UDC 004.934

Inna Humeniuk,
Olexander Markovskyi,
Olga Shevchenko

METHOD TO IMPROVE THE EFFICIENCY OF ELECTRONIC DICTIONARIES WITH CONTENT SEARCH

Інна Гуменюк,
Олександр Марковський,
Ольга Шевченко

СПОСІБ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ЕЛЕКТРОННИХ СЛОВНИКІВ З КОНТЕКСТНИМ ПОШУКОМ

In the article, the method of accelerating the work of electronic dictionaries of computer translation systems is offered. This method is based on using perfect hash-addressing and cryptographic transformations as a hash function.

Keywords: perfect hash addressing, electronic dictionary, cryptographic transformation.

Tabl.: 1. Fig.: 0. Bibl.: 5.

У статті запропоновано спосіб прискорення роботи електронних словників систем комп'ютерного перекладу, за допомогою організації пошуку на основі perfect хеш-адресації та використання криптографічних перетворень в якості хеш-функції.

Ключові слова: perfect хеш-адресація, електронний словник, криптографічне перетворення.

Табл.: 1. Рис.: 0. Бібл.: 5.

Target setting. The last two decades marked a rapid redistribution of centers of production and scientific activity. The countries of the East increasingly take leading positions in many areas of scientific, especially technological, researches.

These processes undoubtedly increase the size of scientific and technical information exchange between East and West. However, this process is inhibited by the language barrier, which is due to the huge linguistic difference between the languages of East and West. Traditional method of language acquisition by a wide range of industry specialists does not give the desired effect [1].

The most promising way to overcome the language barrier, in the context of the exchange of scientific-technical information, is the using better-advanced computer and computerized translation systems. Modern advanced computer translation technologies based on analyzing a large number of translation options, which requires

multiple access to electronic dictionaries. That dictates new requirements for the speed and efficiency of search in electronic dictionaries [2].

Thus, the scientific task of increasing the speed of search in electronic dictionaries is relevant and important for the current stage of the development of information technology.

Analysis of available solutions. To date, there are a number of approaches of organizations for searching in electronic dictionaries, oriented for using in computer translation systems.

There is part of the e-dictionaries based on the principles of databases [1]. The advantage of this approach is the ability to use existing technologies and software packages for working with databases. The main disadvantage is the low speed of search in terms of computer translation systems. E-dictionaries based on tree structures are the most widespread. They are represented by many variations and modifications of search trees [3, 4]. In the e-dictionaries of this class, the compromise between the search speed and the memory resources is resolved at an acceptable level. The number of requests to memory logarithmically depends on the number of words in the dictionary. But this speed is not enough in terms of modern computer translation technologies.

Potentially, the highest search speed is achieved with hash addressing. Until recently, its widespread use was constrained by the existing of collisions and a high level of redundancy memory utilization [5].

In modern conditions, the cost of hardware memory is reduced. As a result, the significance of the second disadvantage is reduced too. The most serious problem is collisions because resolving them requires complex mechanisms which need significant memory resources. It also does not allow to find wanted data by one access to memory.

Thus, existing electronic dictionaries do not provide the necessary speed of searching.

The research objective. The purpose of the research is to increase the speed of the contextual search in electronic dictionaries, for using them as a part of advanced computer translation systems.

The statement of basic materials. To achieve this goal the analysis of the features of the context search in electronic dictionaries has been performed. The Context Dictionary model consists of a search array of keywords and context data that included translation options and word subsets of context language constructs associated with a particular translation option [2].

Potentially, the fastest key search technology is hash addressing. It provides the independence of the search time from the size of the search array. Thus, only hash addressing is able to resolve the compromise between speed of search and size of e-dictionaries on an acceptable level for practical using.

The feature of the hash search in e-dictionaries is that words with the same root

or close to each other addressed in different areas of memory. This problem can be solved by the allocating roots or bases, using computer morpheme analysis [5] or stemming algorithms. The roots or bases are proposed to be used as keywords, and the word and its modifications are stored as context data.

To solve another problem - existing of collisions, it is proposed to use perfect hash addressing and, as a perfect hash function, symmetric cryptographic encryption algorithms, such as DES, AES. The cipher block key is used as a hash configuration code. Thus, the search keyword is introduced to the cipher block input, and the certain ciphertext or part of it is used as a hash address.

That all determine the feasibility of using a two-stage search to increase the efficiency of electronic dictionaries. In the first stage, the using of hash addressing for finding context data by keyword is proposed. The context search for the most relevant translation option takes place on the second stage. Technological realization of this process is carried out with the intelligent technologies of structural-linguistic and lexical-semantic computer analysis.

General structure. Implementation of the proposed approach provides that in the memory cell addressed by a hash transformation of a keyword, the address link to the certain contextual information in the hash memory is stored. Thus, filling in the hash memory is proposed to be carried out in two stages. At first, memory is reserving for primary address links. In the second stage, context data in size $w_1, w_2, \ldots, w_m$, of m keywords is recorded between m primary address links $a_1, a_2, \ldots, a_m$, in a way that for each word the primary address link and context data can be read to the cache by a minimum number of swap cycles. To store data it is necessary to divide memory for two parts: hash and overflow memory. The average value of each keyword's context information is w.

To implement this approach, it is proposed to use four formats of data organization in the memory cell, depending on the stored information:

• format A - free memory cells marked with a marker symbol M_1;

• format B - all bytes of this kind of memory cells are filled with context data;

• format C - for storing primary address links. The memory cells consist of three fields: the first - token M_2; the second - address link to the beginning of the context data of a certain word; the third - address of the last memory cell of the relevant context data.

• format D - for memory cells that contain address link to the rest context data of a certain keyword in the overflow memory. Memory cells of this format are marked by token M_3.

The size of the memory cells is determined by the format B, as $n/4 + 1$ byte. Tokens must be characters that are not used in the dictionary.

The context information of the dictionary words records in the ordering of their perfect hash-addresses and consists of the following action sequence:

1. The first byte of all hash cells is indicated according to the format A.

2. For all keywords, perfect hash addresses are calculated. The hash memory cells, which was addressed, mark according to the format C.

3. Let set j = 1, b = 2n + 1.

4. For the j keyword, a A-format memory cell searched starting with the address $a_j - v/2$.

5. Recording information is carried out in all bytes of a cell changing it to the format B. The address value increments: $a = a + 1$.

6. If the entire size of the j word context data has written to the hash, the address $a - 1$ is fixed in the third field of the memory cell a_j. Go to the step11.

7. If a addressed B-format memory cell and $a \leq a_j$, the address a is incremented by one: $a = a + 1$. Go to step 5

8. If $a = a_j + v/2$ or the memory cell at $a + 1$ is accorded to format B and at the same time $a > a_j$, the a memory cell is marked according to the format D. The current value of b is written to a memory cell. Go to step 10.

9. Go to step 5.

10. The rest of the j word context information is successively written to the overflow memory starting from the address b. The address of the last completed memory cell is written to the third field of the a_j memory cell, this value, incremented by one, is fixed in b.

11. The j increments: $j = j + 1$. If j is less than m, go to step 4.

12. Context information for all words has stored.

The first stage of searching is carried out in the following order:

1. The perfect hash address of keyword s is calculating: $a = h(s)$.

2. Block of memory cells, which size is v, is loaded to the cash from the hash memory, starting with the address $a - v/2$. The address of the first byte of the context information is q.

3. In the cache, the address of the begin s word context data from is hash memory is read at the address $q + (v/2)\cdot(n/4+1) + 1$ to the variable W. The address B, that is adress of the context data begin in the cache, is calculated as $B = q + (W - a + v/2) \cdot (n/4 + 1)$.

4. In the cache memory, the end context data address of the s word in the hash memory is read to the U variable. That is, the value at address $q + (v/2)\cdot(n/4 +1)$ is recorded to U. If $U < a + v/2$, the address E is calculated as: $E = q + (U - a + v/2) \cdot (n/4+1)$, else: $E = 0$.

5. For composing, i is setting as q, j— B.

6. In the cache, the byte addressed by j is forwarded into the byte, addressed by i. After that, both addresses are incremented: $i = i + 1$, $j = j + 1$.

7. If the byte addressed in the cache j contains the M_2 token, then $j = j + n/4 + 1$.

8. If $j - 1 = E$, go to step 12

9. If the byte addressed j contains an M_3 token, the rest of the s keyword

context data is in the overflow memory. From the address $j + 1$, the address of context data continuation is read to the variable D. The size of the continuation of context data is $r = U - D$. Go to step 11.

10. Go to step 6.

11. From the overflow memory to the cache, a block of r memory cells is loaded, starting with address D. g is the address of the first byte of the block in the cache. The end address of this block in the cache is calculated as $E = g + \mathrm{r} \cdot (n/4+1)$, j is set as g. Go to step 6.

12. The s keyword context data is found, loaded into the cache memory and is composed. The block address in the cache memory is q, the end address is $j - 1$.

Experiments. The effectiveness of the electronic dictionary, as components of computer translation systems, can be evaluated for speed of context search and level of memory utilization.

For modern computer systems, the search time $T_в$ is calculated as:

$T_в = h \cdot t_\rho + t_n$,

where t_ρ — the execution time of one swap cycle, t — the time of context search, h — the number of swap cycles.

An analysis of the computational processes on which the search for modern electronic dictionaries is based indicates that t_ρ is greater than t_n. That is, the speed of the dictionary can be evaluated by the average number of ρ swap cycles needed to access the keyword context data.

In addition to linguistic information, all electric dictionaries contain service data, by which access to keyword context data is made. This means that the size u of real dictionaries is always larger than the size y of actual linguistic information, herewith $y = w \cdot m$.The effectiveness of using memory can be estimated by comparing full size u of e-dictionaries, which save the same size of linguistic data.

For the experimental part of performance evaluation of the proposed e-dictionary organization, a software complex of the statistical simulation was developed. For the experiment, it was considered that $m = 1000$, $w = 450$ bytes. These parameters were determined by statistical researches of translated and explanatory dictionaries of computer terms.

The first cycle of research is aimed at detecting the influence of v on the ρ and t_ρ. However, experimental results showed no significant influence of v on ρ, so it allows to choose the value of v equal to w.

The main parameter that determines the effectiveness of the proposed vocabulary organization is α. The performed statistical research showed that the increase of value α reduces memory redundancy, but the value of ρ increases.

For example, for $\alpha = 0.013$, the $\rho = 1.45$. The number of hash memory cells can be calculated as m/α, which for this example is equal to $769 \cdot 10^3$. The memory cell size is 6 bytes, so the total amount of memory u is $8074 \cdot 10^3$ bytes.

Performance evaluation of the proposed electronic dictionary organization can be done by comparing with known developments.

In table 1 performance indicators of electronic dictionaries based on a tree with the storage of data in its nodes, based on a tree with separate storage addresses and context data, based on a hash search with collisions and separate storage of addresses and context data and proposed electronic dictionary organization, are shown.

Table 1

Performance evaluation of e-dictionary organizations

Performance indicator	***E-dictionaries based on:***			***Proposed e-dictionary based on a perfect hash addressing***
	a tree with the storage of data in its nodes	***a tree with separate storage addresses and context data***	***a hash search with collisions and separate storage of addresses and context data***	
ρ	13.29	5	2	1.45
$u \cdot 10^6$, байт	4.6	4.66	$4.686 \cdot 10^6$	8.074

The main advantage of the proposed organization of electronic dictionaries is the increasing speed of access to the keyword context data. The resulting effect is achieved by less efficient of using memory. Today hardware memory is becoming cheaper, so the proposed organization of electronic dictionaries is quite justified.

Conclusion. As a result of the research, a way to increase the speed of electronic dictionaries is proposed. It is based on the perfect hash addressing with taking into account the multilevel memory of modern computer systems.

To achieve the goal, the organization of recording and searching data in electronic dictionaries have developed. It can be used as components of high-speed intelligent computerized translation systems.

References

1. Марчук Ю. Н. Компьютерная лингвистика.- АСТ, Восток-Запад, 2007.-165 с.

2. Агапова Н. А. О принципах создания электронного словаря лингво-культорологического типа: к постановке проблемы / Н. А. Агапова, Н. Ф. Картофелева // Вестник Томского государственного университета. № 382 -2014.- С.6-10.

3. Кашеварова И. С. Электронный словарь как новый этап в развитии лексикографии // Молодой ученый. — 2010. — №10. — С. 145-147.

4. Марковский А. П. Интерактивно-шаблонный метод компьютерного перевода научно-технических публикаций / А. П. Марковский, О. Н. Шевченко, Фань Чуньлэй // Вісник Національного технічного університету України ”КПІ” Інформатика, управління та обчислювальна техніка, – Київ: ВЕК+ – 2013. – № 59. - С. 86-97.

5. Выдрин Д. В., Поляков В.Н. Реализация электронного словаря с использованием *n* грамм / Д.В. Выдрин, В.Н. Поляков // Штучний інтелект. № 4 – 2002. – С.180-183.

Autors

Humeniuk Inna – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: humeniuk.inna@gmail.com

Гуменюк Інна Олександрівна – студентка, кафедри обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Markovskyi Olexander – docent (Associated Professor), Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: markovskyy@i.ua

Марковський Олександр Петрович – доцент, кафедри обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Shevchenko Olga – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: ostl@ukr.net

Шевченко Ольга Миколаївна – student, кафедри обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

Інна Гуменюк,
Олександр Марковський,
Ольга Шевченко

СПОСІБ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ЕЛЕКТРОННИХ СЛОВНИКІВ З КОНТЕКСТНИМ ПОШУКОМ

Актуальність теми дослідження. Останні два десятиліття ознаменували себе стрімким перерозподілом центрів виробничої і наукової діяльності. Країни Сходу дедалі частіше займають передові місця у багатьох галузях наукових, особливо технологічних досліджень, що призводить до збільшення інформаційного метаболізму між Сходом та Заходом. Основною перешкодою для обміну науковою технічною інформацією стає мовний бар'єр.

Найперспективнішим шляхом подолання мовного бар'єру є використання більш ефективних систем комп'ютерного перекладу.

Постановка проблеми. Наявні технології досягнення семантичної адекватності комп'ютерного перекладу базуються на різнорівневому аналізі альтернативних варіантів, що потребує багатократного звернення до електронних словників. Як наслідок, розвиток комп'ютерного перекладу ставить якісно нові вимоги до швидкості пошуку в електронних словниках.

Аналіз наявних рішень. На сьогоднішній день існує велика кількість способів організації електронних словників, найпоширенішими з яких є словники на основі: деревних структур, баз даних, хеш-пошуку з колізіями.

Постановка задачі. Мета досліджень полягає в підвищенні швидкості контекстного пошуку в електронних словниках, задля забезпечення ефективної роботи систем комп’ютерного перекладу.

Викладення основного матеріалу. Проведено теоретичні та експериментальні дослідження роботи електронних словників орієнтованих на використання в складі систем комп'ютерного перекладу. Визначено спосіб організації електронних словників, який забезпечує щонайменше в 2 рази швидший пошук в порівнянні з існуючими.

Висновки. В результаті проведених досліджень, запропоновано новий спосіб організації електронних словників систем комп'ютерного перекладу, який базується на perfect хеш-адресації.

Запропонована розробка забезпечує суттєве пришвидшення пошуку, що дозволяє використати її як компоненту швидкодіючих інтелектуалізованих систем комп'ютерного перекладу.

Ключові слова: perfect хеш-адресація, електронний словник, криптографічне перетворення.