Облікова картка дисертації

I. Загальні відомості

Державний обліковий номер: 0824U001800

Особливі позначки: відкрита

Дата реєстрації: 03-05-2024

Статус: Наказ про видачу диплома

Реквізити наказу МОН / наказу закладу: № НСВС/63/24 від 31.07.2024

II. Відомості про здобувача

Власне Прізвище Ім'я По-батькові:

1. Клещ Кирило Олегович
2. Kyrylo Kleshch

Кваліфікація:

Ідентифікатор ORCID ID: 0009-0006-8133-3086

Вид дисертації: доктор філософії

Аспірантура/Докторантура: так

Шифр наукової спеціальності: 122

Назва наукової спеціальності: Комп’ютерні науки

Галузь / галузі знань: інформаційні технології

Освітньо-наукова програма зі спеціальності: Комп'ютерні науки

Дата захисту: 10-07-2024

Спеціальність за освітою: Комп'ютерні науки

Місце роботи здобувача: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

III. Відомості про організацію, де відбувся захист

Шифр спеціалізованої вченої ради (разової спеціалізованої вченої ради): ДФ 26.002.169; ID 5604

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

IV. Відомості про підприємство, установу, організацію, в якій було виконано дисертацію

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

V. Відомості про дисертацію

Мова дисертації: Українська

Коди тематичних рубрик: 20.23, 20.23.19, 20.01.07

Тема дисертації:

1. Підвищення ефективності алгоритмів нечіткого пошуку з використанням таблиці подібності символів.

2. Improving the efficiency of fuzzy search algorithms with the usage of symbols similarity table.

Реферат:

1. Дисертаційне дослідження присвячене підвищенню швидкодії алгоритмів і розробці методу нечіткого пошуку з використанням таблиці подібності символів, для пошуку найбільш релевантних документів до пошукової фрази. В роботі створено метод нечіткого пошуку, який складається з 9 послідовних кроків та потрібен для швидкого пошуку співпадінь на великому наборі текстових даних. За допомогою цього методу була створена система нечіткого пошуку, яка дозволила вирішити задачу пошуку найрелевантніших документів до пошукової фрази з набору таких документів. Розроблений метод нечіткого пошуку комбінує переваги алгоритмів на основі детермінованих скінченних автоматів та алгоритмів на основі динамічного програмування для підрахунку відстані Дамерау-Левенштейна. Така комбінація дозволила впровадити таблицю подібності символів оптимальним чином. В рамках роботи запропоновано підхід та спосіб

створення таблиці подібності символів та розроблено приклад такої таблиці для символів з англійського алфавіту, що дозволило знаходити міру подібності поміж двома символами з константною асимптотикою та перетворювати поточний символ в його базовий аналог. Для фільтрування й сортування документів було розроблено метод оцінювання відповідності текстових даних до пошукової фрази на основі метрики, яка одночасно враховує кількість знайдених і незнайдених символів та кількість знайдених і незнайдених слів. Алгоритм Дамерау-Левенштейна дозволяє знаходити відстань редагування поміж двома словами, враховуючи помилки наступних типів: заміна, видалення, додавання та транспозиція символів. В рамках роботи була запропонована модифікація цього алгоритму за допомогою використання таблиці подібності для більш точної оцінки відстані редагування між двома словами. На основі розробленого методу була створена система нечіткого пошуку, яка дозволить знаходити шукані результати швидше та підвищить релевантність отриманих результатів шляхом їхнього сортування відповідно до значень розробленої метрики подібності тестових даних. Також у роботі було досліджено, проаналізовано та надано рекомендації, яким чином можна інтегрувати особливості та потужності таблиці подібності символів з алгоритмом нечіткого пошуку Дамерау-Левенштейна. Дослідження алгоритмів нечіткого пошуку в тексті є важливою темою в галузі обробки тексту та інформаційного пошуку. Це обумовлено зростаючим обсягом текстової інформації і ймовірністю помилок через вплив людського фактору при написанні тексту та створенні текстового контенту. Нечіткий пошук використовує алгоритми для пошуку даних в тексті, які приблизно відповідають шаблону. Це досягається шляхом зіставлення та порівняння рядків або ключових слів, які можуть бути схожими між собою, але не ідентичними. У першому розділі дисертаційної роботи були розглянуті та проаналізовані різні алгоритми нечіткого пошуку. Такі як: алгоритм Дамерау-Левенштейна, алгоритм N-грам, алгоритм Джаро-Вінклера, алгоритм Bitap, звичайний алгоритм Левенштейна, алгоритм SoundE та алгоритми на основі скінченних автоматів. У ролі оптимального алгоритму пошуку відстані редагування між двома словами було обрано алгоритм Дамерау-Левенштейна, бо він дозволяє впровадити таблицю подібності оптимальним чином. Також були розглянуті алгоритми на основі скінченних автоматів, а саме: автомат на основі префіксного дерева, автомат на основі таблиці та автомат на основі хешування. Перші два виявились неефективними через певні недоліки, а останній виявився оптимальним та найбільш універсальним з точки зору швидкодії роботи та часу побудови, а також об'єму витраченої пам'яті. У другому розділі дисертаційної роботи було розроблено та покроково описано метод нечіткого пошуку, який дозволяє знаходити найрелевантніші документи до пошукової фрази. Також були розглянуті переваги та недоліки застосування таблиці подібності символів, підходи та способи її побудови. Було створено приклад таблиці подібності для символів з англійської мови за допомогою групування символів у JSON файлі. Використання таблиці подібності покращує отримані результати, особливо при використанні мов зі спеціальними символами. Це дозволяє знаходити набагато більше релевантних результатів, проте швидкодія алгоритму може зменшитись. За допомогою використання такої таблиці підвищується релевантність відповідних документів навіть при наявності орфографічних помилок, скорочень, слів-синонімів або інших форм неточностей у запиті. Підхід із використанням таблиці подібності символів може бути використаний у системах перевірки орфографії та автоматичної корекції, системах автозавершення та автодоповнення, а також у реалізації функцій з виявлення дублікатів даних та плагіату. Ключові слова: нечіткий пошук, таблиця подібності символів, алгоритм Дамерау-Левенштейна, скінченний автомат, метрика подібності, обробка текстових даних, модель, об'єкт, аналіз, експертна система, нечітка логіка, пошук за зразком, пошук даних, пошук відповідностей, бенчмарки.

2. The dissertation research is dedicated to improving the speed of the algorithms and developing a fuzzy search method with the usage of symbols similarity table to find the most relevant documents for a search phrase. The work presents a fuzzy search method consisting of 9 sequential steps necessary for quickly finding matches in a large set of text data. Using this method, a fuzzy search system was created and it can solve the problem of finding the most relevant documents for a search phrase from a set of such documents. The developed fuzzy search method combines the advantages of the algorithms based on deterministic finite automata and algorithms based on dynamic programming for calculating the Damerau-Levenshtein distance. This combination allowed to

implement the symbols similarity table optimally. The work proposed an approach and method for creating a symbols similarity table, so the example of this table for symbols from the English alphabet was created, allowing for the measurement of similarity between two symbols with the constant asymptotics and transforming the current symbol into its base analog. For filtering and sorting documents, a method for evaluating the correspondence of text data to a search phrase based on a metric was developed, which simultaneously considers the number of found and unfound symbols and the number of found and unfound words. The Damerau-Levenshtein algorithm allows to find the editing distance between two words, considering the errors of the following types: substitution, deletion, addition, and transposition of symbols. In the course of the work, a modification of this algorithm was proposed using a similarity table for a more accurate estimation of the editing distance between two words. The developed method allowed to create a fuzzy search system that would help to find the desired results faster and increase the relevance of the obtained results by sorting them according to the values of the developed similarity metric of test data. Additionally, the work investigated, analyzed, and provided recommendations on how to integrate the features and capabilities of the symbols similarity table with the Damerau-Levenshtein fuzzy search algorithm. Research on fuzzy search algorithms in text is an important topic in the field of text processing and information retrieval. The reason is the increasing volume of textual information and the likelihood of errors due to the influence of human factors in writing text and creating textual content. Fuzzy search uses algorithms to search for data in text that approximately match the pattern. This is achieved by comparing and matching strings or keywords that may be similar but not identical. In the first chapter of the dissertation, various fuzzy search algorithms were considered and analyzed, such as the Damerau-Levenshtein algorithm, the N-gram algorithm, the Jaro-Winkler algorithm, the Bitap algorithm, the standard Levenshtein algorithm, the SoundEx algorithm, and algorithms based on finite automata. The Damerau-Levenshtein algorithm was chosen as the optimal algorithm for searching the editing distance between two words because it allows for an optimal implementation of the similarity table. Algorithms based on finite automata were also considered, namely: an automaton based on a prefix tree, an automaton based on a table, and an automaton based on hashing. The first two options were found to be inefficient due to certain drawbacks, while the latter proved to be optimal and the most versatile in terms of the operation speed, construction time, and memory consumption. In the second chapter of the dissertation, a fuzzy search method was developed and described step-by-step, allowing for finding the most relevant documents for a search phrase. The advantages and disadvantages of using a symbols similarity table, approaches, and methods of its construction were also discussed. An example of a symbols similarity table for symbols from the English language was created using symbols grouping in a JSON file. The usage of the symbols similarity table improves the obtained results, especially when using languages with special symbols. This allows for finding significantly more relevant results, although the speed of the algorithm may decrease. Using this table enhances the relevance of the corresponding documents even in the presence of spelling mistakes, abbreviations, synonyms, or other inaccuracies in the query. The approach using a symbols similarity table can be used in spelling check and automatic correction systems, auto-completion and auto-suggestion systems, as well as in implementing functions for detecting data duplicates and plagiarism. Keywords: fuzzy search, symbols similarity table, Damerau-Levenshtein algorithm, finite automaton, similarity metric, text data processing, model, object, analysis, expert system, fuzzy logic, pattern matching, data search, match search, benchmarks.

Державний реєстраційний номер ДіР:

Пріоритетний напрям розвитку науки і техніки: Інформаційні та комунікаційні технології

Стратегічний пріоритетний напрям інноваційної діяльності: Розвиток сучасних інформаційних, комунікаційних технологій, робототехніки

Підсумки дослідження: Теоретичне узагальнення і вирішення важливої наукової проблеми

Публікації:

- Kleshch, K., & Shablii, V. (2023). Comparison of fuzzy search algorithms based on Damerau-Levenshtein automata on large data. Technology Audit and Production Reserves, 4(2(72), 27–32. https://doi.org/10.15587/2706-5448.2023.286382
- Клещ, К. О., & Царьов, М. О. (2023). МОДИФІКАЦІЯ АЛГОРИТМІВ НЕЧІТКОГО ПОШУКУ ДЛЯ ВИКОРИСТАННЯ ТАБЛИЦІ ПОДІБНОСТІ СИМВОЛІВ. Таврійський науковий вісник. Серія: Технічні науки, (3), 21-28. https://doi.org/10.32782/tnv-tech.2023.3.3
- Kleshch, K. (2024). Development of fuzzy search method for creating an efficient information search system in text data. Technology Audit and Production Reserves, 1(2(75), 20–24. https://doi.org/10.15587/2706-5448.2024.298425

Наукова (науково-технічна) продукція: програмні продукти, програмно-технологічна документація

Соціально-економічна спрямованість: забезпечення промисловості чи населення новим видом інформаційно-комунікаційних послуг

Охоронні документи на ОПІВ:

Впровадження результатів дисертації: Впроваджено

Зв'язок з науковими темами: 0120U103046, 0121U110624, 0122U002655

VI. Відомості про наукового керівника/керівників (консультанта)

Власне Прізвище Ім'я По-батькові:

1. Петренко Анатолій Іванович

2. Anatolii Petrenko

Кваліфікація: д. т. н., професор, 05.13.05

Ідентифікатор ORCID ID: 0000-0001-6712-7792

Додаткова інформація:

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

VII. Відомості про офіційних опонентів та рецензентів

Офіційні опоненти

Власне Прізвище Ім'я По-батькові:

1. Завгородній Валерій Вікторович

2. Valerii Zavgorodnii

Кваліфікація: д. т. н., професор, 05.13.06

Ідентифікатор ORCID ID: 0000-0002-8347-7183

Додаткова інформація:

Повне найменування юридичної особи: Державний університет інфраструктури та технологій

Код за ЄДРПОУ: 41330257

Місцезнаходження: вул. Кирилівська, буд. 9, Київ, 04071, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Власне Прізвище Ім'я По-батькові:

1. Снитюк Віталій Євгенович

2. Vitaliy Snytyuk

Кваліфікація: д. т. н., професор, 05.13.06

Ідентифікатор ORCID ID: 0000-0002-9954-8767

Додаткова інформація:

Повне найменування юридичної особи: Київський національний університет імені Тараса Шевченка

Код за ЄДРПОУ: 02070944

Місцезнаходження: вул. Володимирська, буд. 60, Київ, 01033, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Рецензенти

Власне Прізвище Ім'я По-батькові:

1. Рогоза Валерій Станіславович

2. Walery Rogoza

Кваліфікація: д. т. н., професор, 01.05.02

Ідентифікатор ORCID ID: 0000-0003-2327-156X

Додаткова інформація:

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Власне Прізвище Ім'я По-батькові:

1. Шаповалова Світлана Ігорівна

2. Svitlana I. Shapovalova

Кваліфікація: к. т. н., доц., 05.13.12

Ідентифікатор ORCID ID: 0000-0002-3431-5639

Додаткова інформація:

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

VIII. Заключні відомості

Власне Прізвище Ім'я По-батькові голови ради	Глоба Лариса Сергіївна
Власне Прізвище Ім'я По-батькові головуючого на засіданні	Глоба Лариса Сергіївна
Відповідальний за підготовку облікових документів	Клещ Кирило Олегович
Реєстратор	УкрІНТЕІ
Керівник відділу УкрІНТЕІ, що є відповідальним за реєстрацію наукової діяльності	Юрченко Тетяна Анатоліївна