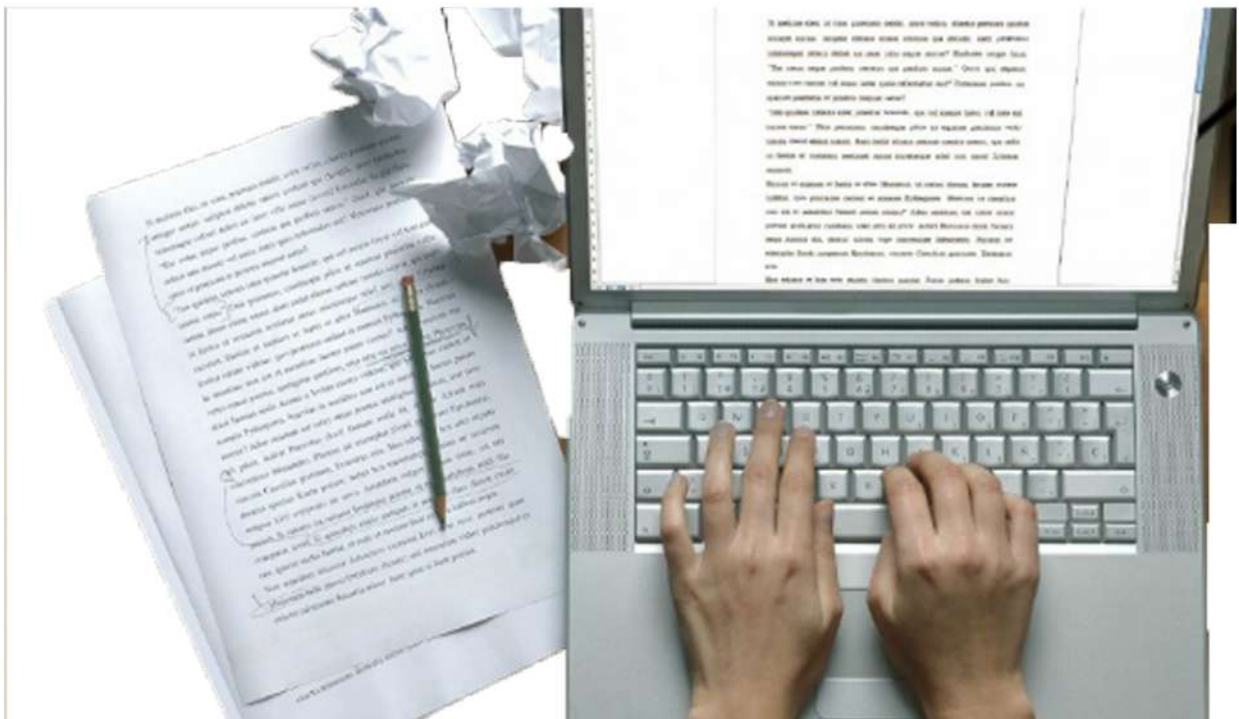


ЛІЗУНОВ П.П.
БЛОЩИЦЬКИЙ А.О.
КУЧАНСЬКИЙ О.Ю.
БЛОЩИЦЬКА С.В.
АНДРАШКО Ю.В.

**РОЗРОБКА КОМБІНОВАНИХ МЕТОДІВ
ІДЕНТИФІКАЦІЇ НЕПОВНИХ ДУБЛІКАТІВ ТА
ВИЯВЛЕННЯ ПОВНОТИ ВИСВІТЛЕННЯ НАУКОВИХ
РЕЗУЛЬТАТІВ ДИСЕРТАЦІЙНИХ ДОСЛІДЖЕНЬ,
ОПУБЛІКОВАНИХ АВТОРОМ**

Монографія



Київ 2020

П.П. Лізунов, А.О. Білощицький, О.Ю. Кучанський, С.В. Білощицька,
Ю.В. Андрашко

**РОЗРОБКА КОМБІНОВАНИХ МЕТОДІВ ІДЕНТИФІКАЦІЇ
НЕПОВНИХ ДУБЛІКАТІВ ТА ВИЯВЛЕННЯ ПОВНОТИ
ВИСВІТЛЕННЯ НАУКОВИХ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЙНИХ
ДОСЛІДЖЕНЬ, ОПУБЛІКОВАНИХ АВТОРОМ**

Монографія

Київ КНУБА 2020

УДК 004.912

ББК 73

Л 55

*Рекомендовано до друку Вченою радою
Київського національного університету будівництва і архітектури
(протокол № 40 від 24 грудня 2020 р.)*

РЕЦЕНЗЕНТИ:

В.Д. Гогунський

д-р техн. наук, проф., заслужений діяч науки і техніки України,
завідувач кафедри управління системами безпеки життєдіяльності
Одеського національного політехнічного університету

Ю.П. Рак

д-р техн. наук, проф., заслужений працівник освіти України,
завідувач кафедри управління проектами, інформаційних технологій та телекомунікацій
Львівського державного університету безпеки життєдіяльності

Лізунов П.П., Білощицький А.О., Кучанський О.Ю., Білощицька С.В., Андрашко Ю.В. Розробка комбінованих методів ідентифікації неповних дублікатів та виявлення повноти висвітлення наукових результатів дисертаційних досліджень, опублікованих автором: монографія. Київ: КНУБА, 2020. 100 с.

Монографія присвячена розробці моделей та методів визначення неповних дублікатів у текстових даних, графічних зображеннях, математичних формулах та таблицях, виявлення повноти висвітлення наукових результатів дисертаційних досліджень. метод виявлення контекстно-залежних значень та індексації текстових даних, а також рекомендації щодо його застосування в програмному комплексі визначення неповних дублікатів в електронних документах; розроблено модифікації методу N-грам аналізу, який дозволяє знаходити ступінь подібності електронних документів з різним типом контенту та їх фрагментів; розроблено методу та схеми адаптації алгоритмів виявлення неповних дублікатів до специфіки документів, зокрема наявності в контенті даних різного типу: графічні дані, математичні формули, таблиці, тощо.

Призначена для наукових працівників, викладачів, аспірантів, студентів вищих технічних навчальних закладів.

УДК 004.912

ББК 73

ЗМІСТ

| | |
|--|-----------|
| ВСТУП | 3 |
| РОЗДІЛ 1. РОЗРОБКА КОМБІНОВАНИХ МОДЕЛЕЙ АНАЛІЗУ ТЕКСТОВИХ ДАНИХ ДЛЯ ЗАДАЧІ ПОШУКУ НЕПОВНИХ ДУБЛІКАТІВ | 5 |
| 1.1. Алгоритми та прикладне програмне забезпечення для пошуку неповних дублікатів | 5 |
| 1.2. Метод виявлення контекстно-залежних значень та індексації текстових даних | 12 |
| 1.3. Застосування методу індексації даних в програмному комплексі визначення неповних дублікатів в електронних документах..... | 21 |
| 1.4. Метод N-грам аналізу для знаходження ступеня подібності електронних документів з різним типом контенту | 24 |
| РОЗДІЛ 2. ЗАСТОСУВАННЯ ЙМОВІРНІСНОГО ЛАТЕНТНОГО СЕМАНТИЧНОГО АНАЛІЗУ ДЛЯ ОЦІНЮВАННЯ ПОВНОТИ ВИСВІТЛЕННЯ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЙНИХ ДОСЛІДЖЕНЬ | 29 |
| 2.1 Ймовірнісна тематична модель представлення текстових документів з врахуванням n-грамів..... | 29 |
| 2.1 Застосування ймовірнісної тематичної моделі для встановлення повноти висвітлення матеріалів дисертації в публікаціях..... | 36 |
| РОЗДІЛ 3. АДАПТАЦІЯ АЛГОРИТМІВ ВИЯВЛЕННЯ НЕПОВНИХ ДУБЛІКАТІВ ДО СПЕЦИФІКИ ЕЛЕКТРОННИХ ДОКУМЕНТІВ . | 39 |
| 3.1. Адаптація методів виявлення неповних дублікатів до типу контенту електронних документів..... | 39 |
| 3.2 Застосування методів виявлення неповних дублікатів з врахуванням розроблених методів у програмному комплексі визначення неповних дублікатів | 75 |
| ВИСНОВКИ | 85 |
| СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ | 88 |