International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

Vol. 31 No. 1 February 2022, pp. 304-310

Optimization Of Control And Correction Of Spelling Of Texts Of Electronic Documents

Jumanov Isroil Ibragimovich, Tolipov Asliddin Erkinovich

Doctor of Technical Sciences, Professor, Department of Information Technologies, Samarkand State University, Samarkand, Uzbekistan

Graduate student, Department of Information Technologies, Samarkand State University, Samarkand, Uzbekistan

Abstract – A technique for designing a hypernetwork for searching natural language word forms with various graph models and implementing it in a spelling control and correction system based on the use of soft computing, in particular neural networks, fuzzy models, and fuzzy inferences, is proposed. The parameters of fuzzy graph models are optimized, numerical results are obtained. Knowledge bases with fuzzy rules and databases, n-grams of frequency dictionaries, dictionaries of word forms are implemented.

Keywords – Control And Correction Of Text Spelling, Automated Document Flow, Search Hypernetwork, Neural Network.

I. RELEVANCE OF THE TOPIC

In practice, the solution of various optimization problems is associated with the construction of a fuzzy search hypernetwork, the use of fuzzy rules for checking the reliability of text information, a database (DB), frequency dictionaries of n -grams of structured models of natural language grammar, a knowledge base (KB), fuzzy inference algorithms and neural networks [1,2]. A fuzzy semantic hypernetwork is based on extracting information properties, patterns of n-fold errors, as well as using the properties of self-adaptation and self-organization of a neural network (NN). Fuzzy control rules are formed in such a way as to minimize the search time (or total time), which vary within certain limits [3,4].

This paper proposes constructive approaches, principles, mechanisms for controlling and correcting spelling errors aimed at creating an automated search, image recognition, using databases, knowledge bases, and the apparatus of neuro-fuzzy networks (NFN), where object identifiers are considered in various forms of representations [5,6].

II. SEARCH ENGINE BASED ON A FUZZY HYPERNET

The search image, which is fixed on the model, has a different importance, is evaluated according to the criteria represented by fuzzy numbers. The solution of the problem is carried out by the transition from the original model to the fuzzy one. The sets of vertices and edges of the graph are fixed, which generally describe the network structure of the search. The constructed, oriented fuzzy search graph is based on the extraction and use of fuzzy rules for controlling a distorted word form, the formation of alternative word forms in order to correct errors in texts. To optimize the structure of a fuzzy graph model, we consider three variants of the network model architecture according to fuzzy criteria, in which graph vertices, graph edges, graph vertices and graph edges are considered as variable parameters [7-9].

Corresponding Author: Jumanov Isroil Ibragimovich

Option 1. A fuzzy graph model of the form $\widetilde{G}_1 = (\widetilde{X}, F)$ is designed, in which $\widetilde{X} = \{< \mu(x_i) / x_i >\}$, $i \in I = \{1,2,...,n\}$ are a fuzzy set of graph vertices; F - set of graph edges; $\mu(x_i)$ - value of FP μ for vertex $x_i \in X$, $\mu : X \to [0,1]$.

Option 2. The distances between the vertices of the graph are given by fuzzy numbers. At the same time, fuzzy graph models of the form $\widetilde{G}_2 = (X, \widetilde{F})$ are designed, in which $X = \{x_i\}$, $i \in I = \{1,2,...,n\}$ are the set of graph vertices; $\widetilde{F} = \{< \mu_{\widetilde{F}} < x_i, x_j > / < x_i, x_j >>\}$, $x_i, x_j \in X$ – fuzzy set of graph edges; $\mu_{\widetilde{F}} < x_i, x_j >$ is the value of the membership function (MF) $\mu_{\widetilde{F}}$ for the edge $< x_i, x_j >$ and $\mu_{\widetilde{F}} : X^2 \to [0,1]$.

Option 3. A fuzzy graph model of the form $\widetilde{G}_3 = (\widetilde{X}, \widetilde{F})$ is projected, in which $\widetilde{X} = \{< \mu(x_i) / x_i >\}$, $i \in I = \{1,2,...,n\}$ are a fuzzy set of graph vertices, $\mu(x_i)$ is the value of the FP u for the vertex $x_i \in X$, $\mu : X \to [0,1]$.; $\widetilde{F} = \{< \mu_{\widetilde{F}} < x_i, x_j > / < x_i, x_j >>\}$, $x_i, x_j \in X$ is a fuzzy set of graph edges, $\mu_{\widetilde{F}} < x_i, x_j >$ is the value of the MF $\mu_{\widetilde{F}}$ for the edge $< x_i, x_j >$ и $\mu_{\widetilde{F}} : X^2 \to [0,1]$.

To determine the parameters of graph models, we use the following definitions. Search path - ρ is represented by a sequence of arcs $(a_1, a_2, ..., a_q)$, its length is taken as a fuzzy number $\widetilde{l}(p)$, equal to the sum of the lengths of all arcs included in ρ, i.e.

$$\widetilde{l}(p) = \sum_{(x_i, x_j) \in p} \widetilde{c}_{ij}$$

where $\widetilde{c}_{ij}$ is a fuzzy number representing the length of the arc connecting vertices x_i and x_j.

To determine the shortest fuzzy path between vertices x_i and $x_k \in X$, a fuzzy number $\overline{p}$ is taken, for which the condition

$$\widetilde{l}(\overline{p}_{ik}) = \min_r \widetilde{l}_r(p_{ik}),$$

where p_{ik} is the path between vertices $x_i, x_k \in X$; $r = 1,2,..., s$;

s is the number of different paths between vertices $x_i, x_k \in X$ of the graph $\widetilde{G}$.

The divergences of paths between vertices are those paths that differ from each other by at least one edge included in the path. The paths between the vertices are represented as fuzzy numbers, and when determining the membership function, they are segmented [10,11].

III. Optimization mechanism for fuzzy graph search models

Model optimizations are minimax search object placement problems. Consider a fuzzy graph built according to option 1.

A path p in a fuzzy graph is represented by a sequence of arcs $(a_1, a_2, ..., a_q)$, its fuzzy length is taken to be a fuzzy number $\widetilde{l}(p)$ equal to the sum of the lengths of all arcs included in p, i.e.

$$\tilde{l}(p) = \sum_{(x_i, x_j) \in p} \tilde{c}_{ij},$$

where $\tilde{c}_{ij}$ is a fuzzy number representing the length of the arc connecting vertices x_i and x_j.

For the shortest path $\bar{p}_{ik}$ between vertices x_i and $x_k \in X$, the condition

$$\tilde{l}(\bar{p}_{ik}) = \min_r \tilde{l}_r(p_{ik}),$$

where p_{ik} is the path between vertices $x_i, x_k \in X$; $r = 1,2,...,s$, s is the number of different paths between vertices $x_i, x_k \in X$ of the graph $\tilde{G}$.

The solution of the problem is based on operations on interval numbers. Let two intervals $A = [d_{s1}, d_{l1}]$ and $B = [d_{s2}, d_{l2}]$ be given. The sum of two interval numbers $A = [d_{s1}, d_{l1}]$ and $B = [d_{s2}, d_{l2}]$ is determined by the formula

$$A + B = [d_{s1} + d_{s2}, d_{l1} + d_{l2}],$$

where the value $w(A) = d_{l1} - d_{s2}$ estimates the width of the interval $A = [d_{s1}, d_{l1}]$.

The center of interval $A = [d_{s1}, d_{l1}]$ is calculated as

$$m_A = (d_{s1} + d_{l1})/2.$$

We propose the following ways to optimize the graph model based on the comparison of intervals [12,13].

Method 1. Comparison of the left boundaries of the intervals. Let two intervals $A = [d_{s1}, d_{l1}]$ and $B = [d_{s2}, d_{l2}]$ be given. Then $A < B$ if $d_{s1} < d_{s2}$.

Method 2. Comparison of the right boundaries of the intervals. Let $A = [d_{s1}, d_{l1}]$ and $B = [d_{s2}, d_{l2}]$. Then $A < B$ if $d_{l1} < d_{l2}$.

Note that these methods do not take into account the length of the interval and only one of the boundaries is used in the comparison. Moreover, if in the first case $d_{s1} = d_{s2}$, and in the second - $d_{l1} = d_{l2}$, then it is necessary to compare the width of the intervals.

Method 3. Comparison of interval centers: $A < B$ if $m_A < m_B$.

This method better takes into account the size of the intervals and their left and right boundaries, but a situation may arise when $m_A = m_B$. In accordance with the above, we propose a generalized method for comparing intervals.

Method 4. To determine the minimum of two intervals, the following condition is checked.

If $d_{s1} < d_{s2}$ and $d_{l1} < d_{l2}$, then $A < B$.

If it is not satisfied, then it is necessary to check the second condition

$$A < B, \text{ if } m_A < m_B,$$

where $m_A = (d_{s1} + d_{l1})/2$ and $m_B = (d_{s2} + d_{l2})/2$.

If $m_A = m_B$, then the third condition is checked: $A < B$ and $w(A) < w(B)$, where $w(A) = d_{l1} - d_{s1}$ and $w(B) = d_{l2} - d_{s2}$.

Thus, first it is checked whether conditions 1 and 2 are satisfied simultaneously, then condition 3 is checked. If this is not enough, the width of the intervals is compared.

Optimization of graph model parameters. For each vertex $x_i \in X$ of the reduced graph $\widetilde{G}$, we define two fuzzy numbers - the external and internal separation of the vertex x_i. Then we get the following expressions for calculating fuzzy numbers of external and internal separation [14,15]:

$$\widetilde{s}_0(x_i) = \max_{x_j \in X}[\widetilde{d}(x_i, x_j)],$$

$$\widetilde{s}_t(x_i) = \max_{x_j \in X}[\widetilde{d}(x_j, x_i)].$$

Vertex x_0^*, for which $\widetilde{s}_0(x_0^*) = \min_{x_i \in X}[\widetilde{s}_0(x_i)]$, represents the outer center of the graph $\widetilde{G}$. And the vertex x_t^*, for which

$$\widetilde{s}_t(x_t^*) = \min_{x_i \in X}[\widetilde{s}_t(x_i)],$$

represents the interior center of graph $\widetilde{G}$.

The vertices x_0^*, which are the outer center, represent the fuzzy inner radius of the graph $\widetilde{p}_t = \widetilde{s}_t(x_t^*)$.

When the fuzzy number of external and internal separation of a vertex x_i is given by

$$\widetilde{s}_{0,t}(x_i) = \max_{x_j \in X}\{\widetilde{d}(x_i, x_j) + \widetilde{d}(x_j, x_i)\},$$

then it is considered that at the vertex $x_{0,t}^*$, the minimum of the expression is reached

$$\widetilde{s}_{0,t}(x_{0,t}^*) = \min_{x_i \in X}[\widetilde{s}_{0,t}(x_i)],$$

which represents the outer-inner center of the graph, and the value $\widetilde{\rho}_{0,t} = \widetilde{s}_{0,t}(x_{0,t}^*)$ is the fuzzy outer-inner radius of the graph $\widetilde{G}$.

Assume that the vertices of the graph are given as intervals

$$[d_s(x_i, x_j), d_l(x_i, x_j)],$$

where $d_s(x_i, x_j)$ and $d_l(x_i, x_j)$ are, respectively, the distances between the closest and most distant points of polygons x_i and x_j, respectively. Then the formulas for calculating fuzzy numbers of external and internal separation will take the form:

$$\tilde{s}_0(x_i) = \max_{x_j \in X} \{[d_s(x_i, x_j), d_l(x_i, x_j)]\},$$

$$\tilde{s}_t(x_i) = \max_{x_j \in X} \{[d_s(x_j, x_i), d_l(x_j, x_i)]\}.$$

Since there are interval numbers on the right-hand sides of the given equalities, the values of the numbers of external and internal separation will also be intervals. In this case, the outer radius of the graph is defined as follows [16,17]:

$$\tilde{s}_0(x_0^*) = [s_{0s}(x_0^*), s_{0l}(x_0^*)] = \min_{x_i \in X} \{[s_{0s}(x_i), s_{0l}(x_i)]\},$$

and the inner radius is defined as

$$\tilde{s}_t(x_t^*) = [s_{ts}(x_t^*), s_{tl}(x_t^*)] = \min_{x_i \in X} \{[s_{ts}(x_i), s_{tl}(x_i)]\}.$$

Fuzzy numbers of external and internal separation of vertices are entered in the distance matrices, which are located in the last column and in the last row of the matrices. As can be seen from these tables, when comparing intervals along their left boundaries, the centers of the graph will be the x_2, x_4 vertices. In this case, the value of the graph radius is [5, 22].

Let us consider the case when the average values of the distances between the vertices on the model should be taken into account. Let on the edge (x_i, x_j) of the graph, the length of which is equal to $\tilde{c}_{ij}$, the point y be located

$$\tilde{s}_0(y) = \max_{x_i \in X}[\tilde{d}(y, x_i)], \qquad \tilde{s}_t(y) = \max_{x_i \in X}[\tilde{d}(x_i, y)].$$

Point y_0^* for which

$$\tilde{s}_0(y_0^*) = \min_y[\tilde{s}_0(y)],$$

represents the absolute outer center of the graph, and the y_t^* point for which

$$\tilde{s}_t(y_t^*) = \min_y[\tilde{s}_t(y)],$$

absolute interior center of the graph.

Let y be a point, the location of the edge (x_1, x_4) of the graph $\tilde{G}$ in the middle, the distance from the vertex x_1 to the point y, and the distance from the point y to the vertex x_4 are [3, 5].

The shortest fuzzy distances from the point y to all other vertices of the graph, as well as the fuzzy numbers of external and internal separation, are calculated, and their results are shown in Table.1.

If we use their average values as a criterion for comparing intervals, then we get that the point is the center of the graph $\widetilde{G}$ for this example, because the radius value equal to [11, 19] is smaller than all other values of $\widetilde{s}_0(x_i)$ and $\widetilde{s}_t(x_i)$, $x_i \in X$, $i = 1,2,3,4$. In minimax placement problems, the center of the graph can be located both at a vertex and on an edge.

Table 1.

	x_1	x_2	x_3	x_4	y	$\widetilde{s}_0(x_i)$
x_1	0	[4, 12]	[11, 23]	[6, 10]	[3, 5]	[11, 23]
x_2	[4, 12]	0	[7, 11]	[10, 22]	[7, 17]	[10, 22]
x_3	[11, 23]	[7, И]	0	[8, 14]	[11, 19]	[11, 23]
x_4	[6, 10]	[10, 22]	[8, 14]	0	[3, 5]	[10, 22]
y	[3, 5]	[7, 17]	[11, 19]	[3, 5]	0	[11, 19]
$\widetilde{s}_t(x_i)$	[11, 23]	[10, 22]	[11, 23]	[10, 22]	[11, 19]	

IV. Conclusion

Thus, the used models, algorithms and mechanisms based on soft computing allow optimizing the fuzzy graph of search, recognition, control and correction of spelling errors in natural language texts. Why can be given fuzzy weights of vertices, characterizing their importance. Fuzzy weights of vertices or edges of a graph can be represented as fuzzy triangular numbers, numbers with a bell-shaped membership function of linguistic variables.

References

[1] Холмонов, С. М., & Абсаломова, Г. Б. Методы и алгоритмы повышения достоверности текстовой информации электронных документов. Science and world, 43.

[2] Холмонов, С. М., & Абсаломова, Г. Б. (2020). Повышение достоверности текстов на основе логических критериев и базы знаний электронных документов. In ТЕХНИЧЕСКИЕ НАУКИ: ПРОБЛЕМЫ И РЕШЕНИЯ (pp. 15-19).

[3] Жуманов, И. И., & Шарипова, М. (2013). Оптимизация контроля орфографии узбекского языка основе моделей стохастического поиска. In СОВРЕМЕННЫЕ МАТЕРИАЛЫ, ТЕХНИКА И ТЕХНОЛОГИЯ (pp. 129-133).

[4] Isroil, J., & Khusan, K. (2020, November). Increasing the Reliability of Full Text Documents Based on the Use of Mechanisms for Extraction of Statistical and Semantic Links of Elements. In 2020 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-5). IEEE.

[5] Jumanov, I. I., & Karshiev, K. B. (2020, May). Mechanisms for optimization of detection and correction of text errors based on combining multilevel morphological analysis with n-gram models. In Journal of Physics: Conference Series (Vol. 1546, No. 1, p. 012082). IOP Publishing.

[6] Jumanov, I. I., Karshiev, K. B., & Tishlikov, S. A. (2019). Examination of the efficiency of algorithms for increasing hereliability of information on criteria of harness and the cost of processing electronic documents. International Journal of Recent Technology and Engineering, 8(2), 4133-4139.

[7] Jumanov, I. I., Akhatov, A. R., & Djumanov, O. I. (2007, September). An effective quality control of textual information on the basis of statistical redundancy in distributed mobile IT systems and e-applications. In 2007 3rd IEEE/IFIP International Conference in Central Asia on Internet (pp. 1-5). IEEE.

[8] Жуманов, И. И., & Каршиев, Х. Б. (2019). Основе базы электронных документов и особенностей правил контроля базы знаний. Проблемы вычислительной и прикладной математики, (3), 57-74.

[9] Jumanov, I. I., & Islomov, A. B. (2016). Optimizasiya obrabotki izobrajeniy mikroobyektov na osnove rekurrentnogo obucheniya neyronnoy seti i implikativnogo otbora informativnix priznakov. Problemi informatiki i energetiki, 4, 12.

[10] Жуманов, И. И., & Ахатов, А. Р. (2010, October). Нечеткая семантическая гиперсеть контроля достоверности информации в системах электронного документооборота. In 4-th International Conference on Application of Information and Communication Technologies, Tashkent (pp. 12-14).

[11] Akhatov, A. R., & Jumanov, I. I. (2006, September). Improvement of text information processing quality in documents processing systems. In 2006 2nd IEEE/IFIP International Conference in Central Asia on Internet (pp. 1-5). IEEE.

[12] Jumanov, I. I., Akhatov, A. R., & Tursinxanov, N. M. (2006, September). Methods and algorithms of input information protection in electronic document processing systems. In 2006 2nd IEEE/IFIP International Conference in Central Asia on Internet (pp. 1-5). IEEE.

[13] Jumanov, I. I., & Karshiev, K. B. (2018). Expanding the possibilities of instruments to improve the information reliability of electronic documents of industrial management systems. Chemical Technology, Control and Management, 2018(3), 146-150.

[14] Ibragimovich, J. I., & Azamat o'g'li, A. J. (2022). Optimization of Text Processing In Documents of Automated Office Work Systems. EUROPEAN JOURNAL OF INNOVATION IN NONFORMAL EDUCATION, 2(2), 374-378.

[15] Ibragimovich, J. I. ., & Erkinovich, T. A. . (2022). Control of the Reliability of Textual Information in Documents Based on Neuro-Fuzzy Identification. Middle European Scientific Bulletin, 21, 144-149. Retrieved from https://cejsr.academicjournal.io/index.php/journal/article/view/1075

[16] Ibragimovich, J. I. ., & Abdusalyamovich, D. B. . (2022). Optimization of Neural Network Identification of a Non-Stationary Object Based On Spline Functions. International Journal of Innovative Analyses and Emerging Technology, 2(2), 49–55. Retrieved from http://openaccessjournals.eu/index.php/ijiaet/article/view/1021

[17] Ibragimovich, J. I., & Baxromovna, M. M. (2022). Adaptive Processing of Technological Time Series for Forecasting Based on Neuro-Fuzzy Networks. International Journal of Human Computing Studies, 4(2), 30-35. Retrieved from