміни вважаються як би рівними за важливістю один одному, але відомо, що, прийменники зустрічаються дуже часто, хоча практично не впливають на зміст тексту. Нам потрібно порахувати IDF – це вважається як логарифм від загальної кількості документів, поділеній на кількість документів, в яких зустрічається терміні [34].

Отже, TF / IDF насамперед використовується для аналізу текстового контенту в великих потоках даних. До цього відношенню вдаються пошукові алгоритми, щоб визначити релевантність конкретної сторінки (перш за все, тексту, який на ній знаходиться), призначеної для користувача запиту в пошуку. Також даний статистичний показник дозволяє визначити близькість

різних документів (текстів) один одному, що може бути використано при їх угрупованню (кластеризації).

2.2.3 Алгоритм Стеммер Портера

Алгоритм, розробив Мартин Портер в 1979 році для англійської мови. В основі цього було створено проект Snowball і на основі оригінального алгоритму були написані стеммери для більшої індоевропейської мови. Основна ідея стеммера Портера включає в себе то, що існує обмежене число формових і словотворених суфіксів, і стеммінг слова відбувається без використання будь яких баз тільки багато існуючих суфіксів при цьому складні суфікси розбиваються просто вручну. На рисунку 2.4 зображено нормалізацію вхідних даних за алгоритмом Стеммера Портера [35].

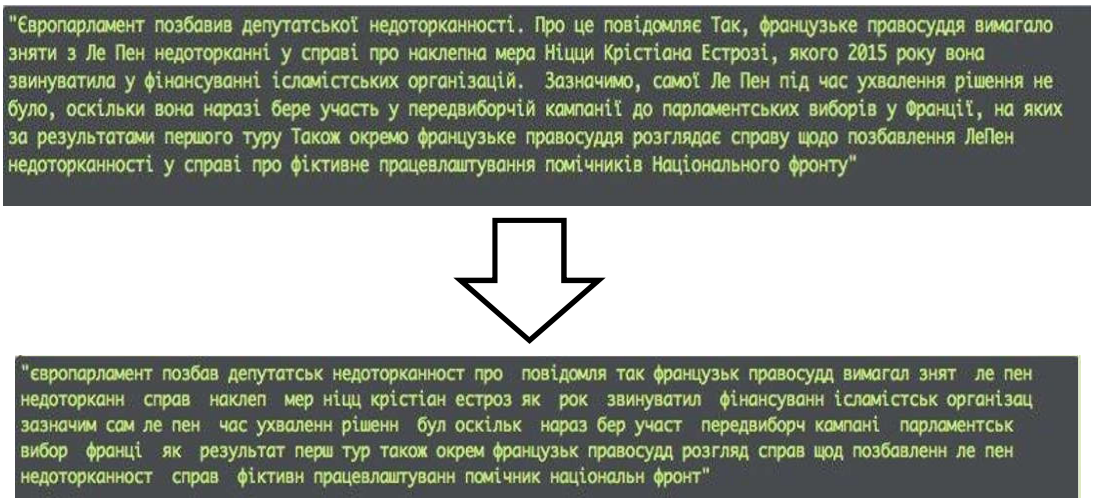


Рисунок 2.4 – Вхідні дані за алгоритмом Стеммера Портера

Алгоритм складається з п'яти кроків. На кожному кроці відсікається формо або слово утоворюючий суфікс і решта перевіряється на відповідність правилам (наприклад, для українських слів основа повинна містити не менше однієї голосної). Якщо отримане слово задовольняє правила, відбувається перехід на наступний крок. Якщо немає − алгоритм вибирає інший суфікс для відсікання. Згідно з офіційним сайтом проекту, на першому кроці відсікається

максимальний формоутворювальний суфікс, на другому − буква «і», на третьому – слово утворюючий суфікс, на четвертому – суфікси чудових форм, "ь" і одна з двох «н» [36].

Стеммер Портера не використовує ніяких словників і баз основ, є плюсом для швидкої дії і спектра застосування, він непогано справляється з неіснуючими словами. Алгоритм часто обрізає слово, що ускладнює синтез нормальної форми по отриманому стеммеру: грибниця − гриба (при цьому реально незмінна частина слова − гриб, але стеммер обрізає найбільш довгу морфему) і не справляється з випадаючими голосними в корені: кішки − кішок, кішками − кішк. До мінусів Портера часто відносять людський фактор те що правила для перевірки задаються вручну і пов'язані з граматичними особливостями мови, а це збільшує ймовірність помилки.

2.3 Алгоритм формування новин на основі попередніх запитів користувачів

Пошукові запити складаються з декількох слів, так як одним словом описати те що треба знайти дуже важко. Для того щоб підвищити якість пошуку у системах Google, використовується спеціальний алгоритми, що розпізнає запити користувачів, їх обробляє і показує користувачу. Незалежно від того, в якій формі вживається слово в запиті, практично всі пошукові системи враховують морфологію і здатні проводити пошук по всіх йо го формам. Класифікацію запитів можна поділити на такі пункти (рис. 2.1).

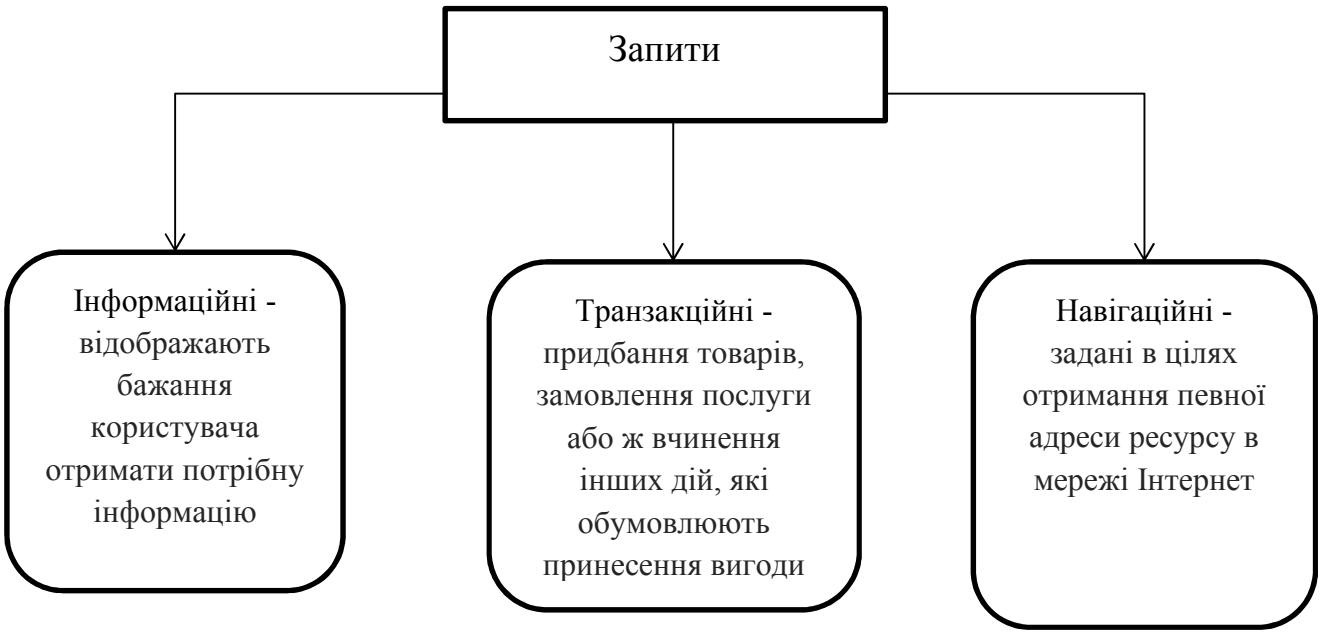


Рисунок 2.2 – Цільове призначення запитів

Широкі можливості для впливу на цільову аудиторію надають оптимізаторам та рекламодавцям інформацію про використання слів у пошуку. Чималий інтерес запити користуваців представляють і для держави. У США пошукові системи зобов'язані пред'являти державним органам докладну статистику на законодавчому рівні. Проте статистика тут показується у вигляді графіків, на яких можна дізнатися про ступінь конкуренції рекламодавців за конкретним запитом, переглянути історію трафіку для ключових слів або отримати дані про їх вартість при участі в системі контекстної реклами.Google підбив підсумки по найпопулярніших запитах користувачів, туди ввійшли: фільми онлайн,Yotube,Facebook,укрзалізниця,телепередачі,серіали,техніка Apple,OLX,Avtoria,інтернет магазини.

Графічне представлення алгоритму формування новин на основі попередніх запитів користувачів наведено у додатку А.

Алгоритм формування новин на основі попередніх запитів користувачів в системі наступний:

Крок 1. Вибираємо категорію новин (музика,туризм,спорт)

Крок 2. Налаштовуємо критерії пошуку. До критеріїв відносимо: дату створення новин, загальну кількість згенерованих новин,кількість переглядів новин.

Крок 3. Підключаємось до бази даних. У БД до кожного користувача зберігається набір ключових слів. Набір формується на основі новинних повідомлень які користувач вводив у пошуковій системі. Наприклад: машина «toyota rav 4 hibryd», звідси ключовими словами будуть відображатись «машина», «Toyota», «rav4», «hibryd».

Крок 4. Згодом вибираємо з новин ключові слова,які зустрічаються в БД.

Крок 5. Якщо користувач читав певну статтю то ймовірність,що з цього сайту виведеться новина збільшується.

Отже, у даному підрозділі на основі результатів дослідження пошукових запитів користувачів розроблено алгоритм формування новин на основі попередніх запитів користувачів. В результаті описано послідовність дій користувача під час пошуку новин в системі та у додатку А наведено алгоритм формування новин на основі попередніх запитів користувачів.

2.4 Структура бази даних

Людина в процесі інформаційної діяльності збирає й накопичує відомості про навколишній світ. До появи обчислювальної техніки вся інформація зберігалася в письмовому або друкованому виді. Однак чим більше були обсяги інформації, з якими приходилось оперувати людині, тим гостріше вставало питання збереження інформації та її обробки [37].

Базою даних є систематизоване та централізоване сховище інформації певної предметної області, до якої мають доступ різні користувачі. За вмістом інформації БД бувають фактографічні (зберігають тільки опис об’єктів) та документальні (зберігають в комп’ютерній формі самі об’єкти). Наприклад, БД “Сучасні поети” може містити тільки перелік коротких відомостей про поетів та їх твори, в цьому випадку вона буде фактографічною. Якщо ж в ній присутні також самі тексти творів, біографічні дані про поетів, то її можна вважати

документальною. За системою доступу до конкретної інформації БД класифікують на ієрархічні, мережені та реляційні. Прикладом ієрархічної БД може бути дискета, на якій користувач зберігає власну інформацію. Щоб дістатись до файлу, треба відкрити певну послідовність папок – пройти більш високі рівні систематизації інформації. Прикладом мереженої БД може бути БД будь-якого банку про вклади клієнтів. Дістати гроші з рахунку клієнт може з будь-якого банкомату, тому що зв'язки між інформаційними вузлами можуть бути прямими. Реляційні бази даних мають вигляд таблиць, в яких інформація впорядкована по стовпчикам (полям) та рядкам (записам). Для створення та опрацювання БД існують спеціальні програми – системи управління БД. Найпоширенішою сучасною СУДБ є програма з пакета Microsoft Office – MS Access.

База даних – впорядкований набір взаємопов'язаних між собою даних, які використовуються спільно та застосовуються для задоволення потреб користувачів . Головне завдання – збереження великих обсягів інформації та надання доступу до неї прикладній програмі або користувачу.У базі даних основними об'єктами є запити, таблиці, звіти, форми, макроси, модулі та сторінки доступу до даних. Інформація, збережена в комп'ютері й об'єднана у взаємозалежну сукупність за рядом ознак, називається базою даних [38]. Щоб керувати даними, які складають базу, необхідна окрема програма. Програми, які управляють зберіганням, обробкою й пошуком інформації в базі даних, називаються системами керування базами даних [51].

Система керування базами даних – це програма, призначена для організації зберігання, обробки й пошуку інформації в базі даних. На рисунку 2.3 зображено основні можливості СКБД.

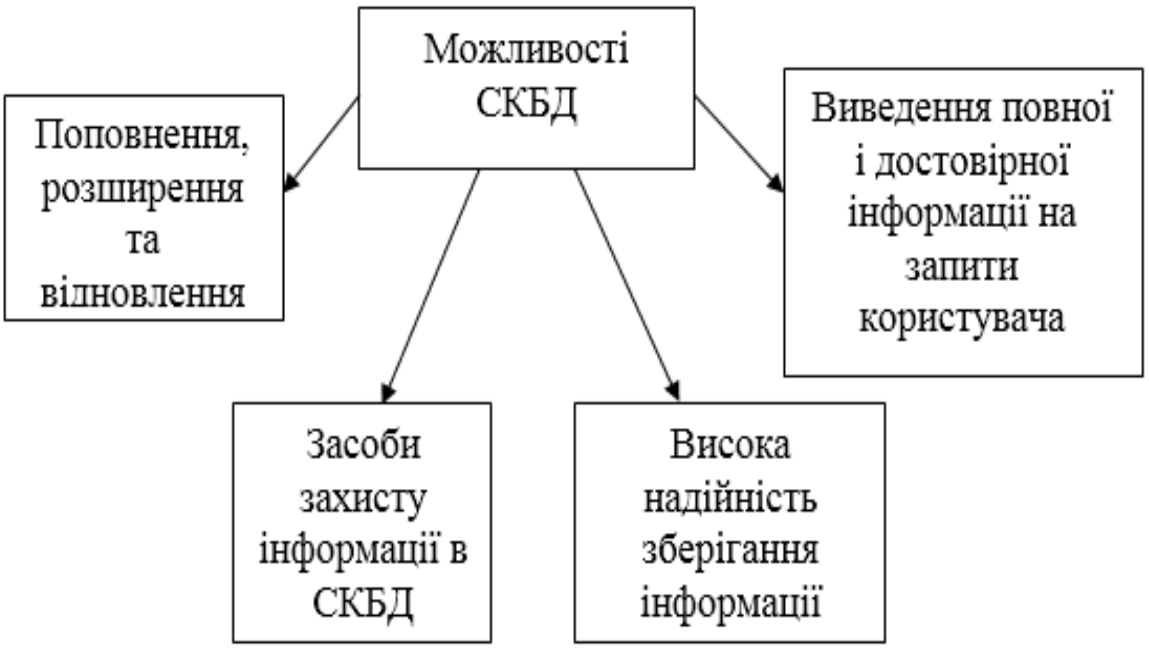


Рисунок 2.3 – Можливості системи керування базою даних

У базі даних основними об’єктами є запити, таблиці, звіти, форми, макроси, модулі та сторінки доступу до даних. Дані у базі організовують відповідно до моделі організації даних. Таким чином, сучасна база даних, крім даних, містить їх опис та може містити засоби для їх обробки. Бази даних класифікують за різними критеріями. На рисунку 2.4 наведено класифікацію бази даних за моделлю даних [39].

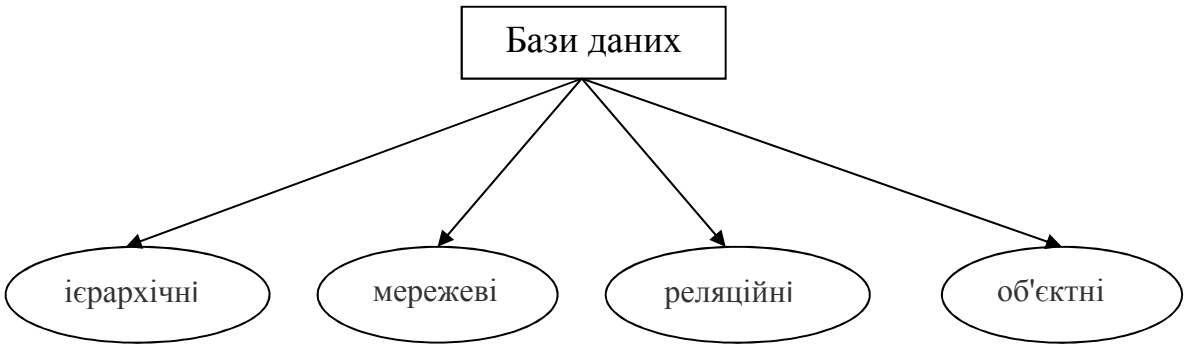


Рисунок 2.3 – Класифікація баз даних за моделлю

Особливістю ієрархічних баз даних є те, що вони відображаються у деревоподібному вигляді . Між об'єктами існують зв'язки, кожен об'єкт включає

в себе об'єкти низького рівня. Дані об'єкти знаходяться у відношенні предка до нащадка.

Мережеві бази даних схожі з ієрархічних, однак, їх особливістю є те, що в них є покажчики в обох напрямках, котрі об’єднують споріднену інформацію. До мережевих баз даних відносять такі поняття: рівень, вузол, зв'язок.

Реляційна база даних − це база даних, заснована на реляційній моделі представлення даних. Для роботи з реляційними базами даних застосовують реляційні СКБД. Інакше кажучи, реляційна база даних − це база даних, яка сприймається користувачем як набір нормалізованих відношень різного ступеню. Метою нормалізації є усунення недоліків структури баз даних, які призводять до шкідливої надмірності в даних, яка в свою чергу потенційно призводить до різних аномалій і порушень цілісності даних [40].

Особливістю об’єктної бази даних є те, що вона не потребує модифікації ядра при додаванні нового типу даних, так як це використовується у реляційних базах даних. У зовнішню структуру бази даних потрапляють нові класи та їх екземпляри. Система керування ними залишається без значних змін. Бази даних класифікують за таким ступенем:

− централізована − це яка повністю підтримується на одному комп'ютері;

− розподілена база даних − складові частини якої розміщуються в різних вузлах комп'ютерної мережі;

− неоднорідна – фрагменти якої підтримуються засобами більше однієї СУБД;

− однорідна − фрагменти розподіленої БД в різних вузлах мережі підтримуються засобами однієї і тієї ж СУБД;

− фрагментована методом розподілу даних є фрагментованість, вертикальна або горизонтальна;

− тиражова розподіл даних відбувається способом тиражування.

У базі даних основними об’єктами є запити, таблиці, звіти, форми, макроси, модулі та сторінки доступу до даних. Розроблена система

передбачає наявність декількох рівнів доступу до системи, тому у базі даних передбачено наявність таблиць для зберігання інформації про користувачів.

Також база даних має 9 типів для зберігання даних, а саме:

− текстовий – використовується для зберігання звичайного тексту;

− числовий – призначений для зберігання дійсних чисел;

− дата/час – для зберігання дати та часу;

− грошовий – для зберігання грошових сум;

− логічний – для зберігання логічних даних (типу «так», «ні»);

− поле МЕМО – спеціальний тип для зберігання великих обсягів тексту;

− гіперпосилання − спеціальне поле для зберігання адреси URL;

− поле об’єкту OLE – призначений для зберігання об’ємів OLE,наприклад мультимедійних;

− рахівник – використовується для запису чисел.

Розробка бази даних відіграє велику роль у проектуванні веб-сайтів чи веб-модулів.

Таблиця 2.1 – Етапи розробки бази даних

№	Назва етапу	Опис
1	Постановка завдання	Формуються основні завдання щодо розробки БД, описується склад БД, призначення та цілі її створення.
2	Аналіз об’єктів	Формується склад об’єктів БД, їхні параметри та типи цих параметрів.
3	Синтез моделі	Вибір моделі бази даних.
4	Програмна частина	Вибір СКБД.

Модель "сутність−зв'язок" базується на важливій змістовній інформації про реальний світ та застосовується для логічного відображення інформації. Дана модель визначає значення даних в контексті їх зв'язку із іншими даними. Із моделі "сутність−зв'язок" породжені такі моделі бази даних: мережева, ієрархічна, об'єктна, реляційна. Будь-який фрагмент предметної області може бути представлений як набір сутностей, між якими існують зв'язки. Загальний опис області дослудження ще називають онтологією предметної області [41].

База даних для модуля агрегації новин складається із таких основних сутностей:

- сутність, де описується сховище даних про новини;
- сутність, де зберігаються дані користувачів таїхні попередні запити;
- сутність, де описується сховище даних про новини, наприклад,назва новини.

У додатку Б наведено структуру бази даних модуля агреггації новин. Основне призначення даної бази даних – зберігання інформації про запити, пошукові запити попередніх користувачів.

У таблиці 2.2 представлено структуру та опис таблиці «Users». Дана таблиця призначена для зберігання інформації про користувачів системи.

Таблиця 2.2 − Структура таблиці «users»

№	Назва поля	Опис
1.	Id	Первинний ключ (Ідентифікатор користувача)
2.	Name	Ім’я
3.	Fathername	По-батькові
4.	Interest	Інтереси
5.	Age	Вік

Щоб пошукова система відображала ключові слова з бази даних розглянемо таблицю 2.3 «keywords». У даній таблиці зберігаються ключові слова кожного користувача за якими буде здійснюватись пошук.

Таблиця 2.3 – Структура таблиці «keywords»

№	Назва поля	Опис
1.	Id	Первинний ключ (Ідентифікатор користувача)
2.	User_id	Ідентифікатор користувача (автора), що додав статтю в базу даних
3.	Text	Повний текст статті (відображається при завантаження сторінки кожної статті)
4.	Sites	Поле,відповідає за сайти які найчастіше переглядаються

Структуру таблиці «sites_roting» наведено у таблиці 2.4. Дана таблиця, окрім полів які відповідають за сайти та вміщує їхню маршрутизацію.

Таблиця 2.4 – Структура таблиці «sites_roting»

№	Назва поля	Опис
1.	Id	Первинний ключ (Ідентифікатор користувача)
2.	User_id	Ідентифікатор користувача (автора), що додав статтю в базу даних
3.	Title	Заголовок
4.	Cout	Лічильник

У таблиці 2.5 знаходиться інформацію про структуру таблиці «Out».

Таблиця 2.5 – Структура таблиці «users»

№	Назва поля	Опис
1.	Id	Первинний ключ (Ідентифікатор користувача)
2.	Login	Логін
3.	Password	Пароль
4.	Outh_key	Ключ авторизації
5.	Created_at	Дата та час створення запису
6.	Updated_at	Дата та час редагування запису
7.	User_id	Унікальний ідентифікатор користувача

2.5 Висновки до другого розділу

Отже, у даному підрозділі на основі дослідження основних етапів розробки баз даних було описано алгоритм формування новин та описано базу даних, розроблено структуру бази даних для функціонування пошуку на запитах попередніх користувачів та описано структуру чотирьох таблиць.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ АЛГОРИТМУ АГРЕГАЦІЇ ПОШУКУ НОВИН

3.1 Структура програмного модуля

3.1.1 Серверна частина

Більшість інформації обробляється на серверній частині веб – сайту, зокрема вибірка інформації з бази даних, пошук по базі даних за ключовими словами, вибірка інформації із випадаючого списку, збереження записів користувачів із форми зворотного зв’язку тощо. Робота більшості веб-сайтів включає в себе клієнтську та серверну частину. Здебільшого всі обчислення та опрацювання інформації введеної користувачем, відбувається на стороні сервера.

Веб-сервер Apache використовується у нас на хостингу як основний. Apache є основою більшості веб-серверів світу і являється найбільш розповсюдженим веб-сервером. Причинами його популярності є такі факти як легка і широка можливість конфігурації, швидкодія, підтримка всіх нині відових протоколів. Спеціально для Apache були створені версії популярних мов програмування як Perl та PHP, а також цей веб-сервер легко інтегрується з розповсюдженими СУБД як MySQL. Користувачі нашого хостингу мають можливість власноруч редагувати конфігурації Apache за допомогою відповідних дериктив в файлі .htaccess. Таким чином можна вирішити більшість задач конфігурації веб-сервера в умовах shared-хостингу. Apache HTTP-сервер − відкритий веб-сервер. Індексний файл - це той файл, який відкривається за замовчуванням при звертанні клієнта до каталогу, а не до конкретного файлу,в першому випадку звертання йде до каталогу сайту, а в другому - до каталогу admin, який знаходиться безпосередньо в каталозі сайту, тобто файл в запиті не вказано. В такому випадку буде запущено на виконання індекний файл, результат його виконання й буде відправлено клієнту на його запит.

За замовчуванням індесним файлом є: index.html, index.htm, index.php, index.php3, index.phtml, index.shtml, default.htm або default.html. Для реалізації серверу було використано API, який є інтерфейсом прикладного програмування. Прикладний програмний інтерфейс API − це набір визначень підпрограм, протоколів взаємодії та засобів для створення програмного забезпечення. Спрощено − це набір чітко визначених методів для взаємодії різних компонентів. Програмний інтерфейс API надає розробнику засоби для швидкої розробки програмного забезпечення. Програмний інтерфейс API може бути для веб-базованих систем. Мова програмування JavaScript є динамічною, об'єктно-орієнтованою прототипною мовою програмування. Найчастіше використовується для створення сценаріїв веб-сторінок, що надає можливість на стороні клієнта (пристрої кінцевого користувача) взаємодіяти з користувачем, керувати браузером, асинхронно обмінюватися даними з сервером, змінювати структуру та зовнішній вигляд вебсторінки. Мову JavaScript класифікують як прототипну (підмножина об'єктноорієнтованої), скриптову мову програмування з динамічною типізацією. Окрім прототипної, JavaScript також частково підтримує інші парадигми програмування (імперативну та частково функціональну) і деякі відповідні архітектурні властивості, зокрема: динамічна та слабка типізація, автоматичне керування пам'яттю, прототипне наслідування, функції як об'єкти першого класу. При розробці великих і нетривіальних веб-застосунків з використанням JavaScript, критично важливим є доступ до інструментів відлагодження. Оскільки браузери, від різних виробників, дещо відрізняються у поведінці JavaScript і реалізації Об'єктної моделі документа, необхідно мати відлагоджувач для кожного браузера, якщо веб-застосування орієнтовано на нього. При розробці великих і нетривіальних веб-застосунків з використанням JavaScript, критично важливим є доступ до інструментів відлагодження. Оскільки браузери, від різних виробників, дещо відрізняються у поведінці JavaScript і реалізації Об'єктної моделі документа, необхідно мати відлагоджувач для кожного браузера, якщо веб-застосування орієнтовано на нього.

Існує безліч версій сервера Apache, однак при розробці розроблювального модуля тестування знань було обрано Apache 2. Основним призначенням Apache 2 є передача через HTTP статичних та динамічних веб-сторінок у всесвітній павутині. Система конфігурації Apache 2 заснована на конфігураційних файлах та має три умовних рівня конфігурації [42]:

− конфігурація сервера (httpd.conf);

− конфігурація віртуального хоста;

− конфігурація рівня директорії (.htaccess).

Структуру програмного модуля тестування знань наведено на рисунку 3.1. У директорії CSS розміщуються файли каскадної таблиці стилів, які відповідають за оформлення блоків сторінок.

Рисунок 3.1 − Структура програмного модуля тестування знань

Файл «config.php» − знаходяться налаштування сайту у вигляді «ключ –> значення», зокрема налаштування підключення до бази даних, дозволені формати завантаження файлів.

Файл «index.php» − головний файл – завантажувач сайту. При зверненні до сайту першим підключається «index.php», який надалі підвантажує файл «config.php» та інші налаштування.

Файл «test.php» − функціонал для тестування алгоритмів пошуку однакових ключових слів.

Файл «test_result.php» − вивід результату роботи сайту підбору новинних повідомлень.

Вимоги до апаратного забезпечення сервера для нормальної роботи розробленого модуля тестування знань наведено у таблиці 3.1. Враховуються такі параметри, як об’єм жорсткого диску, тип процесора, встановлена операційна система, інтерпретатор PHP та веб-сервер [43].

Структура програмного модулю реалізації алгоритму LSA наведено на рисунку 3.2.

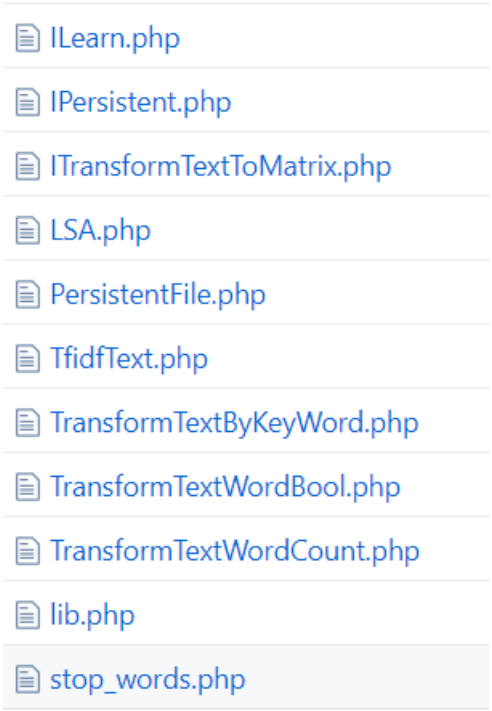


Рисунок 3.2 − Структура програмного модулю реалізації алгоритму LSA

«Ilearn.php» − програмний інтерфейс, який має такі методи: fit, transform, fittransform, save, load.

«ItransformTextToMatrix.php» − програмний інтерфейс, який має такі методи:

− public function transform(array $arDocuments):array;

− public function save(IPersistent $persistent);

− public function load(IPersistent $persistent).

Клас «LSA.php» володіє набором функцій, що безпосередньо реалізують пошук подібних слів в наборі тексту. Зокрема, реалізовано такі методи:

– public function addTextMatrixTransformer(ILearn $matrixTransformer);

– public function fitTransform(array $arDocuments):array;

– public function fit(array $arDocuments);

– public function transform(array $arDocuments):array;

– public function save(IPersistent $persistent);

– public function queryByIndex($index, array $trans).

Конструктор класу LSA:

```
    /**
         * LSA constructor.
         * @param int $nFeatures
         * @param int $nMaxDocuments
         * @param int $nMaxWords
         */
        public function __construct($nFeatures = 5,
$nMaxDocuments = 1000, $nMaxWords = 100)
        {
            $this->nFeatures = $nFeatures;
            $this->nMaxDocuments = $nMaxDocuments;
            $this->nMaxWords = $nMaxWords;
    }
```

Веб – сайт реалізовано з допомогою шаблону проектування «MVC», зображеного на рисунку 3.3.

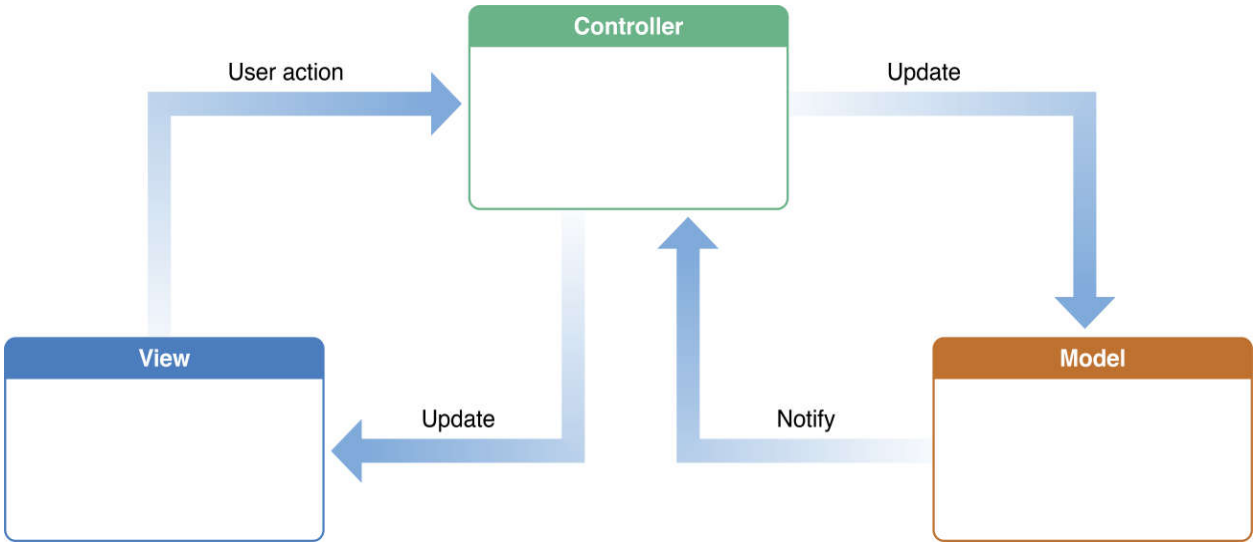


Рисунок 3.3 – Шаблон проектування «MVC»

Модель відповідає за зберігання, обчислення, формування, отримання результатів. Наприклад, запит до бази даних реалізується саме у класі − моделі. У моделях також відбувається формування результату обчислень, стрічок, масивів, об’єктів тощо.

View (Вигляд) призначений для відображення результатів на екран користувача. Наприклад, в результаті дій користувача на сайті відбувається запит до бази даних у моделі, а потім повертається масив. У «View» у циклі відбувається перебір елементів масиву, їх вивід та задання певних стилів (шрифту, фону, кольору тексту).

Контролер є свого роду координатором між моделлю та виглядом. Наприклад, після дій користвача на сайті, на контроллер відправляється запит, контролер передає цей запит моделі. Модель, у свою чергу здійснює SQL − запит до бази даних та повертає результат у формі масиву контролеру. Контролер передає цей масив далі у view. View, у свою чергу виводить цей результат у зручного дял користувача форматі.

Методи класу Shingles навдено на рисунку 3.4.

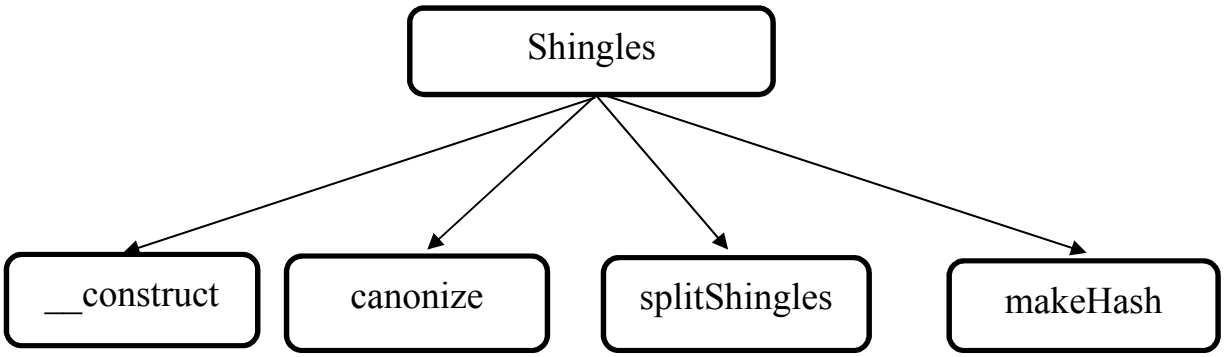


Рисунок 3.4 – Методи класу Shingles

На рисунку 3.4 позначено:

− «__construct» − конструктор класу, викликається при створенні об’єкту класу Shingles.

− «canonize» − вхідним параметром є змінна content, на виході отримуємо масив даних.

− «splitShingles» − вхідним параметром є масив words, вихідним – масив даних array.

− «makeHash» − абстрактний метод, вхідним праметром якого є стрічка shingle. Даний метод призначений для генерування хешу.

Системні вимоги до сервера наведено у таблиці 3.1.

Таблиця 3.1 – Системні вимоги до сервера

Найменування	Параметри
ЦП	2xXeon 3.0 ГГц
ОЗП	2 Гб
Об’єм жорсткого диску	145 Гб
Операційна система	Windows XP/2000/2003 Linux/FreeBSD
Веб-сервер	Apache 1.3/2.0/2.2 із включеним mod_rewrite
Інтерпретатор PHP	PHP версії 4.3 і вище (підійде і PHP 5)

3.1.2 Клієнтська частина

Дизайн сайту відіграє важливу роль та може значно збільшити або зменшити аудиторію користувачів. Для розробки графічного інтерфейсу сайту було обрано фреймворк Twitter Bootstrap. Структуру фреймворку Twitter Bootstrap наведено на рисунку 3.5.

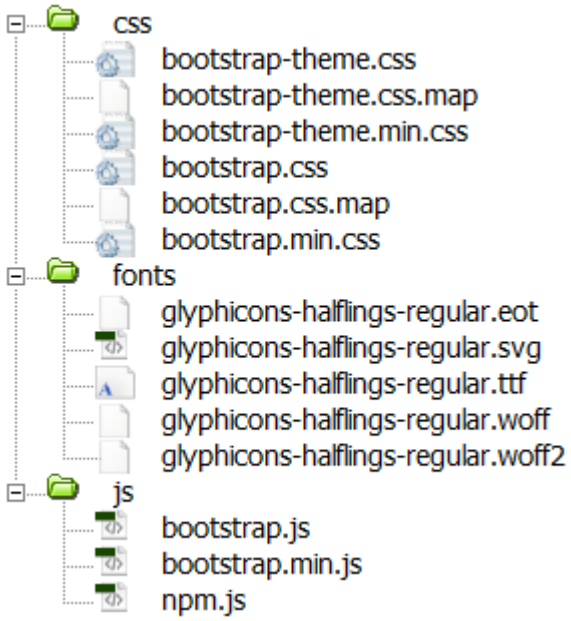
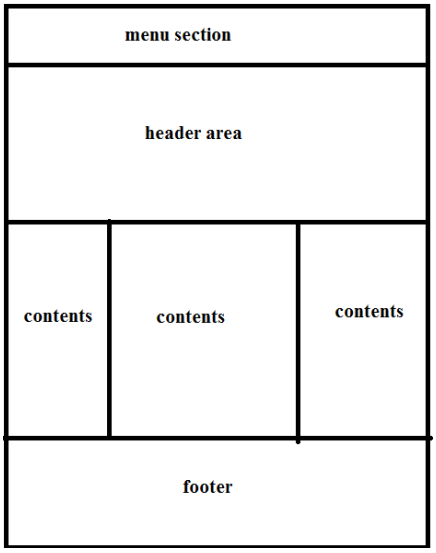


Рисунок 3.5 - Структуру фреймворку Twitter Bootstrap

Папка "css" вміщує таблицю стилів для адаптивного і неадаптивного дизайну. Дана папка необхідна для того, щоб швидше зробити сайт. Папка "js" вміщує bootstrap.js. JavaScript файли, які містять різні компоненти. В папці "img" міститься два набори однакових іконок, за винятком їх фонового кольору. Зображення halflings були надані glyphicons, і надаються безкоштовно в проекті Twitter Bootstrap [44].

Для того щоб активувати фреймворк Twitter Bootstrap, необхідно включити всі відповідні файли і створити HTML структуру. Ми встановлюємо кодування UTF-8, тому що в нашому проекті ми будемо використовувати спеціальні символи, і нам потрібно, щоб браузер коректно їх розпізнавав. Після цього встановлюємо звичайні теги HTML: <html>, <head> і <body>. З міркувань продуктивності веб-сторінки, всі javascript скрипти повинні розташовуватися прямо перед закривається тегом </ BODY>.

Приклад структури сторінки з використанням twitter bootstrap наведено на рисунку 3.6.



Рисунко 3.6 − Структура сторінки з використанням twitter bootstrap

До таблиці стилів CSS може звертатися декілька документів, що значну спрощує роботу програміста. Принцип взаємодії декількох документів із одним файлом CSS наведено на рисунку 3.7.

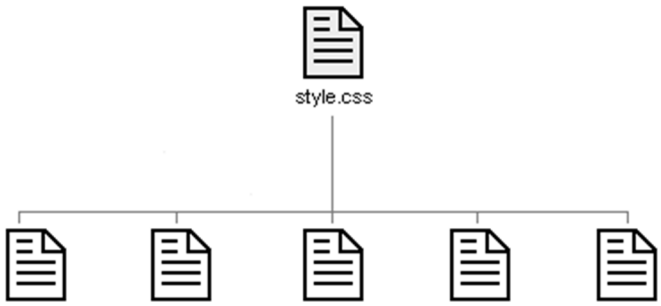


Рисунок 3.7 – Взаємодія декількох документів з одним файлом CSS

Основним призначенням інтерфейсу користувача є обробка та відображення інформації максимально зручно для клієнта. У графічних системах інтерфейс користувача реалізовується багатовіконним режимом, змінами кольору, розміру, видимості вікон, їхнім розташуванням, сортуванням елементів вікон. Інтерфейс відіграє важливу роль у взаємодії користувача з системою, адже він може прискорювати час прийняття рішень та покращувати або погіршувати якість роботи [45]. На рисунку 3.8 зображено інтерфейс головної сторінки сайту.

Рисунок 3.8 – Графічний інтерфейс головної сторінки сайту

У графічному інтерфейсі який зображено на рисунку 3.3 знаходяться посилання на блок пошуку новин, соціальні мережі, категорії новин, рейтинг користувачів також присутня форма зворотного зв’язку.

Блоки головного вікна сайту наведено на рисунку 3.9.

Рисунок 3.9 – Блоки головного вікна сайту. (1) – заголовок сайту; (2) – блоки новин,кожен включає в себе: заголовок,титульне зображення,короткий опис; (3) – блок пошуку новин по базі даних.

Інтерфейс головного вікна є простим та зрозумілим у користуванні. Робота з системою не викликає у користувача ускладнень у пошуках необхідних елементів інтерфейсу. До того ж, передбачено введення

користувачем тільки мінімальної інформації. У стрічці пошук є категорія підказок, користувач клікнувши на неї отримує новини які запитував раніше. Натиснувши на потрібну новину його перекидає на бажану сторінку.

Графічний інтерфейс пошуку новин наведено на рисунку 3.10.

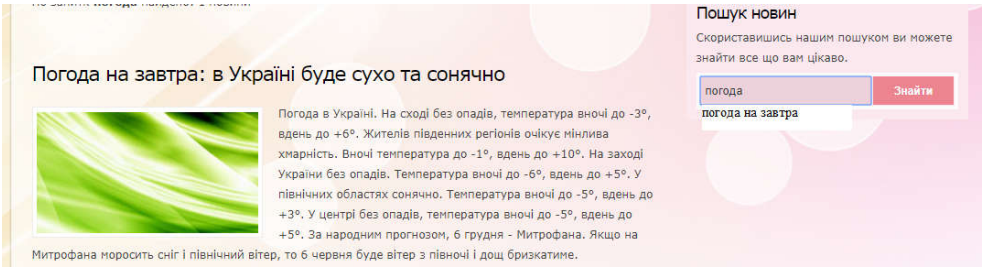


Рисунок 3.10 – Графічний інтерфейс блоку пошуку новин на сайті

У блоці пошук новин вводимо слово погода, з БД підтягується попередні пошуки користувача і відповідно вибираємо те що цікавить.

3.2 Модуль роботи з базою даних

База даних – це інструмент, використовуючи який, можна збирати й упорядковувати інформацію. Під час роботи з великими обсягами інформації значну роль приділяють організації самої бази даних. Наразі найбільш широко застосовується ієрархічна, мережева та реляційна моделі бази даних. Для розробки даного сайту було використано реляційну модель бази даних. Реляційна модель даних являє собою набір двовимірних таблиць, які складаються зі стовпців (полів) і рядків (записів), а також мають ім'я, унікальне в межах даної БД [46].

Для роботи з базою даних у даній дипломній роботі було використано систему керування базою даних MySQL. MySQL – вільна система керування реляційними базами даних. Характеризується високою швидкістю, стійкістю і

простотою використання. Зараз MySQL – одна з найпоширеніших систем керування базами даних. Вона використовується, в першу чергу, для створення динамічних веб – сторінок, оскільки має чудову підтримку з боку різноманітних мов програмування.

Для адміністрування сервера MySQL, перегляду та редагуванню вмісту таблиць баз даних було використано PhpMyAdmin. PhpMyAdmin дозволяє через браузер здійснювати адміністрування сервера MySQL, запускати запити SQL, переглядати та редагувати вміст таблиць баз даних. Ще однією вагомою перевагою використання PhpMyAdmin є можливість додавати, видаляти, редагувати вміст таблиць без використання SQL команд [47].

Графічний інтерфейс вікна СУБД phpmyadmin для таблиці «articles» наведено на рисунку 3.11.

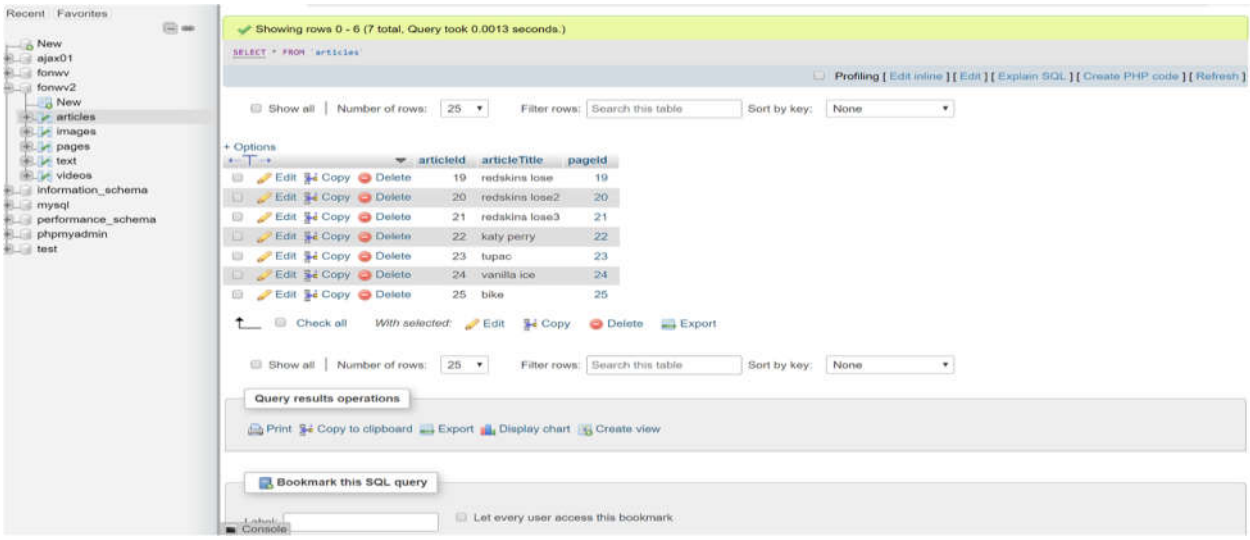


Рисунок 3.11 – Графічний інтерфейс вікна СУБД phpmyadmin для таблиці «articles»

Приклад коду для налаштування доступу до БД:

```
'mysql' => [
            'driver' => 'mysql',
            'host' => env('DB_HOST', 'localhost'),
            'port' => env('DB_PORT', '3306'),
```

```
            'database' => env('DB_DATABASE', 'forge'),
            'username' => env('DB_USERNAME', 'forge'),
            'password' => env('DB_PASSWORD', ''),
            'charset' => 'utf8',
            'collation' => 'utf8_unicode_ci',
            'prefix' => '',
            'strict' => true,
            'engine' => null,
        ],
```

Приклад файл для формування запиту до бази даних такий:

```
class out extends Model
{
    protected $table = 'news';
    protected $fillable = [
        'id',   'login',   'password',   'out_key',   'created_at',
'updated_at', 'user_id'
}
```

Клас «Out» унаслідує клас «Model». У масиві $fillable знаходиться список полів таблиці бази даних 'out', доступних для запиту та виводу на екран.

Лістинг файлу для отримання інформації з БД в контролері такий:

```
     public function index()
         {
             $news = news::paginate(10);
             return view('news.index', compact('news'));
         }
```

Метод paginate() відповідає за вибірку кількості записів з БД. У даному випадку – 10. За допомогою цієї функції можна здійнити посторінковий вивід інформації для зручності. У метод compact() передається масив з БД. На стороні користувача у циклі відбувається вивід інформації за певним ключем [48].

3.3 Порівняльний аналіз розробленого сайту з аналогами

Інтерфейс розробленого модуля новин агрегації є водночас простим та зрозумілим. Робота з системою не викликає у користувача ускладнень у пошуках необхідних елементів інтерфейсу. До того ж, передбачено введення користувачем тільки мінімальної інформації. У пошуку передбачено наявність підказок, для зручної та зрозумілої роботи. Саме тому, щоб забезпечити максимальний комфорт було створено адаптивний інтерфейс сайту. Зовнішній інтерфейс сайту на моніторах із різними розширеннями наведено на рисунку 3.12.

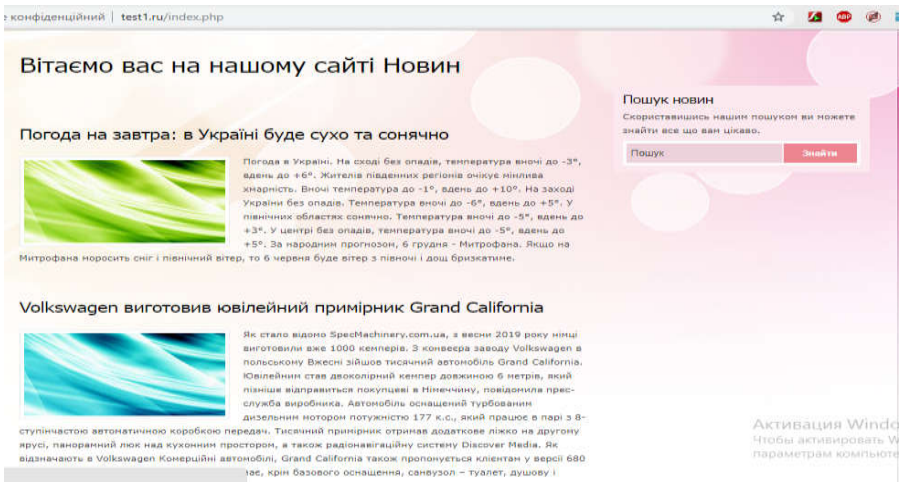


Рисунок 3.12 – Зовнішній вигляд сторінки новин розроблювального модуля

Ще однією перевагою даного сайту є асинхронні запити до сервера. При асинхронному запиті користувач може продовжувати переглядати контент

сайту, поки сервер все ще обробляє запит [52]. Браузер не перезавантажує web-сторінку і дані посилаються на сервер без візуального підтвердження. Не менш важливою перевагою даного сайту у порівнянні з аналогами є наявність рейтингу користувачів. Адміністратор має можливість виставляти бали користувачам за запрошених друзів на сайт і таким чином формувати рейтинг. Також реалізовано сортування за оцінками [53]. Крім цього адміністратор може переглядати зареєстрованих користувачів та здійснювати розсилку електронних повідомлень. Порівняльний аналіз розробленого сайту із аналогами наведено у таблиці 3.2.

Таблиця 3.2 – Порівняльний аналіз розробленого сайту із аналогами

Критерій	Сайт Ukr.net	Сайт Avtoria	Розроблений сайт
Адаптивний інтерфейс	–	+	+
Асинхронні запити до сервера	–	–	+
Перегляд зареєстрованих користувачів	+	+	+
Формування рейтингу користувачів	–	–	+
Розсилка електронних повідомлень	+	+	+

Як видно з таблиці 3.2 розроблений веб – ресурс не поступається аналогам, а у більшості випадків переважає їх [54].

Порівняльний аналіз використання алгоритмів LSA та W − shingle наведено у таблиці 3.3

Таблиця 3.3 − Порівняльний аналіз алгоритмів LSA та W−shingle

К-сть записів	LSA	W-shingle
50	92	90
100	89,5	72
150	85	70,5
200	83	65

Графічне представлення наведено на рисунку 3.13

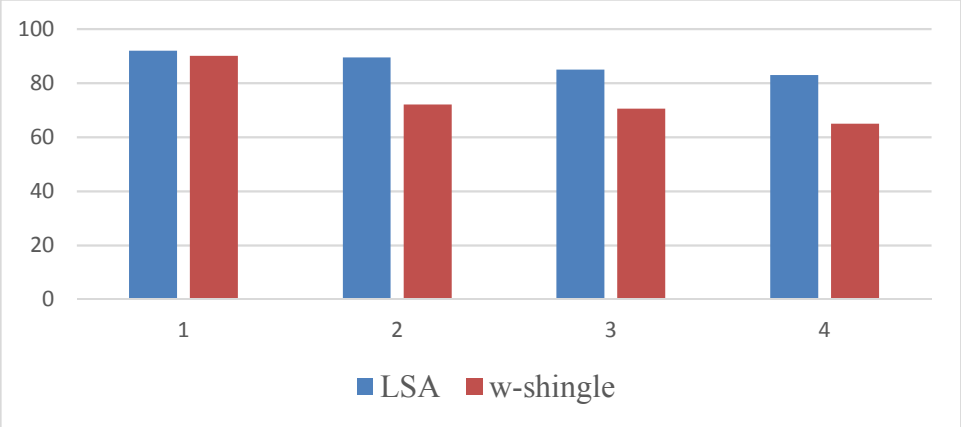


Рисунок 3.13 – Графік порівняння алгоритмів LSA та W− shingle

Отже, у даному розділі представлено структуру серверної та клієнтської частини алгоритму агрегації, графічний інтерфейс сторінки адміністрування, та головної сторінки сайту, модуль роботи бази даних, а також визначено переваги та недоліки розроблювального сайту. Для цього було здійснено порівняльний аналіз за кількома параметрами розроблювального сайту із уже існуючими [55].

3.4 Висновки до третього розділу

Отже, у даному розділі представлено структуру серверної та клієнтської частини модуля агрегації новин, інтерфейс користувача та послідовність дій при роботі з модулем та визначено переваги та недоліки розроблювального у даному дипломному проекті. Для цього було здійснено порівняльний аналіз за кількома параметрами розроблювального модуля агрегації новин із уже існуючими. До основних переваг даного модуля можна віднести адаптивний інтерфейс.

ВИСНОВКИ

Отже, в результаті написання магістерської роботи можна зробити наступні висновки:

1. Проведено аналіз підходів до алгоритму агрегації контенту. На основі аналітичного підходу проведено порівняльний аналіз існуючих агрегаторів контенту, виділено їх переваги та недоліки, що дозволить врахувати їх під час розробки власного веб – ресурсу [49].

2. Розроблено алгоритм формування новин на основі попередніх запитів користувачів з використанням технології vue.js, laravel та бази даних mysql. Розроблено структуру таблиць бази даних для зберігання інформації про користувачів системи, новин. Запропонована структура таблиць дозволяє здійснювати швидкий пошук по блоках які є на розробленому сайті

3. Представлено структуру клієнтської та серверної частини алгоритму агрегації пошуку новин.

4. Здійснено розробку модуля агрегації по базі даних.

5. Здійснено порівняльний аналіз розробленого сайту із існуючими аналогами, що дозволило виділити переваги та недоліки кожного сайту. До переваг розробленого сайту можна віднести: адаптивність до дисплеїв із різним розширенням, наявність модулів пошуку та фільтрації інформації про новини в режимі реального часу [50].

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Методичні вказівки до оформлення курсових проектів, звітів про проходження практики, випускних кваліфікаційних робіт для студентів спеціальності «Комп'ютерна інженерія» / І.В. Гураль, Л.О. Дубчак / Під ред. О.М. Березького. Тернопіль: ТНЕУ, 2019. 33 с.

2. Методичні вказівки до оформлення курсових проектів, звітів про проходження практики, випускних кваліфікаційних робіт для студентів спеціальності «Комп'ютерна інженерія» / І.В. Гураль, Л.О. Дубчак / Під ред. О.М. Березького. Тернопіль: ТНЕУ, 2018. 42 с.

3. Електронна енциклопедія Вікіпедія: веб − сайт. URL: https://uk.wikipedia.org/wiki/internet (дата звернення: 31.09. 2019).

4. Поняття про фейк та його види у ЗМІ : веб − сайт. URL: http://publications.lnu.edu.ua/collections/index.php/teleradio/article/viewFile/694/699 (дата звернення: 31.09.2019).

5. Дезінформація: веб − сайт. URL: http://texty.org.ua/pg/article/editorial/read/85998/Jak_vidriznaty_spravzhni_novyny_vid_brehni_manipulacij (дата звернення: 31.09.2019).

6. Електронна енциклопедія Вікіпедія: веб − сайт. URL: https://uk.wikipedia.org /wiki/dezinformaciya (дата звернення: 31.09.2019).

7. Бібліотека і доступність інформації у сучасному світі: електронні ресурси науці, культурі та освіті: (Підсумки 10-ї міжнар. конф. «Крим-2003») / Л.І. Костенко, А.О. Чекмарьов, А.Г. Бровкін, І.А. Павлуша // Бібл. Вісн. – 2003. – № 4. – С. 43. URL: до журн.:http://www.nbuv.aov.ua/articles/ 2003/03klinko.htm (дата звернення 31.09.2019).

8. Усе про фейки: веб – сайт. URL : https://ranok.ictv.ua/ua/videos/fejki-v-interneti-yaku-zagrozu-voni-nesut-ta-yak-yim-protistoyati/ (дата звернення: 31.09.2019).

9. Товарні агрегатори: веб – сайти. URL: https://www.insales.ru/blogs /university/tovarnyy (дата звернення: 31.09.2019).

10. Що таке сайти агрегатори?: веб – сайт. URL: http://slaidik.com.ua/shho-take-sajti-agregatori-i-chomu-voni-v-topi-zamist-vashih-proektiv-eksperiment-po-stvorennyu-i-prosuvannyu-agregatora/(дата звернення: 31.09.2019).

11. Критерії відбору новин: веб – сайт. URL: https://stud.com.ua/39037/sotsiologiya/kriteriyi_vidboru_novin (дата звернення: 31.09.2019).

12. Дайджести як головний формат регіональної журналістики: веб – сайт. URL: http://ms.detector.media (дата звернення 31.09.2019).

13. Житомирські медіа: про стандарт достовірності та місцеву комунікацію: веб – сайт. URL : http://ms.detector.media/monitoring/regional / (дата звернення 31.09.2019).

14. Регіональні сайти – для кого і на кого працюють: веб – сайт. URL: http://ms.detector.media/monitoring/regional_news (дата звернення 31.09.2019)

15. Ставлення населення до ЗМІ та споживання різних типів ЗМІ в Україні: веб − сайт. URL: https://internews.in.ua/wpcontent (дата звернення 31.09.2019).

16. Пошуковий запит: веб – сайт. URL:https://igroup.com.ua/seo-articles/zapyt (дата звернення 31.09.2019).

17. Класифікація пошукових запитів: веб – сайт. URL: https://igroup.com.ua/seo-articles/klasyfikatzapytiv (дата звернення 31.09.2019)

18. Найпопулярніші пошукові запити користувачів за 2017 рік: веб – сайт.

URL: https://www.unn.com.ua/uk/news (дата звернення 31.09.2019).

19. Дейт К. Руководство по реляционной СУБД DB2. М.: Финансы статистика, 1988. 320 с.

20. Мейер М. Теория реляционных баз данных. М.: Мир, 1987. 608 с.

21. Семантична модель: База даних: веб –сайт.URL: http://citforum.ru/database/advanced_intro/27 (дата звернення 10.10.2019).

22. Риккарди Грег. Системы баз данных. Теория и практика использования в Internet. М.:Вильямс, 2001. 240с.

23. Дейт К. Дж. Введение в системы баз данных. СПб: Изд-во "Пітер", 2005. 1315с.

24. Роберт І.В. Сучасні інформаційні технології в освіті: дидактичні проблеми, перспективи використання . М.: Школа-Пресс, 1994 . 205с.

25. Мюллер Р.Дж. Базы данных и UML. Проектирование / Первод. с англ. Е. Молодцова. М.: Издательство “Лори”,2002. 432с.: ил.

26. Компонентне або модульне тестування:веб – сайт. URL: http://www.protesting.ru/testing/levels/component (дата звернення 10.10.2019).

27. Електронна енциклопедія Вікіпедія:Список кодів стану HTTP :веб – сайт. URL: https://uk.wikipedia.org/wiki/список (дата звернення 10.10.2019)

28. Шапошников І. Web-сайт своїми руками./ І. Шапошников – СПб: Изд-во "Пітер", 2002 – 390с.

29. Гаевский А. Ю. Самоучитель по созданию Web − страниц: HTML, JavaScript, Dynamic HTML / А. Ю. Гаевский, В. А.Романовский. К.: А.С.К., 2002 472с.

30. Кнут Д. Искусство программирования на ЭВМ: В 4т. / Д.Кнут/ М.: Мир, 1977. Т.2: Получисленные алгоритмы . 728 с.

31. Рост Р.Д. OpenGL. Трехмерная графика и язык программирования шейдеров / Р.Д. Рост/ Пер. с англ. СПб.: Питер, 2005. 428 с.

32. Англо − російсько-український тлумачний словник з комп'ютерної графіки та обробки зображень [Текст] / укл. Р. М. Паленичка, В. В. Грицик ; ред.-лексикограф В. М. Брицин. К. : Наук. думка, 1994. 286 с.

33. Фролов А. В. Практика применения PERL, PHP, APACHE и MySQL для активных Web-сайтов / А. В. Фролов, Г. В.Фролов.М.: Русская редакция, 2002. 534с.

34. Шлосснейгл Д.Л. Профессиональное программирование на PHP. М.: Русская редакция, 2006. 624с.

35. Федорчук А.І. Як створюються Web − сайти. СПб.: Пітер, 2000. 224с.

36. Дейт К.Н. Руководство по реляционной СУБД DB2. М.: Финансы и статистика, 1988. 320 с.

37. Електронна енциклопедія Вікіпедія: Реляційні ресурс] БД: веб – сайт. URL:http://ru.wikipedia.org/wiki/ Реляційна БД (дата звернення 20.10.2019).

38. Дейт К. Дж. Введение в системы баз данных. СПб: Изд-во "Пітер", 2005. – 1315с.

39. Коннолли Т.А. Базы данных. Проектирование, реализация и сопровождение. Теория и практика / Т.А. Коннолли, К.М Бегг. 3 е изд. М.: Вильямс, 2003. 1436 с.

40. Коннолли Т.А. Базы данных: проектирование, реализация и сопровождение / Т.А Коннолли, К.М Бегг, А.О Страчан. 2 − е изд. М. : Вильямс, 2000. – 1120 с.

41. Камер Д. Компьютерные сети и Internet М. : Вильямс, 2002. 640 с.

42. Андон Ф.И. Информационные системы / Ф.И. Андон, В.П. Резниченко, У.У. Яшунин К., 2001. 396 с.

43. Дронов В. PHP 5/6, MySQL 5/6 и Dreamweaver CS4. Разработка интерактивных Web−сайтов. БХВ-Петербург. Москва, 2009. 544 с.

44. Петюшкин А.В. HTML экспресс – курс. М.: СПб: БХВ. Петербург, 2004. − 250 с.

45. Поломошнов О.Б. Быстро и легко создаем, программируем и раскручиваем Web – сайт. М.: Эксмо, 2011. 352 с.

46. Курочкин А.С. Организация производства, Учебн.пособие для студентов вузов. / К.: МАУП, 2001. 216с.

47. Багрова І.В. Організація виробництва. Київ, ЦНЛ, 2005. 248с.

48. Герасимчук В.Г. Економіка і організація виробництва / В.Г. Герасимчук, А.Е.Розенплентер, В.І. Кривда. К.: Політехніка, 2007р.

49. Круш П.В. Організація виробництва / П.В. Круш, В.І.Подвігіна, В.О.Гулевич. К.: ЦУЛ, 2008. 348 с.

50. Гаевский А.Ю. 100% самоучитель. Создание Web − страниц и Web-сайтов. HTML и JavaScript / А.Ю. Гаевский, В.А. Романовский. М.: СПб. [и др.] : Питер, 2008. 464 с.

51. Новосад Х.В., Піцун О.Й. Алгоритм агрегації новин на основі попередніх запитів користувачів: зб. матеріалів доп. учасн. Наук. − практ. конф. Тернопіль: Тернопіль, 2019. с. 39.

52. Новосад Х.В., Троць І.І. База даних модуля агрегаціїновин на сайті: зб. матеріалів доп. учасн. ІІ Наук. − практ. конф. Тернопіль : Тернопіль, 2019. с. 22.

53. Електронна енциклопедія Вікіпедія: Список кодів стану: веб−сайт. URL: https://uk.wikipedia.org/wiki/список (дата звернення: 02.12.2019).

54. Шапошников І. Web-сайт своїми руками./ І. Шапошников – СПб: Изд-во "Пітер", 2002. 390с.

55. Гаевский А. Ю. Самоучитель по созданию Web-страниц: HTML, JavaScript, Dynamic HTML / А. Ю. Гаевский, В. А.Романовский. К.: А.С.К., 2002. 472с