ISSN NO: 2770-4491
Date of Publication:30-11-2022

Increase of Efficiency of Word-Forms Search and Processing for the Control and Correction of Spelling Mistakes in the Electronic Texts

Tishlikov Sultonjon Abduraimovich
(Gulistan State University)

Abstract: The principles, methods and algorithms are developed for control and correction of spelling mistakes in the electronic texts. Basic mechanisms are procedures of word-form chains association, allocation of correct morphemes and correction of monogram and combined mistakes. The developed algorithms are realized by parallel calculations technologies.

Keywords: electronic documents, electronic texts, monogram

Introduction. The developing of methods, algorithms and systems for controlling and correction of the text information in electronic documents represents the large scientific interest. The information in the electronic texts is deformed while transferring into communication, because of failures and refusals of input, scanning, recognition devices. Reasons of deformations are also the influence of various handicaps in communication channels, mistakes of the operators [1,2].

As the directions of construction for systems which check and correct the spelling mistakes we can show using the morphological Uzbek words dictionary and the algorithms of search, coding and processing of the texts on the basis of parallel calculations technologies.

The present paper is devoted to results of research and development the algorithms of mistakes detection and correction by morphological dictionary on the basis of methods which analyze and search the correct word-forms chains. In the paper we state also the results of realization the algorithms by parallel calculation technologies.

Principle of mistakes search and correction on the basis of morphemes. Let's consider, that on an input of the information monitoring system the alphabetic chain in letter W moves and we designated positions of discrepancy as D. As d we designated nearest letter. As w we designated the letters in a chain W on a position D. The property of a position D is that any changes of an initial chain W can not result in an allowable chain if these changes leave constant first D letters in W.

The basic principle of the information checking consists in getting a position of discrepancy D with the nearest letter, greater then d. The system includes algorithms with functions of search a word in the dictionary and looking through the lists of morphemes.

The procedure of search in the separate list, both in the dictionary, and in the list of morphemes, defines values D and d with the necessary amendments providing readout D from a beginning of whole analyzed chain W. Let's note that in practice the address in a computer memory is used as D.

The algorithm of morphological analysis at first carries out search of initial sub-chain for analyzed word in the dictionary by viewing the dictionary. After this algorithm search initial sub-chain of the rest part of word in first list of morphemes etc.

While construction of algorithm for correction of mistakes we determine product of letters chains sets $M = M_1 \times M_2$. Here initial part of chains coincides with some chain from M_1, and rest part coincides with chain from M_2. It is required to establish, whether the found element of such product is correct word-form. For this purpose the rule of mistakes correction was constructed on the basis of use the weaker values instead of values D and d. They are defined as the values greater then D and smaller then d for given chain W.

Now let's consider an analyzed chain W at $M = M_1 \times M_2 \times \ldots \times M_n$. We designated through D' and d' values defined for set M' in a view $\{w\} \times M_k$, where $\{w\}$ is sub-chain of correct word-forms, $k < n$.

ISSN NO: 2770-4491
Date of Publication:30-11-2022

We accepted following conditions:

$D = \max\{D'\}$ and $d = \min\{d'\}$ for $\forall d \in \{D\}$.

The above given condition is fair for all algorithms making decomposition of a word W on sub-chains, belonging to $M_1,\ldots,M_n$.

Thus we considered the performance of word-form as element of product of the dictionary and lists of morphemes. Such performance assumes existence only one formal parts of speech. So it is necessary to examinee the performance for several formal parts of speech.

Correction of mistakes on the basis of generalized product of sets. In this case it is considered, that the set of correct word-form for one part of speech p contains in product M^p of the appropriate subset of dictionary and appropriate lists of morphemes. The set of all word-forms for rest parts of speech contains in association of sets M^p, which is used for definition of the weakened values D and d.

As sets in a view M' are accepted any of sets constructed on each of M^p separately.

The algorithm is based on use of the morphological analysis model and for analysis of text from left to right algorithm is guided at each moment by the current status of list with numbers of lines from the morphemes tables.

As the generalized product M of sets $M_1,\ldots,M_n$ is accepted the set of all alphabetic chains, represented as catenation of sub-chains $w_1w_2\ldots w_n$ so, that everyone subset w_k is either element of M_k, or empty chain, depending on value of some $(k-1)$- predicate $P_k(w_1,w_2,\ldots,w_{k-1})$.

The logic rule of the control is:

If $P_k(w_1,w_2,\ldots,w_{k-1}) = True$, then w_k is looked through from all M_k,

Else w_k is empty chain, where $w_k \in M_k$.

Let's note, that predicates $P_k(\ldots)$ are determined by the way of construction the list with numbers of lines from the morphemes tables. Value D' is maximal among values determined by procedures of viewing the dictionary blocks. Value D is maximal among values determined by procedures of viewing the lists of morphemes and value D'.

The stated principles of analysis the word-forms was used for construction of algorithm which corrects monogrammed and combined mistakes.

Algorithms of correction the monogram and combined mistakes. It is supposed, that for any chain the number of possible hypotheses of combined mistakes correction is much less than number of hypotheses of monogram mistakes correction.

The rule of mistakes control and correction consists in replacement of the letter from initial chain by the letter v. The algorithm of detection and correction of monogram and combined mistakes consists in sorting of correction hypotheses for one varied position V in the alphabetic order or in sorting of varied positions V from the end of a word to the beginning. As a hypothesis of mistakes correction for given position V and letter v can be chosen any letters chain conterminous to an initial word on positions which are less V, and having on a position V the letter v.

The developed algorithm includes the following steps and procedures.

Step 1. D is gotten out as an adapted position V.

Step 2. The next varied position V is gotten out. After it is chosen, the list of simple hypotheses for this position is made.

Step 2.1. One letter leaves from the found word. The next letters are rearranged. The hypotheses are derivated for correcting of monogram and combined mistakes conterminous with an initial chain in accuracy on first V letters. On the average such list contains two-three hypotheses.

ISSN NO: 2770-4491
Date of Publication:30-11-2022

Step 3. The letter for replacement v is fixed. Two hypotheses are derivated by replacement of the letter at position V on v and insert of the letter v at position V.

Step 3.1. From the list made on step 2, the hypotheses having at position V the given letter v are gotten out. Then the hypotheses are checked in alphabetic order.

Step 3.2. Any of values getting during check of hypotheses is assigned as new values D and d. The choice of new value v is made. The cyclic shift of the alphabetic order is carried out.

Step 3.3. If v is fixed as a symbol of the end of alphabet, then the value v is replaced to first letter of the alphabet.

Step 3.4. If v it is more, than w or is equal to a symbol of the end of alphabet, then the value v is replaced to a symbol of the end of alphabet.

Step 3.5. If $D_1 < V$, then transition to the first letter of the alphabet is not made.

Step 4. Following values of v or V is gotten out.

Step 4.1. If v is a symbol of the end of alphabet, then the number of a varied position V is decreases and the transition to step 2 is carried out.

Step 4.2. If $V = 0$, then the work of algorithm is finished.

Realization of algorithms of mistakes control and correction. Let's note, that the stated algorithm requires additional memory for storage the list of simple hypotheses, and the size of additional memory on the average makes $3N$, where N is the minimum of length of presented word and of length of longest word from dictionary. Even if all necessary hypotheses are calculated on step 3 each time anew, it is not required the additional memory.

Hypotheses of monogram mistakes correction, such as rearrangement and the inserts, give in addition no more $(2N-1)$ checks, that makes a small part from their common number. Hypotheses of the combined mistakes correction do not increase number of checks essentially, because their number also is proportional N.

As results of testing of developed algorithms shows, the number of addressing to disk memory is increased less than twice in comparison with algorithms of mistakes correction on the basis of dictionary.

In this connection, we offer realization of algorithms on a basis of paralleled information processing for searching on the executive blocks of video chips [3]. The interface of the programmed monitoring system of spelling is realized on the C++ with applications CUDA. System provides the access to memory, shared between flows by the size 16 Кб for a multiprocessor. Such memory is used for organization cache with a wide passband. Besides the system provides more effective data transfer between system and video memory.

The processing of the texts by system is based on principles of splitting the information on blocks, which hold in shared memory. Each block is processed by the block of flows. The subblock is booting in shared memory from global. Above the data in shared memory the appropriate calculations is carried out. The results are copied from shared memory back in global.

The basic process of the application CUDA works on the universal processor (host). It starts some copies of kernel processes on a videomap. The code for CPU makes the following: initializes GPU, distributes memory on a videomap and system, copies constants in memory of a videomap, copies the received result from videomemory, releases memory and finishes work.

Thus, the offered principles, models and algorithms of construction of systems for controlling the authenticity of information allow to combine: the mechanisms of dictionary, statistical and hach-coding; the ways, algorithms, procedures of parallel calculations on the basis of CUDA technology at construction of the system for controlling and correction of spelling mistakes in the texts of electronic documents.

ISSN NO: 2770-4491
Date of Publication:30-11-2022

References

1. Монахов, М. Ю., Семенова, И. И., Полянский, Д. А., Монахов, Ю. М. Особенности среды обеспечения достоверности информации в информационно-телекоммуникационных системах // Фундаментальные исследования. - 2014. - № 9.
2. Бессонов С.В. Оптимизация электронного документооборота в корпоративных системах: Дис. . канд. экон. наук. Москва. 2001. 187с.
3. Гудов А. М. Об одной модели оптимизации документопотоков, реализуемой при создании системы электронного документооборота // Вычислительные технологии. -2006. - том 11, специальный выпуск. - С. 53 - 65
4. Jumanov I.I., Karshiev Kh. B Expanding the possibilities of instruments to improve the information reliability of electronic documents of industrial management systems // «Chemical technology. Control and management» (WCIS-2018)– P.146-150
5. Jumanov I.I., Tishlikov S. A. Method of stochastic search in the system to monitoring and correcting of spelling in electronic texts // 2013 International Conference in Central asia on Internet (ICI), Tashkent, 8-10 october 2013, Section 7, IEEE. – Tashkent, 2013.
6. Jumanov I.I., Akhatov A.R. Fuzzy Semantic Hypernet for Information Authenticity Controlling in Electronic Document Circulation Systems // 4-th International Conference on Application of Information and Communication Technologies, 12-14 october 2010, Section 2, IEEE. - Tashkent, 2010. - p.21-25.
7. Jumanov I.I., Tishlikov S. A. Control of integrity and authenticity of electronic documents on the basis of genetic principles of tests formation and generation // In proceedings of the Eight World Conference on Intelligent Systems for Industrial Automation, 25-27 November, 2014. – Tashkent, Uzbekistan, 2014. – P.242-246
8. Jumanov I.I., Karshiev Kh. B., Tishlikov S.A Examination of the Efficiency of Algorithms for Increasing the Reliability of Information on Criteria of Harness and the Cost of Processing Electronic Documents. International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, September 2019, ISSN: 2277-3878
9. Жуманов И.И., Ахатов А.Р. Оценка эффективности программного комплекса контроля достоверности текстовой информации систем электронного документооборота // «Химическая технология. Контроль и управление». № 2, С. 46-52
10. Jumanov I.I., Karshiev Kh. B Expanding the possibilities of instruments to improve the information reliability of electronic documents of industrial management systems // «Chemical technology. Control and management» (WCIS-2018)– P.146-150
11. Akhatov A.R., Jumanov I.I. Improvement of text information processing quality in documents processing systems // 2nd IEEE/IFIP International Conference In Central Asia On Internet ICI-2006, September 19-21, International Hotel Tashkent, Uzbekistan.
12. Akhatov A.R., Jumanov I.I., Djumanov O.I. An Effective Quality Control of Textual Information on the Basis of Statistical Redundancy in Distributed Mobile IT Systems and e-Applications //3-d International Conference in Central Asia on Internet, Tashkent, 2007.
13. Ney H. and Kneser R. Improved clustering techniques for class-based statistical language modelling. In European Conference on Speech Communication and Technology (Eurospeech), 1993. pages 973–976, Berlin.