MINISTERUL EDUCAȚIEI ȘI CERCETĂRII AL REPUBLICII MOLDOVA

Universitatea Tehnică a Moldovei

Facultatea Calculatoare, Informatică și Microelectronică

Departamentul Ingineria Software și Automatică

Admis la susținere
Șef departament:
Ion Fiodorov, conf. univ., dr.
_____
”____”_____2022

Căutarea semantică bazată pe rețele neuronale

Teză de master

Student: **Stratulat Ștefan, TI-201M**

Conducător: **Chirev Pavel, lector univ., dr.**

Consultant: **Cojocaru Svetlana, lector univ. mag.**

Chișinău, 2022

Rezumat

Tema acestei lucrări este axată pe cercetarea metodelor de căutare semantică și are ca scop utilizarea rețelelor neuronale la căutarea semantică în documente, lucrarea este justificată de necesitatea soluțiilor inteligente de căutare, în care se va procesa semantica documentelor, dar nu numai structura acestora, pentru a putea căuta documente în web în baza interogărilor formate cu ajutorul limbajului natural și a primi rezultatele dorite. Prin urmare, lucrarea dată pune accentul pe utilizarea rețelelor neuronale la căutarea semantică, din considerente că soluțiile actuale din webul semantic nu sunt automatizate complect.

Lucrarea are ca obiective analiza domeniului de căutare structurală și semantică în documente, compararea metodelor de căutare semantică și structurală în documente, cercetarea utilizării rețelelor neuronale artificiale la căutarea structurală și semantică în documente, cercetarea soluțiilor existente de căutare semantică bazată pe rețele neuronale, elaborarea unui algoritm de indexare semantică a documentelor, elaborarea unui algoritm de căutare semantică a documentelor, elaborarea unui motor de căutare semantică în documente în baza algoritmilor elaborați, implementarea unei aplicații de testare a motorului de căutare semantică.

Structura lucrării conține: introducere, 3 capitole, concluzii, bibliografie cu 26 titluri, 61 pagini text de bază, 48 figuri și 5 formule.

Capitolul 1: În acest capitol este realizată analiza domeniului, analiza motoarelor de căutare, analiza metodelor de căutare a documentelor și anume metodele clasice de căutare structurală și semantică, analiza metodelor de căutare structurală și semantică utilizînd inteligența artificială.

Capitolul 2: În acest capitol este realizată analiza motoarelor de căutare semantică precum Swoogle și Watson, definirea elementelor primare a unui motor de căutare semantică în documente bazat pe rețele neuronale, mecanismul de indexare semantică a documentelor și cum are loc formarea interogării și căutarea documentelor.

Capitolul 3: În acest capitol este descrisă realizarea motorului de căutare semantică, ce instrumente și tehnologii au fost utilizate, cum este realizat mecanismul de indexare semantică și de căutare semantică a documentelor, descrierea interfeței grafice și a procesului de evaluare a motorului de căutare semantică elaborat.

Abstract

The topic of this paper is focused on the research of semantic search methods and aims to use neural networks for semantic search in documents, the paper is justified by the need for intelligent search solutions, which will process the semantics of documents, but not only their structure, to could search documents on the web based on natural language queries and get the results they wanted. Therefore, this paper focuses on the use of neural networks in semantic search, considering that current solutions in the semantic web are not fully automated.

The paper aims to analyze the field of structural and semantic search in documents, compare methods of semantic and structural search in documents, research the use of artificial neural networks in structural and semantic search in documents, research existing solutions for semantic search based on neural networks, developing a document semantic indexing algorithm, elaboration of a semantic document search algorithm, elaboration of a semantic search engine in documents based on elaborated algorithms, implementation of a semantic search engine testing application.

The structure of the paper contains: introduction, 3 chapters, conclusions, bibliography with 26 titles, 61 pages of basic text, 48 figures and 5 formulas.

Chapter 1: In this chapter is performed the analysis of the field, the analysis of search engines, the analysis of document search methods, namely the classical methods of structural and semantic search, the analysis of structural and semantic search methods using artificial intelligence.

Chapter 2: This chapter analyzes semantic search engines such as Swoogle and Watson, defines the primary elements of a document search engine based on neural networks, the semantic indexing mechanism of documents, and how querying and document searching takes place.

Chapter 3: This chapter describes the implementation of the semantic search engine, what tools and technologies have been used, how the semantic indexing and semantic search mechanism of documents is performed, the description of the graphical interface and the evaluation process of the developed semantic search engine.

CUPRINS

INTRODUCERE

În zilele noastre, este aproape imposibil de lucrat cu numărul imens de date din web, fără a utiliza un motor de căutare. Posibilitatea de a găsi o anumită informație a jucat un rol crucial în utilizarea internetului în ultimul deceniu. Motoarele de căutare majore pot efectua căutări foarte rapide și precise aproape pe întregul web, oferind interfețe simple de utilizator bazate pe cuvinte cheie. Căutarea bazată pe cuvinte cheie s-a dovedit a fi foarte eficientă într-o colecție de conținut textual nestructurat, de ex. pagini web. Cu toate acestea, dacă utilizatorii doresc să găsească informații într-un conținut structurat, de ex. într-o bază de date, căutarea de bază a cuvintelor cheie eșuează. O soluție simplă, dar suficientă pe web, poate fi furnizată de o interfață de utilizator bazată pe formular. Cu acest tip de interfață, utilizatorul poate combina de obicei cuvinte cheie cu alte restricții, în funcție de structura de domeniu specifică. Cu toate acestea, interfețele de utilizator bazate pe formular sunt mai complexe decît căutarea simplă a cuvintelor cheie.

Un pas dincolo de abordările tradiționale menționate mai sus sunt interfețele de limbaj natural (NLI). Caracteristica cheie a unei astfel de interfețe este că utilizatorii pot căuta informațiile solicitate prin formularea întrebărilor lor folosind un limbaj natural care permite formularea precisă a nevoilor lor de informații. Deoarece NLI poate funcționa atît pe conținut structurat cît și pe cel nestructurat, ajută la unificarea accesului la date și permite un nou mod de experiență a utilizatorului. Nevoia de a înțelege limbajele naturale pe computer a jucat un rol semnificativ în multe domenii de cercetare în ultimele cîteva decenii. Unele dintre rezultate au fost deja transferate cu succes din modele teoretice, conceptuale în sfera comercială și în industrie.

Cu toate acestea, combinația de înțelegere a limbajului natural (NLU) și recuperarea informațiilor (IR) aduce o mulțime de sarcini noi provocatoare. O altă direcție promițătoare în evoluția Web, Web-ul semantic, a adus multe concepte interesante în modelarea domeniului și partajarea datelor. Deoarece fundamentele sale se bazează pe o descriere semantică foarte precisă, multe aplicații web existente pot beneficia de încorporarea tehnologiilor web semantice. Mai mult, dezvoltarea interfețelor de limbaj natural către web-ul semantic (NLISW) a arătat că reducerea decalajului dintre web-ul semantic și interfețele de limbaj natural poate descoperi noi provocări de cercetare. Problema de bază a Web-ul semantic este că necesită structurarea și organizarea documentelor în web, fapt ce în ziua de astăzi se face manual de către experți în domeniul Web-ul semantic și a domeniului din care informațiile sunt plasate în web.

Scopul acestei lucrări este de a utiliza inteligența artificială, rețelele neuronale pentru a efectua căutarea semantică pe informații nestructurate și neadnotate manual.

BIBLIOGRAFIE

[1] Statistica motoarelor de căutare. Disponibil: https://www.internetlivestats.com/total-number-of-websites/

[2] Indexarea și raitingul motorul de căutare Google. Disponibil: https://www.google.com/search/howsearchworks/crawling-indexing/

[3] Dik L Lee, Huei Chuang, and Kent Seamons. "Document ranking and the vector-space model". In: IEEE software 14.2 (1997), pp. 67–75.

[4] Akiko Aizawa. "An information-theoretic perspective of tf–idf measures". In: Information Processing & Management 39.1 (2003), pp. 45–65.

[5] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. "Introduction to information retrieval". In: Natural Language Engineering 16.1 (2010), pp. 100–103.

[6] Mikolov, Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space".

[7] Gruber, T. R. (1991). The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases. In J. A. Allen, R. Fikes, & E. Sandewall (Eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference, Cambridge, MA, pages 601-602, Morgan Kaufmann.

[8] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. "Natural language processing: an introduction". In: Journal of the American Medical Informatics Association 18.5 (2011), pp. 544–551.

[9] Kassim, J. M., & Rahmany, M. (2009, August). Introduction to semantic search engine. In 2009 International Conference on Electrical Engineering and Informatics (Vol. 2, pp. 380-386). IEEE.

[10] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: arXiv preprint arXiv:1810.04805 (2018).

[11] Jürgen Schmidhuber. "Deep learning in neural networks: An overview".In: Neural networks 61 (2015), pp. 85–117.

[12] Ashish Vaswani et al. "Attention is all you need". In: Advances in neural information processing systems. 2017, pp. 590–608.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[14] Nils Reimers and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: arXiv preprint arXiv:1908.10084 (2019).

[15] Davide Chicco. “Siamese Neural Networks: An Overview”. In: Artificial Neural Networks. Springer, pp. 73–94.

[16] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. “Siamese neural networks for one-shot image recognition”. In: ICML deep learning workshop. Vol. 2. Lille. 2015.

[17] Sowa, J.F.: Conceptual graphs summary. In: Nagle, T.E., Nagle, J.A., Gerholz, L.L., Eklund, P.W. (eds.) Conceptual Structures: Current Research and Practice, pp. 3–51. Ellis Horwood, New York (1992) ISBN:0-13-175878-0

[18] Chen, H., Perich, F., Finin, T., Joshi, A.: SOUPA: standard ontology for ubiquitous and pervasive applications. In: Proceedings of the First International Conference on Mobile and Ubiquitous Systems: Networking and Services (Mobiquitous 2004), Boston (2004)

[19] Bizer, C.: The emerging web of linked data. IEEE Intell. Syst. 24, 87–92 (2009)

[20] Bizer, C., Heath, T., Berners-Lee, T.: Linked data, the story so far. Int. J. Semantic Web Inf. Syst. 5(3), 1–22 (2009)

[21] Androutsopoulos, I., Ritchie, G.D., Thanisch, P.: Natural language interfaces to databases: an introduction. Nat. Lang. Eng. 1(1), 29–81 (1995) 698 16 Semantic Web Search Engines

[22] Harth, A., Umbrich, J., Decker, S.: Multi-crawler: a pipelined architecture for crawling and indexing semantic web data. In: Proceedings of the Fifth International Semantic Web Conference (ISWC 2006), Athens, GA. Lecture Notes in Computer Sci- ence, vol. 4273, pp. 258–271. Springer, Berlin (2006)

[23] Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., Kolari, P.: Finding and ranking knowledge on the semantic web. In: Proceedings of the Fourth Inter- national Semantic Web Conference (ISWC 2005), Galway. Lecture Notes in Computer Science, vol. 3729, pp. 156–170. Springer, Heidelberg (2005)

[24] Alani, H., Brewster, C., Shadbolt, N.: Ranking ontol- ogies with aktiverank. In: Proceedings of the Fifth International Semantic Web Conference (ISWC 2006), Athens, GA. Lecture Notes in Computer Sci- ence, vol. 4273, pp. 1–15. Springer, Berlin (2006)

[25] d’Aquin, M., Euzenat, J., Le Duc, C., Lewen, H.: Sharing and reusing aligned ontologies with cupboard. In: Demo, Proceedings of the Fifth International Conference on Knowledge Capture (K-CAP 2009), Los Angeles (2009)

[26] Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., and Sachs, J.: Swoogle: a search and metadata engine for the semantic web. In: Proceedings of the 13th ACM International Conference on Information and Knowledge Management (CIKM 2004), Wash- ington, DC, pp. 652–659 (2004)