Automatic extraction of man-made objects from aerial and space images (II), *Birkhäuser Verlag, Basel, 1998, pp. 213-222. 4. Canny J. F.,* Finding edges and lines in images, *Artificial Intelligence Laboratory Tech. Rep. 720, Massachusetts Institute of Technology, Cambridge, MA, 1983; 5. Torre V., Poggio T.* On edges detection *// IEEE Trans. Pattern Anal. Machine Intell. Vol.8, pp. 147–163, 1986 6. Бертеро М., Поджо Т. А., Торре А.* Некорректные задачи в предварительной обработке визуальной информации *// ТИИЭР, 1988. – Т.76, №8. С. 17–39. 7. Lowe D. G.,* Three-dimensional object recognition from single two-dimensional images *// Artificial Intelligence, Vol.31, pp. 366–395, 1987. 8. Rosin P. L. and West G. A. W.* Non-parametric Segmentation of Curves into Various Representations *// IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 17, №. 12, Dec. 1995, pp. 1140–1153.*

УДК 004.424.43

Л.С. Квурт, Н.Г. Даниленко
Національний університет “Львівська політехніка”
кафедра “Електронні обчислювальні машини”

ЗВОРОТНІ ФАЙЛИ ПОШУКОВИХ СИСТЕМ ІЗ ПРІОРИТЕТОМ СЛІВ

Описано метод пошуку інформації за зворотними файлами з призначенням та використанням пріоритетів слів, наводяться рекомендації щодо визначення тривалості часу обслуговування слів заданого пріоритету.

Method of information retrieval by backup files with assigning and usage of priority of words here is described and guidelines on definition of duration of a service time of words on given priority are resulted.

Вступ. Питання створення автоматизованих систем навчання, пошук методів інформаційного забезпечення таких систем, визначення шляхів підвищення продуктивності роботи з кожним днем стають все актуальнішими. Одним із методів прискорення пошуку інформації в автоматизованих системах є використання пріоритетності слів за деякими загальними ознаками.

Пошук інформації за пріоритетами слів. Серед найсуттєвіших вимог, які автоматизована система навчання (АСН) висуває до пошукових систем, є зменшення часу обслуговування замовлень. Одним із методів вирішення подібних вимог є організація пошуку з використанням пріоритетності відображення слів. При цьому система може розглядатись як система обслуговувань із відносними пріоритетами замовлень (переривань в обслуговуванні не відбувається). Робота учнів в автоматизованих системах навчання базується на запитах, які надходять під час роботи від них в АСН. Забезпечення ефективної роботи при цьому пов’язується із зменшенням реакції системи на запити.

Вважається [1], що затримка в АСН на введений запит, більша за три секунди, збиває темп роботи учня в системі, викликає втрату корисного часу учня, зменшує інтенсивність

розв'язання ним задач. Реакцію системи на запит можна подати через допустимий час обслуговування в системі ($T_{обсл.доп.}$):

$$T_{обсл.доп} >= T_{виб.} + T_{обсл.}, \text{ де} \quad (1)$$

$T_{виб.}$ – час, що визначається моментом появи запиту і моментом початку обслуговування запиту. $T_{обсл.}$ – тривалість безпосереднього виконання запиту.

У сучасних автоматизованих системах навчання функції терміналів виконують персональні комп'ютери, які мають потужні обчислювальні ресурси і які дозволяють учневі вникати в суть відповіді системи фактично з моменту початку обслуговування запиту (інформація послідовно видається на екран). Отже, основним часом, з яким порівнюється $T_{обсл.доп}$, є $T_{виб.}$, тобто час пошуку інформації.

Одним із методів організації пошуку інформації в автоматизованих системах навчання є використання зворотних файлів [2]. Зазначимо, що зворотний файл утворюється системою автоматично. Для зменшення часу доступу до найбільш важливої інформації можна застосувати ранжування слів за допомогою системи пріоритетів. Пріоритети відображення слів у файлі присвоюються автоматично. Наприклад, можна присвоювати пріоритети залежно від форми написання слів. При цьому словам, що записані великими літерами, присвоюється вищий пріоритет, а потім, за зменшенням пріоритету, йде група слів, що розпочинаються з великої літери, потім слова, що пишуться малими літерами тощо. Різне розміщення слів у файлі одного пріоритету утворює чергу. Тобто зворотний файл являє собою сукупність черг різного пріоритету. Обслуговування завжди розпочинається із черги слів вищого пріоритету, і слово нижчого пріоритету розглядаються після повного перегляду слів вищих пріоритетів. Введення пріоритетності слів дозволяє швидше знайти необхідну інформацію за запитом учня.

Враховуючи, що значна частина інформації спеціально готується для АСН, стає доцільним під час формування матеріалу виділяти слова за написанням із врахуванням їх важливості у контексті, що і обумовить швидкий доступ до них під час роботи в АСН.

Автоматизовану систему навчання, в якій одночасно працює N учнів і використовується база даних (БД) певного навчального предмету, можна вважати системою масового обслуговування. В такій системі для розрахунків характеристик можна використовувати відповідні залежності [3]. Допускаючи, що в АСН надходить М потоків замовлень з інтенсивностями λ_1 ..., $\lambda_м$, тривалість обслуговування яких ν_1, ..., $\nu_м$ та інші початкові моменти $\nu_1^{(2)}$, ..., $\nu_м^{(2)}$, середній час очікування замовлень к-го пріоритету (к=1,2,...,М) визначається:

$$\omega_к = \frac{\sum_{i=1}^{м} \lambda_i \nu_i^{(2)}}{2(1-R_{к-1})(1-R_к)} \quad (2)$$

При цьому із врахуванням того, що кількість замовлень на пошук в АСН відповідає кількості слів у зворотному файлі, у виразі (2) позначення набувають таких значень: М – кількість пріоритетних категорій слів, ν_i – тривалість обслуговування слів і-го пріоритету, $R_к$ – завантаження процесора потоком замовлень від вищого пріоритету до пріоритету рівня к.

$$R_{к-1} = \rho_1 + \rho_2 + \rho_3 + ... + \rho_{к-1}\ R_к = R_{к-1} + \rho_к\ , \text{ де} \quad (3)$$

ρ_i – завантаження процесора і-тим потоком замовлень.

Другі початкові моменти можна визначити за залежністю [3]:

$$\nu_i^{(2)} \approx 2\nu_i^2 \qquad (4)$$

Знайдена на сервері інформація надходить до ПК учня, від якого формувався запит, і використовується з темпом, характерним для цього учня. Робота в системі розпаралелюється.

Із врахуванням можливості існування черг слів пріоритетних категорій час чекання обслуговування t_v для слів V пріоритетної категорії, який за значенням характеризує $T_{виб.}$ із залежності (1), можна визначити [5]:

$$t_v = \sum_{K_1=0}^{M} P_{к1} K_1 t_з + \sum_{K_V=0}^{n-(M+M_1)} P_{кV} K_V t_з \qquad (5)$$

При цьому допускається: M – кількість слів, які мають вищий пріоритет; m_1 – кількість слів із нижчих пріоритетних категорій; n – загальна кількість слів зворотного файла; $P_{к1}$ $P_{кv}$ – ймовірності того, що при пошуку слова у відповідній групі пріоритетності (рівня 1 чи v) воно буде відповідати запиту після обробки K_1 чи відповідно K_V слів; $t_з$ – середній час процедур оцінки відповідності чергового слова запиту.

Для зменшення числа слів, які передують реалізації пошуку, і, як наслідок, збільшують $T_{виб}$, можна застосувати для ранжування набір ключових слів (групи слів). Наведені відношення (1) – (5) можуть використовуватись до групи слів. Ускладнення при цьому стосується виявлення і ототожнення синонімів, тривалість процедур обробки яких залежить від довжини ключа. Але такий аналіз може проводитись попередньо під час створення інформаційних файлів та реєстрації результатів аналізу у відповідних таблицях.

Однією із різновидностей даного пошуку є обробка словосполучень. Задача в загальному ускладнюється. Для кожного із слів словосполучення потрібно скласти свою таблицю пріоритетів. Такий підхід дозволяє розширити клас пріоритетів, доповнюючи виділені пріоритети пріоритетами усіх слів у словосполученні. При цьому утворюється нова таблиця, елементами якої є поля таблиць, що були створені для кожного слова. Як і при обробці одного слова, найвищий пріоритет будуть мати комбінації слів, написаних великими літерами. При цьому беруться до уваги усі слова заданого словосполучення за попередньо створеними таблицями.

Враховуючи процес створення таблиці словосполучень, час t_t, витрачений на утворення таких таблиць, можна знайти із залежності:

$$t_t \approx \frac{nt}{2}(n+1)\text{, де} \qquad (6)$$

n – кількість слів у заданому словосполученні, t – час, необхідний для утворення таблиці для одного слова.

Ця залежність вказує, що на утворення таблиці пріоритетів для словосполучень витрачається значно більше часу, ніж на обробку одного слова, через необхідність створення і аналізу декількох таблиць. Загалом же загальний пошук може значно прискоритись за рахунок перегляду матеріалу за контекстним словосполученням, тому що не витрачається час на перегляд тієї інформації, яка стосується кожного із слів цього словосполучення і яка не створює шуканого словосполучення.

Відповідно до викладеного підходу розроблено алгоритм і систему пошуку матеріалу за пріоритетом відображення слів у файлі. Пріоритети створені на основі орфографічних

особливостей написання слів. При цьому враховуються не тільки регістри букв у словах, а й розділові знаки, які належать до цих слів. Наприклад, слова, які написані великими літерами, мають вищий пріоритет ніж слова, написані малими літерами, але нижчий порівняно з такими ж словами, після яких стоїть знак оклику. При такому визначенні пріоритетів на останні позиції потрапляють слова, які написані маленькими літерами і знаходяться у дужках (це скорше за все пояснення до головного тексту).

За основу розробленого алгоритму взято такі положення. Вхідні дані аналізуються і трансформуються у таблицю. Кожен запис цієї таблиці вміщує слово, а також його координати у файлі, тобто рядок і позицію, починаючи з першої літери в документі. Пошук здійснюється перебором цілої таблиці від першої позиції до останньої. При цьому утворюється нова таблиця, у якій знаходяться тільки шукані слова. Слова утворюють чергу, обслуговування якої завжди розпочинається із слова, що має менші значення координат (тобто стоїть у цій черзі першим). Кінцевий пошук використовує тільки цю новостворену таблицю.

Такий підхід спрощує і прискорює пошук даних за рахунок того, що весь інформаційний файл переглядається тільки один раз, а надалі використовується тільки його частина – таблиця із шуканими даними.

За цим алгоритмом розроблена та відлагоджена програма, яка демонструє ефективність пошуку інформації при використанні пріоритетів написання слів.

Висновок. Прискорити пошук інформації в пошукових системах, в автоматизованих системах навчання можна за рахунок присвоєння вищого пріоритету виділеним у файлі ділянкам тексту. Ці ділянки визначаються словами, а для слів ділянки визначаються пріоритети. Пріоритети залежать від форми відображення слів у контексті, відповідно також від їх комбінацій у словосполученнях. Слова записуються у таблиці, які використовуються для пошуку інформації.

На основі розробленого алгоритму створено програму, що дозволяє заданий інформаційний файл перетворити у зворотний файл і на його основі здійснювати пошук із застосуванням принципу пріоритетності слів.

1. Савельєв А.Я., Новиков В.А., Лобанов Ю.И. Подготовка информации для автоматизированных обучающих систем. – М.: Высш. школа, 1986. – 176с. 2. Leonid Kvurt, Anatoliy Melnyk. Organization of automated tutor system based on education department local area network. Proceedings of the VII-th International Conference CADSM 2003. – Lviv, Publishing House of Lviv Polytechnic National University, 2003. –p.298 – 300. 3. Майоров С.А., Новиков Г.И., Алиев Т.И. и др. Основы теории вычислительных систем. – М.: Высш. школа 1978. – 408с. 4. Квурт Л.С. До питання реалізації дисципліни „Вибір за пріоритетом”// Вісник Львівського політехнічного інституту. – 1977. – №118. – С. 69 – 74.