O METODĂ DE REZOLVARE A PROBLEMEI DE SUMARIZĂRE A INFORMAŢIEI TEXTUALE

[1]Victoria LAZU, [1]Stanislav SANDUŢA, [1]Valeria UNGUREANU,
[2]Ghenadie SAFONOV

[1]Universitatea Tehnică a Moldovei, [2]Academia Militară a Forţelor Armate „Alexandru cel Bun”

Abstract: *În lucrarea de faţă sunt prezentate rezultatele dezvoltării unei metode de rezolvare a problemei de sumarizare a informaţiei textuale bazată pe selectarea cuvintelor cheie care sunt întâlnite cel mai frecvent în text şi selectarea propoziţiilor care includ aceste cuvinte pentru modelarea informaţiei de sumarizare a textului.*

Cuvinte cheie: *Logica Fuzzy, teoria mulţimilor, sumarizarea textului, procesarea limbajului natural, calcul natural, calcul membranar, P-systems.*

Introducere

Procesul de sumarizare este o operaţie efectuată de un sistem hardware-software cu scopul extragerii unui rezumat, care include punctele importante, cu un conţinut ce descrie esenţa documentului original [1,2]. Complexitatea algoritmică ale sistemelor dedicate pentru crearea unui rezumat includ operaţii ce depind de lungimea, stilul şi sintaxa textului procesat. Sinteza unui rezumat face parte din domeniul inteligenţei artificiale şi include operaţii din *machine learning* şi *data mining*.

Există două abordări generale ale metodelor de sumarizare automată: sumarizare prin extragere şi sumarizare prin abstractizare. Metodele bazate pe extragere funcţionează prin selectarea unui subset de cuvinte, fraze sau propoziţii existente în textul original pentru a forma rezumatul. Pe când, metodele bazate pe abstractizare construiesc o reprezentare semantică internă şi apoi folosesc tehnici de generare a limbajului natural pentru a crea un rezumat mai aproape de gândirea umană. Un astfel de rezumat ar putea include inovaţii verbale obţinute în rezultatul auto-învăţării [2,3].

Una din metodele cele mai eficiente pentru sporirea performanţelor ale sistemelor dedicate pentru procesarea limbajelor naturale este utilizarea procesoarelor specializate cu procesare paralelă a datelor. Totodată, dezavantajul acestor sisteme este: domeniul limitat de utilizare, lipsa unor modele matematice bazate pe paralelism, cost mare pentru dezvoltarea aplicaţiilor hardware şi software [4].

Un avantaj deosebit poate fi oferit de utilizarea metodelor netradiţionale de procesare a datelor care după conţinut şi structură impun un paralelism în procesul de calcul.

În lucrare se propune dezvoltarea unui sistem specializat destinat pentru rezolvarea problemei de sumarizare a informaţiei textuale bazat pe paralelismul oferit de calculul natural în special calculul celular (P-systems) [5,6].

1. Dezvoltarea modelului matematic pentru rezolvarea problemei de sumarizare a informaţiei textuale

Fie este dat textul $\boldsymbol{T^I} = \bigcup_{i=1}^{I} (\boldsymbol{c_i})$, unde $\boldsymbol{c_i}$ este mulţimea de cuvinte şi $\boldsymbol{I}$ numărul total de cuvinte în text. Textul $\boldsymbol{T^I}$ este structurat în aliniate $\boldsymbol{T^I} = \bigcup_{j=1}^{J} (\boldsymbol{A_j})$ şi respectiv, fiecare aliniat $\boldsymbol{A_j}$ este structurat în propoziţii $\boldsymbol{A_j} = \bigcup_{l=1}^{L_j} (\boldsymbol{P_{j,l}})$. Pentru fiecare propoziţie $\boldsymbol{P_{j,l}}$ este definită mulţimea de cuvinte $\boldsymbol{P_{j,l}} = \bigcup_{i=1}^{I_{j,l}} (\boldsymbol{c_{j,l,i}})$, unde $\forall \boldsymbol{c_{j,l,i}} \in \boldsymbol{T^I}$.

În scopul obţinerii textului sumarizat $\boldsymbol{T^S}$ din textul $\boldsymbol{T^I}$ sunt definite următoarele operaţii de procesare a datelor:

$\boldsymbol{T^I} \xrightarrow{Q} \boldsymbol{T^K}$ - extragerea cuvintelor cheie, unde: $\boldsymbol{Q}$ - este algoritmul de extragere a cuvintelor cheie; $\boldsymbol{T^K} = \bigcup_{m=1}^{M} (\boldsymbol{c_m^K})$ şi $\forall \boldsymbol{c_m^K} \in \boldsymbol{T^I}$;

$T^I \xrightarrow{G(T^K)} T^S$ - extragerea rezumatului, unde $G(T^K)$ este algoritmul de extragere a textului pentru sumarizare în baza cuvintelor cheie T^K.

Modelul matematic al algoritmului de extragere a cuvintelor cheie.

Algoritmul de extragere a cuvintelor cheie T^K din textul T^I s-a elaborat în baza modelului matematic:

$$T^K = \bigcup_{m=1}^{M} \left(c_i \left| \max^{N \geq N^*} \left\{ \sum \left(c_i^R \right), \forall i = \overline{1, I} \right\} \right. \right),$$

unde: $\sum(c_i^R)$ - numărul de rădăcini ale cuvântului c_i în textul T^I; $\max^{N \geq N^*}$ - selectarea cuvintelor c_i care se întâlnesc mai frecvent de N^* ori în textul T^I.

Modelul matematic al algoritmului de sumarizare în baza cuvintelor cheie.

Algoritmul de sumarizare $G(T^K)$ în baza cuvintelor cheie T^K s-a elaborat în baza modelului matematic:

$$T^S = \bigcup_{l=1}^{L_S} \left(\max^{N \geq N^*} \left\{ \bigcup_{h=1}^{H} (P_h) \left| c_m^K \in P_h \right. \right\} \right),$$ unde: $\bigcup_{h=1}^{H} (P_h)$ - reuniunea propoziţiilor care includ cuvântul cheie c_m^K; $\max^{N \geq N^*}$ - selectarea propoziţiilor care includ cuvântul cheie c_m^K de mai multe ori ca N^*.

Menţiuni

Cercetările efectuate fac parte din tematica tezelor de doctorat planificate în cadrul Departamentului Informatică şi Ingineria Sistemelor, FCIM, UTM. Modelările şi testările experimentale au fost efectuate în baza dispozitivelor oferite de CSCT „Hard & Soft” şi ORANGE Cafee.

Bibliografie

1. Marcu, Daniel. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press Cambridge, MA, USA, 2000, 248p., ISBN 0-262-13372-5.
2. Mani, Inderjeet. *Automatic Summarization.* John Benjamins Publishing, 2001, 285p., ISBN 1-58811-060-5.
3. Louis, A. & Nenkova, A., *Performance Confidence Estimation for Automatic Summarization.* Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 541–548, Athens, Greece, 30 March – 3 April, 2009.
4. Н.И. Червяков, П.А. Сахнюк, А.В. Шапошников, С.А. Ряднов. *Модулярные параллельные вычислительные структуры нейропроцессорных систем.* - М.: ФИЗМАТЛИТ, 2003. - 288 с., ISBN: 5-9221-0327-X.
5. Gh. Păun. *A quick introduction to membrane computing*. The Journal of Logic and Algebraic Programming, 79, 2010, pp. 291–294.
6. G. Zhang, J. Cheng, T. Wang, X. Wang, J. Zhu. *Membrane Computing: Theory and Applications*, Science Press, Beijing, China, 2015.
7. ABABII, V.; SUDACEVSCHI, V.; LAZU, V.; MUNTEANU, S.; UNGUREANU, V. Distributed computing system based on mobile program code. *In The XIV International Conference "Measurement and Control in Complex Systems" – MCCS-2018, October 15-17, 2018, Vinnytsia, Ukraine.* { http://ir.lib.vntu.edu.ua//handle/123456789/22769}, file: 145.pdf.
8. АБАБИЙ, В.; СУДАЧЕВСКИ, В.; ЛАЗУ, В.; МУНТЯНУ, С.; УНГУРЯНУ, В. Организация распределенных вычислительных процессов для решения задачи многокритериального поиска информации в неструктурированных текстовых документах. *Proceedings of the Seventh International Conference on Informatics and Computer Technics Problems PICT-2018, 11-14 October, 2018, Chernivtsi, Ukraine.* pp. 71-74.
9. ABABII,V.; SUDACEVSCHI, V.; LAZU, V.; MUNTEANU, S.; UNGUREANU, V. Distributed data processing method for extracting knowledge. *Book of Abstracts of the 26th Conference on Applied and Industrial Mathematics - CAIM-2018, 20-23 September, 2018, Technical University of Moldova, Chisinau. Moldova*. pp. 114.