[1]Лізунов П.П., *професор*
Кафедра основ інформатики
Білощицький А.О., *професор*
Київський національний університет ім. Т. Шевченка
[1]Кучанський О.Ю., *доцент*
Кафедра інформаційних технологій
[1]Київський національний університет будівництва і архітектури
Чала Л.Е., *доцент*
Кафедра штучного інтелекту
Харківський національний університет радіоелектроніки

ГІБРИДНИЙ МЕТОД ВИЗНАЧЕННЯ НЕПОВНИХ ДУБЛІКАТІВ У ТАБЛИЦЯХ

В дослідженні розглядається задача знаходження неповних дублікатів в табличних даних, які представляють результати наукових експериментів і представляються в текстах дисертаційних робіт та наукових публікацій.
Ключові слова: *неповний дублікат, таблиця, подібність.*

Дослідження з побудови методів визначення неповних дублікатів є актуальними, оскільки можуть бути цінним інструментом для уникнення плагіату в сфері вищої освіти. Метод, який пропонується в роботі може бути використано для побудови модуля програмного комплексу для виявлення неповних дублікатів в дисертаційних та дипломних роботах, а також наукових публікаціях. Методи пошуку неповних дублікатів в текстових даних на основі локально-чутливого хешування розглянуті в роботах [1, 2]. Задача знаходження неповних дублікатів в числових даних пов’язана з ідентифікацією подібностей в часових рядах на основі зіставлення зі зразком з використанням методу найближчих сусідів з заданою метрикою [3].

Задача знаходження неповних дублікатів в таблицях представляє собою процес ідентифікації таблиць, які найбільш подібні одна до одної. Подібність в цьому випадку виражається деяким функціоналом F, що задає відстань між таблицями. Якщо ця відстань не перевищує деякого порогового значення, то таблиці вважаються подібними, а отже в даних цих таблиць наявні неповні дублікати. Завдання полягає у визначенні такої множини таблиць з даної бази, для яких відстань до вхідної таблиці мінімальна.

Алгоритм виявлення неповних дублікатів в таблицях відносно вхідної таблиці складається з таких кроків:

1. Виокремити з вхідної таблиці графічні об’єкти: зображення та формули. Ці об’єкти досліджуються окремо. Для порівняння формул може бути застосований метод, який базується на порівнянні зразків або шаблонів. Метод знаходження неповних дублікатів у математичних формулах наведено в [4].

2. У випадку наявності у вхідній таблиці комірок з даними типу дата, пропонується привести всі дати у відповідності з єдиним форматом і розглядати дані комірки як комірки з контентом текстового типу.

3. Весь контент комірок числового типу привести до єдиного типу: десятковий розділювач відображати у вигляді коми «,».

4. Видалити стовпець «№ п/п», що відображає нумерацію рядків.

5. Побудувати на основі контенту вхідної таблиці послідовність з текстових даних та послідовність з числових значень. При розподілі звертати увагу на комірки з комплексним контентом: у випадку, якщо в певній комірці вказано і числові, і текстові дані, то текст даної комірки та числові дані розподіляються по окремим елементам послідовностей.

6. Побудувати для текстової послідовності вхідної таблиці послідовність зі слів в канонізованій формі та за методом локально-чутливого хешування побудувати елементи індексу.

7. Далі розрахувати відстані Хеммінга від елементів кожного індексу послідовностей вхідної таблиці до елементів індексу послідовностей тих таблиць, які знаходяться в базі. Якщо ця відстань для певної таблиці з бази не перевищує деякого порогового значення, то це означає, що ця таблиця подібна до вхідної, а отже неповний дублікат ідентифіковано.

8. Побудувати для числової послідовності набір підпослідовностей, які утворені з даної числової послідовності методом плинного вікна.

9. Представити побудовані підпослідовності у вигляді векторів та розрахувати відстані за формулами Евкліда, Мінковського або міської метрики. Якщо для деякого порогового значення розрахована відстань мінімальна, то таблиця, яка подібна до вхідної таблиці є ідентифікованою.

В результаті дослідження побудовано гібридний метод для виявлення неповних дублікатів на основі методу локально-чутливого хешування та методу найближчого сусіда. Цей метод може бути використано в антиплагіат-системах, а також системах, які призначені для проведення інтелектуального аналізу на ідентифікацію подібностей інформації, яка презентована у табличному вигляді.

Література

1. Білощицький, А. О. Оптимізація системи пошуку збігів за допомогою використання алгоритмів локально чутливого хешування наборів текстових даних [Текст] / А.О. Білощицький, О.В. Діхтяренко // Управління розвитком складних систем. – 2014. – № 19.– С. 113 – 117.
2. Білощицький, А. О. Метод вилучення помилкових збігів текстів в електронних документах [Текст] / А. О. Білощицький, С. Д. Криштоф, С. В. Білощицька, О. В. Діхтяренко // Управління розвитком складних систем. – 2015. - № 22(1). – С. 144 - 150.
3. Кучанський, О. Прогнозування часових рядів методом селективного зіставлення зі зразком / О. Кучанський, А. Білощицький // Eastern-European Journal of Enterprise Technologies. – 2015. – Vol. 6, N 4(78). - P. 13–18.

Лізунов, П. П. Гібридний підхід до аналізу та розпізнавання математичних формул з метою виявлення в них подібностей [Текст] / П. П. Лізунов, А. О. Білощицький, Л. Е. Чала, С.В. Білощицька, О. Ю. Кучанський, С. Г. Удовенко // Управління розвитком складних систем. – 2016. – № 27. – С. 145 – 155