

# ANDDigest: An AI-enhanced web-tool for finding information in the biological literature

Ivanisenko T.V.<sup>1,2\*</sup>, Demenkov P.S.<sup>1,2</sup>, Kolchanov N.A.<sup>1,3</sup> Ivanisenko V.A.<sup>1,2,3</sup>

<sup>1</sup> *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

<sup>2</sup> *Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

<sup>3</sup> *Novosibirsk State University, Novosibirsk, Russia*

\* [itv@bionet.nsc.ru](mailto:itv@bionet.nsc.ru)

**Key words:** text-mining, ANDDigest, ANDSystem, knowledge retrieval, machine learning, neural network

*Motivation and Aim:* Finding relevant information is an extremely important issue in almost any scientific research. Well-known scientific search engines such as Google Scholar, Scopus, PubMed, etc. [1–3] allow finding relevant documents based on user-defined keywords, but do not provide the user with tools for the automated extraction and analysis of information from search results, which can reach thousands of documents. Also, such queries do not take into account the synonyms for the objects of interest and their linkage with external databases.

The vast amounts of published scientific literature made the task of searching for the information relevant to the field of the researcher's study one of the most important problems. Well-known search engines, such as Google Scholar, Web of Science, PubMed, Scopus, and others, are highly effective in finding documents; however, they do not allow the use of an ontology of the subject area [1–3]. Programs based on automatic text analysis provide the extraction of knowledge and its presentation in the form of semantic networks [4]. Of particular interest among them are systems providing a full cycle of knowledge engineering, including their automatic extraction, integration, presentation in the form of semantic networks, visualization, and analysis. Examples of such software products are STRING [5], Pathway Studio [6], MetaCore [7, 8], ANDSystem [9–12], etc.

Earlier, we developed the ANDDigest tool [13] designed for the finding of the information from the textual sources based on the ANDSystem ontology, presented in the form of the dictionaries. However, due to the use of the dictionary-based approach one of the main drawbacks of this system, was a high number of the false positive recognitions of the molecular-genetics objects in texts. The most of such errors were associated with incorrect recognition of the short names (5 or less symbols).

*Methods and Algorithms:* To deal with a described problem we used a fine-tuning strategy [14] to the pretrained BioBERT transformer model [15], previously trained on the large number of the PubMed abstracts and full-text articles. The fine-tuned models were used as binary classifiers for the 5 groups of objects from the ontology of ANDSystem. Each classifier was trained on the separated set of abstracts, generated from the mapped data, where only the long names of objects were considered.

*Results:* The developed new version of the ANDDigest system designed for the intelligent search, analysis, and graphical representation of the information in the scientific literature. The system was improved with five deep learning classification models, providing the automated context-based filtration of the short names in natural language texts.

The models were created for the classification of short names of objects of the following types: genes/proteins, metabolites/drugs, cell components, diseases/side effects, and pathways. For the assessment of the quality of each model, a verification dataset was created. Each of the sets consisted of about 15,000 of sentences, containing short names of objects from the described group, mapped by the dictionary. Each object was manually assigned as positively or negatively recognized. The AUC values appeared to be the following: diseases/side effects classification model – 0.96; genes/proteins classification model – 0.94; cell components model – 0.91; drugs/metabolites model – 0.91; pathways model – 0.88.

The new version of ANDDigest is available by <http://anddigest.sysbio.ru>

**Acknowledgements:** Work was funded by the Ministry of Science and Higher Education of the Russian Federation project “Kurchatov Center for World-Class Genomic Research” No. 075-15-2019-1662 from 2019-10-31.

## References

1. Beel J., Gipp B. Google Scholar's Ranking Algorithm: An Introductory Overview. In: In Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09). 2009;1:230-241.
2. McEntyre J., Ostell J. The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information (US). 2002.
3. Jacso P. As we may search - Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Curr Sci.* 2005;89:1537-1547.
4. McEntyre J., Ostell J. The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information (US). 2013.
5. Szklarczyk D. et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* 2017;45(D1):D362-D368.
6. Nikitin A. et al. Pathway studio – the analysis and navigation of molecular networks. *Bioinformatics.* 2003;19(16):2155-2157.
7. Nikolsky Y. et al. Biological networks and analysis of experimental data in drug discovery. *Drug Discov Today.* 2005;10(9):653-62.
8. Ekins S. et al. Algorithms for network analysis in systems-ADME/Tox using the MetaCore and MetaDrug platforms. *Xenobiotica.* 2006;36(10-11):877-901.
9. Demenkov P.S. et al. ANDVisio: A new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biol.* 2012;11(3, 4):149-161.
10. Ivanisenko V.A. et al. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Syst Biol.* 2015;9(S2):S2.
11. Ivanisenko V.A. et al. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics.* 2019;20(1):34.
12. Saik O.V. et al. Prioritization of genes involved in endothelial cell apoptosis by their implication in lymphedema using an analysis of associative gene networks with ANDSystem. *BMC Med Genomics.* 2019;12(2):47.
13. Ivanisenko T.V. et al. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics.* 2020;21(11):1-21.
14. Ji Z. et al. (2020) Bert-based ranking for biomedical entity normalization. In: AMIA Summits on Translational Science Proceedings. 2020:269-277.
15. Lee J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* 2020;36(4):1234-1240.