БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ ACADÉMIE BULGARE DES SCIENCES
ИЗВЕСТИЯ НА МАТЕМАТИЧЕСКИЯ ИНСТИТУТ
BULLETIN DE L'INSTITUT DE MATHÉMATIQUES
Том (Tome) XIV

СЪЗДАВАНЕ НА ЧЕСТОТЕН РЕЧНИК НА ДУМИТЕ В ПИСМЕНИЯ ЕЗИК

Румяна Киркова

Анализът на честотните характеристики на писмените текстове представлява интерес както за лингвистиката, така и за някои приложни области, като например набиране на печатарски текстове, определяне оптималното разположение на символите в различни клавиатури, създаване на оптимални стенографски знаци и др.

За получаването на по-точни резултати се налага да се обработва по-голямо количество текстов материал. Когато това обработване се извършва ръчно, необходими са много време и голям брой технически сътрудници, като има и риск за натрупване на грешки. Значително по-големи предимства има машинното обработване на текстов материал, тъй като изчислителните машини са едно от техническите средства, което облекчава труда на човека с универсалността, бързодействието и надеждността си. Чрез подходящи алгоритми машините лесно биха могли да се научат да „четат".

В България са правени опити както за ръчно обработване [1], така и за машинно обработване [2] на текстова информация. Докато в [1] се публикуват резултати от изследване на честотата на отделни букви в писмения български език, в [2] са получени вече честоти на отделните букви, на всички възможни двойки от букви или буква и интервал, на избрани тройки и четворки от букви или букви и интервали, на дължините на думите, при което материалът е пет пъти по-голям по обем от този в [1]. Освен това при машинното изследване е било обърнато внимание и на положението на съответното съчетание в думата, т. е. дали се намира на първо, второ и т. н. място.

При изследване честотните характеристики в даден текст се налага многократното му проследяване. Целта на настоящата работа е разработване и реализиране на алгоритми, с чиято помощ да може значително да се съкрати изследваният текст, но по такъв начин, че да не се загуби съдържащата се в него информация. Освен това се иска резултатите да се получават за възможно най-кратко време. За целта се предлага да се създаде честотен речник на думите от текста, в който всяка дума да се среща по един път и да е снабдена с параметър B, означаващ броя на повторенията ѝ в текста. Думите се разглеждат като поредици от символи и не се взима под внимание граматичната форма, т. е. дали думата

е в единствено или множествено число, дали е в първо или второ лице, дали е членувана или не и т. н. По този начин думите „стол", „стола" „столове" са различни и биха фигурирали в речника.

Думите в речника биха могли да се наредят по азбучен ред или във възходящ или низходящ ред по параметъра B.

За всички по-нататъшни изследвания се използва така създаденият речник. Така например от него могат да се получат характеристики за всички съчетания от символи с дължина $n = 1, 2, \ldots$ както общо за текста, така и по позиции в думата. Въз основа на тези характеристики би могло да се пресметне и ентропията на езика.

За създаването на такъв речник беше разработен алгоритъм, ориентиран към машини със сравнително малък обем оперативна памет. По изготвения алгоритъм беше разработена и цялостна система за обработване на текстова информация, наречена TEXT. Системата беше експериментирана в Математическия институт с Изчислителен център при Българската академия на науките върху изчислителната машина МИНСК-22.

Към основните процедури на системата TEXT спадат:

Процедура 1 — осигуряване верността на текста, подготвен на входен носител така, както е бил напечатан. Това е подготвителна процедура, при която машината се използва за отстраняване на евентуалните грешки, допуснати при пренасяне на текста върху входния носител (в нашия случай това е перфолента). Освен това се извършва записване на верните порции информация във външната памет (в случая магнитна лента).

Процедура 2 — линеализиране на текста — също е подготвителна процедура за привеждане на текста във вид на един единствен ред, т. е. премахват се тиретата за пренасяне, заради които при по-нататъшните обработки биха се наложили твърде много анализи, а следователно би се изразходвало и твърде много машинно време.

Процедура 3 — стандартизиране на всички думи — също е подготвителна процедура за отделяне и стандартно записване на думите в определени полета в подходящо избран вътрешен код. Кодът е съобразен с възможностите на машината МИНСК-22 за паралелна работа с всички разреди на машинната дума, което дава възможност за ускоряване работата на по-нататъшните процедури.

Процедура 4 — създаване масив M1 от всички думи, започващи с определена начална буква. Тази процедура е първата стъпка за създаване на честотния речник. В резултат от изпълнението ѝ се получава масив M1 от всички думи (еднакви и различни), които започват с определена начална буква. При тази процедура поради ограничения обем на оперативната памет се наложи да се разработят подходящи алгоритми и програми (SORT/MERGE) за работа с големи масиви при използване на магнитна лента като външна памет.

Процедура 5 — Изброяване на различните думи от масива M1 и подреждане на новополучения масив M2 от различни думи, придружени от параметъра B, по азбучен ред. Тази процедура е втората стъпка за създаване на честотния речник. В резултат от изпълнението ѝ се получава нареден по азбучен ред масив M2 от различните думи на масива M1, като всяка дума е придружена от параметъра брой на повторения B. И тук се наложи създаването на серия програми (SORT/MERGE) за ра-

бота с големи масиви. При тази процедура от особено значение беше избраният вътрешен код за представяне на символите, който позволи бързото нареждане по азбучен ред. С оглед ускоряване на работата програмите бяха ориентирани така, че резултатите да се получават за един пас през масива M1.

Ако се приложат последователно процедури 4 и 5 за всички букви от азбуката на дадения език, ще се получи пълният речник на думите, срещащи се в обработваемия текст, като всяка дума е придружена от параметъра B (брой на повторенията).

Процедура 6 — сервиз. Тази процедура се състои от серия програми, които дават възможност за извеждане на получените резултати в определен формат, както и за допълнителни обработки върху получения честотен речник.

Чрез вътрешна управляваща система се осъществява последователното включване на отделните процедури.

Освен тези основни процедури към системата TEXT са предвидени блок за генериране и прекъсване на системата, блок за защита на информацията, блок за протоколиране работата на системата и действията на оператора.

Структурата на системата е такава, че към нея лесно могат да се добавят нови блокове в зависимост от изследването, което предстои да се извърши, както и да се внасят корекции в съществуващите вече блокове.

Освен това системата не е ориентирана специално към обработване на български текстове и поради това с успех може да се приложи и в други страни. За целта трябва каталогът на буквите от процедура 6 да се смени с каталог, отговарящ на съответната азбука. Такова едно предимство прави TEXT универсална система.

Чрез TEXT беше извършена обработка на български текст — романа „Под игото“ от Ив. Вазов, състоящ се от 696 081 символа = 121 270 думи = 19 546 различни думи. Резултатите показаха, че алгоритъмът и създадената въз основа на него система функционират правилно, надеждно и бързо. Последното предимство е особено важно при машини със сравнително малък обем на оперативната памет, както и с немного голяма скорост на изпълнение на операциите.

За извеждане на резултатите бяха подготвени редица програми към процедура 6. Между тях има както такива, които отпечатват само онези думи, чийто параметър B е над някаква отнапред зададена стойност, така и такива, които отпечатват думите в определен ред в зависимост от параметъра B.

Някои от така отпечатаните резултати се привеждат в табл. 1 и 2. В табл. 1 се дават част от думите с брой на повторенията B, по-голям или равен на 100, които са наредени по азбучен ред. В табл. 2 се дават част от думите, наредени в низходящ ред по параметъра B. И в двете таблици информацията е зададена по колони, както следва:

1. колона — самата дума;
2. колона — параметър B за думата;
3. колона — честота на думата спрямо всички думи в текста (в проценти);

Таблица 1

1	2	3	4	5	6
А	0546	0,45023	0,18538	22,75476	41,04950
АЗ	0504	0,41560	0,09420	21,46594	40,58415
АКО	0162	0,13358	0,35170	2,82833	10,36633
БАЙ	0157	0,12946	0,34085	0,27272	1,84077
БЕ	0208	0,17151	0,45157	0,01100	1,73809
БЕЗ	0124	0,10225	0,26920	0,79502	1,61309
БЕХА	0100	0,08246	0,21710	0,44759	0,72023
БЕШЕ	0887	0,73142	0,92570	10,84018	30,99851
БОЖЕ	0122	0,10060	0,26486	0,76606	1,53869
БОЙЧО	0314	0,25892	0,68170	1,54545	10,68154
БОРИМЕЧКАТА	0156	0,12863	0,33868	0,25825	1,80357
БЪДЕ	0170	0,14018	0,36907	0,46091	1,32440
БЪЛГАРИЯ	0152	0,12534	0,32999	0,20034	1,65476
БЯЛА	0139	0,11462	0,30177	0,01215	1,17113
БЯХА	0159	0,13111	0,34519	0,30167	1,91517
В	0832	0,68607	0,80630	11,92088	42,96998
ВЕЧЕ	0154	0,12698	0,33433	0,13199	2,06415
ВИ	0102	0,08410	0,22144	0,07443	1,00353
ВИДЯ	0146	0,12039	0,31697	0,96929	2,59329
ВРЕМЕ	0136	0,11214	0,29526	0,76591	2,00470
ВСИЧКИ	0123	0,10142	0,26703	0,50152	1,23955
ВСИЧКО	0106	0,08740	0,23012	0,15578	1,23896
ГИ	0295	0,24325	0,64045	1,46646	20,18045
ГО	0933	0,76935	0,02557	20,61427	70,15037
ГРАДА	0102	0,08410	0,22144	0,58162	1,66917
ДА	2961	2,44165	5,42843	40,42417	80,30919
ДЕТО	0155	0,12781	0,33651	0,16843	1,20395
ДО	0306	0,25232	0,66433	1,28091	2,29943
Е	1148	0,94664	0,49234	41,59173	51,37131

Таблица 2

1	2	3	4	5	6
И	5118	4,22033	9,11135	50,58311	71,86718
НА	3528	2,90921	6,65940	30,63297	50,40000
СЕ	3472	2,86303	6,53783	20,88881	41,02332
ДА	2961	2,44165	5,42843	40,42417	80,30919
ТОЙ	1620	1,33586	3,51707	12,28029	21,01801
НЕ	1593	1,31359	3,45845	10,83172	20,74714
ОТ	1561	1,28721	3,38898	20,24597	51,17850
СИ	1484	1,22371	3,22181	10,21054	12,67126
С	1441	1,18825	3,12846	2,91468	12,10127
ЗА	1205	0,99365	0,61609	20,98009	71,24522
Е	1148	0,94664	0,49234	41,59173	51,37131
МУ	1103	0,90954	0,39465	20,14241	42,15329
ЩЕ	1078	0,88892	0,34037	62,41403	82,60931
ЧЕ	1043	0,86006	0,26438	30,91259	71,37388
ГО	0933	0,76935	0,02557	20,61427	70,15037
БЕШЕ	0887	0,73142	0,92570	10,84018	30,99851
КАТО	0884	0,72895	0,91919	10,50442	20,80113
В	0832	0,68607	0,80630	11,92088	42,96998
НО	0658	0,54259	0,42854	1,71329	2,40000
ТОВА	0643	0,53022	0,39597	1,25569	10,72381
ПО	0622	0,51290	0,35038	1,89841	41,40853
ТИ	0552	0,45518	0,19841	1,22884	2,20613
А	0546	0,45023	0,18538	22,75476	41,04950
Я	0538	0,44363	0,16801	62,44783	92,81412
ОГНЯНОВ	0514	0,42384	0,11591	1,32506	12,16896
АЗ	0504	0,41560	0,09420	21,46594	40,58415
КАЗА	0459	0,37849	0,99650	1,97345	10,83905
ТЕ	0410	0,33808	0,89012	1,62649	1,83789
ЛИ	0383	0,31582	0,83150	20,64197	92,73890

4. колона — честота на думата спрямо всички думи в текста, започващи със същата начална буква (в проценти);

5. колона — честота на думата спрямо всички думи в текста с $B=100$ (в проценти);

6. колона — честота на думата спрямо всички думи в текста с $B=100$ и започващи със същата начална буква (в проценти).

От таблиците се вижда, че характерът на текста оказва влияние върху резултатите.

ЛИТЕРАТУРА

1. Penkov, B., A. Obretenov, B. Sendov, T. Kirpikova, T. Džukanov. Frequences of letters in written Bulgarian. — C. R. Acad. Bulg. Sci. **15**, 1962, No. 3.
2. Бърнев, П., Д. Добрев, Р. Киркова. Изследвания на някои честоти в писмения български език. — Известия на Мат. инст. на БАН, **13**, 1972, 159.

Постъпила на 16. XI. 1971

СОЗДАНИЕ ЧАСТОТНОГО СЛОВАРЯ СЛОВ В ПИСЬМЕННОМ ЯЗЫКЕ

Румяна Киркова

(*Резюме*)

Рассматривается алгоритм для создания целостной системы машинной обработки текстовой информации, ориентированный на применение на машинах со сравнительно небольшим объемом оперативной памяти, при котором основная цель — ускорение работы машины при значительном сокращении обрабатываемого текста без потери содержания посредством составления и использования частотного словаря.

Приводятся некоторые результаты, полученные в Математическом институте с Вычислительным центром Болгарской академии наук при проведении экспериментов по применению с помощью алгоритма системы на вычислительной машине Минск-22.

ON THE CREATION OF A FREQUENCE DICTIONARY OF OCCURENCE OF THE WORDS IN A WRITTEN LANGUAGE

Rumjana Kirkova

(*Summary*)

An algorithm enabling the creation of a whole system of machine processing of text information is considered. The algorithm is oriented to machines with a comparatively small main storage and it is aimed at fastening the work of the machine by means of shortening the text processed but not on the account of the text contents. The shortening of the text is realized through compiling and usage of a frequence dictionary.

Some results obtained at the Mathematical Institute with Computing Centre of the Bulgarian Academy of Sciences when experimenting on a computer MINSK 22 the system realized on the basis of this algorithm are given in the paper.