

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ
ФАКУЛЬТЕТ КОМП'ЮТЕРНО-ІНФОРМАЦІЙНИХ СИСТЕМ
І ПРОГРАМНОЇ ІНЖЕНЕРІЇ

ШАПОВАЛОВА АНАСТАСІЯ СЕРГІЇВНА

УДК 004.415.2

**МЕТОДИ ТА АЛГОРИТМИ ПОШУКУ ІНФОРМАЦІЇ З
ВИКОРИСТАННЯ МЕТА-ПОШУКОВИХ СИСТЕМ**

123 «Комп'ютерна інженерія»

Автореферат

дипломної роботи на здобуття освітнього ступеня «магістр»

Тернопіль, 2018

Роботу виконано на кафедрі комп'ютерних систем та мереж Тернопільського національного технічного університету імені Івана Пулюя Міністерства освіти і науки України

Керівник роботи: кандидат технічних наук, доцент кафедри комп'ютерних систем та мереж
Шингера Наталія Ярославівна,
Тернопільський національний технічний університет імені Івана Пулюя,

Рецензент: кандидат технічних наук, доцент кафедри програмної інженерії
Кінах Ярослав Ігорович,
Тернопільський національний технічний університет імені Івана Пулюя,

Захист відбудеться 23 лютого 2018 р. о 9⁰⁰ годині на засіданні екзаменаційної комісії №34 у Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56, навчальний корпус №1, ауд. 1-603

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми роботи. На ранніх стадіях еволюції в алгоритмі web-пошукових системах враховувалося мала кількість факторів, що впливають на ранжування у видачі результатів пошуку, тому, знаючи базові принципи роботи пошукових систем, можна було досить легко маніпулювати результатами, що й робили багато компаній, які займалися просуванням Інтернет ресурсів. Для того щоб підтримувати якість пошуку, самого важливого аспекту, пошукові системи були змушені ускладнювати свої алгоритми - кількість чинників, що враховуються зросли в сотні, і навіть тисячі разів, стали все частіше з'являтися різні алгоритми і фільтри.

При подальшому розвитку кількість таких чинників зростала в геометричній прогресії, постійно збільшувався рівень конкуренції між пошуковими системами, виводячи на пошуковий ринок тільки ті ресурси, які надавали користувачам релевантні результати незалежно від вміння seo-фахівців просувати сайти.

Ресурси Інтернету давно перестали бути просто розвагою, перетворившись в незамінний інструмент для повсякденної роботи людей багатьох професій. Швидке зростання кількості інформації в мережі зробили його океаном найрізноманітніших даних, важливість яких зростає пропорційно їх обсягу. За оцінкою фахівців обсяг інформації, що передається по каналах всесвітньої павутини, подвоюється кожні півроку. Щодня в мережі з'являються сотні тисяч нових документів, і природно, що без систем пошуку вони в значній мірі залишилися б не актуальними, або не знаходилися б взагалі. Виникла необхідність створення таких засобів, які дозволили б легко орієнтуватися в інформаційних ресурсах глобальних мереж, швидко і надійно знаходити потрібні відомості. В інтернеті з'явилися спеціальні пошукові засоби.

Ще декілька років тому існувала така думка: «в Інтернеті є все, але знайти там нічого неможливо». Однак з появою і швидким розвитком пошукових каталогів, пошукових машин, і різноманітних пошукових програм ситуація змінилася, і тепер в мережі необхідну інформацію іноді можна знайти швидше, ніж в книзі, що лежить на столі.

Таким чином, актуальність роботи полягає у проведенні наукового дослідження в області інформаційного пошуку в Інтернеті, результати якого націлені, перш за все, на створення нових пошукових систем або на удосконалення алгоритмів роботи вже існуючих засобів інформаційного пошуку.

Мета і задачі дослідження. Мета роботи полягає у дослідженні існуючих алгоритмів та методів пошуку в сучасних пошукових системах для виявлення основних факторів, що впливають на ранжування сайтів в результатах пошуку.

Досягнення поставленої мети передбачає розв'язання наступних завдань:

- провести аналіз сучасних пошукових систем, використовуваних в українському і світовому сегменті мережі Інтернет;
- розробити систему факторів, що беруть участь в формулах ранжування пошукових систем;
- здійснити огляд алгоритмів пошукових систем;

- розглянути теоретичні підходи до обґрунтування проблеми пошукової оптимізації та ранжування веб-сайтів;
- з'ясувати сутність пошукової оптимізації та ранжування як предметів дослідження, охарактеризувати фактори ранжування;
- надати рекомендації щодо розробки програмного модуля на основі власного алгоритму пошукової системи;
- розробити власну мета-пошукову систему.

Об'єкт дослідження – мета-пошукова система, створена на платформі .NET на базі аналогів Wikipedia, Bing та DuckDuckGo.

Предмет дослідження – методи та алгоритми пошукової оптимізації та ранжування досліджуваної системи з врахуванням існуючих альтернативних варіантів.

Методи дослідження. Для вирішення поставлених завдань використані методи порівняльного аналізу, об'єктно-орієнтованого аналізу, розрахована спрощена формула для визначення релевантності сторінки сайту, за допомогою схем, графіків, діаграм та спостережень.

Наукова новизна отриманих результатів:

- запропоновано структурну модель мета-пошукової системи на базі аналогів Wikipedia, Bing та DuckDuckGo з аналізом отриманих даних та пошуком входжень за допомогою регулярних виразів;
- обґрунтовано організацію алгоритмів пошуку, фільтрації, ефективності ранжування результатів, а також інших показників якості, що впливають на пошук інформації в веб-пошукових системах;
- розроблено мета-пошукову систему, що враховує результати видачі та забезпечує ефективний і релевантний пошук необхідного інтернет ресурсу безпосередньо користувачеві у вигляді відсоткових діаграм;
- набуло подальшого розвитку застосування алгоритмів та методів представлення результату в системах інформаційного пошуку для аналізу та оцінки якості пошуку.

Теоретична цінність роботи полягає в можливості використання отриманих науково-технічних результатів при експлуатації, дослідженні, що вимагають відносного порівняння альтернативних алгоритмів і методів.

Практична цінність дослідження полягає в підвищенні ефективності функціонування мета-пошукових систем на основі організації програмного модуля, що забезпечує ранжування видань за їх інформативністю та створення адаптивного інтерфейсу користувача.

Апробація. Окремі результати дослідження апробовано на науково-практичних конференціях Тернопільського національного технічного університету міста Тернопіль у вигляді тез конференцій.

Структура роботи. Робота складається з розрахунково-пояснювальної записки та графічної частини. Розрахунково-пояснювальна записка складається із вступу, 7 частин, висновків, списку використаних джерел та додатків. Обсяг роботи: розрахунково-пояснювальна записка – 111 арк. формату А4, графічна частина – 8 аркушів формату А1.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність дослідження, мету роботи, задачі, об'єкт, предмет, наукову новизну, практичну та теоретичну цінність, апробацію та публікації дипломних досліджень.

У першому розділі дипломної роботи «Походження веб-пошукових систем в мережі Інтернет» було розглянуто історію виникнення та розвиток пошукових систем. Здійснено опис механізму пошуку мережових інформаційних ресурсів Інтернет та загальних принципів роботи пошукових систем. Проведено огляд існуючих пошукових систем, а саме: найпопулярніших пошукових систем, мета-пошукових систем, електронних бібліотек, інтелектуальних та нестандартних пошукових систем та обґрунтовано сутність мета-пошукової системи.

У другому розділі дипломної роботи «Алгоритми пошуку і фільтрація веб-пошукових систем» визначено та обґрунтовано структуру пошукових систем їх фактори ранжування та принципи індексації сайтів. Здійснено опис методів і алгоритмів сортування результатів пошуку. Детально досліджено фактори ранжування в пошукових системах Bing, DuckDuckGo та онлайн енциклопедії Wikipedia.

У третьому розділі дипломної роботи «Аналіз та порівняння якості пошуку існуючих веб-пошукових систем» проведено аналіз та порівняння якості пошуку існуючих веб-пошукових систем за допомогою таблиць, графіків та діаграм, в тому числі дослідження світових показників та в Україні. А також здійснено детальний аналіз пошукових систем у створеній програмі.

У четвертому розділі дипломної роботи «Реалізація мета-пошукової системи» реалізована мета-пошукова система на платформі .NET на базі аналогів Wikipedia, Bing та DuckDuckGo. Обґрунтовано вибір програмного середовища для реалізації власної мета-пошукової системи. Здійснено детальний опис архітектури програми, в тому числі проведено оцінювання функціональності роботи розробленої мета-пошукової системи.

У п'ятому розділі «Обґрунтування економічної ефективності» виконано обчислення показників економічної ефективності для реалізації мета-пошукової системи, що забезпечило можливості для обґрунтування доцільності проведення НДР.

У шостому розділі дипломної роботи «Охорона праці та безпека в надзвичайних ситуаціях» проведено аналіз вимог з охорони праці та техніки безпеки в процесі використання комп'ютерної техніки, в тому числі розглянуто фактори, що впливають на функціональний стан користувачів комп'ютерів.

У сьомому розділі дипломної роботи «Екологія» розглянуто ряд питань, пов'язаних із екологічними аспектами навколишнього середовища. Зокрема, звернуто увагу на: енергозбереження і його роль у вирішенні екологічних проблем та форматування бази статистичних даних в екології.

У загальних висновках щодо дипломної роботи описано одержані в процесі виконання дипломної роботи магістра результати, що відображають сучасний стан пошуку із використанням можливостей Інтернету та отримані

наукові та практичні результати, запропоновані автором.

В додатках до пояснювальної записки приведено копії наукових публікацій автора та лістинги коду реалізованої мета-пошукової системи.

В графічній частині до дипломної роботи магістра наведено тему і актуальність, мету, завдання, об'єкт, предмет, наукову новизну роботи, аналіз існуючих пошукових систем, загальну архітектуру роботи, алгоритм роботи мета-пошукової системи, пошук інформації 3 різними методами, висновки.

ВИСНОВКИ

В даній дипломній роботі проведено дослідження методів та алгоритмів пошуку інформації з використанням мета-пошукових систем. Результати аналізу пошукової галузі дозволили визначити основні фактори та принципи ранжування сайтів і окремих веб-документів в пошуковій видачі.

В процесі виконання роботи були вирішені наступні завдання:

- проведено аналіз сучасних пошукових систем, використовуваних в українському та світовому сегменті мережі Інтернет;
- розроблено систему факторів, що беруть участь у формулах ранжування пошукових систем;
- виявлені показники, що впливають на якість алгоритмів пошукових систем;
- розглянуто теоретичні підходи до обґрунтування проблеми пошукової оптимізації та ранжування web-сайтів;
- з'ясовано сутність пошукової оптимізації та ранжування, як предметів дослідження, охарактеризовано чинники ранжування;
- розроблено програмний модуль на основі власного алгоритму пошукової системи;
- реалізовано власну мета-пошукову систему.

Зростання інформації нескінченна, а тому немає межі вдосконалення пошукових машин. Найважливішим завданням розробників пошукових систем є поліпшення якості пошуку, покращення ефективності та зручності у використанні системи. З цією метою постійно змінюються пошукові алгоритми, фільтри, створюються додаткові сервіси, допрацьовується і оптимізується дизайн. Також удосконалюється галузь SEO-оптимізації, де фахівці в цій галузі намагаються домогтися високих результатів в пошуковій видачі.

Висновки проведені під час дослідження в області глобального пошуку інформації були корисні для розробки власної мета-пошукової системи.

На основі отриманих даних розроблена мета-пошукова система враховує результати видачі та забезпечує ефективний і релевантний пошук необхідного Інтернет ресурсу безпосередньо користувачеві у вигляді відсоткових діаграм. Її переваги та новизна полягають в особливій системі ранжування та подання результатів. Реалізована система має унікальний дизайн, оригінальну видачу вмісту пошуку, а також широкий функціонал.

Грунтуючись на проведених обчисленнях техніко-економічних показників ефективності доведено економічну доцільність та ефективність розробки для реалізації мета-пошукової системи, а також прийнято рішення щодо подальшого розвитку і впровадження даної розробки. Розраховане значення економічної ефективності становить 0,56, що є високим значенням. Аналогічно нормальним є термін окупності. Для даного дослідження він становить 1,8 року.

Проведено аналіз вимог з охорони праці, електро- та пожежної безпеки на робочому місці, дотримання санітарних норм та техніки безпеки в процесі використання комп'ютерної техніки, в тому числі для проведення досліджень та експлуатації системи автоматизації процесів як апаратної і програмної системи. Також, розглянуто фактори, що впливають на функціональний стан користувачів комп'ютерів.

Розглянуто ряд питань з екології, пов'язаних із екологічними аспектами навколишнього середовища. Зокрема, звернуто увагу на енергозбереження і його роль у вирішенні екологічних проблем. Проаналізовано питання форматування бази статистичних даних в екології.

СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

1. Шаповалова А. Дослідження методів і алгоритмів пошуку в web-пошукових системах / Шаповалова А. // Збірник тез X Всеукраїнської студентської науково-технічної конференції «Природничі та гуманітарні науки. Актуальні питання», 25-26 квітня 2017 року. — Т. : ТНТУ, 2017. — Том 1. — С. 99–100. — (Секція: Інформаційні технології).
2. Шаповалова А. С. Розробка архітектури мета-пошукової системи «Їжачок» / А. С. Шаповалова // Збірник тез доповідей VI Міжнародної науково-технічної конференції молодих учених та студентів «Актуальні задачі сучасних технологій», 16-17 листопада 2017 року. — Т. : ТНТУ, 2017. — Том 2. — С. 186–187. — (Комп'ютерно-інформаційні технології та системи зв'язку).
3. Шаповалова А. С. Реалізація мета-пошукової системи на платформі .NET / А. С. Шаповалова // Збірник тез доповідей V науково-технічній конференції «Інформаційні моделі, системи та технології», 1-2 лютого 2018 року. — Т. : ТНТУ, 2018. — Том 1. — С. 87. — (Комп'ютерні системи та мережі).

АНОТАЦІЯ

Мета дипломної роботи полягає у дослідженні існуючих алгоритмів та методів пошуку в сучасних пошукових системах для виявлення основних факторів, що впливають на ранжування сайтів в результатах пошуку.

Актуальність дипломної роботи полягає у проведенні наукового дослідження в області інформаційного пошуку в Інтернеті, результати якого націлені, перш за все, на створення нових пошукових систем або на

удосконалення алгоритмів роботи вже існуючих засобів інформаційного пошуку.

Об'єктом дослідження є мета-пошукова система, створена на платформі .NET на базі аналогів Wikipedia, Bing та DuckDuckGo.

Предмет дослідження – методи та алгоритми пошукової оптимізації та ранжування досліджуваної системи з врахуванням існуючих альтернативних варіантів.

Мета дипломної роботи полягає у дослідженні існуючих алгоритмів та методів пошуку в сучасних пошукових системах для виявлення основних факторів, що впливають на ранжування сайтів в результатах пошуку.

Для реалізації мета-пошукової системи була обрана платформа .NET з використанням .net framework 4.7, об'єктно-орієнтована мова програмування C# та декларативна мова розмітки XAML.

Ключові слова: МЕТА-ПОШУКОВА СИСТЕМА, РАНЖУВАННЯ, МЕТОДИ ПОШУКУ, АЛГОРИТМИ ПОШУКУ, ФУНКЦІОНАЛЬНІСТЬ, ПЛАТФОРМА .NET FRAMEWORK, DUCKDUCKGO, WIKIPEDIA, BING

ABSTRACT

The purpose of the thesis is to explore existing algorithms and search methods in modern search engines to identify the main factors affecting the ranking of sites in the search results.

The topicality of the thesis is to conduct scientific research in the field of information search on the Internet. The results are primarily aimed on creating new search engines or improving the algorithms of existing information search tools.

The object of the study is a meta-search engine created on .NET platform based on the analogues Wikipedia, Bing and DuckDuckGo.

Subject of the research - methods and algorithms of search optimization and ranking of the studied system, taking into account existing alternatives.

The purpose of the thesis is to explore existing algorithms and search methods in modern search engines to identify the main factors affecting the ranking of sites in the search results.

To implement the meta-search engine, a .NET platform was created using .net framework 4.7, an object-oriented C # programming language, and a declarative markup language XAML.

Keywords: META-SEARCH SYSTEM, RANGING, SEARCH METHODS, SEARCH ALGORITHMS, FUNCTIONALITY, .NET FRAMEWORK PLATFORM, DUCKDUCKGO, WIKIPEDIA, BING