УДК 004.421

К.т.н., доцент Петрашенко А.В., магістрант Подолян О.Д.

Національний технічний університет України
«Київський політехнічний інститут»

ЗАСОБИ КОНТРОЛЮ ОРИГІНАЛЬНОСТІ ТЕКСТІВ У НЕСТРУКТУРОВАНОМУ ТЕКСТОВОМУ СХОВИЩІ

Abstract

Andrii V. Petrashenko, assoc. prof., PhD; Oleksii Podolian, student
Tools for controlling texts' originality in an unstructured data warehouse

This paper concerns the task of creation tools for controlling texts' originality, which can be applied in an unstructured data warehouse model. The shingling algorithm is described briefly and used as a basis for the research. New operation mode, based on uniform set of functions is proposed. The ways for further research are proposed as well.

Вступ

На даний момент людство знаходиться на етапі переходу до інформаційного суспільства. Цей процес пов'язаний із генеруванням великої кількості нової інформації, щорічні обсяги якої постійно збільшуються [1].

В цих умовах гостро постає питання подальшого розвитку засобів обробки інформації для забезпечення можливості ефективної роботи в інформаційному просторі. Під час дослідження проблеми було запропоновано узагальнений підхід до формалізованого подання алгоритмів обробки неструктурованої текстової інформації [2].

У багатьох областях застосування інформаційно-аналітичних систем існує потреба в засобах контролю оригінальності текстів. В даній статті розглянуто створення таких засобів на основі описаного в [2] моделі уніфікованого набору операцій.

Постановка задачі

Метою дослідження є розробка засобів контролю оригінальності текстів у неструктурованому текстовому сховищі на основі описаного уніфікованого набору операцій.

Метод «шинглів»

Для перевірки оригінальності тексту використовуються різні підходи, одним з яких є метод «шинглів», що набув широкого розповсюдження, зокрема в пошукових машинах. В даній роботі був використаний саме цей метод. Він полягає у розбитті тексту на частини, розрахунку їхніх контрольних сум та виборі певних сум, які характеризуватимуть даний документ. При порівнянні двох документів шукають однакові контрольні суми [3]. У цьому дослідженні текст розбивався на ланцюжки по десять слів, ланцюжки перекривались (два сусідніх містять дев'ять однакових слів), контрольні суми обирались кратні двадцяти п'яти.

Атрибут «Hash» документів сховища

Реалізація алгоритму «шинглів» передбачає створення набору контрольних сум, який характеризує даний документ. Набір утворюється на основі нормалізованого тексту документа, який відрізняється від початкового тексту відсутністю знаків пунктуації, стоп-слів та спеціальних символів. Можливе також приведення слів до початкової форми та використання інших засобів, які мінімізують можливість створення різних контрольних сум на основі подібних ланцюжків символів. У межах моделі неструктурованого текстового сховища набір контрольних сум доцільно зберігати як атрибут документа, який назвемо «Hash».

Реалізація режиму встановлення рівня оригінальності конкретного документа

Нехай є сховище $W_A = \{D_1\}$ та сховище W . Необхідно визначити, чи містить W документи які цілком, або частково подібні до D_1. Для цього створюємо нове сховище

$W_1 = W.where(A(\text{«Hash»}) = V_1)$, де

V_1 - перший елемент множини $V = W_1 A.values(A(\text{«"hash»)"})$.

Тоді якщо $W_1 \neq \emptyset$ - знайдено частковий збіг D_1 із якимось документом із W . В залежності від розміру D_1 та налаштувань алгоритму (зокрема, кількості слів в ланцюжку, на основі якого будується контрольна сума), на цій підставі можна стверджувати про копіювання частини тексту.

Більше інформації можна отримати, створивши аналогічним чином сховища W_2, W_3 і т.д. На рис. 1 запропоновано узагальнену схему функції перевірки оригінальності тексту, критерієм в якій виступає кількість будь-

яких збігів з іншими документами. Проте, якщо характер D_1 передбачає широке використання цитат і механізми їхнього видалення не передбачені

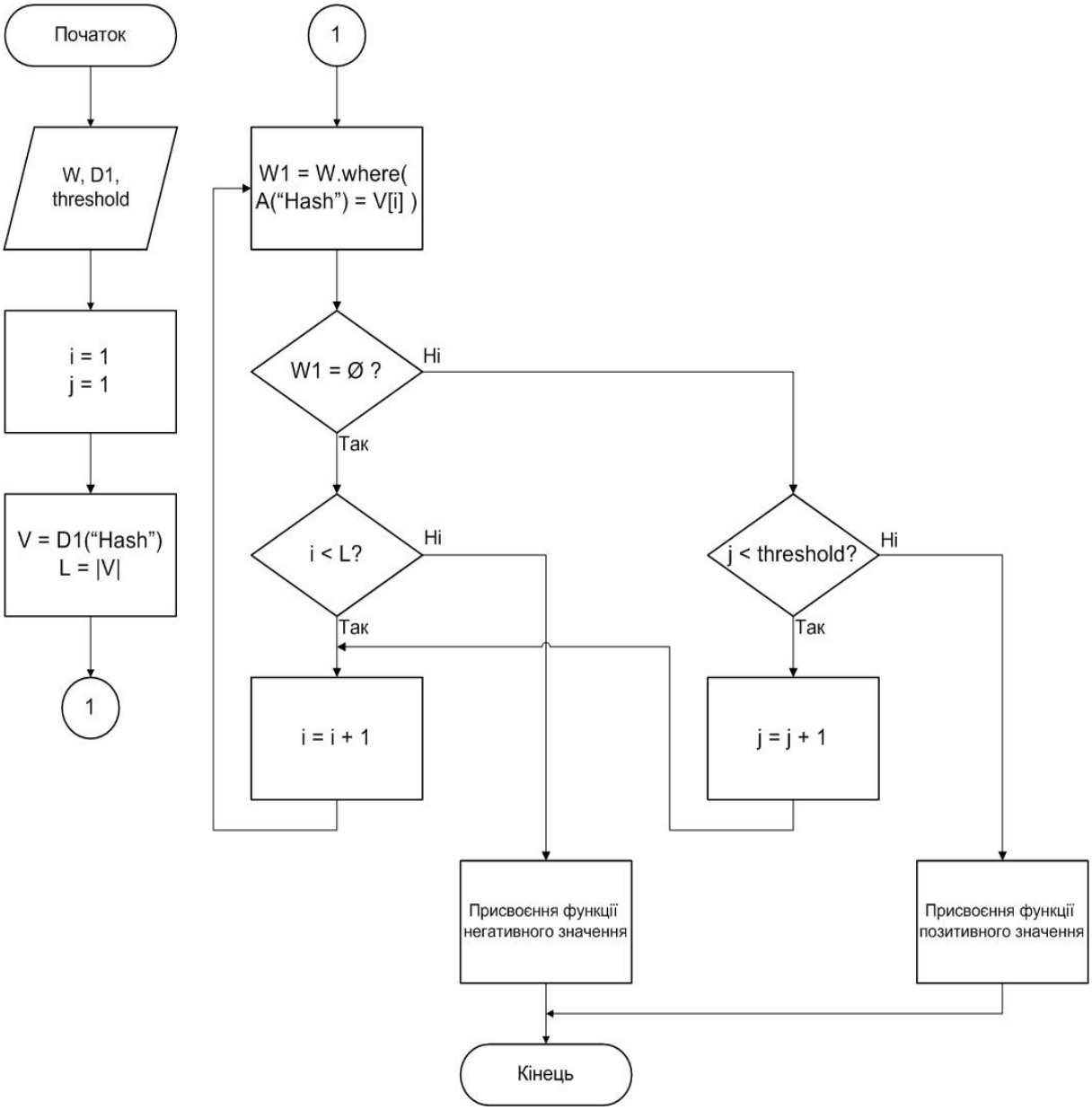

Рис. 1. Узагальнена схема функції перевірки оригінальності тексту

в процедурі нормалізації тексту, то такий підхід може виявитися неефективним. В цьому разі прийнятним критерієм виступатиме розмір сховища $W' = W_i \cap W_j \cap \cdots \cap W_n : \forall W \in \{W_i, \ldots, W_n\}\, W \neq \emptyset$, де $W_i, \ldots, W_n$ будуються аналогічно до W_1. Існування непустого W' означає збіг D_1 з іншим документом по двом, чи більше контрольним сумам і є вагомою підставою взяти під сумнів оригінальність D_1. Слід зазначити, що вибір

конкретного критерію тісно пов’язаний з предметною областю, в якій буде застосовуватись інформаційно-аналітична система і особливостей реалізації нормалізації текстів в сховищі.

Перспективи практичних застосувань

Описаний алгоритм дозволяє виконувати певні завдання (перевірка рівня оригінальності тексту, пошук цитат) у неструктурованому текстовому сховищі. Незначні модифікації забезпечать можливість пошуку подібних документів у межах одного сховища та створення нового, яке містить несхожі між собою документи. Така функція також має широкий ряд практичних застосувань, зокрема в пошукових машинах.

Висновки

У ході дослідження визначено засоби контролю оригінальності текстів у неструктурованому текстовому сховищі, які базуються на методі «шинглів». Такі засоби можуть використовуватись для контролю оригінальності документа (боротьба із плагіатом, захист авторських прав), пошуку схожих документів (оптимізація результатів пошукового запиту), пошуку цитат і таке інше. Розроблений алгоритм і описані можливі модифікації дозволяють виконувати ряд поставлених завдань. В подальшому доцільно дослідити вплив окремих параметрів алгоритму та аспектів реалізації деяких процедур (наприклад, нормалізації тексту) на якість результатів використання методу в конкретній предметній області.

Література

1. "As the Economy Contracts, the Digital Universe Expands", by John Gantz and David Reinsel, IDC, Multimedia White Paper.
2. Mykhailyuk A., Zamiatin D., Petrashenko A. Unstructured Data Warehouse Processing System Based on an Uniform Set of Functions // Proceedings of the 4-th International Conference ACSN-2009 “Advanced Computer System and Networks: Design and Application”. – Lviv. – 2009. – P. 117-119.
3. Andrei. Z. Broder, Steven. C. Glassman, and Mark. S. Manasse. Syntactic Clustering of the Web. In Proceedings of the Sixth World Wide Web Conference, 1997.