

# ПРИМЕНЕНИЯ СЕМАНТИЧЕСКОЙ СЕТИ ДЛЯ ПОИСКА НОВЫХ ТЕРМИНОВ И ЗАКОНОМЕРНОСТЕЙ В МНОГОМОДАЛЬНЫХ ДАННЫХ АИС НА БАЗЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

## APPLICATION OF SEMANTIC NETWORK FOR SEARCH OF NEW TERMS AND PATTERNS IN MULTIMODAL AIS DATA BASED ON ARTIFICIAL INTELLIGENCE

V. Yurchak

*Summary.* With the introduction of artificial intelligence into the analysis and processing of multimodal data, it has become possible to extract information from unstructured texts, to perform automatic clustering, Categorization and classification of information with subsequent use in various scientific and production sectors. Automated analysis allowed researchers not only to collect and study the amount of material that could not be analyzed manually, but also to identify patterns and relationships between different terms. Automated analysis based on international computer systems, The dictionary and database set itself the task of searching for new concepts and terms in already analyzed multidimensional data sets. This article addresses the problem of automated analysis of multimodal data, namely: identification of common terms and concepts at the junction of one or more subject areas (thematic clusters). In particular, the study describes approaches to building a semantic network at the intersection of several subject areas (thematic clusters) based on the terms of the international knowledge base of medical publications (PubMed) and the international dictionary of medical terms (UMLS). The approaches to building a semantic network at the intersection of several subject areas (thematic clusters) considered in this article include mechanisms for forming thematic clusters, The identification of semantic links between terms and the visualization of new terms. The result described in this article is the creation of a semantic network (graph) «Diseases and their mutations» on the basis of PubMed libraries (PMC) and UMLS, which will help to define new terms for disease mutations and understand the reasons why they occur.

*Keywords:* semantic network(graph), multimodal data, automated analysis, UMLS, PubMed, BERTopic, thematic clusters.

Юрчак Владимир Александрович

Аспирант, Российский новый университет (РосНОУ)  
yurchak777x@gmail.com

*Аннотация.* С внедрением искусственного интеллекта в процессы анализа и обработки многомодальных данных стало возможно извлекать информацию из неструктурированных текстов, проводить автоматическую кластеризацию, категоризацию и классификацию информации с последующим ее использованием в различных научных и производственных секторах. Автоматизированный анализ позволил исследователям не только собирать и изучать объем материала, анализ которого вручную невозможен, но и выявлять закономерности и взаимосвязи между различными терминами. Автоматизированный анализ на основе международных компьютерных систем, словарей и баз знаний поставил перед собой задачу поиска новых понятий и терминов в уже проанализированных массивах многомерных данных. В рамках данной статьи рассматривается проблема автоматизированного анализа многомодальных данных, а именно: выделения общих терминов и понятий на стыке одной или нескольких предметных областей (тематических кластеров). В частности, в исследовании описываются подходы к построению семантической сети на стыке нескольких предметных областей (тематических кластеров) на базе терминов международной базы знаний медицинских публикаций (PubMed) и международного словаря медицинских терминов (UMLS). Рассматриваемые в рамках данной статьи подходы к построению семантической сети на стыке нескольких предметных областей (тематических кластеров) включают механизмы формирования тематических кластеров, выделение семантических связей между терминами и визуализацию новых терминов. Результатом, описанным в рамках данной статьи, становится создание семантической сети(графа) «Заболеваний и их мутаций» на базе библиотек PubMed (PMC) и UMLS, которая позволит определять новые термины мутаций заболеваний и понимать причины их появления.

*Ключевые слова:* семантическая сеть (граф), многомодальные данные, автоматизированный анализ, UMLS, PubMed, BERTopic, тематические кластеры.

### Введение

В последнее время наблюдается значительный рост объема текстовой информации. Этот огромный объем данных представляется в различных формах, таких как записи в социальных сетях, записи осмотра пациентов, данные медицинского страхования, статьи новостных агентств, отчеты о работе технических устройств. Весь этот огромный объем данных формиру-

ет целые кластеры, содержащие информацию из различных источников таких как: базы данных, озера данных (DataLake) и данные Dashboards. Вместе все эти огромные объемы информации составляют многомодальные данные.

*Определение 1.* Многомодальные данные — это данные, которые представлены в нескольких модальностях: текст, изображения, аудио и видео. Слияние многомо-

дальних данных — это синергетическое использование информации, полученной из разных модальностей (каналов взаимодействия) [6].

С внедрением искусственного интеллекта в процессы анализа и обработки многомодальных данных стало возможно извлекать информацию из неструктурированных текстов, проводить автоматическую кластеризацию, категоризацию и классификацию информации с последующим ее использованием в различных научных и производственных секторах [16].

Автоматизированный анализ позволил исследователям не только собирать и изучать объем материала, анализ которого вручную невозможен, но и выявлять закономерности и взаимосвязи между различными терминами.

В настоящее время существует проблема выделения общих терминов и понятий на стыке одной или нескольких предметных областей (тематических кластеров). Автоматизированный анализ на основе международных компьютерных систем, словарей и баз знаний ставит перед собой задачу поиска новых понятий и терминов в уже проанализированных массивах многомерных данных. Целями автоматизированного анализа является нахождение закономерной и связей новых терминов с существующими понятиями, содержащими в международных компьютерных системах и базах знаний таких как: UMLS, PubMed, Babelnet, DBpedia, WordNet и других [7–8].

В рамках данной статьи будут освещены основные подходы к поиску новых понятий и терминов в международной базе знаний медицинских публикаций PubMed и международном словаре медицинских терминов UMLS. Главными задачами, описанными в рамках исследования, будут:

- механизм формирования тематических кластеров в распознанных и размеченных многомерных данных [1];
- механизм выделения семантических связей терминов в рамках нескольких тематических кластеров;
- визуализация новых терминов и понятий в рамках нескольких тематических кластеров.

Результатами исследования, описанными в данной статье, служат создание семантической сети(графа) «Заболеваний и их мутаций» на базе библиотек PubMed (PMC) и UMLS. Использование международной базы знаний медицинских публикаций PubMed и международного словаря медицинских терминов UMLS в качестве базового набора данных может служить ключом для поиска новых заболеваний, помогая исследователям проследить динамику изменения взаимосвязей между

заболеваниями. Данное исследование играет важную роль в определении новых мутаций заболеваний и понимания причин их появления [5].

Статья организована следующим образом. В разделе 1 содержится информация о методах извлечения многомодальных данных из автоматизированной информационной системы на базе искусственного интеллекта.

В разделе 2 описан подход к формированию общих тематических кластеров на примере международной базы знаний медицинских публикаций PubMed и международного словаря медицинских терминов UMLS.

В разделе 3 приводится информация о методе выделения семантических связей терминов на примере автоматизированной информационной системы на базе искусственного интеллекта, использующей в исследовании набор из международной базы знаний медицинских публикаций PubMed и международного словаря медицинских терминов UMLS.

Результаты исследования и их обсуждение приводятся в разделе 4 Заключение.

### 1. Методы извлечения многомодальных данных из автоматизированной информационной системы на базе искусственного интеллекта

Автоматизированная информационная система на базе искусственного интеллекта основана на обращении к электронным ресурсам для хранения структурированной информации, включая научные публикации, обзоры научных исследований и т.д.

В рамках данного функционала ИС выполняется обращение к международной базе медицинских публикаций PubMed и международному словарю UMLS, которые позволяют хранить научные статьи со следующими полями: тип публикации, заголовок, авторы, аннотация, ключевые слова, год публикации, разделы документа. Для работы с этой информацией используется база данных Sqlite, в которой были созданы соответствующие таблицы с полями для хранения данных. Выгрузка данных из библиотек UMLS и PubMed (PMC) ведется с использованием Entrez утилит [7–8]. В информационной системе формируется запрос, который далее направляется PubMed (PMC). Затем в информационной системе формируется запрос, который далее направляется UMLS. В результате выполнения поискового запроса на сервере в базе данных формируется структурированный по разделам корпус документов. Также автоматизированная информационная система предоставляет функционал построения семантической сети в виде структурированного графа, формируемого на основе медицинских терминов из международного

словаря UMLS и распознанных разделов корпуса текста каждой публикации PubMed. Медицинский термин в паре с идентификатором статьи задают topic (тематику) публикации [7–8]. В рамках одной публикации алгоритмы искусственного интеллекта позволяют выделить несколько topic(тем). Благодаря алгоритмам кластеризации из нескольких topic формируется один общий тематический кластер, который вместе с номером публикации в международной базе данных и мерой косинусного сходства отображается на узловых точках графа [13]. Построенные в виде семантического графа topic по каждой публикации сравниваются с терминами, полученными из библиотек UMLS и Pubmed (PMC) с целью поиска потенциально новых терминов в разделах публикации. Динамика изменения медицинских терминов отражается в семантической сети, представленной в виде графа «Заболеваний и их мутаций» на базе библиотек PubMed (PMC) и UMLS для определения потенциально новых заболеваний и причин их появления [7–8].

В рамках функционала автоматизированной информационной системы на базе искусственного интеллекта реализован метод извлечения медицинских терминов из публикаций PubMed и словаря UMLS. Целью предлагаемого метода извлечения терминов является изучение массива терминов заболеваний из корпуса текстов. Перед извлечением терминов корпус текстов подвергается предварительной обработке.

#### *А. Формирование корпуса документов:*

Выгрузка данных из библиотек PubMed(PMC) и UMLS осуществляется с помощью Entrez утилит. В автоматизированной информационной системе формируется запрос, который затем отправляется в PubMed(PMC) и UMLS. В результате поискового запроса на сервере формируется корпус документов, загруженных из библиотеки PubMed и список размеченных медицинских терминов из библиотеки UMLS. Каждый документ включает название статьи, информацию об авторе, тип документа, дату публикации и аннотацию [7–8].

#### *Б. Предварительная обработка:*

В процедуре извлечения терминов из медицинских текстов в автоматизированной информационной системе на базе искусственного интеллекта используется стандартный набор стоп-слов (включал только союзы, междометия и т. д). Для борьбы с проблемой наличия «шума» (большого количества слов или словосочетаний из немедицинских терминов) расширяется набор стоп-слов. Данное решение позволяет снизить количество «шума» при извлечении медицинских терминов из научных документов посредством автоматической обработки [7–8].

#### *В. Извлечение термина:*

Извлечение терминов в автоматизированной информационной системе на базе искусственного интеллекта происходит с использованием следующих грамматических категорий: простое существительное, существительные во множественном числе, прилагательное, герундий, иностранное слово. В результате обработки публикаций корпуса текстов идентифицируют термины в виде n-грамм, выполняется лемматизация (приведение слов текста к нормальной форме) [7–8].

### **2. Подход к формированию общих тематических кластеров на примере международной базы знаний медицинских публикаций PubMed**

В основе выделения тематических кластеров лежат методы кластеризации многомодальных данных. Под кластеризацией многомодальных данных понимают процесс группировки информации по их семантической схожести. Любое сходство может быть оценено численно, например, расстояние между терминами из двух слов. Кластеры представляют собой не только набор значений объектов с числовым сходством, но и группу объектов, которые представляют одно и то же понятие друг для друга. Высокая эффективность и точность кластеров данных — две основные и важные цели кластеризации [3].

В рамках функционала автоматизированной информационной системы искусственного интеллекта в разделе построения семантической сети, представленной в виде графа

«Заболеваний и их мутаций» на базе библиотек PubMed(PMC) и UMLS используется перекрывающаяся и нечеткая кластеризация медицинских терминов.

Перекрывающаяся и нечеткая кластеризация позволяет определить принадлежность термина к нескольким тематическим группам, находящимся между двумя кластерами. Кроме того, каждый медицинский термин относится к кластеру с весом (вероятностью) от 0 до 1, где 0 — абсолютно не принадлежит, 1 — полностью принадлежит [14].

В рамках функционала информационной системы искусственного интеллекта в разделе построения семантической сети, представленной в виде графа «Заболеваний и их мутаций» на базе библиотек PubMed(PMC) и UMLS используется метод тематического моделирования BERTopic [14].

Метод тематического моделирования BERTopic использует возможности языковой модели BERT, основанной на архитектуре трансформер. Подобно Top2Vec,

BERTopic использует UMAP для уменьшения размеров и HDBSCAN для кластеризации многомодальных данных. Встраивание слов создается на основе следующих моделей, а именно: LaBSE, парафраза-MiniLM-L3-v2, all-MiniLM-L6-v2, all-MiniLM-L12-v2, all-mpnet-base-v2 [14].

Метод тематического моделирования BERTopic в отличие от алгоритма LDA (Латентного размещения Дирихле), предоставляет форму непрерывного тематического моделирования, а не дискретного. После расчета модели исследователи извлекают наиболее важные темы [14].

Метод тематического моделирования BERTopic позволяет выполнять поиск по конкретным ключевым словам и получать наиболее релевантные темы на основе их показателей сходства. Они также могут углубляться в отдельные темы, изучая связанные с ними ключевые слова [14].

В рамках функционала информационной системы искусственного интеллекта в разделе построения семантической сети, представленной в виде графа» Заболеваний

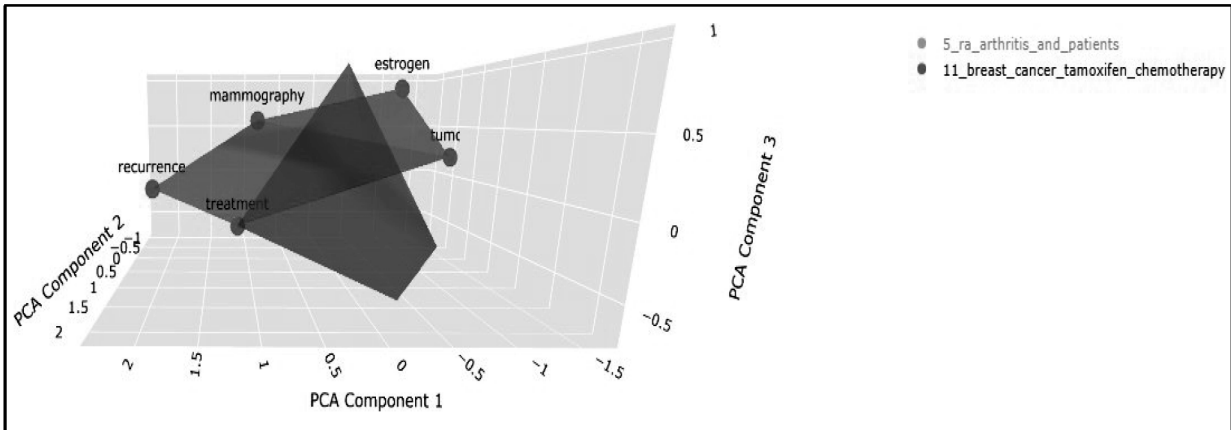


Рис. 1. Результаты визуализации медицинских терминов: интерактивная карта межтемных расстояний

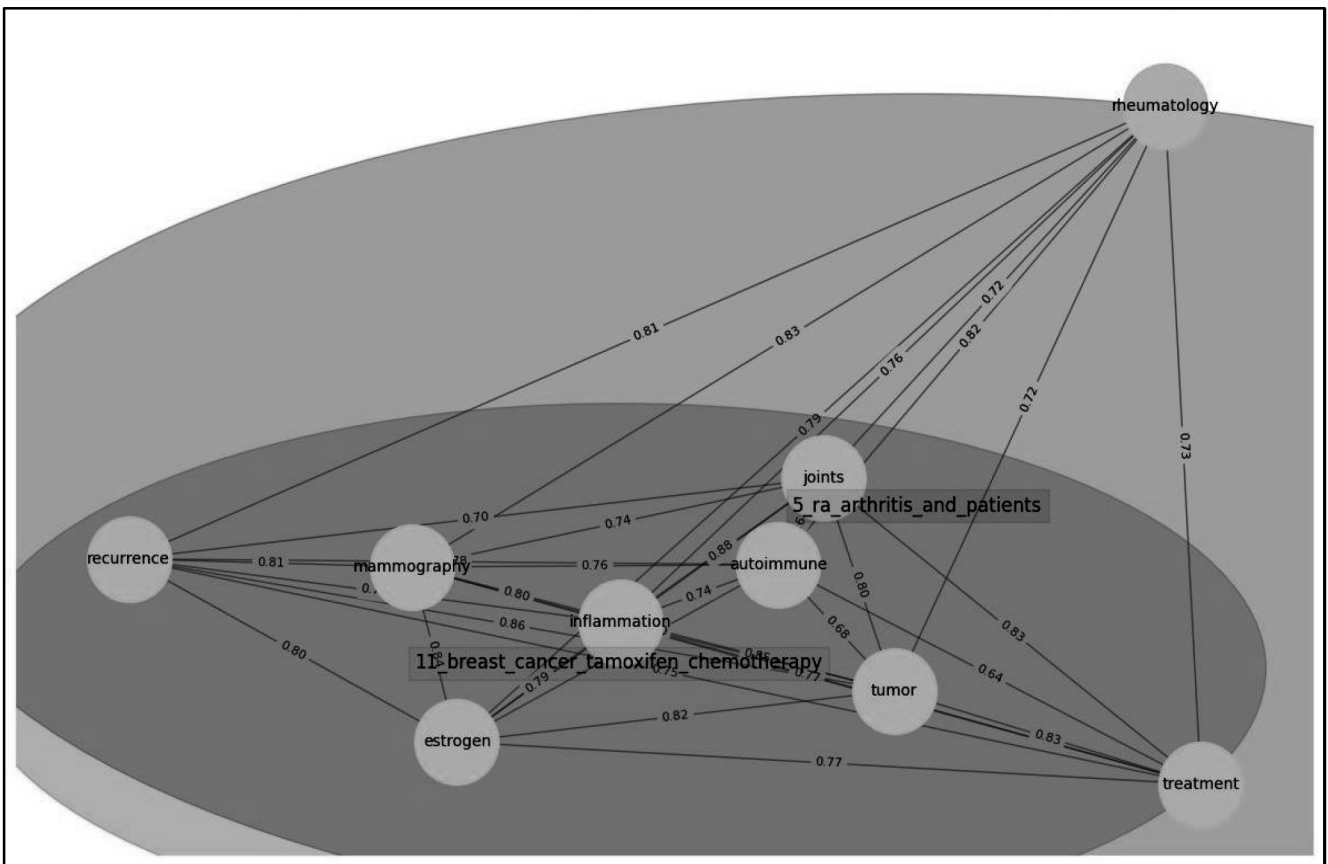


Рис. 2. Результаты визуализации медицинских терминов: тематические кластеры с автоматическим размещением на основе семантического сходства медицинских терминов



index	Термины	Термины из тематического кластера "Arthritis and patients"					Термины из тематического кластера "Breast cancer tamoxifen chemotherapy"					Термины
		inflammation	joints	autoimmune	rheumatology	treatment	mammography	tumor	estrogen	treatment	recurrence	
2							0.87	0.90	0.84	0.88	0.88	inflammation
3							0.84	0.88	0.83	0.87	0.85	joints
4							0.79	0.83	0.86	0.76	0.86	autoimmune
5							0.91	0.86	0.87	0.87	0.87	rheumatology
6							0.87	0.86	0.80	1.00	0.84	treatment
7	mammography	0.87	0.84	0.79	0.91	0.87						
8	tumor	0.90	0.88	0.83	0.86	0.86						
9	estrogen	0.84	0.83	0.86	0.87	0.80						
10	treatment	0.88	0.87	0.76	0.87	1.00						
11	recurrence	0.88	0.85	0.86	0.87	0.84						

Рис. 3. Результаты по семантическому схождению медицинских терминов 2-ух тематических кластеров

и их мутаций» на базе библиотек PubMed(PMC) и UMLS метод тематического моделирования BERTopic предоставил интерактивную карту межтемных расстояний по каждому медицинскому. Результаты визуализации медицинских терминов представлены на рис. 1 и рис. 2.

Представленные выше рисунки отображают пересечение двух различных тематических кластеров, построенных на основе метода BERTopic. Пересечение тематических кластеров происходит в узловых точках терминов [14]. Каждый термин является ключевым понятием в рамках тематического кластера. Однако при наложении терминов в рамках 2 и более тематических кластеров ключевых понятий отображается числовая мера приближения, выраженная в вероятности семантического схождения данного термина из 1-го тематического кластера к тому же термину из 2-го тематического кластера. Результаты по семантическому схождению медицинских терминов представлены на рис. 3.

В процессе построения интерактивной карты межтемных расстояний по каждому медицинскому было выявлено, что ключевую роль при построении ассоциативных связей между терминами на основе многомодальных данных играют отношения синонимии и определения, которые позволяют проецировать известные свойства на новые сущности.

**3. Метод выделения семантических связей терминов на примере международной базы знаний медицинских публикаций PubMed и словаря медицинских терминов UMLS**

Концепция семантических связей (сетей) была впервые предложена Тимом Бернерсом-Ли, основателем Всемирной паутины, в 2001 году. Согласно концепции Тима Бернерсома-Ли, семантическая связь (сеть) является расширением и дополнением всемирной паутины, которая обеспечивает взаимосвязь данных в Интернете таким образом, что их семантика может быть автоматически понятна машинам.

*Определение 1.* Семантическая связь в основном предоставляет механизм для семантического представления информации, конечной целью которого является реализация совместного использования информации и семантической совместимости [12].

Семантическая сеть опирается на онтологии, чтобы управлять обменом и повторным использованием знаний.

*Определение 2.* Онтология — это стандартная концептуальная система в семантической сети, удобная для компьютеров. Онтология обеспечивает основную технологию построения взаимосвязей объектов и является ключом к реализации семантической сети [4].

В данном разделе описывается метод построения семантической сети, представленной в виде графа «Заболеваний и их мутаций» на базе библиотек PubMed(PMC) и UMLS, которая включает в себя онтологию заболевания, онтологию мутаций заболеваний и связи между ними, помогая исследователям проследить динамику изменения взаимосвязей между заболеваниями. Данное исследование играет важную роль в определении новых мутаций заболеваний и понимания причин их появления.

Метод создания семантической сети, представленной в виде графа «Заболеваний и их мутаций» на базе библиотек PubMed(PMC) и UMLS предполагает интеграцию онтологии заболевания и онтологии мутации заболеваний. Процесс создания онтологий мутаций заболеваний проходит следующие этапы:

А. Сбор корпуса текстов: получение корпуса текстов о мутациях заболеваний из авторитетных баз знаний в области медицины, таких как PubMed(PMC) и UMLS.

Б. Определение словарного запаса мутаций заболеваний: Выявление потенциальных слов для включения в словарь, описывающих мутации заболеваний.

В. Вычисление семантического сходства: Вычисление показателей семантического сходства между словами-мутаций заболеваний в наборе многомодальных данных для количественной оценки их концептуального родства.

Г. Формирование онтологии симптомов: Слияние слов-мутаций заболеваний со значительным семантическим сходством для создания уточненной онтологии мутаций заболеваний.

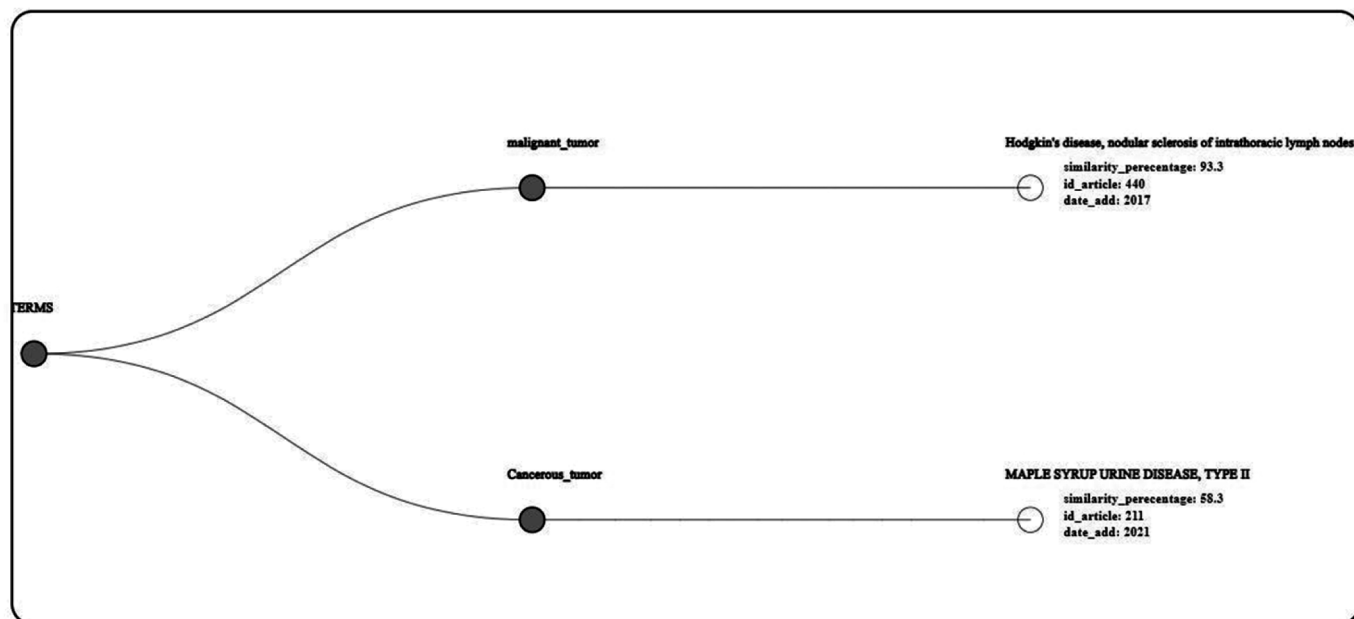


Рис. 4. Результаты процесса создания онтологий мутаций заболеваний в виде семантического графа «Заболеваний и их мутаций»

Результаты процесса создания онтологий мутаций заболеваний представлены в виде семантического графа «Заболеваний и их мутаций» на рис. 4.

Семантический граф «Заболеваний и их мутаций», основанный на иерархической структуре содержит в своей основе синонимичные пары терминов «болезнь — мутация болезни», объединенные в рамках общих тематических кластеров, построенных на основе ранее описываемого метода тематического моделирования BERTopic. Синонимичные пары терминов «болезнь — мутация болезни» представлены на графе в виде узлов, где между понятиями, описывающими похожие заболевания устанавливается семантическая связь на основе их корней и других лексических особенностей семантической категории [12]. Взаимосвязь между заболеваниями и мутациями заболеваний, отраженная на семантическом графе «Заболеваний и их мутаций» является строго односторонней. При некоторых заболеваниях наряду с редкими мутациями заболеваний могут присутствовать и распространенные. Помимо установления базовой связи «болезнь — мутация болезни», в данном семантическом графе отражается вес каждой мутации заболевания являющийся числовой мерой приближения данного термина к исходному термину заболевания [12].

Метод, использующийся при построении семантического графа «Заболеваний и их мутаций» строит синонимичные связи на основе словаря терминов из библиотеки UMLS, описывающего вес каждой мутации заболевания. Затем термины из библиотеки UMLS, описывающий вес каждой мутации заболевания, находятся в текстах корпуса, а понятия — заболевания в том же

предложении извлекаются на основе расширенного набора мутаций заболеваний. Этот процесс определяет вес как числовую меру приближения мутации заболевания к конкретному заболеванию, облегчая установление взаимосвязей между заболеванием и его мутацией.

### Заключение

В настоящее время автоматизированный анализ позволил исследователям не только собирать и изучать объем материала, анализ которого вручную невозможен, но и выявлять закономерности и взаимосвязи между различными терминами. Целями автоматизированного анализа стал поиск закономерной и связей новых терминов с существующими понятиями, содержащимися в международных компьютерных системах и базах знаний таких как: UMLS, PubMed, Babelnet, DBpedia, WordNet и других.

В рамках данной статьи были освещены подходы к поиску новых понятий и терминов в международной базе знаний медицинских публикаций PubMed и международном словаре медицинских терминов UMLS. Главными задачами, рассмотренными в рамках исследования, были:

- механизм формирования тематических кластеров в распознанных и размеченных многомерных данных;
- механизм выделения семантических связей терминов в рамках нескольких тематических кластеров;
- визуализация новых терминов и понятий в рамках нескольких тематических кластеров.

С помощью методов обработки естественного языка и интеллектуального анализа текста в библиотеках PubMed(PMC) и UMLS были получены онтологии заболеваний, онтологии мутаций заболеваний и связи между ними, помогая исследователям проследить динамику изменений взаимосвязей между заболеваниями. Также в международном словаре медицинских терминов UMLS был получен расширенный набор слов-кандидатов мутаций заболеваний. Этот набор включал в себя взаимосвязи между заболеванием и последствием заболевания (мутацией заболевания). По мере развития биоинформатики и расширения диапазона соответствующих источников данных будущие исследования будут направлены на обогащение библиотек PubMed(PMC) и UMLS дополнительными биомедицинскими характеристиками для повышения семантической точности и полноты моделей семантической сети.

В рамках данного исследования были описана модель семантической сети (графа) «Заболеваний и их мутаций» на базе библиотек PubMed(PMC) и UMLS. Использование международной базы знаний медицинских публикаций PubMed и международного словаря медицинских терминов UMLS в качестве базового набора данных послужило ключом для поиска новых заболеваний. Главным атрибутом модели семантической сети(графа) «Заболеваний и их мутаций» на базе библиотек PubMed(PMC) и UMLS стали онтологии. В модели семантической сети(графа) именно онтологии обеспечили основную технологию построения взаимосвязей объектов и стали ключом к реализации семантической сети.

## ЛИТЕРАТУРА

1. Юрчак В.А., Инструменты решения проблем распознавания и кластеризации данных из документов методами машинного обучения/Золотарев О.В., Юрчак В.А.// ИВД. — 2023. — № 2 (98). — С. 156–164.
2. Корней А.О., Семантико-статистический алгоритм определения категорий аспектов в задачах сентимент-анализа /Корней А.О., Крючкова Е.Н.// Известия ЮФУ. Технические науки. — 2020. — №6 (216). — С. 66–74.
3. Юрчак В.А., Решение проблемы кластеризации документов методами машинного обучения/Юрчак В.А., Маликов А.В., Т.-А.А.Хаитов// Журнал «Наука и бизнес: пути развития». — 2024. — №3 (153). — С. 12–15.
4. Клименко С.В., Использование онтологического подхода для анализа текстов естественного языка/ Клименко С.В., Золотарев О.В. Шарин М.М.//Вестник российского нового университета. Серия: сложные системы: модели, анализ и управление. — 2017. — С.67–71.
5. Хакимова А.Х., Подходы к созданию многоязычного лексического ресурса для семантометрической оценки межъязыкового семантического подобию текстов/ Хакимова А.Х., Золотарев О.В., Шарнин М.М. // Нижегородский государственный архитектурно-строительный университет, Научно-исследовательский центр физико-технической информатики. — 2019. — С. 319–324.
6. Иванова О.Н., Классификация мультимодальных данных о заболеваниях легких на основе позднего слияния модальностей/Иванова О.Н., Кумар Сэчин, Цымблер М.Л., Иванова Е.В.//Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. — 2024. — №1. — С. 74–86.
7. Программа выделения терминов из корпуса текстов/ Золотарев О.В. [и другие]// Автономная некоммерческая организация высшего образования «Российский новый университет». — 2023. — С. 1–2.
8. Программа построения структурированного корпуса текстов на основе электронных баз публикаций/Золотарев О.В. [и другие]//Автономная некоммерческая организация высшего образования «Российский новый университет». — 2023. — С. 1–2.
9. Модель и технология извлечения новых терминов из медицинских текстов/Золотарев О.В. [и другие]//Информатика и ее Применения. — 2022. — С. 80–86.
10. Мера подобию текстов как инструмент оценки интертекстуальности при анализе больших коллекций документов/Золотарев О.В. [и другие]//Вестник российского нового университета. Серия: сложные системы: модели, анализ и управление. — 2016. — С. 62–71.
11. Методика построения ассоциативно-иерархического портрета предметной области: иерархия категорий/Клименко С.В. [и другие]//Автономная некоммерческая организация «Институт физико-технической информатики». — 2017. — С. 251–260.
12. Y. Li, W. Wang, X. Yan, M. Gao, M. Xiao. Research on the Application of Semantic Network in Disease Diagnosis Prompts Based on Medical Corpus//International Journal of Innovative Research in Computer Science and Technology (IJIRCST). — 2024,1–9. Available from: <https://doi.org/10.55524/ijircst.2024.12.2.1>
13. Тутубалина Е.В. Модели и методы автоматической обработки неструктурированных данных в биомедицинской области: дис. доктора компьютерных наук. — Казань: Казанский (Приволжский) федеральный университет, 2023. — 225 с.
14. Бручес Е.П. Методы и алгоритмы распознавания и связывания сущностей для построения систем автоматического извлечения информации из научных текстов: дис. на соискание учёной степени кандидата технических наук. — Новосибирск: Федеральное государственное бюджетное учреждение науки Институт систем информатики им. Ершова, 2021. — 112 с.
15. Корней А.О. Методы и алгоритмы аспектного анализа тональности на основе гибридной семантико-статистической модели естественного языка: дис. на соискание ученой степени кандидата технических наук. — Барнаул: Федеральное государственное бюджетное образовательное учреждение высшего образования «Алтайский государственный технический университет им. И.И. Ползунова», 2021. — 134 с.
16. Дударин П.В. Исследование и разработка моделей и методов нечеткой кластеризации коротких текстов: дис. на соискание учёной степени кандидата технических наук. — Ульяновск: «Ульяновский государственный технический университет», 2021. — 141 с.