УДК 81.322

И.Н. Ефремова, канд. техн. наук, доцент, ФГБОУ ВО «Юго-Западный государственный университет» (Курск, Россия) (e-mail:efremova-in@inbox.ru)

В.В. Ефремов, ст.преподаватель, ФГБОУ ВО «Юго-Западный государственный университет» (Курск, Россия) (e-mail:efremova-in@inbox.ru)

Н.А. Емельянова, студент, ФГБОУ ВО «Курский государственный медицинский университет» (Курск, Россия) (e-mail:efremova-in@inbox.ru)

СПОСОБ ПОСЛЕДОВАТЕЛЬНОГО ПОИСКА ВХОЖДЕНИЙ В ТЕКСТЕ С УЧЕТОМ ВОЗМОЖНЫХ КОЛЛИЗИЙ

Одной из фундаментальных задач современных компьютерных информационных систем является обработка символьной информации, объем которой превалирует в общем объеме всей информации. В настоящее время применительно к задачам обработки символьной информации эффективно используется продукционный подход. В работе рассматриваются вопросы специфики текстового поиска с применением продукционного подхода. Основная суть подхода заключается в поиске вхождений образца в текст и возможном осуществлении подстановки (модификации текста). Между тем, при реализации поиска вхождений могут возникать различного рода коллизии, которые необходимо учитывать для корректного решения поставленных задач. Алгоритмы последовательного сопоставления слов могут, например, сталкиваться с коллизиями, которые заключаются в возможности пропуска позиций вхождения образца в слово при некоторых их структурных особенностях. В работе описывается разработанный авторами способ поиска с учетом возможных коллизий, а также алгоритмические и автоматные модели способа. Разработанный способ заключается в разметке образца и задании последовательности его просмотра в виде схемы алгоритма. Разработаны также три алгоритма (варианта реализации) способа. Алгоритмы отличаются тем, к каким позициям образца и текста будет осуществляться переход в зависимости от результата сопоставления (равенство или неравенство текущих символов образца и текста). Разработана автоматная модель способа. Способ последовательного сопоставления с образцом с устранением коллизий повышает эффективность вычислительной системы при реализации поисковых процедур и обработки символьной информации. Предлагаемый способ может быть использован в системах обработки символьной информации.

Ключевые слова: *поиск вхождений, текстовый поиск, символьная информация.*

DOI: 10.21869/2223-1560-2017-21-4-68-74

Ссылка для цитирования: Ефремова И.Н., Ефремов В.В., Емельянова Н.А. Способ последовательного поиска вхождений в тексте с учетом возможных коллизий // Известия Юго-Западного государственного университета. 2017. Т. 21, № 4(73). С. 68-74.

***

Введение

Информация является стратегическим ресурсом общества. Средства ее обработки во многом определяют скорость и качество принимаемых решений в различных сферах человеческой деятельности. Одной из фундаментальных задач современных компьютерных информационных систем является обработка символьной информации, объем которой превалирует в общем объеме всей информации, циркулирующей в системах обработки данных. Следует заметить, что в настоящее время применительно к задачам обработки символьной информации эффективно используется продукционный подход. Основная суть подхода заключается в поиске вхождений образца в текст и возможном осуществлении подстановки (модификации текста). Между тем, при реализации поиска вхождений могут возникать различного рода коллизии, которые необходимо учитывать для корректного решения поставленных задач.

Постановка задачи

Пусть имеются слова S1, S2 в алфавите B={b_i}, i=1,n, n- мощность алфавита, и требуется найти позицию вхождения образца (искомого слова) S2 в слово S1. При этом условимся считать:

Слова считаются графически равны тогда, когда они состоят из одних и тех же символов и одинаков порядок их следования. Запись S1 ≙ S2 означает графическое равенство слов S1 и S2.

Под фрагментом слова условимся понимать слово S2, полученное из представления S1 ≙ Q1 S2 Q2 при наличии хотя бы одного непустого Q. При Q1=^, S2 является начальным фрагментом S1, Q2=^ -оконечным.

Слово S2 входит в слово S1 только тогда, когда верным является графическое равенство: S1 ≙ Q1 S2 Q2, где S1, S2, Q1, Q2- произвольные слова в алфавите B,

Начальной и конечной позициями вхождения условимся считать позиции первого и последнего символов, соответственно, слова S2 в слове S1.

Будем считать, что слово S1 больше слова S2 тогда, когда выполняется условие: [(S1≙S2 Q1) ∨ ((S1≙Q1 γ Q2)&(S2≙Q1 ξ Q3))]=1, где Q3- произвольное слово в алфавите B; ξ, γ- символы того же алфавита, согласно линейной метрики которого γ входит в алфавит позже ξ.

Под сопоставлением (сравнением) слов условимся понимать процесс и алгоритм, результатом которых является заключение о их отношениях (графическое равенства слов, или одно слово больше другого, либо одно слово входит в другое).

Под продукцией (формулой подстановки) будем понимать слова вида:

S2 → + T, где S2, T- графически не равные слова в алфавите B, «+», «→» не принадлежат B.

Работа продукции заключается в сопоставлении слова S1 и слова S2 для обнаружения вхождения и при положительном исходе замене первого слева фрагмента слова, графически равного S2 словом T (подстановкой). Специальный символ «+» принимает значение «*» и тогда подстановка осуществляется только один раз (заключительная формула) или пустого множества и тогда подстановка осуществляется столько раз, сколько обнаруживаются первые вхождения, при этом слово читается слева направо каждый раз сначала. Левая часть подстановки называется антецедентом, правая- консеквентом (модификатором).

Алгоритмы последовательного сопоставления слов могут сталкиваться с коллизиями, которые заключаются в возможности пропуска позиций вхождения образца S2 в слово S1 при некоторых их структурных особенностях.

Так если образец S2 имеет структуру:

S2≙P R3 P R4, где P, R3, R4- произвольные слова в алфавите B, то при сопоставлении с некоторыми словами, например, S1≙P R3 P R3 P R4, вхождение образца не будет обнаружено тогда, когда после неудачного сопоставления с первой позицией R3 образца S2 начинать следующую итерацию в соответствии с алгоритмом с текущей позиции слова и начальной буквой образца. В общем случае образец, приводящий к коллизиям, имеет вид:

S2≙P_1 …P_n R3 {P_i} R4…{P_n}…Rk, где P и R- слова в алфавите B.

Способ поиска вхождений в тексте с учетом возможных коллизий

Разработанный способ заключается в разметке образца и задании последовательности его просмотра в виде схемы алгоритма. Вариантами способа 3 являются алгоритмы: А1, который заключается в просмотре образца с начала и одновременном сопоставлении текущего символа образца и i+1 позиции рекурсии (для неначальных рекурсий) и переходе после каждого удачного сопоставления с текущей позицией образца на следующую, на i+2 позицию образца в случае равенства буквы слова i+1 позиции рекурсии (для неначальных рекурсий) или на начальную позицию образца в противном случае. Алгоритм А2 осуществляет следующим за i не начальной рекурсией символом на символ, следующий за i начальной рекурсией с повторением сопоставления с текущим символом текста. Отличается от А1 тем, что для i неначальной рекурсии сопоставление осуществляется только с текущей позицией образца и при отрицательном результате осуществляется переход на i+1 позицию образца и сопоставление производится с той же позицией слова. Алгоритм А3 отличается от А2 тем, что увеличение позиции сопоставляемой буквы слова осуществляется на каждом шаге сопоставления, но зато необходим второй просмотр i+1позиций слова.

Для примера зададим образец в виде слова “асае”, тогда граф-схема алгоритма (ГСА) А1- В1, размеченная как автомат Мура, приведена на рис.1, где Р – состояния автомата, Y1 – выходной сигнал нахождения вхождения.

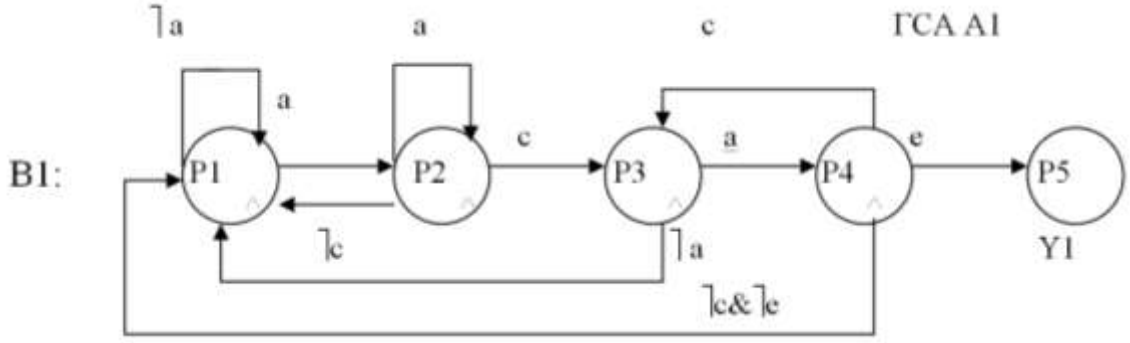

Рис. 1

Алгоритм преобразования образца в таблицу переходов автомата, реализующего способ сопоставления с устранением коллизий, приведен на рис.2, где обозначено:

N – длина образца S2; I, J- указатели позиции символа образца; X1: S2[1]… S2[I]=S2[J]…S2[J+I-1]; M1, M2- массивы значений позиций перехода по положительному и отрицательному результату сопоставления с соответствующим символом образца, соответственно; RD- массив признаков разрешения (1) или запрета (0) чтения очередного символа текста при отрицательном результате сопоставления; ВХ- массив значений результата поиска вхождения, выдаваемых при положительном сопоставлении с очередным символом.

Автомат Мура для алгоритма А1 задан в виде таблицы переходов. Граф В1 (рис.1) и таблица переходов представляет

собой автоматную модель исследуемых процессов для конкретного образца. Представим автоматную модель с помощью Марковских алгоритмов на основании того, что автоматы эквивалентно представимы машиной Тьюринга, а машины Тьюринга и НА эквивалентны.

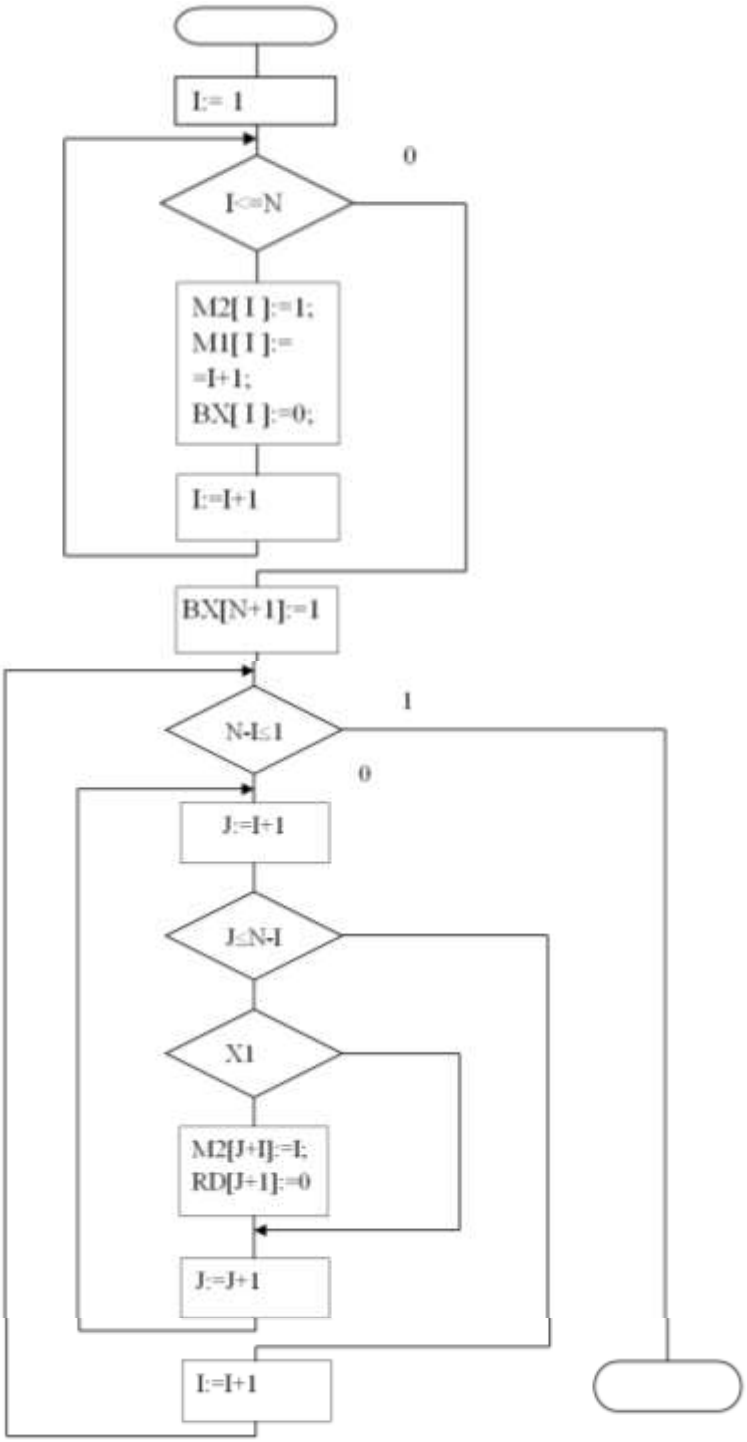

Рис. 2. Алгоритм преобразования исходного представления образца в автоматную модель

Переходы автомата

P_i	Условия X_i	Y_{i+1}	P_{i+1}
P_1	X1=⌉a	^	P_1
P_1	X2=a	^	P_2
P_2	X3=c	^	P_3
P_2	X2=a	^	P_2
P_2	X4=⌉c&⌉a	^	P_1
P_3	X2=a	^	P_4
P_3	X1=⌉a	^	P_1
P_4	X5=e	Y1	P_5
P_4	X6=⌉e&⌉c	^	P_2
P_4	X3=c	^	P_3

Алгоритм преобразования табличного представления в продукционную автоматную модель имеет следующий вид.

{“^” →* ” \”} ⇒ {“\”PX→ “;” PX “→”} ⇒ {“;” →* “^”}, где “- служебный для алгоритма символ, заключающий обрабатываемые символы.

Результаты работы алгоритма преобразования для таблицы, представляющие продукционную автоматную модель способа А1 для конкретного образца имеют следующий вид, при этом исходным словом является $P_1 X_1$.

$$\begin{cases} P_1X_1 \rightarrow \wedge P_1; \\ P_1X_2 \rightarrow \wedge P_2; \\ P_2X_3 \rightarrow \wedge P_3; \\ P_2X_2 \rightarrow \wedge P_2; \\ P_2X_4 \rightarrow \wedge P_1; \\ P_3X_2 \rightarrow \wedge P_4; \\ P_3X_1 \rightarrow \wedge P_1; \\ P_4X_5 \rightarrow * Y1\ P_5; \\ P_4X_6 \rightarrow \wedge P_2; \\ P_4X_3 \rightarrow \wedge P_3. \end{cases}$$

Заключение

Таким образом, описан способ последовательного сопоставления с образцом с устранением коллизий, повышающий эффективность вычислительной системы при реализации поисковых процедур и обработки символьной информации, разработана автоматная модель способа.

Предлагаемый способ может быть использован в системах обработки символьной информации, описанных, например, в [1-10].

Список литературы

1. Ефремова И.Н., Ефремов В.В. Способ сопоставления символьной информации с множеством образцов // Известия Юго-Западного государственного университета. 2012. №3 (42). Ч.1. С.50-53.

2. Ефремова И.Н., Ефремов В.В. Способ аннулирования коллизий при сопоставлении слов // Известия Юго-Западного государственного университета. 2013. №1 (46). С.20-22.

3. Ефремова И.Н., Ефремов В.В. Способы и устройства обработки символьной информации. Курск, 2014. 182 с.

4. Информационные системы обработки и сжатия текста / В.В. Ефремов, И.Н. Ефремова, В.В. Серебровский, А.А. Черепанов // Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика. 2014. Т. 29. № 1-1 (172). С. 182-184.

5. Ефремова И.Н., Ефремов В.В. К вопросу повышения эффективности автоматической обработки текстов. Современное общество, образование и наука:

сборник научных трудов по материалам Международной научно-практической конференции: в 9 ч. М., 2014. С. 22-23.

6. Серебровский В.В., Ефремова И.Н., Ефремов В.В. К вопросу представления семантики естественно-языковых текстов // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. 2014. № 2. С. 37-41.

7. Серебровский В.В., Ефремова И.Н., Ефремов В.В. К вопросу учета смысловой составляющей текста в информационно-поисковых системах // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. 2015. № 2 (15). С. 8-12.

8. Ефремова И.Н., Ефремов В.В. Способ неточного поиска в тексте, содержащем ошибки антропогенного характера // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. 2015. № 2 (15). С. 54-61.

9. К вопросу учета смысловой составляющей текста в информационно-поисковых системах в медицине / И.Н. Ефремова, В.В. Ефремов, Н.А. Емельянова // Научные механизмы решения проблем инновационного развития: сборник статей Международной научно-практической конференции. М., 2016. С. 229-230.

10. Разработка концепции информационной системы построения информационно - образовательного мультимедийного интерактивного пространства / В.И. Шнырков, В.В. Ефремов, И.Н. Ефремова, Н.Н. Бочанова //Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. 2012. № 2-3. С. 16-20.

Поступила в редакцию 12.07.17

UDC 81.322

I.N. Efremova, Candidate of Engineering Sciences, Associate Professor**,** Southwest State University (Kursk, Russia) (e-mail: efremova-in@inbox.ru**)**

V.V. Efremov, Senior Lecturer, Southwest State University (Kursk, Russia) (e-mail:efremova-in@inbox.ru)

N.A. Emelianova, Student, Kursk State Medical University (Kursk, Russia) (e-mail:efremova-in@inbox.ru)

A METHOD OF SEQUENTIAL SEARCHING OF OCCURANCES IN TEXT WITH THE ACCOUNT OF POSSIBLE COLLISIONS

One of the fundamental tasks of modern computer information systems is processing of symbol information, the amount of which prevails in the total amount of information. At present, rules-based approach is effectively applied to the tasks of processing symbol information. The paper deals with the peculiarities of text search applying rules-based approach. The main essence of the approach is to find pattern occurrences in the text and possible implementation of substitution (text modification). Meanwhile, when implementing search for occurrences, various kinds of collisions may arise. They should be taken into account to solve the set tasks correctly. For example, algorythms of sequential word matching can run into collisions which involve the possibility of skipping positions of pattern occurrence in a word with some structural peculiarities.

The paper presents a method of searching taking into account possible collisions developed by the authors, as well as algorithmic and automatic models of the method. The developed method involves patterm markup and setting a sequence of its viewing in the form of algorithm diagram. Three algorythms (implementation variants) of the method have been developed. Algorithms differ in the possibility to carry out transition to this ок that position of the pattern and the text depending on the result of matching (equality or inequality of the current symbols of the patten and text). An automation model of the method has been developed. The proposed method of sequential matching

with the pattern with collisions elimination increases the effectiveness of the computer system when implementing search procedures and symbol information processing. The method can be used in the systems of symbol information processing.

Key words: *search of occurrences, text search.*

DOI: 10.21869/2223-1560-2017-21-4-68-74

For citation: Efremova I.N., Efremov V.V., Emelianova N.A. A Method of Sequential Searching of Occurances In Text with the Account of Possible Collisions. Proceedings of the Southwest State University, 2017, vol. 21, no. 4(73), pp. 68-74 (in Russ.).

***

Reference

1. Efremova I.N., Efremov V.V. Sposob sopostavlenija simvol'noj informacii s mnozhestvom obrazcov. Izvestija Jugo-Zapadnogo gosudarstvennogo universiteta, 2012, no. 3 (42), ch.1, pp.50-53.

2. Efremova I.N., Efremov V.V. Sposob annulirovanija kollizij pri sopostavlenii slov. Izvestija Jugo-Zapadnogo gosudarstvennogo universiteta, 2013, no. 1 (46), pp. 20-22.

3. Efremova I.N., Efremov V.V. Sposoby i ustrojstva obrabotki simvol'noj informacii. Kursk, 2014. 182 p.

4. Informacionnye sistemy obrabotki i szhatija teksta / V.V. Efremov, I.N. Efremova, V.V. Serebrovskij, A.A. Cherepanov. Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta. Serija: Jekonomika. Informatika, 2014, vol. 29, no. 1-1 (172), pp. 182-184.

5. Efremova I.N., Efremov V.V. k voprosu povyshenija jeffektivnosti avtomaticheskoj obrabotki tekstov. Sovremennoe obshhestvo, obrazovanie i nauka: sbornik nauchnyh trudov po materialam Mezhdunarodnoj nauchno-prakticheskoj konferencii: v 9 ch. M., 2014, pp. 22-23.

6. Serebrovskij V.V., Efremova I.N., Efremov V.V. K voprosu predstavlenija semantiki estestvenno-jazykovyh tekstov. Izvestija Jugo-Zapadnogo gosudarstvennogo universiteta. Serija: Upravlenie, vychislitel'naja tehnika, informatika. Medicinskoe priborostroenie, 2014, no. 2, pp. 37-41.

7. Serebrovskij V.V., Efremova I.N., Efremov V.V. K voprosu ucheta smyslovoj sostavljajushhej teksta v informacionno-poiskovyh sistemah. Izvestija Jugo-Zapadnogo gosudarstvennogo universiteta. Serija: Upravlenie, vy-chislitel'naja tehnika, informatika. Medicinskoe priborostroeni, 2015, no. 2 (15), pp. 8-12.

8. Efremova I.N., Efremov V.V. Sposob netochnogo poiska v tekste, soderzhashhem oshibki antropogennogo haraktera. Izvestija Jugo-Zapadnogo gosudarstvennogo universiteta. Serija: Upravlenie, vychislitel'naja tehnika, informatika. Medicinskoe priborostroenie, 2015, no. 2 (15), pp. 54-61.

9. Efremova I.N., Efremov V.V., Emel'janova N.A. K voprosu ucheta smyslovoj sostavljajushhej teksta v informacionno - poiskovyh sistemah v medicine. Nauchnye mehanizmy reshenija problem innovacionnogo razvitija: sbornik statej Mezhdunarodnoj nauchno-prakticheskoj konferencii. M., 2016, pp. 229-230.

10. Shnyrkov V.I., Efremov V.V., Efremova I.N., Bochanova N.N. Razrabotka koncepcii informacionnoj sistemy postroenija informacionno - obrazovatel'nogo mul'timedijnogo interaktivnogo prostranstva. Izvestija Jugo-Zapadnogo gosudarstvennogo universiteta. Serija: Upravlenie, vychislitel'naja tehnika, informatika. Medicinskoe priborostroenie, 2012, no. 2-3, pp. 16-20.