RU EN ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В МЕДИЦИНЕ

УДК 004.652.3, 616-079.4
10.25209/2079-3316-2023-14-3-59-94

Построение этиопатогенетического образа концептов метатезауруса UMLS с использованием графовых метрик

Павел Андреевич **Астанин**[1✉], Светлана Евгеньевна **Раузина**[2], Татьяна Васильевна **Зарубина**[3]

Российский национальный исследовательский медицинский университет им. Н. И. Пирогова, Москва, Россия

[1✉]*med_cyber@mail.ru*

Аннотация. Разработка средств информационной поддержки принятия клинических решений (ППКР) является актуальной задачей медицинской информатики. Довольно часто в системах ППКР используются информационно-поисковые алгоритмы, важным этапом проектирования которых служит создание средств автоматического распознавания этиопатогенетического образа заболеваний при работе с неструктурированным текстом. В настоящей статье произведены обзор и сравнительная характеристика аналитических метрик, применимых для построения образа концептов метатезауруса Unified Medical Language System (UMLS), представленного в виде графовой информационной модели. Предложен собственный вариант графовой метрики, показавший наибольшую эффективность при решении данной задачи.

Ключевые слова и фразы: медицинская информационная система, информационно-поисковый алгоритм, база знаний, теория графов, UMLS

Благодарности: работа выполнена за счет средств стратегического проекта «Приоритет-2030» на базе Института цифровой трансформации медицины (ИЦТМ) ФГАОУ ВО «Российский национальный исследовательский медицинский университет имени Н. И. Пирогова» Минздрава России.

Для цитирования: Астанин П. А., Раузина С. Е., Зарубина Т. В. *Построение этиопатогенетического образа концептов метатезауруса UMLS с использованием графовых метрик* // Программные системы: теория и приложения. 2023. **Т. 14**. № 3(58). С. 59–94. https://psta.psiras.ru/read/psta2023_3_59-94.pdf

Введение

Unified medical language system (UMLS) является крупнейшим сводом биомедицинских справочников и словарей, применимых в работе с неструктурированными данными [1]. Актуальная версия UMLS (2022AB) обеспечивает терминологический охват свыше 4.6 млн концептов – уникальных междисциплинарных понятий, классифицированных по тематической принадлежности на 127 групп. Семантически близкие концепты соединены связями, однозначно отнесенными к 9 основным (и 2 дополнительным) типам и 992 уточняющим необязательным подтипам (уточнениям). Практически каждый концепт UMLS связан хотя бы с одним другим концептом, что позволяет представить данный свод терминов в виде ориентированного мультиграфа с 98 млн уникальных связей.

Организация данных в виде графовых информационных моделей имеет ряд преимуществ, среди которых следует выделить наличие больших возможностей для оптимизации аналитических операций и существование средств, обеспечивающих наглядную интерпретацию структуры знаний на пользовательском уровне [2]. Однако в настоящее время не существует единого подхода к извлечению релевантных знаний из графовых информационных моделей [3]. Применение простых инструментов автоматического извлечения знаний из UMLS (например, фильтров на типы связей и тематические группы терминов) не приводит к клинически значимому результату по причине недостатка прямых связей между концептами метатезауруса и неоднородности структуры знаний. Необходимо ансамблирование разнородных аналитических инструментов с использованием сложных систем весовых коэффициентов, оптимизированных валидированных аналитических инструментов и метамоделей – унифицированных сводов правил технической реализации моделей знаний.

Одним из элементов системы весовых коэффициентов сущностей UMLS могут стать значения, отражающие степень принадлежности каждого термина к отдельным клиническим профилям (пульмонологии, кардиологии, гастроэнтерологии), а также свидетельствующие о наличии этиологической или патогенетической связи с соответствующими областями. Каждый профиль включает патологические состояния, относящиеся к определенному классу заболеваний. Набор подобных значений для отдельного концепта формирует его этиопатогенетический образ – вектор ненулевых значений функции принадлежности термина к характерным клиническим профилям.

Создание этиопатогенетического образа для концептов UMLS позволит определять степень их значимости в контексте решаемых задач и обеспечивать значительное сокращение ширины поиска при выполнении графовых запросов. Вычисление мер близости между векторами функции

принадлежности сущностей к соответствующим медицинским областям даст возможность ранжирования симптомов и заболеваний по степени их клинического сходства, что может стать основой автоматического формирования дифференциально-диагностических рядов. Автоматическое построение этиопатогенетического образа терминов UMLS позволит сравнить существующие графовые метрики и обосновать их применение при решении задач поиска релевантных знаний. В свою очередь, успешное построение системы весовых коэффициентов для концептов метатезауруса UMLS обеспечит значительное улучшение информационно-поисковых алгоритмов за счет повышения математической контрастности получаемых результатов.

Цель настоящего исследования заключается в построении и оценке этиопатогенетического образа концептов метатезауруса UMLS с использованием графовых метрик для анализа связности узлов.

1. Постановка задач

Для достижения цели вся работа была разделена на несколько последовательных этапов, соответствующих задачам исследования. Первая задача заключалась в подготовке графовой информационной модели UMLS и подключении русскоязычных справочников терминов. Вторая задача состояла в создании семантических и логических правил включения формулировок заболеваний в свод для построения этиопатогенетического образа концептов. В рамках третьей задачи осуществлялись проектирование и программная реализация алгоритма вычисления функций принадлежности концептов к клиническим профилям. Наконец, в рамках четвертой задачи выполнены обзор и сравнительная оценка различных графовых метрик, применявшихся для построения этиопатогенетического образа концептов-нозологий. Также в ходе данного этапа был разработан собственный вариант графовой метрики, предназначенный для оценки связности концептов UMLS.

2. Графовая информационная модель UMLS

Техническая реализация графовой информационной модели UMLS осуществлялась с использованием стандарта Ecore. В построении модели были задействованы исходно переведенные на русский язык термины из справочников MedDRA (Medical Dictionary for Regulatory Activities Terminology), LOINC (Logical Observation Identifiers Names and Codes) и MeSH (Medical Subject Headings). MedDRA позволяет достичь терминологического покрытия большинства возможных вариантов описания клинической картины заболеваний, лабораторных и инструментальных находок, используемых в практике [4]. LOINC является унифицированным

стандартом в области описания клинико-лабораторных исследований [5]. MeSH содержит вертикальную иерархию заголовков и их синонимов, применяемых для распределения научных статей по предметным рубрикам, и пригодных к использованию в качестве словаря для описания биомедицинских областей [6].

Графовая информационная модель, содержащая русскоязычные формулировки концептов UMLS и связи между ними, оказалась значительно меньше исходной англоязычной версии: всего в полученном графе содержалось 144 тыс. (3.1%) концептов и 7.9 млн (8.0%) связей между ними. Все концепты относились к одной или нескольким группам из 124 (97.6%) возможных. Отсутствующие в русскоязычной версии UMLS тематические группы не имеют прямого отношения к описанию возможных вариантов клинических проявлений и инструментально-лабораторных методов исследования.

Для разметки терминов по клиническим направлениям оценивалась графовая связность их англоязычных аналогов с концептами-нозологиями, соотнесенными с кодами справочника международной классификации болезней 10 пересмотра (МКБ-10) в UMLS. Использовано сопоставление кодов МКБ-10 с актуальной версией справочника из реестра нормативно-справочной информации Минздрава России.

3. Подготовка перечня нозологий для разметки концептов

Задача предварительной диагностики не предполагает точной формулировки диагноза, однако привязка симптомов к нозологическим группам требует введения ограничений на их перечень. Подготовлен набор правил отсечения, позволивших снизить количество нозологических единиц для поиска. К их числу относятся наименования, содержащие слова «неуточненный», «классифицированный», «идентифицированный», «другой», а также общие формулировки, содержащие в подстроках другие нозологические единицы. Например, код «K29» – «Гастрит и дуоденит» можно считать избыточным по причине существования кодов «K29.7» – «Гастрит» и «K29.8» – «Дуоденит», служащих для него подстроками.

Для дополнительного сокращения списка потенциальных нозологий экспертным способом были отобраны клинические симптомы и синдромы, не являющиеся самостоятельными заболеваниями (например, K92.0 – «Кровавая рвота»). Далее для всех кодов МКБ-10, с которыми соотнесено более одного узла, осуществлен отбор концептов с наибольшим числом прямых связей. Указанные процедуры позволили снизить степень потенциального смещения оценок принадлежности терминов к классам и сократить перечень заболеваний с 11232 до 1577. Концептам, соотнесенным с кодами МКБ-10, автоматически присваивалась метка принадлежности к определенному клиническому профилю, согласно таблице 1.

Таблица 1. Основные клинические профили, использованные для построения этиопатогенетического образа концептов UMLS

№	Класс МКБ-10 и диапазон кодов	Наименование профиля
1	V(F00-F99)	Психиатрический
2	VI(G00-G99)	Неврологический
3	VII(H00-H59)	Офтальмологический
4	VIII(H60-H95)	Сурдологический
5	IX(I00-I99)	Кардиологический
6	X(J00-J99)	Пульмонологический
7	XI(K00-K93)	Гастроинтестинальный
8	XII(L00-L99)	Дерматологический
9	XIII(M00-M99)	Ортопедический
10	XIV(N00-N99)	Урогенитальный

Для разметки не использовались состояния, представленные в классах:

I некоторые инфекционные и паразитарные болезни,

II новообразования,

III болезни крови, кроветворных органов и отдельные нарушения, вовлекающие иммунный механизм,

IV болезни эндокринной системы, расстройства питания и нарушения обмена веществ,

XV беременность, роды и послеродовой период,

XVI отдельные состояния, возникающие в перинатальном периоде,

XVII врожденные аномалии [пороки развития], деформации и хромосомные нарушения,

XVIII симптомы, признаки и отклонения от нормы, выявленные при клинических и лабораторных исследованиях, не классифицированные в других рубриках,

XIX травмы, отравления и некоторые другие последствия воздействия внешних причин,

XX внешние причины заболеваемости и смертности,

XXI факторы, влияющие на состояние здоровья населения и обращения в учреждения здравоохранения и

XXII коды для особых целей.

4. Алгоритм построения этиопатогенетического образа концептов-симптомов UMLS

Построение этиопатогенетического образа для концептов-симптомов осуществлялось итеративным циклом, состоящим из нескольких шагов. Первый шаг заключался в агрегации всех терминов, напрямую или косвенно связанных с корневым концептом одним или несколькими из следующих

типов связей: SIB (sibling relationships – горизонтальные связи между близкородственными концептами), RO (other relationships – вертикальные связи несинонимичных терминов), CHD (child relationships) и RN (narrower relationships) – вертикальные связи родительских терминов с дочерними, RQ (related and possibly synonymous relationships – горизонтальные и вертикальные связи близкородственных или синонимичных терминов) и SY (synonymous relationships – строго синонимичные связи). Краткая характеристика основных типов связей UMLS дана в таблице 2 [7].

Таблица 2. Характеристика основных типов связей между концептами UMLS

Класс связей	Тип связи	Пример связи	
		Корневой термин	Концевой термин
Ассоциативные	RO	Боль в спине	Поясничный радикулит
	RQ	Боль в спине	Ощущения дискомфорта в спине
	RL	Гиперурикемия	Боль в пальце ноги
	AQ	Боль в спине	Диагностический аспект
	QB	Диагностический аспект	Боль в спине
Иерархические	SIB	Боль в спине	Артралгия
	PAR	Боль в спине	Дорсопатия
	CHD	Боль в спине	Боль в пояснице
	RB	Боль в спине	Боль
	RN	Боль в спине	Боль в верхней части спины
Синонимичные	SY	Боль в спине	Боль: спина

Необходимо отметить, что иерархические связи PAR (parent relationships) и RB (broader relationships) – вертикальные связи дочерних терминов с родительскими – ведут к обобщающим терминам, искажающим этиопатогенетический образ. По этой причине данные типы связей не использовались для агрегации терминов. В свою очередь, связи RL («like» relationships – горизонтальные или вертикальные связи между близкородственными терминами), AQ (allowed qualifier – вертикальные связи дочерних терминов с родительскими) и QB (can be qualified by – вертикальные связи родительских терминов с дочерними) не были представлены в справочниках терминов UMLS, переведенных на русский язык.

На втором шаге отбирались термины, принадлежащие к тематическим группам, входящим в надкласс Disorders (расстройства) семантической сети UMLS. Коды и расшифровки групп терминов указанного надкласса представлены в таблице 3. Данное поисковое условие позволило исключить группы промежуточных концептов, не имеющих прямого отношения к этиопатогенезу заболеваний: таксономические и хронологические

Таблица 3. Тематические группы терминов, используемые для поиска связей с нозологиями справочника МКБ-10

№	Код	Наименование тематической группы и адаптированный перевод
1	T019	Congenital abnormality (врожденные аномалии)
2	T020	Acquired abnormality (приобретенные аномалии)
3	T033	Finding (клинические находки)
4	T037	Injury or poisoning (травмы или отравления)
5	T046	Pathologic function (патологические функции)
6	T047	Disease or syndrome (заболевания или синдромы)
7	T048	Mental or behavioral dysfunction (расстройства мышления и психики)
8	T049	Cell or molecular dysfunction (клеточные или молекулярные нарушения)
9	T050	Experimental model of disease (экспериментальные модели заболеваний)
10	T184	Sign or symptom (признаки или симптомы)
11	T190	Anatomical abnormality (анатомические нарушения)
12	T191	Neoplastic process (опухолевые процессы)

сущности, экономико-юридические и географические термины, а также узкоспециализированные понятия из биологии и химии.

На третьем шаге определялось число незамкнутых графовых путей между корневым (симптоматическим) и концевым (нозологическим) концептами, связанными друг с другом не более, чем через два промежуточных узла, относящихся хотя бы к одной тематической группе из таблицы 3. Необходимо отметить, что проведенные ранее исследования продемонстрировали возможность нахождения подавляющего большинства релевантных концептов при указанной глубине поиска [8].

В ходе третьего шага извлекались нозологические концепты (соотнесенные с кодами МКБ-10) и рассчитывались различные метрики их связности с корневым узлом. Далее по каждому клиническому профилю полученных концептов значения графовых метрик суммировались. Полученный числовой ряд сортировался по убыванию и подвергался минимаксной нормализации, где единице соответствовала максимальная близость к соответствующей медицинской области, а нулю – минимальная. Значения сохранялись в базе данных и могли использоваться для построения вектора функций принадлежности этиопатогенетического образа отдельных концептов или набора концептов.

5. Оценка качества построения образа нозологий

Для проведения сравнительной характеристики метрик оценки связности при построении этиопатогенетического образа полученная для концепта метка образа нозологической единицы сравнивалась с фактической (класс МКБ-10). Образ считался корректным, если метка класса

МКБ-10 присутствовала в перечне трех наиболее характерных классов в векторе этиопатогенетического образа концепта. Важно отметить, что образы были построены только для концептов, не соотнесенных с кодами МКБ-10. Для нозологий этиопатогенетические образы строились путем суммирования значений функций принадлежности, напрямую связанных с ними симптоматических концептов. Данный подход позволил интегрально оценить эффективность разметки нозологических концептов и сделать вывод о возможности ее применения при решении задачи поиска наиболее подходящих диагнозов на основании любого исходного перечня симптомов.

Для выполнения запросов с большим количеством операций объединения (в первую очередь, cross join), лежащих в основе расчета метрик связности узлов, применялась графовая СУБД Neo4j. При работе с указанной СУБД использовался декларативный язык запросов Cypher [9]. Для точечного поиска концептов UMLS применялась объектно-реляционная СУБД PostgreSQL, демонстрирующая наилучшие результаты при извлечении информации из хранилищ данных [10].

Для интеграции аналитических процедур в единый вычислительный алгоритм применялись библиотеки языка программирования Python: Psycopg2 (для работы с СУБД PostgreSQL), Neo4j (для работы с СУБД Neo4j), Scipy, Pandas и Numpy (для реализации вычислительных операций и формирования OLAP-срезов).

6. Виды графовых метрик для оценки связности узлов

Универсальным языком разметки графовых моделей знаний является Eclipse Modeling Framework (EMF) – набор инструментов и стандартов, обеспечивающих основу для взаимодействия создаваемых метамоделей с другими программными продуктами. Согласно стандарту Ecore, использующему EMF, в зависимости от системы типизации узлов графовые модели делятся на два класса. Так, модель знаний считается монотипной (one-dimensional), если в графе отсутствует деление узлов на типы. При существовании справочника, соотносящего узлы графа с конкретными типами (из множества представленных в модели типов), граф считается многотипным (multidimensional) [11]. Следует уточнить, что исходно в UMLS представлено 127 тематических групп концептов, которые правомерно могут считаться типами. Для многотипных моделей знаний существует ряд общепринятых правил, допущений и ограничений, которые необходимо учитывать при абстрактно-логическом анализе. Прежде всего, каждая семантическая сущность (формулировка термина) должна быть сопоставлена с соответствующим узлом графа. Помимо этого, каждый узел графа должен принадлежать хотя бы к одному из выделяемых

в структуре знаний типов. Наконец, между любой парой узлов графа должен существовать как минимум один путь.

Для монотипных и многотипных графовых моделей существуют различные виды метрик оценки связности узлов [12]. Однако важно отметить, что многотипная модель всегда может быть приведена к монотипной при опущении информации о выделяемых типах узлов. Поскольку UMLS полностью удовлетворяет требованиям стандарта Ecore для классических многотипных графовых моделей знаний, в настоящей статье рассмотрены метрики для обоих классов.

Важным дополнительным атрибутом графовых метрик является наличие или отсутствие нормализации их вычисляемых значений. Для всех нормализованных метрик интервал допустимых значений лежит в строгих пределах от 0 до 1, где нулевое значение метрики указывает на полное отсутствие связности узла с подграфом, а единичное – на достижение максимальной связности [13]. Очевидным преимуществом нормализованных метрик является возможность проведения сравнительных оценок мощности и валидности. Помимо этого, нормализованные метрики могут быть объединены в единый экспертный ансамбль для повышения эффективности решения задач извлечения знаний из графовых моделей. Тем не менее, ненормализованные метрики также могут представлять практическую значимость, благодаря возможности их адаптации под индивидуальные требования работы с отдельными графовыми моделями знаний, а также за счет доступности их модификации путем проведения нормализации с использованием собственных аналитических подходов.

Наиболее информативными структурами UMLS являются узлы (термины), в то время как связи между ними служат в качестве вспомогательных элементов, хранящих сведения о природе взаимоотношений соответствующих понятий. В связи с этим особую значимость представляют аналитические метрики, основанные на оценке связности узлов в подграфе (таблица 4) [14].

Таблица 4. Графовые метрики для оценки связности узлов в подграфе

№	Оригинальное название	Смысловой перевод на русский язык	Тип модели	Наличие нормализации
1	Clustering coefficient	Коэффициент кластеризации	А	Да
2	Dimensional clustering coefficient	Коэффициент многотипной кластеризации	Б	Да
3	Node activity	Коэффициент типовой принадлежности	Б	Нет
4	Multiplex participation coefficient	Коэффициент взвешенной типовой принадлежности	Б	Да
Примечания: А – Монотипная графовая модель, Б – Многотипная графовая модель				

Согласно данным из таблицы 4, к числу метрик для оценки связности узлов в монотипных графовых моделях знаний относится коэффициент

кластеризации (КК). КК предусматривает нормализацию выходных значений [15] и определяется по формуле

$$\text{КК}_n = \frac{2 \cdot C_n}{(R_n^{\text{в}} + R_n^{\text{из}}) \cdot (R_n^{\text{в}} + R_n^{\text{из}} - 1)}, \tag{1}$$

где C_n – количество графовых (геометрических) контуров, в образовании которых участвует узел n; $R_n^{\text{в}}$ – число рёбер (связей), входящих в узел n, а $R_n^{\text{из}}$ – число рёбер (связей), исходящих из узла n.

К метрикам оценки значимости узлов в многотипных графовых моделях знаний относятся коэффициент многотипной кластеризации, коэффициент типовой принадлежности и коэффициент взвешенной типовой принадлежности.

Коэффициент многотипной кластеризации (КМК) является обобщенной версией рассмотренного ранее КК для монотипных моделей, также рассчитывается по формуле (1) и имеет три возможных вариации. Первый вариант КМК предполагает включение в расчет узлов только того типа, к которому относится корневой узел. Второй вариант данной метрики отличается от предыдущего включением в расчет узлов исключительно тех типов, к которым не относится корневой узел. Третий вариант КМК не предполагает введения ограничений на соответствие или несоответствие типов корневого узла и связанных с ним узлов.

Коэффициент типовой принадлежности (КТП) равняется числу уникальных типов узлов, с которыми связан корневой узел. Необходимо отметить, что КТП не предполагает нормализации значений и не учитывает структуру встречаемости различных типов связанных узлов.

Коэффициент взвешенной типовой принадлежности (КВТП) устраняет проблемы, связанные с применением классического КТП, и определяется по формуле

$$\text{КВТП}_n = \frac{T}{T-1} \cdot \left[1 - \sum_{d \in D} \left(\frac{R_n(d) + R_n^{\text{из}}(d)}{R_n(D) + R_n^{\text{из}}(D)} \right)^2 \right], \tag{2}$$

где T – число уникальных типов узлов в подграфе для узла n; $R_n(d)$ – число прямых связей (исходящих или входящих) между корневым узлом n и другими узлами того же типа (d); $R_n(D)$ – число прямых связей (исходящих или входящих) между корневым узлом n и другими узлами любого типа (D).

7. Эмпирический закон «ранг–частотность»

Необходимо отметить, что структура связей между концептами UMLS является неоднородной. При большом числе представленных типов узлов (тематических групп), а также многообразии типов и атрибутов связей

между терминами целесообразным является применение наиболее мощных многотипных нормализованных графовых метрик – третьего варианта КМК и КВТП. Применение этих аналитических инструментов для работы с огромными массивами данных может стать причиной длительного поиска знаний, в связи с чем необходимо заниматься параллельной разработкой более простых способов оценки связности терминов в семантической сети. К числу таких способов следует отнести подсчет числа прямых связей между исследуемым корневым узлом и другими узлами, а также подсчет геометрических контуров, в которых присутствует корневой узел.

При построении монотонно убывающей функции распределения частоты встречаемости узлов формируется классическая кривая Парето (рисунок 1) [16]. Согласно данным из рисунка, для любой точки на кривой можно определить количество концептов UMLS с числом прямых связей не менее заданного. Подобное распределение плотности структуры графовых элементов встречается при анализе языковых моделей и известно как закон Ципфа [17]. Настоящее исследование показывает, что схожая обратная закономерность характерна для графовой информационной модели UMLS.

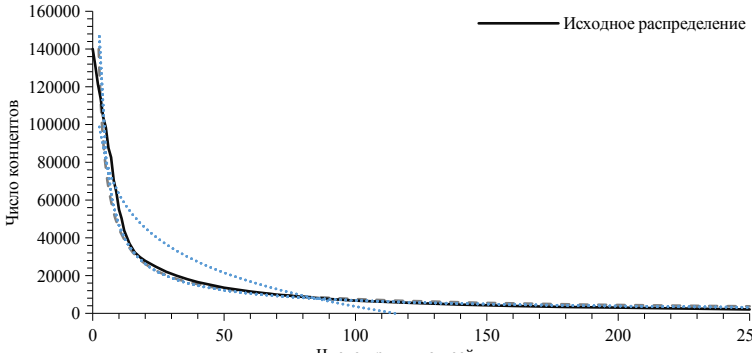


Рисунок 1. Кривая зависимости минимального числа прямых связей и числа концептов UMLS с соответствующим числом прямых связей

Математическое выражение, аппроксимирующее описанную ранее зависимость, представлено формулой

$$y = 2 \cdot 10^6 \cdot x^{-1.290}, \tag{3}$$

где x – число прямых связей термина, y – количество концептов с числом прямых связей не менее x. Коэффициент детерминации (R^2) для данного эмпирического закона составил 0.939, что свидетельствует о высоком качестве аппроксимации ($p < 0.001$).

8. Взвешенный коэффициент кластеризации

Вышеупомянутый закон может стать фундаментальной основой для создания нормализованных метрик ранжирования концептов графовой модели UMLS. Подобные математические инструменты позволят учесть плотность прямого окружения концептов при решении задачи поиска релевантных знаний и избежать потенциального смещения оценок при ранжировании узлов графа.

С использованием формулы (3) было получено выражение для расчета взвешенного коэффициента кластеризации термина (ВКК) внутри извлекаемого подграфа UMLS

$$\mathrm{ВКК}_n = \frac{C_n}{{R_n}^{1.290}}, \tag{4}$$

где C_n – число геометрических контуров или незамкнутых путей, в образовании которых участвует узел n, R_n – число прямых связей между узлом n и любыми другими вершинами графа.

Для масштабирования значений ВКК, рассчитанных по формуле (4), необходимо применить минимаксную нормализацию

$$\mathrm{ВКК}'_n = \mathrm{ВКК}_n - \frac{\min(\mathrm{ВКК}N)}{\max(\mathrm{ВКК}N) - \min(\mathrm{ВКК}N)}, \tag{5}$$

где $\mathrm{ВКК}_n$ – значение ВКК для узла n; $\mathrm{ВКК}_N$ – набор рассчитанных значений ВКК для набора узлов N, для которого необходимо произвести ранжирование.

9. Варианты графовых контуров и их применение в анализе структуры UMLS

Необходимо отметить, что геометрическими контурами в теории графов принято называть любые замкнутые фигуры – структуры, в которых один и тот же узел является корневым и концевым. На рисунке 2 представлены некоторые возможные типы геометрических контуров в абстрактном подграфе. Наиболее сбалансированным и оптимизированным по вычислительным затратам является использование поиска графовых треугольников. Именно этот тип контуров рассчитывается в классических версиях КК и КМК, рассмотренных ранее. В зависимости от количества включаемых узлов в каждый уровень подграфа могут быть использованы различные виды геометрических контуров. В настоящем исследовании предлагается использовать три разновидности в зависимости от размера и числа слоев извлекаемого подграфа терминов UMLS. Так, подсчет числа внешних треугольников (у которых одна вершина обращена к корневому

концепту, а две – к концевому) целесообразно при условии

$$N_{\text{внутр.}} < N_{\text{внешн.}}, \tag{6}$$

где $N_{\text{внутр.}}$ – число концептов в слое, расположенном ближе к корневому концепту, а $N_{\text{внешн.}}$ – число концептов на более внешнем (терминальном) слое по отношению к $N_{\text{внутр.}}$. Подобная ситуация характерна для подграфа, представленного на рисунке 2, поскольку в его промежуточном слое насчитывается всего 4 концепта, а в терминальном слое – 7.

При изменении знака неравенства в формуле (6) на противоположный рекомендуется использовать подсчет внутренних треугольников – контуров, у которых одна вершина обращена к терминальному слою концептов, а две – к корневому узлу). Подсчет числа графовых ромбов целесообразен при приблизительном равенстве размеров слоев при необходимости учета наличия непрямых связей между терминами. Использование варьирования различных типов графовых контуров в зависимости от размера слоев извлекаемого подграфа позволит повысить контрастность расчетов и улучшить качество ранжирования концептов.

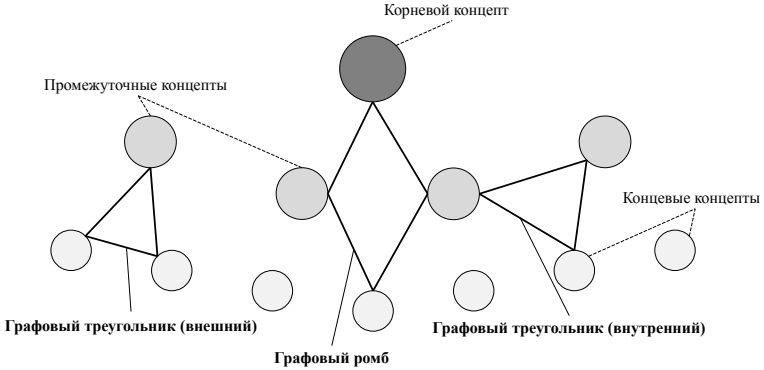


Рисунок 2. Варианты геометрических контуров

10. Результаты оценки качества разметки нозологических концептов UMLS

При использовании метрик из таблицы 4, а также эмпирически полученного взвешенного коэффициента кластеризации удалось построить этиопатогенетический образ для 134 тыс. симптоматических (93.9%) концептов русскоязычной версии UMLS (оставшиеся концепты не имели прямых и непрямых связей с нозологиями МКБ-10 при глубине поиска третьего уровня).

Далее с использованием образов симптоматических концептов осуществлена оценка качества построения образов для нозологических

концептов при использовании различных графовых метрик (таблица 5).

Таблица 5. Сравнительная характеристика результатов классификации нозологических концептов UMLS ($n = 1577$)

Тип графовой метрики	Тип графовых контуров		P
	гамильтоновы	негамильтоновы	
Коэффициент кластеризации (КК)	63 ± 1.2	64 ± 1.2	0.361
Коэффициент многотипной кластеризации (КМК)	72 ± 2.2	74 ± 2.2	0.297
Коэффициент типовой принадлежности (КТП)	24 ± 2.4	24 ± 2.4	0.965
Коэффициент взвешенной типовой принадлежности (КВТП)	84 ± 1.6	88 ± 1.6	0.024
Взвешенный коэффициент кластеризации (ВКК)	86 ± 1.7	91 ± 1.4	0.009

В ячейках таблицы представлены значения долей концептов МКБ-10, для которых эталонная метка присутствовала в перечне трех наиболее характерных классов в векторе этиопатогенетического образа ($p \pm m$, где p – значение доли, m – 95% доверительный интервал). Согласно данным, представленным в таблице, наилучший результат показал подход с использованием эмпирически полученного ВКК. Значение метрики качества классификации составило $86 \pm 1.7\%$ при использовании только гамильтоновских графовых контуров (со строго не повторяющимися гранями) [18]. Использование негамильтоновских графовых контуров (с возможностью повторения одних и тех же граней в контуре дважды) позволило статистически значимо повысить качество классификации и добиться значений метрики качества в $91 \pm 1.4\%$ ($p = 0.009$).

Клиническая интерпретация этиопатогенетического образа терминов UMLS может быть продемонстрирована на следующем примере. Так, для концепта-симптома с русскоязычной формулировкой «боль в спине» при использовании алгоритма классификации, основанного на применении взвешенного коэффициента кластеризации, был получен следующий этиопатогенетический образ: [«Ортопедический профиль»: 1.000, «Неврологический профиль»: 0.921, «Кардиологический профиль»: 0.652, «Урогенитальный профиль»: 0.560]. Указанный термин имеет наиболее выраженную связь с патологиями костно-мышечной (1.000) и нервной (0.921) систем, сердечно-сосудистыми патологиями (0.652) и заболеваниями урогенитального тракта (0.560). Полученный образ согласуется с современными данными об этиопатогенезе боли в спине [19–21]. Связь с ортопедическим и неврологическим профилями может указывать на наличие неспецифических (деструктивных, дисфункциональных или дистрофических) изменений опорно-двигательного аппарата с возможностью вторичного повреждения смежных структур периферической нервной системы. Связь с заболеваниями сердечно-сосудистой системы заключается в возможности возникновения боли с иррадиацией

в спину при остром инфаркте миокарда, когда боль может отдавать в спину под левую лопатку или между лопатками. Наконец, боль в спине может возникать при некоторых заболеваниях почек и проявляться в виде симптома Пастернацкого, что позволяет считать полученный образ корректным.

Заключение

Вопрос о возможности использования графовых информационных моделей в системах ППКР характеризуется низкой степенью научной проработанности [22–25]. Реализация информационной поддержки принятия клинических решений с использованием метатезауруса UMLS требует применения метрик оценки связности концептов для их ранжирования по степени значимости в контексте решаемой задачи. Извлечение релевантных знаний из UMLS может стать возможным при разработке специализированных аналитических метрик и создании системы весовых коэффициентов в структуре графовой информационной модели.

В настоящем исследовании проведена сравнительная характеристика графовых метрик при решении задачи построения этиопатогенетического образа концептов метатезауруса UMLS. Среди найденных в литературе метрик оценки связности узлов в графе допустимые результаты показали коэффициент многотипной кластеризации и коэффициент взвешенной типовой принадлежности. Доли совпадений меток алгоритма с фактическим классом заболевания из справочника МКБ-10 составили 72–74% и 84–88%, соответственно. Наилучшее качество классификации продемонстрировал авторский взвешенный коэффициент кластеризации (86–91%), основанный на использовании эмпирически полученной функции распределения плотности структуры прямых связей узлов. Определение числа негамильтоновских графовых треугольников, образованных целевым концептом, позволяет с высокой степенью точности осуществлять оценку степени принадлежности любого концепта метатезауруса UMLS к соответствующим клиническим направлениям.

Исходя из результатов многоклассовой классификации нозологий МКБ-10 по клиническим направлениям, был сделан вывод о возможности использования взвешенного коэффициента кластеризации для решения задач разметки концептов UMLS. В дальнейшем предполагается использование полученных образов симптоматических концептов при дифференциальной диагностике заболеваний на основании данных, получаемых из неструктурированного текста с использованием алгоритмов извлечения именованных сущностей. Соотнесение формулировок из текстов с клинически размеченными концептами UMLS позволит строить уникальные этиопатогенетические варианты образов по клинической картине и сопоставлять их с образами заболеваний, что сделает возможным осуществление качественной работы информационно-поисковых алгоритмов.

Список литературы

[1] Астанин П. А., Ронжин Л. В., Раузина С. Е., Зарубина Т. В. *Алгоритмы семантического анализа данных и возможности их применения в разработке медицинских информационных систем* // *Цифровая статистика. Новые задачи и траектория движения – 2022*, Материалы IV Съезда медицинских статистиков Москвы (21–23 сентября 2022 года).– 2022.– С. 6–9. ↑60

[2] Кукарцев В. В., Колмакова З. А., Мельникова О. Л. *Системный анализ возможностей по извлечению именованных сущностей с применением технологии Text Mining* // Перспективы науки.– 2019.– Т. **120**.– № 9.– С. 18–20. ↑60

[3] Berlingerio M., Coscia M., Giannotti F., Monreale A., Pedreschi D. *Multidimensional networks: foundations of structural analysis* // World Wide Web.– 2013.– Vol. **16**.– Pp. 567–593. ↑60

[4] Клабукова Д. Л., Давыдовская М. В. *Внедрение международной терминологической базы MedDRA в практику фармаконадзора в Российской Федерации* // Московская медицина.– 2020.– Т. **35**.– № 1.– С. 64–69. ↑61

[5] Кузьмин А. Г., Умаров М. Ф. *Интеграция современных медицинских информационных технологий* // Вестник Вологодского государственного университета. Серия: Технические науки.– 2021.– Т. **12**.– № 2.– С. 32–35. ↑62

[6] Зацман И. М., Золотарев О. В., Хакимова А. Х., Дунсяо Гу *Модель и технология извлечения новых терминов из медицинских текстов* // Информ. и её примен..– 2022.– Т. **16**.– № 4.– С. 80–86. ↑62

[7] Mougin F., Grabar N. *Auditing the multiply-related concepts within the UMLS* // Journal of the American Medical Informatics Association.– 2014.– Vol. **21**.– No. e2.– Pp. e185–e193. ↑64

[8] Астанин П. А. *Применение автоматизированного анализа семантической сети UMLS для решения задачи поиска релевантных знаний о ревматических заболеваниях* // *Математическое моделирование систем и процессов*, Сборник материалов Международной научно-практической конференции (г. Псков, 9-11 ноября 2022 г.), Псков: Псковский государственный университет.– 2022.– ISBN 978-5-00200-102-6.– С. 6–12. ↑65

[9] Ямашкин С. А., Скворцов М. А., Большакова М. В., Ямашкин А. А. *Сравнительный анализ подходов к управлению базами данных для организации хранилища репозитория нейросетевых моделей* // Современные наукоемкие технологии.– 2021.– № 6-1.– С. 108–113. ↑66

[10] Елисеева Е.А., Горячкин Б.С., Виноградова М.В., Черненький М.В. *Оценка времени выполнения поисковых запросов в NoSQL и объектно-реляционной базах данных* // Динамика сложных систем – XXI век.– 2022.– Т. **23**.– № 2.– С. 44–51. ↑66

[11] Еремеев А. П., Мунтян Е. Р. *Разработка онтологии на основе графов с множественными и разнотипными связями* // Искусственный интеллект и принятие решений.– 2021.– № 3.– С. 3–18. ↑66

[12] Nicosia V., Latora V. *Measuring and modeling correlations in multiplex networks* // Physical Review E.– 2015.– Vol. **92**.– id. 032805. ↑67

[13] Battiston F., Nicosia V., Latora V. *Structural measures for multiplex networks* // Physical review E.– 2014.– Vol. **89**.– id. 032804. ↑67

[14] Szárnyas G., Kővári Z., Salánki 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* (Saint-malo, France, October 2–7, 2016), New York: ACM.– 2016.– ISBN 978-1-4503-4321-3.– Pp. 87–94. ↑67

[15] Nascimento M. C. V., Carvalho A. C. P. L. F. *A graph clustering algorithm based on a clustering coefficient for weighted graphs* // Journal of the Brazilian Computer Society.– 2011.– Vol. **17**.– Pp. 19–29. ↑68

[16] Пулькин И. С., Татаринцев А. В. *Достаточная статистика для параметра распределения Парето* // Russian Technological Journal.– 2021.– Т. **9**.– № 3.– С. 88–97. ↑69

[17] Синицын В. Ю., Кашпарова В. С. *Частотные свойства лексики научных текстов и законы Ципфа высших порядков* // Вестник РГТУ. Серия: Информатика. Информационная безопасность. Математика.– 2022.– № 4.– С. 75–91. ↑69

[18] Ghaderpour E., Morris D. *Cayley graphs on nilpotent groups with cyclic commutator subgroup are hamiltonian* // Ars Mathematica Contemporanea.– 2011.– Vol. **7**.– No. 1, Special Issue Bled'11.– Pp. 55–72. ↑72

[19] Каратеев А. Е. *Хроническая боль в спине как проявление остеоартрита позвоночника: обоснование и практика применения симптоматических средств замедленного действия* // Современная ревматология.– 2022.– Т. **16**.– № 4.– С. 88–97. ↑72

[20] Higuchi H., Harada T., Hiroshige J. *Evaluation of the usefulness of costovertebral angle tenderness in patients with suspected ureteral stone* // J. Gen. Fam. Med..– 2023.– Vol. **24**.– No. 1.– Pp. 56–58. ↑72

[21] Се Л., Ду Ч., Ин Ч., Вэй Я. *Острое расслоение аорты с правосторонней болью в грудной клетке и спине, сопровождающейся левосторонней дискинезией конечности* // Кардиология.– 2022.– Т. **62**.– № 6.– С. 74–76. ↑72

[22] Дуга С. В., Труфанов А. И. *Сеть знаний как концепция систем поддержки принятия решения в предварительном следствии* // Безопасность информационных технологий.– 2020.– Т. **27**.– № 3.– С. 54–65. ↑73

[23] Мосалов О. П. *Векторные представления рёбер графа онтологии как инструмент для анализа и генерации новых данных* // Информационно-технологический вестник.– 2021.– Т. **27**.– № 1.– С. 93–101. ↑73

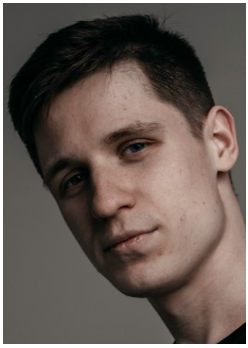
[24] Ананьева Е. А. *Уровни представления маршрутных пассажирских транспортных сетей в виде графовых моделей* // Colloquium-Journal.– 2019.– Т. **35**.– № 11-1.– С. 63–68. ↑73

[25] Близнякова Е. А., Куликов А. А., Куликов А. В. *Сравнительный анализ методов поиска кратчайшего пути в графе* // Архитектура, строительство, транспорт.– 2022.– № 1.– С. 80–87. ↑73

Поступила в редакцию 30.03.2023;
одобрена после рецензирования 05.05.2023;
принята к публикации 18.06.2023;
опубликована онлайн 07.10.2023.

Рекомендовал к публикации *к.т.н. Я. И. Гулиев*

Информация об авторах:



Павел Андреевич Астанин

аспирант кафедры медицинской кибернетики и информатики имени С. А. Гаспаряна, аналитик (лаборатория семантического анализа медицинской информации РНИМУ им. Н. И. Пирогова). Область интересов: обработка естественного языка (NLP), теория графов, базы данных, нечёткая логика, информационная поддержка принятия решений. Автор более 50 научных работ

0000-0002-1854-8686
e-mail: *med_cyber@mail.ru*

Светлана Евгеньевна Раузина

к. м. н., доцент, зав. лаб. семантического анализа медицинской информации (РНИМУ им. Н. И. Пирогова). Область интересов: медицинская информатика, проектирование МИС, разработка систем поддержки принятия решений. Автор более 50 научных работ, в том числе 1 учебник, 2 главы в монографиях, 4 сертификата и свидетельства на программные средства, 3 методических пособия

0000-0002-9535-2847
e-mail: *rauzina@mail.ru*

Татьяна Васильевна Зарубина

д. м. н., профессор, член-корр. РАН, главный внештатный специалист Минздрава России по информационным системам в здравоохранении. Автор более 270 научных работ, в том числе 2 учебника, 5 монографий, 10 сертификатов и свидетельств на программные средства, патенты, методические пособия, сборники под редакцией

0000-0002-4403-8049
e-mail: *t_zarubina@mail.ru*

Все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

EN RU PROGRAM SYSTEMS: THEORY AND APPLICATIONS ISSN 2079-3316
Research Article INFORMATION TECHNOLOGY IN MEDICINE

UDC 004.652.3, 616-079.4
10.25209/2079-3316-2023-14-3-59-94

Computing of umls concepts etiopathogenetic image using graph metrics

Pavel Andreevich **Astanin**[1✉], Svetlana Evgen'evna **Rauzina**[2], Tat'yana Vasil'evna **Zarubina**[3]
Pirogov Russian National Research Medical University, Moscow, Russia
[1✉]*med_ cyber@mail.ru*

Abstract. At present, the development of clinical decision support (CDS) tools is a crucial task in medical informatics. A lot of different information searching algorithms are used in CDS systems. A fundamental step in the design of these algorithms is the creation of an etiopathogenetic image for the analysis of unstructured medical texts. In this paper, we have conducted the literary review and a comparative evaluation of analytical metrics used to compute the etiopathogenetic image of concepts within the graph model of the Unified Medical Language System (UMLS) metathesaurus. Subsequently, we developed and validated our version of a graph metric suitable for the aforementioned task implementation.

Key words and phrases: hospital information system, information searching algorithms, knowledge base, graph theory, UMLS

2020 *Mathematics Subject Classification*: 68T30; 92C50

Acknowledgments: the current study has been performed within the framework of the Federal program «Priority 2030» based on the Healthcare Digital Transformation Institute (HDTI) in the Pirogov Russian National Research Medical University.

For citation: Pavel A. Astanin, Svetlana E. Rauzina, Tat'yana V. Zarubina. *Computing of umls concepts etiopathogenetic image using graph metrics*. Program Systems: Theory and Applications, 2023, **14**:3(58), pp. 59–94. https://psta.psiras.ru/read/psta2023_3_59-94.pdf

Introduction

The Unified medical language system (UMLS) metathesaurus is the biggest set of biomedical terms applicable to unstructured text analysis. The latest version of UMLS includes 4.6 million interdisciplinary concepts classified into 127 different thematic groups. Semantically close concepts are connected by relationships uniquely assigned into nine main (and two additional) types with 992 optional clarifying attributes. Almost each UMLS concept is connected with at least one concept. That is why this metathesaurus is able to be presented as oriented multigraph with about 98 million different relationships.

The approach to organizing data as graph informational models offers numerous advantages, including significant opportunities for optimizing analytical operations and the flexibility to choose various tools for data visualization. This approach enables the creation of user-friendly program modules for interpreting the structure of knowledge. However, currently, there is no unified method for the meaningful aggregation of data from graph informational models. The use of simple analytical options, such as filters on relationship types or thematic groups of concepts, does not yield clinically significant results. This challenge is attributed to the heterogeneity of the UMLS Metathesaurus structure and the lack of directed relationships between clinically relevant concepts. Therefore, it is crucial to develop ensembles of different analytical algorithms and combine them with complex systems of weight coefficients. Additionally, the utilization of optimized and validated tools and metamodels, which are unified sets of rules for the technical implementation of knowledge models, is essential.

One of the key elements of the weight system for UMLS can be values indicating the fuzzy membership degree of every concept in clinical branches (such as pulmonology, cardiology and neurology). These values are intended to demonstrate strength of etiologic and pathogenetic relationships between concepts and branches. The set of values described above forms an etiopathogenetic image, which is a vector with nonzero values indicating the degree of concept membership in clinical branches.

The creation of an etiopathogenetic image for UMLS concepts enables us to quantify their significance within a specific task context. This process ensures a reduction in the breadth of information search when

conducting graph queries. Calculating the distance between etiopathogenetic image vectors of different UMLS concepts allows us to rank symptoms and diseases based on their clinical proximity. This approach can be employed for automatically generating lists of diseases for differential diagnosis. Furthermore, it is important to compare various graph metrics and scientifically validate their usage in the aggregation of clinically relevant data from the UMLS graph model. Additionally, computing etiopathogenetic images for UMLS concepts enhances information retrieval algorithms by increasing the precision of mathematical contrasts.

The aim of this study is to compute and estimate etiopathogenetic image of UMLS metathesaurus concepts using metrics for graph nodes connectivity analysis.

1. Research phases

We have identified sequential key stages in our study. The first stage includes deploying the UMLS graph model and implementing Russian clinical thesauruses. The second stage involves creating semantic and logic rules for extracting atomic formulations. The third stage implies designing and technically implementing algorithms for calculating the membership degree function values for clinical branches of UMLS concepts. Finally, the fourth stage includes reviewing and comparing graph metrics, using nosological concepts as an example, and computing etiopathogenetic images. Additionally, during this stage, we developed a graph metric for evaluating the connectivity of UMLS concepts.

2. Graph model of UMLS

UMLS graph information model has been deployed using ECORE standard and all initially Russian translated concepts. They are presented in MedDRA (Medical Dictionary for Regulatory Activities Terminology), LOINC (Logical Observation Identifiers Names and Codes) and MeSH (Medical Subject Headings) terminological systems. MedDRA is a semantic tool for the majority of common clinical findings and symptoms coverage [4]. LOINC is a unified standard for different variants of medical tests describing [5]. MeSH includes vertical heading hierarchy and their synonyms for clinical articles labeling and rubricating. That is why MeSH is applicable for medical text analysis as atomic formulations thesaurus.

UMLS graph model with Russian translated concepts is significantly smaller than the original metathesaurus. In total, our graph model includes 144 thousand (3.1%) concepts and 7.9 million (8.0%) relationships between them. All concepts belong to 124 (97.6%) thematic types. The three missing concept thematic types are not clinically significant UMLS groups. They may never be applicable in describing clinical presentations or used for diagnostic tests in patients.

The matching of UMLS concepts with various clinical branches have been performed by calculating the connectivity of graph nodes. It has been based on a density analysis of relationships between non-nosological concepts and nosological concepts, which are uniquely mapped to ICD-10 (International Classification of Diseases, 10th Revision) codes in UMLS. In turn, ICD-10 has been mapped to its Russian translated version from the regulatory and reference information register of the Russian Ministry of Health.

3. Preparation of diseases list for data labeling

Early diagnostic tasks do not require the formulation of a medical diagnosis. Nevertheless, mapping symptoms to diagnoses requires the introduction of restrictions on the list of diseases. In the current research, a set of rules for filtering out irrelevant nosological concepts has been formulated. These concepts include atomic terms with lemmas such as "unspecified", "classified", "identified", "other", or terms that contain any diseases as substrings. For example, the code "K29" — "Gastritis and duodenitis" is redundant because there are codes in the Russian version of ICD-10, namely "K29.7" — "Gastritis" and "K29.8" — "Duodenitis", which are substrings of it.

Then, for additional filtering of nosological concepts, UMLS nodes matched with ICD-10 and interpreted as symptoms or signs but not as unique disorders by expert way have been removed from diseases list (for example, "K92.0" — "Hematemesis"). Only concepts with the highest count of directed relationships are used for ICD-10 codes mapped with more than one UMLS concept. These procedures have resulted in a reduction in the length of the diseases list from 11232 to 1577. These measures are important for reducing potential bias in the calculation of clinical branch membership degree values. UMLS nosological concepts mapped with ICD-10 codes have been automatically associated with their clinical branches, as shown in Table 1. There were not used in concepts labeling several ICD-10 classes:

Table 1. Clinical branches used in computing of etiopathogenetic image for UMLS concepts

№	CD-10 class (with codes diapasons)	Clinical branch name
1	V(F00-F99)	Psychiatry
2	VI(G00-G99)	Neurology
3	VII(H00-H59)	Ophthalmology
4	VIII(H60-H95)	Surdology
5	IX(I00-I99)	Cardiology
6	X(J00-J99)	Pulmonology
7	XI(K00-K93)	Gastroenterology
8	XII(L00-L99)	Dermatology
9	XIII(M00-M99)	Orthopedy
10	XIV(N00-N99)	Urology and sexology

I certain infectious and parasitic diseases,

II neoplasms,

III diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism,

IV endocrine, nutritional and metabolic diseases,

XV pregnancy, childbirth and the puerperium,

XVI certain conditions originating in the perinatal period,

XVII congenital malformations, deformations and chromosomal abnormalities,

XVIII symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified,

XIX injury, poisoning and certain other consequences of external causes,

XX external causes of morbidity and mortality,

XXI factors influencing health status and contact with health services and

XXII codes for special purposes.

4. Algorithm for UMLS concepts etiopathogenetic images computing

The computation of etiopathogenetic images for UMLS symptomatic concepts has been implemented using an iterative cycle that includes

Table 2. Brief semantical describing of basic UMLS relationships types

Relationship class	Type	An example of relationship	
		Root concept	Leaf concept
Associative	RO	Back pain	Lumbar sciatica
	RQ	Back pain	Discomfort in the back
	RL	Hyperuricemia	Pain in the toe
	AQ	Back pain	Diagnostic aspect
	QB	Diagnostic aspect	Back pain
Hierarchical	SIB	Back pain	Arthralgia
	PAR	Back pain	Dorsopathy
	CHD	Back pain	Lumbar back pain
	RB	Back pain	Pain
	RN	Back pain	Pain in upper part of back
Synonymous	SY	Back pain	Pain: back

several mathematical steps. In the first step, we have collected all graph nodes related to the targeted concept by considering SIB (sibling relationships), RO (other relationships, — largely, vertical relationships between non-synonymous concepts), CHD (child relationships), RN (narrower relationships, — vertical relationships between parent and child concepts), RQ (related and possibly synonymous relationships), and (SY (synonymous relationships). A brief description of UMLS relationship types is presented in Table 2.

It is important to note that hierarchical PAR (parent relationships) and RB (broader relationships — vertical relationships between child and parent concepts) connect clinical concepts with non-specific UMLS nodes. Use of them skews results of etiopathogenetic images computing. That is why those relationships types were not used in concepts searching. In turn, RL (*like* relationships — different horizontal and vertical relationships between semantically related concepts), AQ (allowed qualifier — technical vertical ascending relations) and (QB (can be qualified by — technical vertical descending relations) did not appear in Russian translated UMLS sources.

In the second step, we have separated concepts non-related to the "Disorders" semantic group in the UMLS network. Thematic types and their names including in the "Disorders" group are described in Table 3.

Table 3. UMLS thematic groups used for aggregating of relationships with ICD-10 nosological concepts

№	Thematic group code (tui)	Name of thematic group
1	T019	Congenital abnormality
2	T020	Acquired abnormality
3	T033	Finding
4	T037	Injury or poisoning
5	T046	Pathologic function
6	T047	Disease or syndrome
7	T048	Mental or behavioral dysfunction
8	T049	Cell or molecular dysfunction
9	T050	Experimental model of disease
10	T184	Sign or symptom
11	T190	Anatomical abnormality
12	T191	Neoplastic process

This limitation allows us to eliminate concept groups that are not related with diseases etiology and pathogenesis. These include economical, juridical, taxonomic and chronological concepts, geographic objects and specified biochemical terms.

In the third step, we have calculated count of graph path between root (symptomatic) and leaf (nosological) concepts related by not more than 2 other nodes mapped with groups in table 3. According to the results of our early studies, this graph path length is sufficient to achieve a relevant concept searching sensitivity of 90% [8].

We have collected nosological UMLS concepts (mapped to ICD-10) and calculated values of different graph metrics. Then their values have been summarized for each clinical branch and were normalized data between 0 and 1 range by using the feature scaling (min-max normalization). Maximal value 1 corresponded to the maximum proximity to the corresponding medical area, and zero corresponded to the minimum. All calculated values have been saved in database and were used for etiopathogenetic images computing for UMLS concepts then.

5. Estimation of etiopathogenetic images quality

In current study, estimation of graph metrics efficiency based on nodes connectivity analysis in etiopathogenetic images computing. Received labels for nosological concepts have been compared with ICD-10 class (actual labels). According to the research condition, the computed image is correct if ICD-10 label is present in top-3 clinical labels for corresponding disease in its etiopathogenetic image vector. It is important to note that images have been computed for non-mapped with ICD-10 concepts only. For concepts mapped to ICD-10 codes we summarized values of membership function for directly related clinical nodes in UMLS graph model. It provides the integral images estimation for nosological concepts to make conclusions about their applicability in multipurpose symptom checkers development.

Queries with union operations (cross-joins) that underlie node connectivity calculations have been performed by the graph DBMS Neo4j. The declarative query language Cypher is used for Neo4j interaction [9]. However, the object-relational DBMS PostgreSQL is more effective in handling specific queries. That is why this DBMS has also been used for data aggregation from UMLS.

Analytical procedures have been integrated into a single script using Python libraries: Psycopg2 (for executing queries in PostgreSQL DBMS), Neo4j (for executing queries in Neo4j DBMS), Scipy, Pandas, and Numpy (for data processing and statistical analysis).

6. Metrics for connectivity analysis of graph nodes

The Eclipse Modeling Framework (EMF) is a set of tools for unifying the development and deployment of metamodels. According to the Ecore standard used by EMF, there are two types of graph models based on node typification. A knowledge graph model is one-dimensional if there is no system for node differentiation by type. Otherwise, the graph model is multidimensional [11]. To clarify, the current version of the UMLS Metathesaurus contains 127 concept types. There are a set of rules, limitations, and assumptions that need to be considered in the abstract-logical analysis of multidimensional graph knowledge models. First and foremost, each semantic entity (atomic formulation) must be associated with the most appropriate graph node. Additionally, each graph node must be mapped to at least one type. Lastly, there must be at least one path between each pair of graph nodes.

Table 4. Graph metrics for subgraph nodes connectivity analysis

№	Original name	Using graph model	Values normalization
1	Clustering coefficient	monodimensional	+
2	Dimensional clustering coefficient	multidimensional	+
3	Node activity	multidimensional	–
4	Multiplex participation coefficient	multidimensional	+

Different one-dimensional and multidimensional graph metrics have been described for analyzing node connectivity in previous research [12]. However, it is important to note that multidimensional graph models can be interpreted as one-dimensional if the distribution of node types is not considered. The UMLS Metathesaurus fully complies with the Ecore requirements for multidimensional graph knowledge models. That is why we have reviewed graph metrics for both model classes in the current research.

One of the most significant attributes characterizing graph metrics is value normalization. As a rule, normalized metrics involve scaling values in the range from zero (minimum node connectivity) to one (maximum node connectivity) [13]. One distinct advantage of normalized metrics is that they can be compared and validated, making them more effective than metrics without value normalization. Additionally, normalized metrics can be integrated into a single expert ensemble to improve the quality of data aggregation. However, unnormalized metrics have practical significance due to their adaptability for specific graph model analyses. They are also applicable for modification and integration into analytical algorithms.

Nodes are the most significant knowledge sources in UMLS, while relationships are addiction elements that contain data about the semantic nature of concepts' connection. For this reason, there is a great interest in graph metrics for the estimation of node connectivity in graphs (Table 4) [14].

According to the Table 4, one of the nodes connectivity metrics for monodimensional graph models is clustering coefficient (CC). It provides values normalization [15] and is calculated using expression

$$\mathrm{CC}_n = \frac{2 \cdot C_n}{(R_n^{\mathrm{in}} + R_n^{\mathrm{out}}) \cdot (R_n^{\mathrm{in}} + R_n^{\mathrm{out}} - 1)} \tag{1}$$

In the Equation 1, C_n notes the count of graph contours included the node n; R_n^{in} is the count of relationships directed to the node n, and R_n^{out} means the count of relationships directed from node n.

Multidimensional graph metrics considered in the current study include dimensional clustering coefficient, node activity and multiplex participation coefficient.

The Dimensional Clustering Coefficient (DCC) is a generalized version of CC for one-dimensional models. It can be calculated using expression 1 in three different variants. The first variant involves using nodes whose type coincides with the root graph node. The second variant of DCC is the opposite of the first because it requires nodes whose type does not coincide with the root graph node. The third type of DCC has no limitations on the types of graph nodes.

The Node Activity (NA) is calculated as the count of unique types of nodes directly related to the root in the graph. It is important to note that NA is an unnormalized metric that does not take into account the frequency of occurrence of different node types.

The Multiplex Participation Coefficient (MPC) does not have any issues related to NA. It can be calculated using the expression

$$\text{MPC}_n = \frac{T}{T-1} \cdot \left[1 - \sum_{d \in D} \left(\frac{R_n(d) + R_n^{\text{out}}(d)}{R_n(D) + R_n^{\text{out}}(D)} \right)^2 \right]. \tag{2}$$

In the expression 2, T notes the count of unique types which include graph node n; $R_n(d)$ is the count of direct relationships between root node n and nodes that type coincides (d) with n; $R_n(D)$ means the count of direct relationships between root node n and any nodes regardless (D) of their types.

7. Empirical law "rank-frequency"

It should be noted that the UMLS semantic network, which includes clinical concepts, is heterogeneous. Given the presence of 127 types of nodes and 992 relationship attributes in UMLS, the most suitable approach is to use multidimensional normalized graph metrics, specifically the third variant of DCC and MPC. The application of these analytical tools for big data analysis has the potential to result in lengthy knowledge aggregation. Therefore, the development of graph tools optimization ways becomes an essential task to address before implementing knowledge bases in hospital

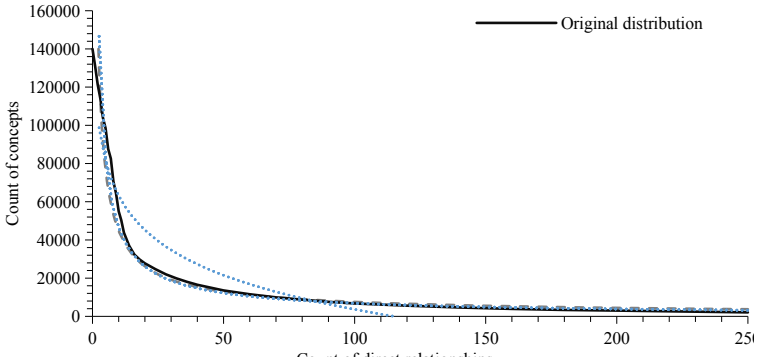


FIGURE 1. Approximation for UMLS concepts direct relationships count and count of nodes with no less neighbors

information systems. This entails the creation of simplified metrics for estimating graph node connectivity, which may involve counting direct relationships of the root node and the number of graph contours involving the root node.

We determined the number of direct relationships for each Russian-translated UMLS concept and constructed an approximation of the number of concepts that have such node neighbors in the graph model. This monotonically decreasing function follows a Pareto distribution [**16**]. According to Figure 1, it is possible to calculate the count of UMLS concepts with a direct relationship count no less than the corresponding value. The same distribution is characterized, as empirical Zipf's law applies to the frequency table of words in a text or corpus of natural language [**17**]. In the current study, we have discovered the application of this law in UMLS graph model analysis.

According to Figure 1, the observed empirical regularity is similar to a Pareto distribution. This pattern is also observed in Zipf's law, which is applicable to semantic models.

This mathematical pattern is statistically significant and highly determined ($R^2 = 0.939$, $p < 0.001$) by the following mathematical expression:

$$Y = 2 \cdot 10^6 \cdot X^{-1.290} \tag{3}$$

In the expression 3, X is the number of graph contours that the appropriate graph node is included in and Y is the actual value of the

concept significance degree. The R-Squared (R^2) for this empirical law is 0.939, indicating the high quality of our approximation function ($p < 0.001$).

8. Weighted clustering coefficient

The aforementioned empirical law can serve as the mathematical foundation for normalized graph metrics used in ranking UMLS concepts. These analytical tools take into account the density of the graph structure when measuring concept relevance. This is important for reducing bias in metric values when ranking graph nodes.

Using equation 3, we have created a mathematical expression for calculating the Weighted Clustering Coefficient (WCC) used in ranking concepts within the aggregated UMLS subgraph:

$$\mathrm{WCC}_n = \frac{C_n}{{R_n}^{1.290}} \tag{4}$$

In the expression 4, C_n is the count of subgraph contours or subgraph paths with node n, R_n is the count of direct relationships between node n and any graph nodes.

WCC (calculated with expression 4) values scaling is carried out with minimax normalization application:

$$\mathrm{WCC}'_n = \mathrm{WCC}_n - \frac{\min\ (\mathrm{WCC}\, N)}{\max(\mathrm{WCC}\, N) - \min\ (\mathrm{WCC}\, N)} \tag{5}$$

In the equation 5, WCC_n denotes the WCC value for node n; WCC_N is the set of WCC values for ranked set of nodes (N).

9. Kinds of graph contours and their usage in UMLS structure analysis

Graph theory maintains that graph contours are any closed geometric shapes, which implies that the root node is also a leaf node in these graph structures. Figure 2 illustrates various types of graph contours. Among these, the most mathematically balanced and computationally optimized are graph triangles. This type of contour is used in the CC and DCC graph metrics described earlier in this study. There are different types of contours that can be used in graph metrics calculations, depending on the number of nodes at each subgraph level. In this study, we have presented three types of graph contours and provided recommendations for their usage.

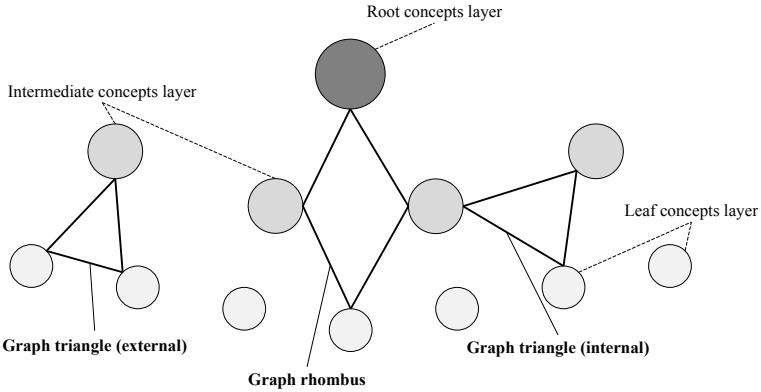


Figure 2. Kinds of geometrical contours

Specifically, it is advisable to aggregate external triangles (those with 1 vertex in the internal nodes layer and 2 vertices in the external layer) from the graph if the following condition is met:

$$N_{\text{internal}} < N_{\text{external}} \tag{6}$$

In the inequality 6, N_{internal} — count of nodes in internal layer (that is closer to root node), N_{external} — count of nodes in external layer (that is closer to leaf node). This condition is met for Figure 2 because in the intermediate subgraph layer, there are only four concepts, while there are seven in the terminal (external) layer.

In the opposite case, we recommend using the count of internal triangles (those with 1 vertex in the external node layer and 2 vertices in the internal layer) for graph metric calculations.

Counting the number of graph diamonds is advisable when the subgraph layer sizes are equal or when it is needed to shortest indirect paths searching between concepts. Utilizing variations of different graph contour types depending on the size of the extracted subgraph provides increasing of mathematical contrast of calculated values and ranking quality of concepts.

10. Quality estimation results for etiopathogenetic images of nosological UMLS concepts

Using metrics from Table 4 and empirically obtained WCC, we have computed an etiopathogenetic image for 134 thousand Russian translated symptomatic (93.9%) UMLS concepts. Other nodes are not connected with

Table 5. Comparison of nosological UMLS concepts classification results ($n = 1577$)

Type of graph metric	Types of graph contours		P
	hamiltonian	non-hamiltonian	
Clustering coefficient (CC)	63 ± 1.2	64 ± 1.2	0.361
Dimensional clustering coefficient (DCC)	72 ± 2.2	74 ± 2.2	0.297
Node activity (NA)	24 ± 2.4	24 ± 2.4	0.965
Multiplex participation coefficient (MPC)	84 ± 1.6	88 ± 1.6	0.024
Weighted clustering coefficient (WCC)	86 ± 1.7	91 ± 1.4	0.009

ICD-10 concepts by direct relationships and indirect paths with length less than 4.

Then we have estimated the quality of etiopathogenetic images calculated by different graph metrics usage for nosological concepts (Table 5).

There are percentages ($p \pm CI95\%$) of concepts for which the basic clinical label (ICD-10 class) is present in the list of the three most inherent classes in the etiopathogenetic image vector presented in Table 5. According to Table 5, the best results have been identified using the empirically obtained WCC. The classification quality metric reached $86 \pm 1.7\%$ when only Hamiltonian graph contours (with non-repeating edges) were used [18]. The use of non-Hamiltonian graph contours (allowing for the possibility of repeated edges in the contour) statistically significantly improved the classification quality to $91 \pm 1.4\%$ (p=0.009).

Clinical interpretation of the etiopathogenetic images for UMLS concepts can be demonstrated using the following example. There is a beforehand computed etiopathogenetic image vector calculated for symptomatic concept "back pain" using WCC in our database: ["Orthopedy": 1.000, "Neurology": 0.921, "Cardiology": 0.652, "Urology": 0.560]. Thus, the specified concept is most strongly related with musculoskeletal pathology (1.000), neurological disorders (0.921) and a little poor with cardiovascular pathologies (0.652) and urogenital diseases (0.560). Based on this we can conclude that the calculated image aligns with contemporary data on the back pain etiopathogenesis [19–21]. Back pain association with orthopedic and neurology may be explained by the presence of this symptom in patients with nonspecific (degenerative, dysfunctional, or dystrophic) changes

in the musculoskeletal system. Also back pain may be associated with the secondary damage to some structures of the peripheral nervous system. The association with cardiovascular diseases may be explained by back pain presence in patients with acute myocardial infarction when it can refer to the back under the left scapula or between the shoulder blades. Finally, back pain may be related to some kidney diseases, which are manifested by this symptom. That is why the etiopathogenetic image automatically calculated for this UMLS concept is quite correct.

Conclusion

The lack of knowledge about the possibility of using graph information models in clinical decision support systems (CDSS) is an actual issue [22–25]. UMLS Metathesaurus integration into CDSS requires the application of graph metrics for concepts significance estimation in the specific task context. Clinically relevant UMLS knowledge extraction may become feasible with the development of specialized analytical metrics and automatically calculated weight coefficients system.

In the current study, we have estimated graph metrics in computing of etiopathogenetic images for UMLS concepts. The acceptable results have been taken by the dimensional coefficient and the multiplex participation coefficient described in early studies by other research teams. The percentages of relevant clinical labels for nosological concepts were 72–74% and 84–88%, respectively. The best classification quality has been demonstrated by the weighted clustering coefficient (86–91%), based on the use of an empirically obtained density distribution function of nodes direct relationships. That is why the calculation of non-Hamiltonian graph triangles included root node may be used to relate concepts to the corresponding clinical branches with high quality. Based on the results of multi-class ICD-10 concepts classification, we have concluded that the weighted clustering coefficient could be used for UMLS concepts marking and in particular etiopathogenetic images computing. In the future, it is anticipated that these images will be used for differential diagnosis of diseases with the use of named entity recognition. Text mapping with UMLS clinical concepts provides generation of disease semantic profiles applicable to their integration into CDSS intended for relevant data aggregation.

References

[1] P. A. Astanin, L. V. Ronzhin, S. E. Rauzina, T. V. Zarubina. "Semantic analysis algorithms for data processing and possibilities of their usage in medical information systems development", *Cifrovaya statistika. Novye zadachi i traektoriya dvizheniya — 2022*, Materialy IV S"ezda medicinskix statistikov Moskvy (21–23 sentyabrya 2022 goda), 2022, pp. 6–9 (in Russian). ↑

[2] V. V. Kukarcev, Z. A. Kolmakova, O. L. Mel'nikova. "System analysis of possibilities to retrieve essentials using text mining technology", *Perspektivy nauki*, **120**:9 (2019), pp. 18–20 (in Russian). ↑

[3] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, Pedreschi D. . "Multidimensional networks: foundations of structural analysis", *World Wide Web*, **16** (2013), pp. 567–593. ↑

[4] D. L. Klabukova, M. V. Davydovskaya. "Implementation of the MedDRA international terminology base into the pharmacovigilance practice Russia", *Moskovskaya medicina*, **35**:1 (2020), pp. 64–69 (in Russian). ↑79

[5] A. G. Kuz'min, M. F. Umarov. "Integration of modern medical information technologies", *Vestnik Vologodskogo gosudarstvennogo universiteta. Seriya: Texnicheskie nauki*, **12**:2 (2021), pp. 32–35 (in Russian). ↑79

[6] I. M. Zaczman, O. V. Zolotarev, A. X. Xakimova, Gu Dunsyao. "Model and technology for discovering new terms in medical texts", *Inform. i eyo primen.*, **16**:4 (2022), pp. 80–86 (in Russian). ↑

[7] F. Mougin, N. Grabar. "Auditing the multiply-related concepts within the UMLS", *Journal of the American Medical Informatics Association*, **21**:e2 (2014), pp. e185–e193. ↑

[8] P. A. Astanin. "Application of automated analysis of the UMLS semantic network to solve the problem of searching for current knowledge about rheumatic diseases", *Matematicheskoe modelirovanie sistem i processov*, Sbornik materialov Mezhdunarodnoj nauchno-prakticheskoj konferencii (g. Pskov, 9-11 noyabrya 2022 g.), Pskovskij gosudarstvennyj universitet, Pskov, 2022, ISBN 978-5-00200-102-6, pp. 6–12 (in Russian). ↑83

[9] S. A. Yamashkin, M. A. Skvorczov, M. V. Bol'shakova, A. A. Yamashkin. "Comparative analysis of approaches to database management for organizing a repository of neural network models", *Sovremennye naukoemkie texnologii*, 2021, no. 6-1, pp. 108–113 (in Russian). ↑84

[10] Eliseeva E.A. , Goryachkin B.S. , M.V. Vinogradova, M.V. Chernen'kij. "Estimating search execution time in NoSQL and object-relational databases", *Dinamika slozhnyx sistem — XXI vek*, **23**:2 (2022), pp. 44–51 (in Russian). ↑

[11] A. P. Eremeev, E. R. Muntyan. "Development of an ontology based on graphs with multiple edges of different types", *Iskusstvennyj intellekt i prinyatie reshenij*, 2021, no. 3, pp. 3–18 (in Russian). ↑84

[12] V. Nicosia, V. Latora. "Measuring and modeling correlations in multiplex networks", *Physical Review E*, **92** (2015), id. 032805. ↑85

[13] F. Battiston, V. Nicosia, V. Latora. "Structural measures for multiplex networks", *Physical review E*, **89** (2014), id. 032804. ↑85

[14] G. Szárnyas, Z. Kővári, A. Salánki, D. Varro. "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* (Saint-malo, France, October 2–7, 2016), ACM, New York, 2016, ISBN 978-1-4503-4321-3, pp. 87–94. ↑85

[15] M. C. V. Nascimento, A. C. P. L. F. Carvalho. "A graph clustering algorithm based on a clustering coefficient for weighted graphs", *Journal of the Brazilian Computer Society*, **17** (2011), pp. 19–29. ↑85

[16] I. S. Pul'kin, A. V. Tatarincev. "Sufficient statistics for the Pareto distribution parameter.", *Russian Technological Journal*, **9**:3 (2021), pp. 88–97 (in Russian). ↑87

[17] V. Yu. Sinicyn, V. S. Kashparova. "Frequency properties of the lexis of scientific texts and Zipf's laws of higher orders", *Vestnik RGTU. Seriya: Informatika. Informacionnaya bezopasnost'. Matematika*, 2022, no. 4, pp. 75–91 (in Russian). ↑87

[18] E. Ghaderpour, D. Morris. "Cayley graphs on nilpotent groups with cyclic commutator subgroup are hamiltonian", *Ars Mathematica Contemporanea*, **7**:1, Special Issue Bled'11 (2011), pp. 55–72. ↑90

[19] A. E. Karateev. "Chronic back pain as a spinal osteoarthritis manifestation: rationale and practice of symptomatic slow acting drugs for osteoarthritis use", *Sovremennaya revmatologiya*, **16**:4 (2022), pp. 88–97 (in Russian). ↑90

[20] H. Higuchi, T. Harada, J. Hiroshige. "Evaluation of the usefulness of costovertebral angle tenderness in patients with suspected ureteral stone", *J. Gen. Fam. Med.*, **24**:1 (2023), pp. 56–58. ↑90

[21] L. Se, Ch. Du, Ch. In, Ya. Vej. "Acute aortic dissection with right-sided chest and back pain accompanied by left-sided limb dyskinesia", *Kardiologiya*, **62**:6 (2022), pp. 74–76 (in Russian). ↑90

[22] S. V. Duga, A. I. Trufanov. "The knowledge graph concept of decision support system in preliminary investigation", *Bezopasnost' informacionnyx texnologij*, **27**:3 (2020), pp. 54–65 (in Russian). ↑91

[23] O. P. Mosalov. "Edge embedding of ontology graphs as a tool for analysis and generation of new data", *Informacionno-texnologicheskij vestnik*, **27**:1 (2021), pp. 93–101 (in Russian). ↑91

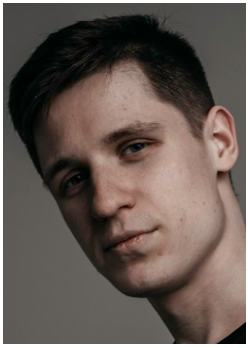
[24] E. A. Anan'eva. "Levels of presentation of passenger transport route networks in the form of graph models", *Colloquium-Journal*, **35**:11-1 (2019), pp. 63–68 (in Russian). ↑91

[25] E. A. Bliznyakova, A. A. Kulikov, A. V. Kulikov. "Comparative analysis of methods for finding the shortest distance in a graph", *Arxitektura, stroitel'stvo, transport*, 2022, no. 1, pp. 80–87 (in Russian). ↑91

Received 30.03.2023;
approved after reviewing 05.05.2023;
accepted for publication 18.06.2023;
published online 07.10.2023.

Recommended by *PhD. Ya. I. Guliev*

Information about the authors:



Pavel Andreevich Astanin

postgraduate student of the medical cybernetics and informatics department named after S. A. Gasparyan, data analyst of the Laboratory for semantic analysis of medical information, (HDTI, RNRMU). Research interests: natural language processing (NLP), graph theory, databases, fuzzy logic, CDSS development. An author of more than 45 scientific papers

0000-0002-1854-8686

e-mail: ***med_cyber@mail.ru***

Svetlana Evgen'evna Rauzina

phD, docent of the medical cybernetics and informatics department named after S. A. Gasparyan, Head of the Laboratory for semantic analysis of medical information (HDTI, RNRMU). Research interests: medical informatics, HIS designing, CDSS development. An author of more than 50 scientific papers includes 1 classbook, 2 monography chapters, 4 certificates confirming the result of intellectual activity, 3 toolkits

0000-0002-9535-2847

e-mail: *rauzina@mail.ru*

Tat'yana Vasil'evna Zarubina

phD, professor, head of the medical cybernetics and informatics department named after S. A. Gasparyan, member of RAS, head of HDTI, chief specialist of the Russian Ministry of Health on CDSS. An author of more than 270 scientific papers includes 2 classbook, 5 monographs, 10 certificates confirming the result of intellectual activity

0000-0002-4403-8049

e-mail: *t_zarubina@mail.ru*

The authors declare no conflicts of interests.