Обработка научно-технической информации в междисциплинарных исследованиях методами математико-лингвистического направленного поиска на примере области изучения биоматериалов для тканевой инженерии

Антонов Е.В., Артамонов А.А., Орлов А.В., Николаев В.С., Захаров В.П., Хохлова М.В., Концевая Ю.С., Бонарцев А.П., Воинова В.В.

Аннотация— **Разработка новых эффективных методов обработки научно-технической информации в междисциплинарных исследованиях с применением математико-лингвистического направленного поиска является актуальной проблемой в таких научных направлениях как изучение биоматериалов для тканевой инженерии из-за необходимости обрабатывать большие объемы информации, разнообразия источников информации и терминологической путаницы. Авторами проведено исследование по анализу входящего потока публикаций на предмет соответствия их тематическому направлению пользователя. В качестве исходного массива взята персональная база данных, собранная за 20 лет профессиональной деятельности научной группой «Медицинские биополимеры», размер которой составляет 3650 статей, который сравнивался с данными, собранными автоматизированным способом с 3-х высокорейтинговых изданий – Acta Biomaterialia, Biomaterials, Materials Today Bio. Для определения схожести текстов использовалась программная библиотека difflib, основанная на алгоритме Ратклифа–Мезнера. По итогам исследования выявлено, что разработанный подход смог адекватно выявить публикации соответствующие интересам лидера научной группы «Медицинские биополимеры», но также был выявлен ряд проблем, которые планируется решить на следующих этапах исследования.**

Ключевые слова— **научные статьи, междисциплинарный, тканевая инженерия, биоматериалы, информационный поиск, схожесть текстов.**

Введение

Современная тенденция развития научных исследований состоит в движении к «мультидисциплинарному» статусу, то есть к получению нового синтезированного знания методами различных научных дисциплин. Особенно активно данный процесс идет в рамках естественных и технических наук. Лидером в данной области следует назвать науки о жизни, так как наибольшее количество новых междисциплинарных программ в рамках укрупненных групп специальностей наблюдается на стыке именно данных областей знания с очень широким кругом других наук [1].

Можно выделить два основных подхода к междисциплинарности. Согласно первому, междисциплинарность понимается как взаимодействие двух или более научных дисциплин, каждая из которых имеет свой предмет, свою терминологию и методы исследования (также называется мультидисциплинарностью). Такое взаимодействие реализуется в форме работы над конкретными исследовательскими проектами, создания междисциплинарных центров при академических организациях, проведения междисциплинарных

Статья получена 28 октября 2022.

Антонов Евгений Вячеславович, Национальный исследовательский ядерный университет МИФИ, ассистент кафедры анализа конкурентных систем, ORCID 0000-0003-1498-9131 (eantonov@kaf65.ru).

Артамонов Алексей Анатольевич, канд. тех. наук, Национальный исследовательский ядерный университет МИФИ, заведующий кафедрой анализа конкурентных систем, ORCID 0000-0002-9140-5526 (aaartamonov@mephi.ru).

Орлов Артемий Владимирович, Государственный научный центр Российской Федерации Институт медико-биологических проблем Российской академии наук, научный сотрудник, ORCID 0000-0003-1290-0113 (orlovartem@mail.ru).

Николаев Вадим Сергеевич, ООО «НГ Логик», руководитель департамента аналитики, ORCID 0000-0003-4741-2208 (vsnikolayev@nglogic.ru).

Захаров Виктор Павлович, канд. филол. наук, Санкт-Петербургский государственный университет, доцент кафедры математической лингвистики, ORCID 0000-00030522-7469 (v.zakharov@spbu.ru).

Хохлова Мария Владимировна, канд. филол. наук, Санкт-Петербургский государственный университет, доцент кафедры математической лингвистики, ORCID 0000-0001-9085-0284 (m.khokhlova@spbu.ru)

Концевая Юлия Марковна, Российский экономический университет им. Г.В.Плеханова, лаборант-исследователь, Научная лаборатория «Перспективных систем хранения и обработки сверхбольших массивов данных», ORCID 0000-0001-8046-2237 (Koncevayayul.Yu@edu.rea.ru).

Бонарцев Антон Павлович, докт. биол. наук, Московский государственный университет им. М.В.Ломоносова, биологический факультет, доцент кафедры биоинженерии, ORCID 0000-0001-5894-9524 (ant_bonar@mail.ru).

Воинова Вера Владимировна, канд. биол. наук, Московский государственный университет им. М.В.Ломоносова, биологический факультет, ст. науч.сотр. кафедры биохимии, 0000-0002-0253-6461 (veravoinova@mail.ru).

конференций, издания проблемно, а не дисциплинарно ориентированных журналов и т. п. Второй подход к междисциплинарности предполагает выявление тех областей знания, которые не исследуются существующими научными дисциплинами. Приставка «меж» в этом случае указывает на наличие некого провала между дисциплинами, «ничейной земли», не являющейся традиционным объектом исследования ни одной из дисциплин. В таком случае на стыке научных дисциплин может возникнуть новая [2].

Существенной проблемой при проведении междисциплинарных исследований, является несовпадение специализированных языков и понятийного аппарата различных дисциплин, что затрудняет как анализ информации на этапах подготовки и проведения исследования, так и последующую экспертизу.

Не меньшую проблему представляет многократно возрастающий объем информации, необходимый на всех стадиях исследовательской работы. Кроме того, развитие информационно-коммуникационных технологий привело к экспоненциальному росту информации по всем направлениям деятельности человека, значительная часть которой представляет собой информационный шум.

Следствием обозначенных проблем становятся неоправданно большие затраты сил и времени профильных специалистов на анализ первичной информации без гарантии получения целевого результата. При этом, даже в случае наличия собранной лидером научной группы подборки публикаций, монографий, результатов экспериментов, отчетов по НИР и других объектов интеллектуальной собственности, как своих, так и коллег, материалы хранятся в виде отдельных файлов. Вследствие этого поиск по ним стандартными средствами операционных систем затруднен, и на поиск целевой информации уходит много времени и сил.

В этих условиях актуальной становится разработка системы, позволяющей в интерактивном режиме осуществлять поиск информации по всей базе знаний, выделять сущности (автор, технология, организация и т. д.), получать статистические данные по запросу в реальном масштабе времени. Важной задачей является также внутренняя кластеризация документов по направлениям деятельности человека, которая позволит более продуктивно осуществлять поиск новых материалов по направлениям деятельности авторов. При этом стоит учитывать важный момент – разработка подобной системы невозможна без участия как специалистов в области разработки поисковых систем и математической лингвистики, так и специалистов в целевой предметной области.

Для коллектива авторов статьи предметной областью работы создаваемой системы является тканевая инженерия и изучение биоматериалов, база научных публикаций создавалась в ходе экспериментальной работы научной группы «Медицинские биополимеры» (http://biopolymers.pro/) кафедры биоинженерии биологического факультета МГУ им. М.В. Ломоносова.

Наука о биоматериалах и тканевая инженерия – ярко выраженные мульти- и междисциплинарные направления в современной науке, в которых пересекаются все естественно-научные дисциплины с различными их более узкими поддисциплинами:

- физика (световая, конфокальная, электронная и атомно-силовая микроскопия, спектроскопия, механика, биомеханика, физика полимеров, молекулярное моделирование, кристаллография, теплофизика),
- химия (химия высокомолекулярных соединений, органическая химия, биоорганическая химия, физическая химия, коллоидная химия, химия композиционных материалов),
- биология (биохимия, биофизика, молекулярная биология, микробиология, клеточная биология, гистология, физиология и др.),
- медицина (регенеративная медицина, травматология, челюстно-лицевая хирургия, стоматология, сердечно-сосудистая хирургия, нейрохирургия, онкология, трансплантология, иммунология, фармакология, токсикология, ветеринария),
- инженерные науки (материаловедение, приборостроение, электротехника, метрология).

Тканевая инженерия объединяет последние достижения в области инженерии, материаловедения, клеточной биологии, биохимии и медицины, предлагая новые подходы для восстановления функций тканей и органов человека. Тканевая инженерия – это принципиально иная парадигма по сравнению с хирургией и трансплантацией. Она основана не на подходах замещения или функциональной компенсации тканей, а на регенерации тканей, при которой организм сам может восстановить поврежденную ткань при наличии соответствующих условий [3].

В частности, инженерия костной ткани предполагает совместное использование клеток (прежде всего, стволовых клеток), скаффолдов (искусственных конструкций из какого-либо биоматериала для выращивания в них клеток) и биоактивных молекул. Вместе они позволят обеспечить межклеточную коммуникацию и необходимое взаимодействие клеток и биоматериала, что будет способствовать достижению наилучшего терапевтического эффекта [4,5].

Скаффолды необходимы в качестве временного субстрата для прикрепления и роста клеток и накопления внеклеточного матрикса. Чтобы наилучшим образом имитировать костную ткань, скаффолды, используемые в тканевой инженерии, должны иметь трехмерную (3D) микроструктуру с высокой пористостью и определенным размером пор, необходимую топографию поверхности, биосовместимость и приемлемые механические свойства. Для создания скаффолдов могут быть использованы различные типы биоматериалов: металлы, керамика, синтетические или натуральные полимеры (биоматериалы природного происхождения не в виде композитов с полимерами, а в чистом виде), некоторые керамические и силикатные биоматериалы,

алломатериалы и ксеноматериалы, получаемые обработкой тканей животных, трупного и биопсийного материала. При этом наиболее предпочтительны для регенерации костной ткани скаффолды на основе биоразлагаемых и биосовместимых полимеров природного или синтетического происхождения. Существуют различные методы изготовления скаффолдов с желаемой микроструктурой: электроформование, выщелачивание, вспенивание, агрегация частиц, сублимационная сушка, термоиндуцированное фазовое разделение, микролитье, микроволокнистое прядение, быстрое прототипирование (включая 3D-печать) [6, 7, 8].

Исследования в области изучения биоматериалов для тканевой инженерии значительно осложняется терминологической путаницей, которая тесно связана с серьезной проблемой неселективного поиска научной информации в этой области. Уже сама мультидисциплинарность приводит к необходимости использовать термины, пришедшие из разных наук, и потому несущие специфическую смысловую нагрузку каждой из них. Другой проблемой является выраженное заимствование в отечественную терминологию англоязычных терминов, точнее калек разной степени подобия с англоязычных терминов. Например, термин «матрикс» имеет множество синонимов, которые используются параллельно: «скаффолд», «скэффолд», «матрица», «подложка», «каркас». В данном случае, только термины «подложка» и «каркас» и в меньшей степени «матрица» являются русскоязычными, которые употреблялись еще до 90-х гг. прошлого века, т.е. до эпохи массового проникновения в отечественную науку англоязычных терминов. Термины же «скаффолд» и «скэффолд» являются полной калькой англоязычного термина “scaffold”, наиболее употребительный в английском языке перевод которого означает «строительные леса».

Проблема заключается в том, что многие ученые вкладывают немного разный смысл в каждый из этих терминов. Например, термин «матрица» может быть отнюдь не тождественен для большой группы ученых калькированному термину «скаффолд», тогда как для другой большой группы специалистов эти термины могут быть полными синонимами. Разумеется, эта ситуация приводит часто к большой путанице в употреблении терминов, и, как следствие, значительно усложняют поиск научной литературы в этой области, особенно, русскоязычной.

Приведенный выше пример показывает, насколько широким может быть охват материалов и методов при проведении междисциплинарных исследований. При этом исследования с подобным охватом требуют постоянного мониторинга научной литературы в самых различных областях знаний, т.к. научно-технологическое развитие здесь происходит очень быстро и постоянно сопровождается пересечением и разветвлением различных направлений. Любой планируемый эксперимент, а тем более начинание какого-либо нового направления исследований, что происходит периодически в этой области, немыслимо без предварительной глубокой проработки и анализа данных из научных литературных источников, как самых свежих, так и архивных, и дальнейшего слежения за развитием научной мысли в этом выбранном направлении.

Исследователь в этой области должен быть постоянно в курсе всех наиболее значимых и менее важных открытий, а также во всех смежных областях; во многих случаях – довольно далеких научных областях, если происходит их пересечение с основной исследовательской проблемой. В противном случае неверный выбор направления исследования может привести к большим потерям времени, ресурсов и финансов, хотя первое несоизмеримо важнее в условиях интенсивного развития науки о биоматериалах и тканевой инженерии, а также смежных областей. Более того, поток информации, который должен отслеживать ученый, не ограничивается научными статьями, книгами и патентами, хотя они наиболее значимый источник информации. В дополнение необходимо анализировать новостные сообщения о науке в СМИ и социальных сетях, научно-популярные статьи, профильная производственная и медицинская информация, выступления ученых и представителей индустрии на конференциях и выставках, научно-популярные фильмы, лекции, выступления ученых и популяризаторов науки на различных видеохостингах и многое другое.

Иногда для правильного распознавания всех основных терминов в том или ином мульти- и междисциплинарном направлении необходимо иметь 15-20-летний опыт активной исследовательской работы в ней. Практически у каждого активно работающего исследователя имеется своя база научной литературы – книг, научных статей, патентов, протоколов и др. Далеко не всегда (практически никогда) эти источники научной информации систематизированы должным образом, тогда как со временем количество статей в таких базах данных может возрастать до 5 тыс. и более. И поиск нужной статьи в ней в какой-то момент становится более сложной задачей, чем новый поиск статей в открытых базах научной литературы, таких как Scopus и PubMed.

Лидером научной группы «Медицинские биополимеры» за последние двадцать лет накоплена база научных материалов из более чем трех с половиной тысяч файлов, которая и стала объектом экспериментальных исследований. При этом была поставлена задача исследовать потокового анализа и доставки целевой научно-технической информации заинтересованным пользователям.

Таким образом, целью авторского коллектива является разработка новых эффективных методов обработки научно-технической информации в междисциплинарных исследованиях с применением математико-лингвистического направленного поиска на примере области изучения биоматериалов для тканевой инженерии.

МЕТОДЫ

Определение схожести текстов, в частности научных статей, одна из ключевых проблем научной деятельности в рамках рекомендательных систем и систем таргетированного автоматизированного поиска данных [9-11]. Многие работы авторов посвящены данной проблемы, где используются различные подходы:

1) евклидово расстояние, расстояние по прямой линии между двумя точками в евклидовом пространстве [12, 13];

2) косинусное расстояние, где сходство вычисляется путем измерения косинуса угла между двумя векторами [14, 15];

3) манхэттен расстояние, рассчитывается как сумма абсолютных разностей между двумя векторами [16, 17].

Также существуют методы, основанные на семантике и использующиеся как для обнаружения плагиата, так и для поиска информации [18, 19]. Может быть проведен предварительный анализ текстов, включающий определение морфологических, синтаксических или семантических признаков, которые далее будут учтены в алгоритмах. Отдельную группу составляют вероятностные модели. Например, можно упомянуть о методе LDA (Latent Dirichlet Allocation — Латентное размещение Дирихле), который используется для тематического моделирования документов.

Авторами в рамках исследования был поставлен эксперимент в части нахождения наиболее близких публикаций для конкретного исследования. В качестве исходного массива взята персональная база данных, собранная за 20 лет профессиональной деятельности, размер которой составляет 3650 статей. Данный набор научных публикаций (в большинстве англоязычных научных статей) собирался постепенно с 2003 года, он организован в виде каскада тематических папок. Основная тематика статей: биоматериалы (преимущественно, биоразлагаемые полимеры и их композиты) для использования в медицине и тканевой инженерии, однако, некоторая часть статей была по близким соприкасающимся тематикам. Выделены следующие тематики основных папок и групп папок:

1) биоразлагаемые полимеры и медицинские изделия из них для инженерии костной ткани,

2) физико-химические свойства биоразлагаемых полимеров,

3) обзоры по различным тематикам в области исследования биоматериалов, 4) книги по различным тематикам в области исследования биоматериалов,

4) получение и исследование свойств композитов биоразлагаемых полимеров с другими материалами (синтетическими полимерами, природными полимерами, ксено- и аллогенными материалами, минеральными веществами, металлами и их оксидами, углеродными наноматериалами),

5) нано- и микрочастицы из биоразлагаемых полимеров для пролонгированного, направленного и/или контролируемого высвобождения и доставки низкомолекулярных лекарственных веществ, белков и генетических конструкций,

6) исследование биодеградации биоразлагаемых полимеров,

7) исследование биосовместимости биоразлагаемых полимеров,

8) исследование биосинтеза бактериальных полимеров,

9) методики и технологии получения различных конструкций из биоразлагаемых полимеров и их композитов (электроформование, быстрое прототипирование, микролитье, лазерная резка, выщелачивание, эмульгирование и др.),

10) скаффолды из биоразлагаемых полимеров, их композитов и других биоматериалов для тканевой инженерии,

11) роль биоразлагаемых полимеров для микробиоты человека и животных.

Анализируемый (входящий массив) получен на основе разработанного программного агента для сбора данных на языке программирования Python 3.8. Анализируемые данные собраны автоматизированным способом с 3-х высокорейтинговых изданий – Acta Biomaterialia, Biomaterials, Materials Today Bio. Итоговый размер собранного набора данных составил 1560 научных статей. Каждая отдельная статья включают в себя структурированные данные со следующими полями: название, аннотация, ключевые слова, данные автором, ключевые слова, год публикации, журнал, авторы, страна, аффилиации.

Также анализируемый набор прошел дополнительный этап предобработки, где вся текстовая информация статей преобразована в единых регистр, удалены числа, удалены знаки препинания и стоп-слова. Кроме этого, с помощью программной библиотеки yake извлечены ключевые слова, которые в дальнейшем использовались для определения релевантности. Библиотека yake использует статистический подход для извлечения ключевых слов и не требует предварительного обучения в отличие от алгоритмов машинного обучения, что также уменьшает время работы алгоритма [20]. Для определения схожести текстов использовалась программная библиотека difflib, в основе библиотеки лежит алгоритм Ратклифа–Мезнера. Алгоритм производит поиск длиннейшей общей подстроки [21].

Для написания алгоритма (общая схема представлена на рис. 1) использован высокоуровневый язык программирования Python 3.8 в среде Jupyter Notebook, которая объединяет код и результаты в виде одного документа, содержащего текст, математические уравнения и визуализации. Основными этапами создания алгоритма являются:

1) сбор данных,

2) предварительная обработка и очистка текста,

3) применение методов библиотек и получение результатов с необходимой точностью.

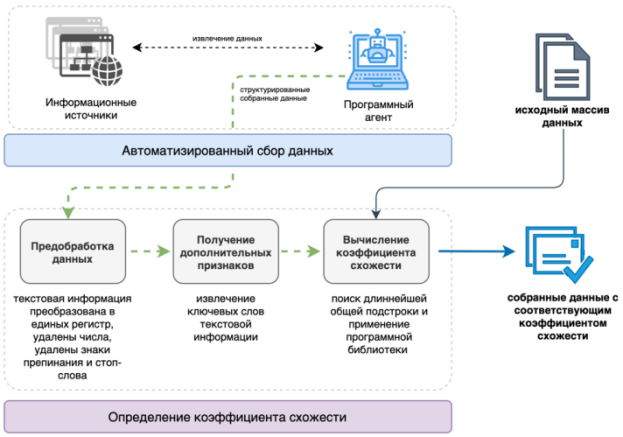

Рис. 1. Общая схема алгоритма

Для каждой статьи из анализируемого набора данных произведено попарное сравнение схожести на исходном наборе данных. Итоговый размер файла составлял 5 694 000 строк. После применения библиотеки на двух «чистых» наборах данных вычислена нормализированный коэффициент схожести статей, данный метод возвращает число в диапазоне [0, 1].

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Первоначально проведен анализ собранной выборки из 1560 статей. По итогам анализа ключевых слов текста статьи составлено облако ключевых слов (рис. 2).

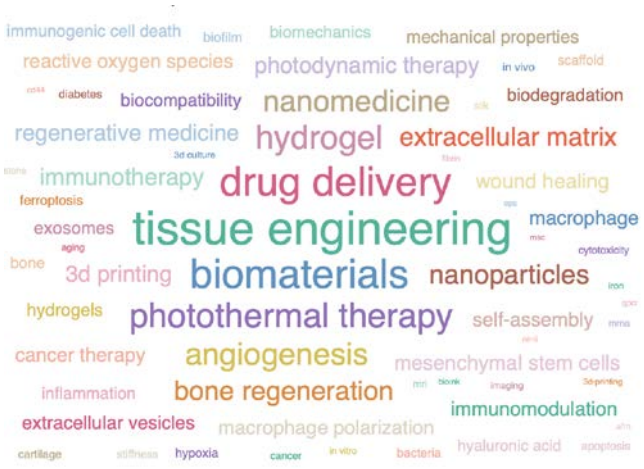

Рис. 2. Облако ключевых слов

На основании данных рисунка можно сделать вывод, что основные направления деятельности изданий связаны с такими темами как тканевая инженерия, биоматериалы, доставка лекарств, что в целом соответствует интересам лидера научной группы «Медицинские биополимеры».

В результате проведенных вычислений с помощью алгоритма, схема которой представлена на Рисунке 1, получено, что: 306 статей имеют коэффициент равный выше 0.5, 1045 статей получили коэффициент в диапазоне [0.4; 0.5], остальные – ниже 0.4. Статьи, с самыми высокими коэффициентами схожести представлены в таблице 1.

Таб. 1. Лучшие 5 статей по итогам отработки алгоритма

	Название статьи	**DOI**	**Коэффициент**
1.	Tumor microenvironment triple-responsive nanoparticles enable enhanced tumor penetration and synergetic chemo-photodynamic therapy	https://doi.org/10.1016/j.biomaterials.2020.120574	0.695
2.	Polyphenols as a versatile component in tissue engineering	https://doi.org/10.1016/j.actbio.2020.11.004	0.672
3.	Strontium regulates stem cell fate during osteogenic differentiation through asymmetric cell division	https://doi.org/10.1016/j.actbio.2020.10.030	0.661
4.	Treating brain diseases using systemic parenterally-administered protein therapeutics: Dysfunction of the brain barriers and potential strategies	https://doi.org/10.1016/j.biomaterials.2020.120461	0.660
5.	Liver donor age affects hepatocyte function through age-dependent changes in decellularized liver matrix	https://doi.org/10.1016/j.biomaterials.2021.120689	0.655

Полученные результаты алгоритма проанализированы автором исходной подборки материалов. Выявлено, что приемлемым коэффициентом схожести является значение коэффициента 0.6. Из 30 статей, характеризующихся такой точностью, все относятся к интересам исследователя.

Анализ всей выборки исследователем выявил следующие проблемы:

1) некоторые статьи обладают высоким коэффициентом схожести, однако не являются релевантными. У большинства таких статей есть общее свойство, а именно ключевое слово по которому их можно сразу же удалять, т.к. ее внетематическая смысловая нагрузка перевешивает все прочие ключевые слова, находящиеся в целевом тематическом поле. К таким ключевым словам можно отнести, например, термины «age-dependent» и «protein therapeutics». Тема изучения старения является собственным довольно обособленным научным направлением; получение, исследование и применение биоактивных рекомбинантных белков также является отдельным направлением в биофармакологии и белковой инженерии;

2) для некоторых статей показан заниженный коэффициент схожести, т.к. некоторые ключевые слова имеют относительно больший вес, по сравнению с прочими (например, «tissue engineering»), который перевешивает по своей значимости другие ключевые слова, которые могут обладать относительно высоким соответствием целевой тематике;

3) наличие в предоставленной базе данных

некоторого числа научных статей, не относящихся к целевой тематике, например, собранных не для научной, а для преподавательской деятельности по другим тематикам, что требует доработки механизма входящего контроля «чистоты» исходной базы данных.

4) тематика журналов (Acta Biomaterialia, Biomaterials, Materials Today Bio), выбранных для первичного анализа, сужает границы базы данных для обучения системы по этим довольно широким по тематике направлениям, для лучшего обучения необходимо использовать более тематически широкие и разнообразные базы данных научных публикаций.

Решение поставленных задач возможно посредством разработки методологии построения специализированных персональных тезаурусов со структурой весовых значений объектов тезауруса. Также для улучшения показателей точности необходимо использовать различные методы, которые были созданы в области обработки естественного языка и активно применяются для решения задач компьютерной лингвистики. Эти методы распространяются как на процедуры предварительной обработки текстов, так и непосредственно на задачу определения схожести текстов. В первой части, возможно, стоит отказаться от удаления знаков препинания, которые часто являются маркерами сегментации текста на семантико-синтаксическом уровне. Также предполагается более дифференцированно подойти к подзадаче формирования списка стоп-слов. Отдельно упомянем метод выделения ключевых слов, который предполагает сравнение целевого (узкотематического) корпуса с нейтральным, в который могут войти тексты более общего научного направления (медицинские или биологические).

ЗАКЛЮЧЕНИЕ

Приведенное исследование показывает, что выбранный подход в целом позволяет проводить направленный поиск новых научных статей в специализированных научных журналах, который удовлетворяет целевым для исследователя тематикам: тканевая инженерия, биоматериалы, доставка лекарств, выбранных на основании анализа тематической базы данных исследователя. Выявленные в ходе исследования проблемы требуют детальной проработки, в том числе необходимо рассмотреть возможность использования методов тематического моделирования, которые могут работать как на входном массиве, так и поверх массива с вычисленными коэффициентами подобия; методов на основе семантических словарей (тезаурусов), которые необходимо разработать.

Использование перечисленных методов потенциально позволит улучшить показатели полноты и точности в выборке научно-технической информации для групп исследователей и конкретных пользователей.

БИБЛИОГРАФИЯ

[1] Лысак, И.В. Междисциплинарность: преимущества и проблемы применения // Современные проблемы науки и образования. 2106. № 5. С. 264.

[2] Книгин А.Н. Междисциплинарность: основная проблема // Вестник Томского государственного университета. Философия. Социология. Политология. 2008. № 3(4). С. 14-21.

[3] Paschos N.K., Brown W.E., Eswaramoorthy R., Hu J.C., Athanasiou K.A.. Advances in tissue engineering through stem cell-based co-culture // J Tissue Eng Regen Med. 2015. No. 9(5). P. 488-503. DOI: 10.1002/term.1870.

[4] Koons G.L., Diba M., Mikos A.G. Materials design for bone-tissue engineering // Nat Rev Mater. 2020. No. 5. P. 584–603. DOI: 10.1038/s41578-020-0204-2.

[5] Qu H., Fu H., Han Z., Sun Y. Biomaterials for bone tissue engineering scaffolds: A review // RSC advances. 2019. No. 9(45). P. 26252-26262.

[6] Eltom, A., Zhong, G., & Muhammad, A. Scaffold techniques and designs in tissue engineering functions and purposes: a review. Advances in materials science and engineering, 2019, 3429527, https://doi.org/10.1155/2019/3429527

[7] Feng Y., Zhu S., Mei D., Li J., Zhang J., Yang S., Guan S. Application of 3D printing technology in bone tissue engineering: a review // Current Drug Delivery. 2021. No. 18(7). P. 847-861.

[8] Chahal S., Kumar A., Hussian, F.S.J. Development of biomimetic electrospun polymeric biomaterials for bone tissue engineering. A review // Journal of biomaterials science, polymer edition. 2019. No. 30(14). P. 1308-1355.

[9] Wang J., Dong Y. Measurement of text similarity: a survey // Information. 2020. Vol. 11. No. 9. DOI: 10.3390/info11090421.

[10] Qurashi A.W., Holmes V., Johnson A.P. Document processing: Methods for semantic text similarity analysis // International Conference on Innovations in Intelligent SysTems and Applications (INISTA). IEEE, 2020. P. 1-6.

[11] Magara M.B., Ojo S.O., Zuva T. A comparative analysis of text similarity measures and algorithms in research paper recommender systems // ICTAS2018. IEEE, 2018. P. 1-5.

[12] Deza M.M., Deza E. Encyclopedia of distances. Berlin, Heidelberg: Springer, 2009 .

[13] Cancho R.F. Euclidean distance between syntactically linked words // Physical Review E. 2004. Vol. 70. No. 5. DOI: 10.1103/PhysRevE.70.056135.

[14] Gunawan D., Sembiring C.A., Budiman M.A. The implementation of cosine similarity to calculate text relevance between two documents // Journal of physics: conference series. 2018. Vol. 978. No. 1. DOI: 10.1088/1742-6596/978/1/012120.

[15] Lahitani A.R., Permanasari A.E., Setiawan N.A. Cosine similarity to determine similarity measure: Study case in online essay assessment // 4th International Conference on Cyber and IT Service Management. IEEE, 2016. DOI: 10.1109/CITSM.2016.7577578.

[16] Jotheeswaran J., Kumaraswamy Y.S. Opinion mining using decision tree based feature selection through manhattan hierarchical cluster measure // Journal of Theoretical & Applied Information Technology. 2013. Vol. 58. No. 1.

[17] Eminagaoglu M. A new similarity measure for vector space models in text classification and information retrieval // Journal of Information Science. 2022. Vol. 48. No. 4. P. 463-476.

[18] Знаменский С.В. Моделирование задачи оптимального выравнивания последовательностей // Программные системы: теория и приложения. 2014. Vol. 5. No. 4(22). P. 257-267.

[19] Бермудес С.Х.Г., Керимова С.У. О методе определения текстовой близости основанном на семантических классах // Инженерный вестник Дона. 2016. Т. 43. №. 4(43).

[20] Campos R. et al. YAKE! Keyword extraction from single documents using multiple local features // Information Sciences. 2020. Vol. 509. P. 257-289.

[21] Wołk K., Marasek K. A sentence meaning based alignment method for parallel text corpora preparation // New Perspectives in Information Systems and Technologies. 2014. Vol. 1. P. 229-237.

Processing of Scientific and Technical Information in Interdisciplinary Research by Methods of Mathematical and Linguistic Directed Search by the Example of the Study of Biomaterials for Tissue Engineering

E.V. Antonov, A.A. Artamonov, A.V. Orlov, V.S. Nikolaev, V.P. Zakharov, M.V. Khokhlova, Yu.S. Kontsevaya, A.P. Bonartsev, V.V. Voinova

***Abstract*— The development of new effective methods for processing scientific technical information in interdisciplinary research using mathematical-linguistic targeted search is an urgent problem in such scientific field as the biomaterials for tissue engineering, because of the need to process large information volume, the diversity of information sources and terminological confusion. The authors have conducted a study to analyze the incoming flow of research papers for their relevance to the user's subject area. A personal database of 3,650 articles collected over 20 years of professional activity by the "Medical Biopolymers" scientific group was taken as an initial dataset. It was compared with the data collected by automated method from 3 highly ranked journals - Acta Biomaterialia, Biomaterials, Materials Today Bio. The software library difflib, based on the Ratcliffe-Mezner algorithm, was used to determine the similarity of the texts. According to the results of the study it was found that the developed approach was able to adequately identify the publications corresponding to the interests of the leader of the scientific group "Medical Biopolymers", but it also revealed a number of challenges, which are planned to solve at the next stages of the study.**

***Keywords*— research papers, interdisciplinary, tissue engineering, biomaterials, information search, text similarity.**

REFERENCES

[1] I.V. Lysak, "Interdisciplinarity: advantages and problems of application", *Modern problems of science and education*, no. 5, p. 264, 2016.

[2] A.N. Knigin, "Interdisciplinarity: the main problem", Vestnik *Tomsk State University. Philosophy. Sociology. Political science*, no. 3(4), pp. 14-21, 2008.

[3] N.K. Paschos, W.E. Brown, R. Eswaramoorthy, J.C. Hu, K.A. Athanasiou, "Advances in tissue engineering through stem cell-based co-culture", *J Tissue Eng Regen Med*, no. 9(5), pp. 488-503, 2015, doi: 10.1002/term.1870.

[4] Koons, G.L., Diba, M. & Mikos, A.G. Materials design for bone-tissue engineering. Nat Rev Mater 5, 584–603 (2020). https://doi.org/10.1038/s41578-020-0204-2

[5] H. Qu, H. Fu, Z. Han, Y. Sun, "Biomaterials for bone tissue engineering scaffolds: A review", *RSC advances*, no. 9(45), pp. 26252-26262, 2019.

[6] A. Eltom, G. Zhong, A. Muhammad, "Scaffold techniques and designs in tissue engineering functions and purposes: a review", Advances in materials science and engineering, 2019, 3429527, doi: 10.1155/2019/3429527.

[7] Y. Feng, S. Zhu, D. Mei, J. Li, J. Zhang, S. Yang, S. Guan, "Application of 3D printing technology in bone tissue engineering: a review", *Current Drug Delivery*, no. 18(7), pp. 847-861, 2021.

[8] S. Chahal, A. Kumar, F.S.J. Hussian, "Development of biomimetic electrospun polymeric biomaterials for bone tissue engineering. A review", *Journal of biomaterials science, polymer edition*, no. 30(14), pp. 1308-1355, 2019.

[9] J. Wang, Y. Dong, "Measurement of text similarity: a survey", *Information*, vol. 11, no. 9. 2020, doi: 10.3390/info11090421.

[10] A.W. Qurashi, V. Holmes, A.P. Johnson, "Document processing: Methods for semantic text similarity analysis", *International Conference on Innovations in Intelligent SysTems and Applications (INISTA)*, IEEE, 2020, pp. 1-6.

[11] M.B. Magara, S.O. Ojo, T. Zuva, "A comparative analysis of text similarity measures and algorithms in research paper recommender systems", *ICTAS2018*, IEEE, 2018, pp. 1-5.

[12] M.M. Deza, E. Deza, Encyclopedia of distances, Berlin, Heidelberg: Springer, 2009.

[13] R.F. Cancho, "Euclidean distance between syntactically linked words", *Physical Review E*, vol. 70, no. 5, 2004, doi: 10.1103/PhysRevE.70.056135.

[14] D. Gunawan, C.A. Sembiring, M.A. Budiman, "The implementation of cosine similarity to calculate text relevance between two documents", *Journal of physics: conference series*, vol. 978, no. 1, 2018, doi :10.1088/1742-6596/978/1/012120.

[15] A.R. Lahitani, A.E. Permanasari, N.A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment", *4th International Conference on Cyber and IT Service Management*, IEEE, 2016, doi: 10.1109/CITSM.2016.7577578.

[16] J. Jotheeswaran, Y.S. Kumaraswamy, "Opinion mining using decision tree based feature selection through manhattan hierarchical cluster measure", *Journal of Theoretical & Applied Information Technology*, vol. 58, no. 1, 2013.

[17] M. Eminagaoglu, "A new similarity measure for vector space models in text classification and information retrieval", *Journal of Information Science*, vol. 48, no. 4, pp. 463-476, 2022.

[18] S.V. Znamenskij, "A model and algorithm for sequence alignment", *Program systems: theory and applications*, vol. 5, no. 4(22), pp. 257-267, 2014.

[19] S.J.G. Bermudez, S.U. Kerimova, "About method of determination of textual proximity, based on the semantics classes", *Engineering journal of Don*, vol. 43, no. 4(43), 2016.

[20] R. Campos et al. "YAKE! Keyword extraction from single documents using multiple local features", *Information Sciences*, vol. 509, pp. 257-289, 2020.

[21] K. Wołk, K. Marasek, "A sentence meaning based alignment method for parallel text corpora preparation", *New Perspectives in Information Systems and Technologies*, vol. 1, pp. 229-237, 2014.