льно зручно. Дана спроба є першим кроком для створення повноцінної системи підтримки рішень. В майбутньому планується автоматизувати даний процес, щоб машина окрім результату та його вірогідності показувала причини даного прогнозу, що допоможе експерту при прийнятті рішення.

Література:

1. Mehrotra D. Basics of Artificial Intelligence & Machine Learning. *Notion Press*. 2020. 80 p.
2. Biau G. Analysis of a Random Forests Model. *Journal of Machine Learning Research*. 2012. Vol. 13. P. 1063-1095.
3. Bühlmann P. Bagging, Boosting and Ensemble Methods. *Handbook of Computational Statistics*. 2011. P. 985–1022. DOI: 10.1007/978-3-642-21551-3_33
4. Дьяконов А.Г. Методы ансамблирования обучающихся алгоритмов. 2015. 41 ст.
5. Campesato O. Python 3 for Machine Learning. *Stylus Publishing, LLC*. 2020. 364 p.

DOI https://doi.org/10.30525/978-9934-588-79-2-1.3

ПРОГРАМНИЙ СПОСІБ ПІДГОТОВКИ ТЕКСТОВИХ ДАНИХ ДЛЯ ЇХ АПАРАТНОЇ ОБРОБКИ З ВИКОРИСТАННЯМ ПЛІС

Голуб Т. В.
аспірант кафедри комп'ютерних систем та мереж
Національного університету «Запорізька політехніка»

Зеленьова І. Я.
кандидат технічних наук,
доцент кафедри комп'ютерних систем та мереж
Національного університету «Запорізька політехніка»

Грушко С. С.
кандидат технічних наук,
доцент кафедри комп'ютерних систем та мереж
Національного університету «Запорізька політехніка»
м. Запоріжжя, Україна

Вступ. Інформація у вигляді текстових даних, зокрема представлених в природомовному виді, використовується у більшості напрямків

діяльності людини – від збору та аналізу статистичних даних до функціонування кіберфізичних систем. У зв'язку з безперервним зростанням обсягу текстової інформації, яка потребує подальшого аналізу, виникає необхідність прискорення процесу її обробки. На даний час автоматизація обробки текстової інформації виконується здебільшого програмно [1]. Одним із шляхів прискорення обробки є перенесення частини функцій зазначеного процесу на апаратну платформу, в результаті чого створюється програмно-апаратний комплекс, спрямований на вирішення певної задачі [2]. В цьому випадку спостерігається певна надлишковість стандартних форматів даних [3], які використовуються в програмній реалізації, для їх подальшої обробки на апаратній платформі. Це зумовлено фіксованим розміром двійкового слова, яке передається в програмно-апаратному комплексі. При цьому фактичний розмір звичайного текстового слова в більшості випадків потребує значно меншого розміру коду, ніж зазначений фіксований. Отже, актуальною є задача розробки оптимізованого формату для узгодженого передавання даних між програмною та апаратною частинами комплексу з метою прискорення даного процесу.

Запропонований спосіб скорочення фіксованої довжини кодування слова. Для кодування природомовного тексту при виборі способу одним із визначних факторів є мова тексту. В даній роботі в якості базової мови було обрано українську. Для кодування зазначеної мови використано розширену таблицю ASCII як таку, що містить в собі символи кирилиці [3]. В даному типі кодування для відображення символів передбачається їх представлення у вигляді восьмирозрядного коду, який є найменшим за розміром із сучасних стандартних систем кодування. Використання мінімальної розрядності для кодування символів дозволяє зменшити обсяг вхідної інформації, представленої в двійковому вигляді, який підлягає подальшій обробці.

Мінімально достатня кількість розрядів для кодування символів будь-якої мови залежить від розміру її алфавіту та визначається за формулою 1.

$$R_\alpha =]\log_2 \alpha[\,, \qquad (1)$$

де α – число символів алфавіту; R_α – розрядність коду символу.

В рамках даної роботи проведено аналіз символів українського алфавіту і метою зменшення необхідної кількості розрядів для кодування літер української мови було обрано частину таблиці ASCII в діапазоні від «11100000» (відповідає рядковому символу «а») до «11111111» (відповідає рядковому символу «я»). Зважаючи на те, що

даний діапазон спрямовано на кодування літер також і російської мови, він включає символи, які не використовуються в українській мові («ъ», «ы», «э»), і при цьому не містить деяких літер української абетки («ґ», «ї», «і», «є»). Тому можна виконати відповідне співставлення даних літер з метою зменшення мінімального переліку кодових значень символів.

Таким чином, для подальшої оптимізації розміру коду символу української мови було виконано кодування літер двійковим кодом у відповідності з наступними співставленнями:

– символ «ґ» кодується двійковим кодом символу «г»;
– символ «ї» кодується двійковим кодом символу «ъ»;
– символ «і» кодується двійковим кодом символу «ы»;
– символ «є» кодується двійковим кодом символу «э».

В цьому випадку втрачається символ «ґ» шляхом заміни його на літеру «г», але цей крок не впливає на якість подальшої обробки тексту.

При використанні описаного перетворення загальна кількість символів українського алфавіту, які потребують кодування, дорівнює 32, а розрядність коду, відповідно, обчислюється як $R_{32} =]\log_2 32[= 5$.

Таким чином, для кодування всіх необхідних символів української мови при фіксованій довжині коду слова достатньо п’яти молодших розрядів. Старші три розряди в даному випадку є однаковими, тому їх можна відкинути. Це дозволяє зменшити кількість розрядів для кодування одного символу українського алфавіту з 8 до 5 бітів.

Адаптація довжини кодування слова. Представлення слова в формі, зручній для обробки за допомогою апаратних ресурсів ПЛІС, супроводжується потребою в розробці формату комірки пам’яті, необхідної для зберігання слова у вигляді двійкового коду.

З метою оптимізації формату представлення термів в двійковому вигляді та зменшення вимог до обсягів пам’яті є доцільним фіксувати лише необхідну кількість символів терму (попередньо підготовленого слова тексту). Для цього необхідно визначити ознаку положення символу в термі. В якості такої ознаки пропонується на початку коду кожного символу додати один інформаційний розряд, який ідентифікує, чи є даний символ першою літерою слова. Якщо символ стоїть на початку слова, то перший розряд його коду визначається як логічне значення «один». В іншому випадку він матиме значення «нуль».

В результаті розрядність поля опису одного символу з урахуванням інформаційного розряду R становить (формула 2):

$$R = R_{\alpha} + 1 . \quad (2)$$

Таким чином, запропонований програмний спосіб підготовки даних в форматі апаратної обробки тексту включає наступні етапи:

1. Кожна літера терму кодується п’ятирозрядним двійковим кодом.

2. Двійковий код кожної літери доповнюється одним старшим розрядом: якщо дана літера є першою в термі – із значенням «1», в іншому випадку – із значенням «0».

Таке представлення слова дає можливість формувати вхідний сигнал у вигляді безперервного потоку слів тексту, який аналізується, уникаючи представлення його в надлишковому вигляді. На рис. 1 наведено приклад представлення текстового слова (терма) у двійковому коді.

Рис. 1. Приклад представлення терма у двійковому вигляді

При виконанні подальшого аналізу на платформі ПЛІС FPGA розглядаються п’ять молодших бітів кожного символу. При цьому старший розряд використовується лише для індикації початку наступного слова. В результаті такого представлення вхідних даних формується масив закодованого вхідного тексту, оптимізований за розміром слів тексту без вмісту надлищкової інформації. Перевагою такого кодування є можливість прискорення обробки на ПЛІС з використанням послідовної передачі даних швидкісними інтерфейсами.

Дослідження. При дослідженні запропонованих способів кодування було обрано текст загальною кількістю 20 унікальних слів різної довжини.

Довжина більшості слів мов слов’янської групи, включаючи українську, менша за 15 літер, а довжина основ цих слів, тобто термів, які підлягають подальшому аналізу, становить 11 символів [1, с. 24]. Тому, для кодування таких слів за допомогою 8-розрядного ASCII – коду необхідно 11·8=88 біт.

Використання запропонованого способу кодування дозволяє перейти до опису символу замість 8-розрядного коду 5-розрядним (при

фіксованій довжині кодування слова), або 6-розрядним (адаптована довжина кодування слова).

Результати порівняння обсягів двійкової інформації при різних типах кодування термів при середній довжині 6-7 символів наведені в таблиці 1.

Таблиця 1

Необхідні обсяги двійкової інформації при різних типах кодування термів

	Кодування слова ASCII кодом	**Фіксована довжина кодування**	**Адаптована довжина кодування**
Середня кількість розрядів, біт	88	55	37,8
Загальна кількість розрядів для кодування 20 слів	1760	110	756
Відсоток необхідної кільк. розрядів, %	100%	63%	43%

Висновки. Використання запропонованого способу кодування з фіксованою довжиною дозволило скоротити кількість необхідних розрядів на 37%, а використання адаптованої довжини – на 57%, що дозволяє зменшити обсяги подальшої обробки тексту на апаратній платформі [2].

Запропоновані способи кодування термів досить просто реалізуються програмно та можуть бути використані для текстів, представлених на будь-якій європейській мові при використанні відповідної таблиці кодування символів.

Література:

1. Боярский К.К., Введение в компьютерную лингвистику: учебное пособие. СПб: НИУ ИТМО, 2013. 72 с.

2. Програмно-апаратний спосіб прискорення процесу класифікації текстових документів / Т.В. Голуб, І.Я. Зеленьова, С.С. Грушко, М.А. Павлішин, А.О. Котенко // міжнар. наук.-практ. конф.: «Technical sciences: history, the present time, the future, EU experience», 27-28 september, 2019: тези доп., – Wlolawek, Republic of Poland, 2019. – P. 90–93.

3. Таблица ASCII кодов символов. URL: https://istarik.ru/blog/programmirovanie/53.html (дата звернення: 18.09.2020).