Healthcare Transformation with Informatics and Artificial Intelligence
J. Mantas et al. (Eds.)

doi:10.3233/SHTI230458

Digital Tools in UMLS Metathesaurus Knowledge Processing

Pavel A. ASTANIN[a, 1], Svetlana E. RAUZINA[a] and Tatyana V. ZARUBINA[a]
[a]*Pirogov Russian National Research Medical University*
ORCiD ID: Pavel A. Astanin – 0000-0002-1854-8686, Svetlana E. Rauzina – 0000-0002-9535-2847, Tatyana V. Zarubina – 0000-0002-4403-8049.

Abstract. Clinical search engines development is actual task for medical informatics. The main issue in this area is to implement high-quality unstructured texts processing. Ontological interdisciplinary metathesaurus UMLS can be used to solve this problem. Currently, there is no unified method to relevant information aggregation from UMLS. In this research, we have presented the UMLS as graph model and performed the spot check of UMLS structure to identify basic problems. Then we created and integrated new graph metric in two created by us program modules for relevant knowledge aggregation from UMLS.

Keywords. UMLS, unstructured text, graph metric.

1. Introduction

Currently, the Unified medical language system (UMLS) is the largest metathesaurus that may be used in unstructured medical text analysis [1]. The latest version of UMLS (2022AB) provides terminological coverage of 11.2 million writing variants (aui) of 4.6 million concepts – unique interdisciplinary terms classified by thematic affiliation into 127 groups (tui). Semantically similar concepts are connected by relationships that are unambiguously assigned into 9 main (and 2 additional) groups and 992 specifying optional subgroups. Each UMLS concept is connected with at least one any concept. That is why this metathesaurus may be visualized as directed multigraph with 98 million unique relationships between 4.6 million graph nodes [2].

Data organization as graph models is characterized by a number of advantages. Among them there are almost unlimited possibility to processing optimization and availability of software tools for visualize knowledge results [3]. At this moment, there is no generally accepted approach to relevant knowledge aggregation from graph models. This approach should be based on using of valid optimized analytical tools and metamodels – unified set of rules and tools that regulates many aspects in models technical realization. However, it is nearly impossible to develop such generally accepted approach. That is why creating of analytical tools for aggregation and visualization of relevant data at least from UMLS remains the current task of medical informatics [4].

[1] Corresponding Author: Pavel Astanin, E-mail: med_cyber@mail.ru.

The primary objective in multigraph processing is to range concepts by significance degree in searching task context. G. Szarnyas and coauthors have evaluated some graph metrics before [5]. In the current research, we have adapted one of the most powerful metrics known as clustering coefficient. Then we used it to design two analytical modules for automatic aggregation data from UMLS. Consequently, the aim of this study was to formalize as scripts two modules for knowledge aggregation from UMLS.

2. Methods

The current study was performed under the grant «Priority – 2030» in the Medicine digital transformation Institute – scientific body of the Pirogov Russian National Research Medical University. A wide specter of technical tools has been used in the current study. They include one of the most used graph database management system (DBMS) Neo4j, relational DBMS PostgreSQL and column-oriented DBMS ClickHouse. In addition, the PyCharm integrated development environment and the Django framework were used for consolidation of analytical scripts as unify system.

3. Results

At the primary stage of this study, the spot check of UMLS structure was performed for five clinically heterogeneous diseases: ischemic heart diseases (I20-I25), gastrointestinal hemorrhage (K92.2), ischemic stroke (I63, G45), malignant neoplasm of breast (C50) and ankylosing spondylitis (M45). Unstructured texts of clinical guidelines for these diseases were used as verified information sources. Three main problems were formulated after expert extraction of terms and their mapping with UMLS concepts have been completed.

The first problem lies in deficiency of direct relationships between clinically significant concepts. So, at least one direct relationship with concept-disease is presented in 19–52% symptomatic terms (an example is demonstrated on the fig. 1). That is why mathematic algorithms based on graph theory are very important in detecting of indirect paths between clinically relevant UMLS nodes.

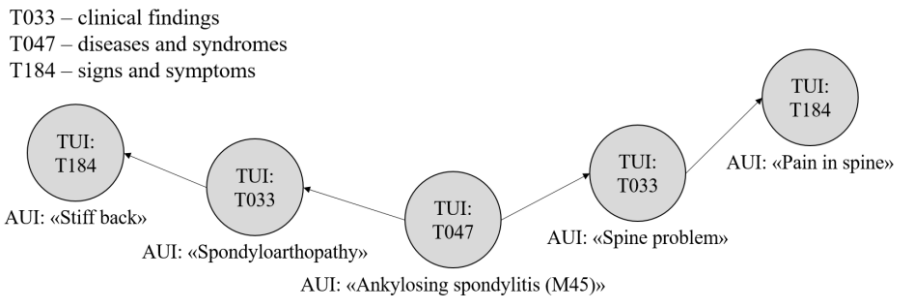

Figure 1. An example of indirect relationships between clinically significant UMLS concepts.

The second issue lies in difficulty of concept searching with using automatic unstructured text analysis algorithms. It is caused by low part of originally Russian and

other Slavic translated clinical terms. Just over 144 thousand or 3% of concepts have Russian wording variants in the current version of UMLS.

Lastly, the third problem lies in strongly pronounced UMLS structure unevenness. The number of direct relationships that demonstrate the complexity of knowledge structure, ranges from 1 to 91730 depending on the concept. It leads to shifting of concept significance degree estimates with using non-specialized mathematical metrics.

Overall, we determined the number of direct relationships for each Russian translated UMLS term and built a distribution function that approximated number of concepts that having such nodes-neighbors in graph model.

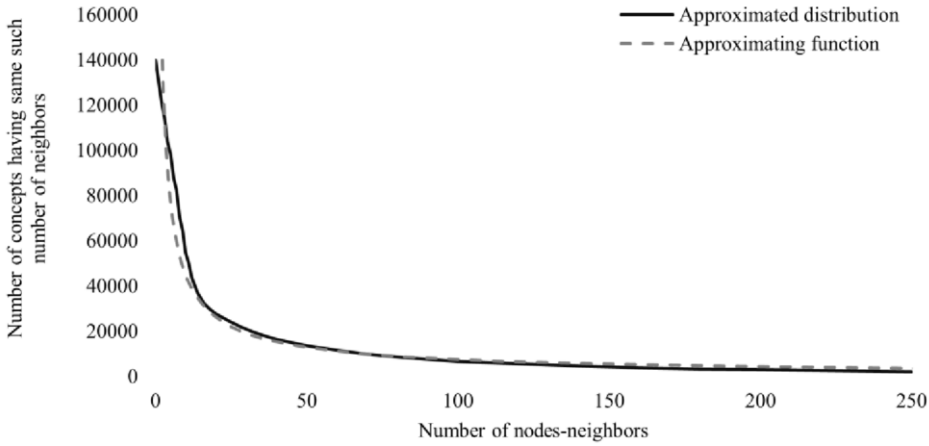

Figure 2. Approximation of the direct relationships amount on the number of nodes with no less nodes-neighbors.

According to fig. 2, the stated empirical regularity is similar to Pareto distribution that also is observed in the famed Zipf's law applicable for semantic models. This pattern is high determinated ($R^2 = 0{,}939$, $p<0{,}001$) by the following mathematical expression:

$$y = 2 \cdot 10^6 \cdot X^{-1.29} \quad (1)$$

In the expression 1, X is the number of graph contours that the appropriate graph node is included in, Y – the actual value of the concept significance degree. We have supposed that combining of this approximating function with counting of graph contours (similarly, as in the clustering coefficient) provides wide possibilities to range UMLS concepts. This new graph metric was named as weighted clustering coefficient (WCC). Then we used this metric in developing two modules for UMLS analysis.

The first module is designed for retrospective seeking of concepts. This searching type means that the information about root and leaf nodes is known initially. It is based on shortest paths between root and lead nodes searching. Aggregated data about nodes and relationships types has been processed and online analytical processing (OLAP) of this data is implemented then. This module allows to formulate searching patterns that can be used in the future. So, user can customize following settings: ID of root and leaf nodes, using types of nodes and relationships and the search depth.

The second module is designed for prospective seeking of concepts. It means that the information about root nodes is known initially, but the information about leaf nodes

is unknown. Consequently, the main task of this module is to find relevant lead nodes with aforementioned searching patterns using. Thus, user can customize following settings: ID of root nodes, using types of intermediate and leaf nodes, types of relationships and graph contours, search depth and size of each layer (number of ranged by significance degree nodes in descending order).

Returning to the mathematical expression 1, we offer to use 2 types of contours in WCC counting in analytical modules: graph triangles and graph rhombs (an example is demonstrated on fig. 3).

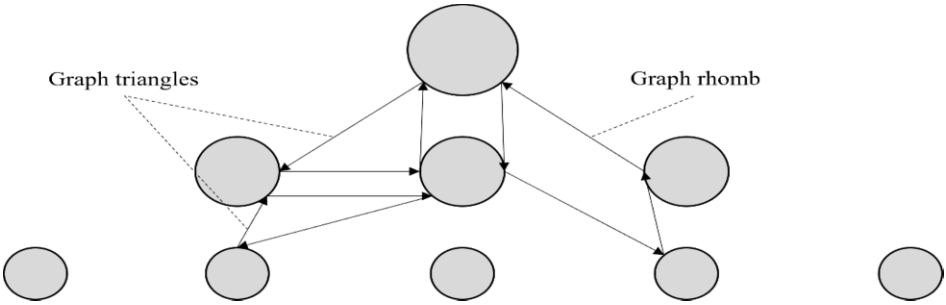

Figure 3. Some types of graph contours that using in UMLS concepts ranging.

4. Conclusion

Graph model of UMLS is effective way to formalized knowledge presentation method. It allows to use this metathesaurus in unstructured medical text analysis. Our primary spot check of UMLS structure for five heterogeneous diseases allowed us to identify main problems in automatic knowledge analysis. Based on using empirical function of relationships density distribution, the weighted clustering coefficient makes it possible to rank graph nodes by significance degree. Retrospective and prospective analysis combining allows to detection of searching patterns that may be used in relevant data aggregation from UMLS metathesaurus. The primary task in subsequent step is to formulate the searching patterns set and use it to develop symptom-checkers and clinical decision support systems based on unstructured text analysis.

References

[1] Humphreys BL, Tuttle MS. Something new and different: The Unified Medical Language System. Information services & use. 2022 Mar; 42(1): 95–106. DOI: 10.3233/ISU-210138.

[2] Wood EC, Glen AK, Kvarfordt LG., Womack F, Acevedo L, Yoon TS et al. RTX-KG2: a system for building a semantically standardized knowledge graph for translational biomedicine. BMC bioinformatics, 2022 Sep; 23(1): 400. DOI: 10.1186/s12859-022-04932-3.

[3] Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023 Feb; 10(1): 67. DOI: 10.1038/s41597-023-01960-3.

[4] Reimer AP, Milinovich A. Using UMLS for electronic health data standardization and database design. Journal of the American Medical Informatics Association: JAMIA. 2020 Sep; 27(10): 1520–1528. DOI: 10.1093/jamia/ocaa176.

[5] Szarnyas G, Kovari Z, Salanki A, Varro D. Towards the characterization of realistic models: evaluation of multidisciplinary graph metrics. Proceedings of the ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems. 2016 Oct; 87–94. DOI: 10.1145/2976767.2976786.