Література

1. Острейковский В. А. Теория надежности / В.А. Острейковский. – М.: Высшая школа, 2003. – 463 с.
2. K.A. Anderson and B. H. Kirouac. A Simple and Free System for Automated Network Backups // In The Third Annual System Adminstration, Networking and Security Conference (SANS III), 1994, p. 63– 68.
3. Introduction to Algorithms. 2-nd edition / T. Cormen, C. Leiserson, R. Rivest, C. Stein. – London: The MIT Press, 2001. – 1180 p.
4. J. da Silva, O. Gudmundsson, and D. Mosse. Performance of a Parallel Network Backup Manager // In USENIX Conference Proceedings, 1992, p. 17 – 26.
5. Воеводин В.В. Параллельные вычисления / В.В. Воеводин, Вл. В. Воеводин. – СПб.: БХВ-Петербург, 2002. – 608 с.

Надійшла 11.4.2011 р.

УДК 004.912

Р.С. ЯРМОЛЮК
Хмельницький національний університет

ЗАДАЧА АНАЛІЗУ ТЕКСТОВИХ АТРИБУТІВ В ЕЛЕКТРОННОМУ КАТАЛОЗІ

Розглянута одна з актуальних проблем: задача аналізу текстових атрибутів запису в електронних каталогах. Проведено огляд та аналіз основних задач, що виникають при обробці та аналізі текстової інформації. Запропоновано ефективні алгоритми розв'язку кожної із задач.

Considered one of the critical problems: the problem of analysis of text attribute record in electronic catalogs. The review and analysis of the main problems arising in the processing and analysis of text information. The efficient algorithms for solving each problem.

Ключові слова: текстовий рядок, зіставлення рядків, відстань між рядками, нечітке порівняння рядків, текстові алгоритми, наївний алгоритм.

Постановка проблеми

Бібліотека у інформаційному суспільстві є невід'ємною складовою життя. Електронний каталог, як інформаційна система, що забезпечує доступ до бібліографічних баз даних бібліотеки повинен відповідати ряду вимог [1,2]. Одним із основних критеріїв якості електронного каталогу бібліотеки є наявність механізмів верифікації та пошуку помилок у бібліографічних записах. Якість електронного каталогу безпосередньо залежить від його інформаційного наповнення (наявності різного типу помилок у записах). Основні атрибути бібліографічного запису (назва, автор, видання тощо) мають текстовий тип даних. Тому основними задачами, що виникають при верифікації записів електронного каталогу, є задачі аналізу та обробки текстової інформації.

Аналіз останніх досліджень і публікацій

Теоретичні і практичні основи проблеми автоматичного пошуку та корекції помилок в записах електронного каталогу розробляли Вершинин М. И., Белоногов Г. Г., Бабко-Малая О. Б., Крауш А. С., Randall B. N., Ballard T, та інші.

Алгоритми аналізу текстових рядків представлені у роботах: Stephen G.A., Knuth D.E., Morris J.H., Pratt V.R., Karp R.M., Rabin, M.O., Boyer R.S., Moore J.S., Fischer M.J., Hirschberg D.S., Hunt J.W., Szymanski T.G., Landau G.M., Vishkin U. Хемминг Р.В., Левенштейн В.И.

Формулювання цілей статті та актуальність досліджень

На даний час існує багато методів та алгоритмів аналізу текстових рядків. Проблема полягає у практичній реалізації даних алгоритмів у системах пошуку та корекції помилок в текстових даних. Аналіз описових можливостей переважної більшості популярних автоматизованих бібліотечних інформаційних систем (АБІС) показав відсутність ефективних засобів для аналізу та обробки текстових даних в атрибутах бібліографічного запису. Тому проблема аналізу текстових даних в системах електронного каталогу бібліотек є актуальною.

Виклад основних матеріалів дослідження

До основних задач аналізу текстових рядків, що виникають при розробці засобів верифікації інформації в електронних каталогах, відносяться:

- задача зіставлення текстових рядків;
- задача розрахунку відстані між текстовими рядками;
- задача нечіткого порівняння текстових рядків;
- задача пошуку найдовшого повторюваного текстового підрядка.

Дамо означення основним поняттям. Під текстовим рядком будемо розуміти послідовність символів певного скінченного алфавіту. Текстовий рядок x довжини $|x| = m$ записується, як $x_1 x_2 \ldots x_m$, де x_i представляє собою i-й символ текстового рядка x. Текстовий підрядок $x_i x_{i+1} \ldots x_j$ рядка x, де $i \le j \le m$,

будемо позначати $x(i,j)$. У випадку, коли $i > j$, підрядок позначимо $x_R(i,j)$.

Зіставлення текстових рядків. Задача зіставлення текстових рядків формулюється наступним чином. Нехай задано зразок x, $|x| = m$ і текст y, $|y| = n$, де $m,n > 0$ і $m \le n$. Якщо x міститься в y, знайти позицію першого входження x в y, тобто визначити найменше i, при якому $y(i,i+m-1) = x(1,m)$ У загальному випадку потрібно знайти всі входження x в y.

Наївний підхід до цієї задачі: просуваючись по тексту символ за символом, порівнювати x з підрядками y. При цьому рядок вхідного тексту доводиться записувати у буфер, тому що у випадку невдачі порівняння, необхідні повернення. Всі підрядки тексту $y(i,i+m-1)$, $1 \le i \le n-m+1$ послідовного порівнюються із зразком x, доки не відбудеться перше входження зразка або не досягнуто кінця тексту [3]. Псевдокод, що реалізує наївний підхід, показано на рис. 1.

```
i = 1;
j = 1;
while (i <= n) and (j <= m)
   if x[j] = y[i]
      i = i + 1
      j = j + 1
   else
      i = i - j + 2
      j = 1
if j > m
   i = i - m
else
   i = 0
```

Рис. 1. Зіставлення текстових рядків наївним методом

Для «складних» випадків наївному алгоритму потрібен час $O(m \cdot n)$ [3], однак на практиці, при роботі з звичайними текстами, його середня ефективність $O(m+n)$. Окрім наївного були розроблені ефективніші алгоритми. До основних алгоритмів зіставлення текстових рядків відносяться:

- алгоритм Кнута-Моріса-Прата [4]:
- алгоритм Рабіна-Карпа [5];
- алгоритм Бояра-Мура [6] та його варіації.

Задача розрахунку відстаней між текстовими рядками. Для заданих текстових рядків x та y, де $|x|,|y| > 0$, і метрики d, що задає відстань між текстовими рядками, обрахувати $d(x,y)$. Поняття функції-виміру відстані, або метрики, використовується в різних областях науки. Такі функції застосовують для визначення сили зв'язку двох ознак, для оцінки подібності двох векторів, в розпізнаванні образів при співставленні шаблонів з різними частинами зображення. Метрика задається функцією d з наступними властивостями:

- невід'ємність $d(x,y) \ge 0 \ \ \forall x,y$;
- властивість нуля $d(x,y) = 0 \Leftrightarrow x = y$;
- симетричність $d(x,y) = d(y,x) \ \ \forall x,y$;
- нерівність трикутників $d(x,z) \le d(x,y) + d(y,z) \ \ \forall x,y,z$.

При обробці послідовності довжини n інколи її визначають вектором в R^n і застосовують звичайні відстані для векторів, але при обробці текстових рядків символи не завжди доцільно вважати числами, хоча коди символів є числа. Крім того, найчастіше потрібно порівнювати текстові рядки різної довжини Тому для порівняння текстових рядків зазвичай використовують метрики, що оцінюють максимальну ціну перетворення одного текстового рядка в інший. В загальному випадку, операціям редагування, а саме заміна символів, їх вставка і видалення, можна надати різну ціну [3].

Відстань Геммінга [7] між двома текстовими рядками однакової довжини визначається, як кількість позицій. в яких символи не співпадають. Це еквівалентно перетворенню першого рядка в другий коли дозволена лише операція заміни з одиничною вагою. Якщо допускається порівняння рядків різної довжини, то, як правило, потрібні також вставка і видалення. Якщо додати їм ту ж вагу, що і заміні, мінімальна загальна ціна перетворення буде дорівнювати одній з метрик Левенштейна [8].

До основних алгоритмів розрахунку відстані між текстовими рядками відносяться:

- алгоритм Вагнера-Фішера [9];
- алгоритм Хіршберга [10];
- алгоритм Ханта-Шиманського [11];
- алгоритм Уконена-Маєрса [12].

Задача нечіткого порівняння текстових рядків. Нехай дано зразок x, $|x| = m$ і текст y, $|y| = n$, де $m,n > 0$ і $m \le n$. Також задано ціле число $k \ge 0$ і функція відстані d. Потрібно знайти всі текстові підрядки s тексту y такі, що $d(x,s) \le k$.

Задача полягає в тому, щоб при заданій функції відстані знайти всі підрядки тексту, віддалені від зразка не більше, ніж на k. Якщо d є відстанню Геммінга, задачу називають зіставленням рядків з k незбігів,

якщо ж d відстань Левенштейна, задачу називають зіставленням рядків з k відмінностями (або помилками).

Наївний підхід до задачі зіставлення рядків, який потребує часу $O(m \cdot n)$, легко адаптувати до задачі k - незбігів. Але, як і для задачі зіставлення рядків, для задач k - незбігів та k - відмінностей були розроблені ефективніші підходи [3].

До основних алгоритмів нечіткого порівняння текстових рядків відносяться:

- алгоритм k - незбігів Ландау-Вишкин [13];
- алгоритм k - відмінностей Ландау-Вишкин [14].

Задача пошуку найдовшого повторюваного текстового підрядка. Задача пошуку найдовшого підрядка, що зустрічається в даному текстовому рядку більше ніж один раз, можна сформулювати наступним чином.

Для даного текстового рядка y, $|y| = n > 0$, знайти найдовший текстовий підрядок, що зустрічається в y більше одного разу. Найдовший текстовий рядок, що повторюється – це найдовший текстовий рядок із рядків максимальної довжини x, такий, що $y = uxvxw$ для довільних u, v, w, де $|u|, |v|, |w| \geq 0$.

Наївний підхід до вирішення цієї задачі базується на наступному [3]: будується матриця M розмірності $n \times n$, така, що $M_{i,j} = 1$ якщо $y_i = y_j$ інакше $M_{i,j} = 0$. За виключенням головної діагоналі, всі діагоналі в матриці, що складаються з серій одиниць, представляють максимальні повторювані текстові підрядки, найдовший з яких є найдовшим повторюваним текстовим підрядком. Оскільки, матриця симетрична відносно головної діагоналі, тоді достатньо обчислити лише її верхню або нижню половину. Наївний підхід до розв'язку цієї задачі потребує квадратичного часу $O(n^2)$. Але найдовший повторюваний текстовий підрядок можливо знайти і за лінійний час, якщо використовувати одну зі структур даних, запропонованих Вейнером [15], що містить компактний індекс для всіх можливих підрядків рядка.

Висновки

Для задач, що виникають при розробці систем верифікації текстової інформації, розроблено чимало ефективних алгоритмів [4-15]. Ці алгоритми можуть застосовуватись у системах верифікації інформації баз даних та засобах пошуку та корекції помилок передачі даних в телекомунікаційних системах та радіоелектроніці. Переважна більшість алгоритмів мають певні обмеження щодо функціональності та структури інформації. Ці недоліки зумовлені відносною давністю алгоритмів, більшість яких розроблені ще у другій половині минулого століття і не враховують можливостей сучасних технологій. Зокрема, застосування концепції паралельного програмування та високопродуктивних обчислень базі GPU архітектури виглядають перспективними.

Література

1. Daniels M.K. The Catalog: Its Nature and Prospects / M. K. Daniels // J. of libr. automation – 1976 – Vol. 9, № 1. – P. 48-66.

2. Шрайберг, Я.Л. Автоматизированные библиотечно-информационные системы России / Я.Л. Шрайберг, Ф.С. Воройский. – М.: Либерея, 1996. – 273с.

3. Stephen G.A. String searching algorithm / G.A. Stephen. – Singapore: World sci. publ., 1994. – 614p.

4. Knuth D.E Fast pattern matching in string / D.E. Knuth, J.H. Morris, V.R. Pratt // SIAM j. on computing. – 1977 – Vol. 6, № 2. – P. 323-350.

5. Karp R.M. Efficient randomized pattern-matching algorithms / R.M. Karp, M.O Rabin // IBM Journal of Research and Development – 1987 – Vol. 3, № 2. – P. 249-260.

6. Boyer R.S. A fast string searching algorithm / R.S. Boyer, J.S. Moore // Communications of the ACM – 1977 – Vol. 20, № 10. – P. 762-772.

7. Хемминг Р.В. Теория кодирования и теория информации: пер. с англ. / Р.В. Хемминг. – М.: Радио и связь, 1983. – 174с.

8. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов / В.И. Левенштайн // Докл. АН СССР. – 1965 – Т.163, № 4. – С. 845-848.

9. Wagner R.A. The string-to-string correction problem / R.A. Wagner, M.J. Fischer // Journal of the ACM – 1974 – Vol. 21, № 1. – P. 168-173.

10. Hirschberg D.S. A linear space algorithm for computing maximal common subsequences / D.S. Hirschberg // Communications of the ACM – 1975 – Vol. 18, № 6, – P. 341-343.

11. Hunt J.W. A fast algorithm for computing longest common subsequences / J.W. Hunt, T.G. Szymanski // Communications of the ACM – 1977 – Vol. 20, № 5, – P. 350-353.

12. Myers E.W. An O (ND) difference algorithm and its variations / E.W. Myers // Algorithmica – 1986 – Vol. 1, – P. 251-266.

13. Landau G.M. Efficient string matching with k mismatches/ G.M. Landau, U. Vishkin // Theoretical Computer Science – 1986 – Vol. 43, – P. 239-249.

14. Landau G.M. Fast parallel and serial approximate string matching/ G.M. Landau, U. Vishkin // Journal of Algorithms – 1989 – Vol. 10, – P. 157-169.

15. Weiner P. Linear pattern matching algorithm / P. Weiner // Proceedings of the 14th IEEE Symposium on Switching and Automata Theory – 1973 – P. 1-11.

Надійшла 3.4.2011 р.

UDC 004.72, 004.73

M. KARPINSKI, M. GIZYCKI, D. SZTAFINSKI
University of Bielsko-Biala, Poland
T. YAREMCHUK
Ternopil National Economic University, Ukraine

CENTRALISED MANAGEMENT OF WIRELESS NETWORK

This paper describes some characteristic features of the optimisation of a wireless network. The results of the research concerning a real wireless network were presented as well as a centralised management of an extended wireless network by means of WLAN controllers.

Описано деякі характерні особливості оптимізації безпровідних мереж. Представлені результати дослідження, що стосуються реальної безпровідної мережі, а також централізоване керування розширеної безпровідної мережі за допомогою контролерів WLAN.

Ключові слова: безпровідна мережа, централізоване керування.

Introduction

In order to receive information about the signal of a given network, it is necessary to take some measures. A professional measuring of a signal power can be done by means of some professional, however, expensive spectrum analysers. We distinguish two types of analysers: stationary analysers such as Antrisu MT8801B or Aeroflex 328X and portable analysers such as PDA made by AirMagnet, EthesSkope or BumbleBee. Measuring attainable for an average user is based on a suitable software. This sort of software can be installed on a laptop fitted with a wireless network card. Exemplary programs of this kind are: Kimset, Wavemoon and Network Stumbler in particular.

Research on a wireless network

The wireless network measures were taken on the basis of Network Stumbler. This program was chosen because it is free of charge, easy to use and it can provide the user with a lot of interesting information about the examined network, MAC address, the channel on which it operates, signal power, signal-to-noise-ratio, signal level, the maximal signal level measured in a particular area, minimal level, etc. (Fig. 1).

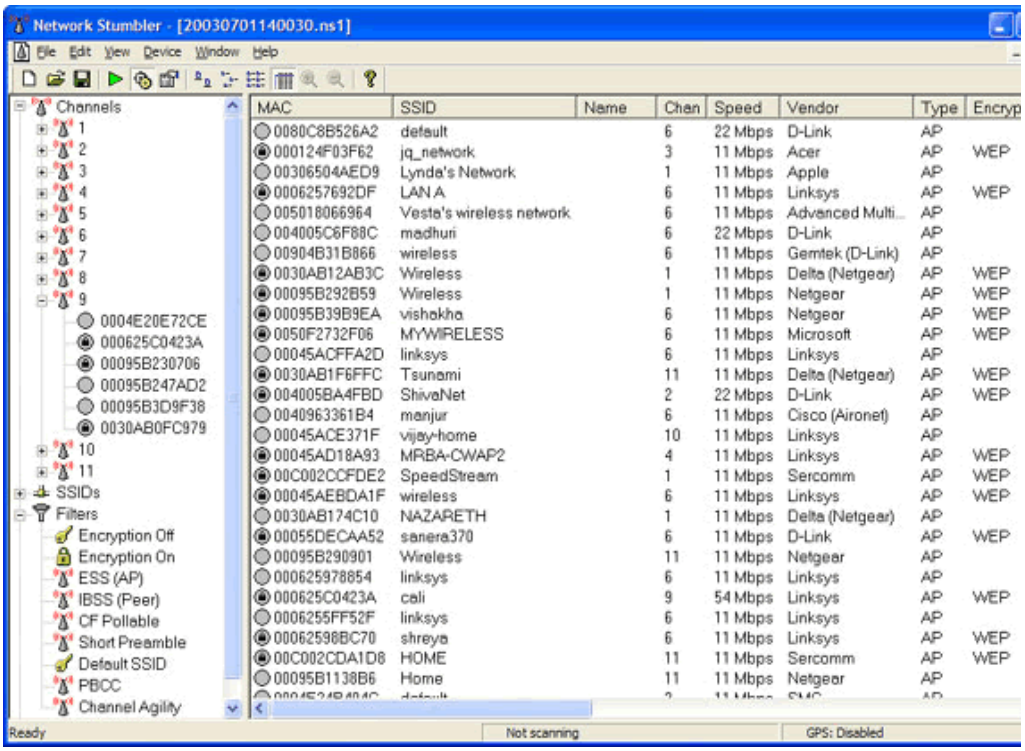

Fig. 1. The information about wireless networks found by Network Stumbler program

Network Stumbler had standard settings and it was installed on an ASUS F3F laptop fitted with a wireless network card I*ntel (R) PRO/Wireless 3945ABG Network Connection*. A maximal power of the wireless network card