A. V. Shvets, D. A. Devyatkin, I. V. Smirnov, I. A. Tikhomirov, K. V. Popov, K. N. Yarygin, Investigation of systems and methods for scientometric analysis of scientific publications,
Artificial Intelligence and Decision Making, 2014, Issue 3, 62–71

А.В. Швец, Д. А. Девяткин, И.В. Смирнов, И.А. Тихомиров, К.В. Попов, К.Н. Ярыгин

Исследование систем и методов наукометрического анализа научных публикаций [1]

Аннотация. В статье приведен обзор методов и систем наукометрического анализа научных публикаций, описаны методы определения перспективных направлений исследований, приведены результаты экспериментального исследования выявления направлений научных исследований различными методами в предметной области «регенеративная медицина». Сделаны выводы о перспективности описанных методов, их недостатках, а также направлениях дальнейших исследований.

Ключевые слова: наукометрический анализ, поддержка научной деятельности, карта науки, регенеративная медицина.

Введение

В настоящее время для оценки перспективности направлений научных исследований используются различные методы экспертной оценки. Однако этот подход не лишен недостатков. Во-первых, академическое знание растет так быстро, что ни один эксперт не может охватить всю структуру конкретной области знаний, поэтому результат экспертизы может оказаться необъективным. Во-вторых, процесс извлечения экспертных знаний занимает много времени, поэтому такой подход применим для длительного стратегического планирования и не подходит для оперативного планирования.

В последнее время активно развиваются компьютерные системы наукометрического анализа научных публикаций, которые призваны облегчить процесс получения информации, необходимой экспертам для определения перспективных направлений в конкретной предметной области. К таким системам относятся цитатные базы научных публикаций, поисковые и аналитические системы. Системы наукометрического анализа позволяют обрабатывать огромные массивы данных, выявлять и структурировать публикации различных научно-исследовательских направлений, проводить аналитику на множестве публикаций, вычислять различные библиометрические показатели. К таким показателям относят, например, суммарные показатели цитирования, хронологическое распределение библиографических ссылок и другие.

Рассмотрим основные существующие системы, предоставляющие данные для проведения наукометрического анализа.

1. Системы наукометрического анализа

1.1. ScienceResearch

Система ScienceResearch представляет собой поисковую машину по научным ресурсам. Она работает примерно с тремя сотнями баз данных, электронных библиотек и других источников научных публикаций [1]. Основные свойства ScienceResearch – применение технологии «объединенного поиска» (Federated Search), а также кластеризации результатов поиска. Технология Federated Search была разработана компанией Deep Web Technologies и реализована в поисковом алгоритме Explorit Research Accelerator. Построенные на ее основе ресурсы не собирают

[1] Работа выполнена при финансовой поддержке РФФИ (грант № 13-04-12051).

собственную индексную базу, а работают в режиме реального времени с большим количеством внешних подключаемых БД.

Для уточнения запросов система ScienceResearch использует функцию кластеризации: найденные публикации делятся на тематические кластера и подкластера до трех уровней вложенности с указанием количества статей в кластерах. Описание кластера строится на основании аннотаций статей, входящих в кластер, и представляется в виде ключевого слова или словосочетания. Такая функция позволяет детально рассмотреть, каким научным направлениям соответствует множество публикаций, релевантных введенному запросу.

1.2. Scopus

База данных Scopus позиционируется издательской корпорацией Elsevier как крупнейшая в мире универсальная реферативная база данных с возможностями отслеживания научной цитируемости публикаций [2]. База включает информацию о публикациях в 21915 научных изданиях, покрывающих все предметное поле науки и техники. В состав базы входят и неанглоязычные журналы, в том числе из России, однако их количество недостаточно для проведения анализа направлений. Все журналы разбиты на 27 широких направлений и 313 узких тематических категорий. Каждый журнал может принадлежать более чем к одному направлению и категории. Встроенный инструмент Journal Analyzer позволяет проводить расширенный анализ научного уровня изданий (в том числе, сравнительный анализ нескольких изданий) по различным показателям активности. Система также предоставляет данные об учреждениях, сотрудники которых опубликовали более одной статьи. Примером таких данных является диаграмма тематического распределения публикаций сотрудников учреждения.

Существуют и другие системы компании Elsevier, например, такие как базы данных ScienceDirect, Embase, PharmaPendium и аналитические средства SciVal и illumin8 [3, 4]. Продукт ScienceDirect является полнотекстовой базой данных научно-технической и медицинской информации и содержит 25% научных публикаций во всем мире, Embase и PharmaPendium – базы данных для специалистов биомедицины и фармакологии.

1.2.1. SciVal

Система SciVal представляет собой комплекс инновационных веб-решений, который обеспечивает поддержку процесса принятия решений в области научных исследований. Наиболее значимыми функциями SciVal являются: создание отчетов для анализа достижений исследователей, команд, отделов или определенных пользователем групп; поиск специалистов или партнеров по терминам или свободному тексту, характеризующему тему исследований; измерение исследовательского потенциала научных организаций с целью определения конкурентных преимуществ и возможностей; выявление сильных междисциплинарных сторон организаций для определения областей для дальнейших инвестиций; измерение эффективности институтов по отношению к другим с помощью сравнения их потенциала; представление трендов исследований в организациях, регионах и государстве для выбора стратегии развития исследований; выявление быстро развивающихся направлений исследований и перспективных областей для дальнейшего инвестирования; оценка продуктивности исследователей и групп и их влияние на других исследователей и потребителей научной информации.

1.2.2. illumin8

Система illumin8 содержит инструменты для изучения новых процессов и технологий, поиска перспективных партнеров, мониторинга конкурентов, а также для выявления возможных рисков и преимуществ при внедрении новых подходов или при выходе на незнакомые рынки. illumin8 индексирует полные тексты статей из ScienceDirect и аннотации из Scopus и предоставляет семантический поиск, который анализирует отношения между словами, встречающимися в предложениях, чтобы определить их значение в зависимости от контекста. Извлеченные сущности затем используются при построении трендов. Например, можно построить графики распределения преимуществ и проблем технологии по годам на основании количества соответствующих формулировок, выделенных в публикациях. Стоит отметить, что среди информационных источников, которыми оперирует система, кроме указанных находятся также миллионы веб-страниц, более 1000 новостных источников и 24 миллиона патентов от 5 международных патентных агентств.

1.3. Web of Science

Компания Thomson Reuters выпустила систему ISI Web of Knowledge, которая является одной из самых крупных в мире по масштабам включаемой информации, представляет собой информационную среду для получения доступа к научной информации практически по всем отраслям знания и сочетает в себе возможности получения научной информации, ее анализа и оценки на основе современных электронных технологий. Наибольший интерес для решения наукометрических задач представляет база Web of Science Core Collection [5], которая по словам разработчиков является наиболее выверенным индексом цитирования (с тщательным сопровождением данных) и предоставляет необходимую информацию для проведения исследовательской работы.

Система реализует функции поиска с учётом метаданных, а также булева (логического) поиска с применением стандартного языка запросов. Функция Create Citation Report выполняет поверхностный анализ цитируемости найденных публикаций и предоставляет графики распределения статей и цитирующих статей по годам при условии, что найдено не более 10000 публикаций (в противном случае система просит уточнить запрос). Еще одна уникальная функция Related records позволяет обнаружить и показать статьи, связанные социтированием. Функция Results Analysis позволяет сгруппировать и отранжировать публикации по различным полям, например, возможно посмотреть распределение публикаций по научным направлениям.

Рассмотрим два модуля системы Web of Knowledge, обеспечивающих механизм оценки и анализа научного содержания – ISI Essential Science Indicators (ESI) и InCites. Модули ESI и InCites позволяют получить информацию о ключевых научных исследованиях в мире, выявлять основные тенденции развития науки. Научные исследования можно ранжировать по странам, журналам, ученым. Возможно создание списков научных коллективов и компаний в соответствии с тематикой проводимых ими исследований. При анализе организации ранжируются предметные области организации (249 заданных предметных областей), выявляются ключевые слова, характеризующие тематическую направленность, строятся графики изменения различных показателей во времени.

1.4. Google Scholar

Наряду со специализированными проектами в области наукометрии имеет смысл пользоваться и вертикальными сервисами универсальных поисковых машин. Наиболее известный ресурс такого плана – проект «Академия Google» [6]. Его бета-версия под оригинальным названием Google Scholar стартовала в октябре 2004 г.

«Академия Google» индексирует ресурсы открытого доступа, интернет-сайты, а также издательские сервисы, предоставляющие доступ к публикациям на коммерческих условиях. Русская версия «Академии Google» по умолчанию включает поиск по электронному каталогу Государственной публичной научно-технический библиотеки (ГПНТБ). В настройках поиска пользователю разрешается добавить до трех собственных ссылок на онлайн каталоги библиотек, поддерживающих названные технологии.

Библиографические ссылки Академии Google позволяют авторам следить за цитированием своих статей. Можно построить множество ссылающихся на определенные публикации, создать диаграмму цитирования и вычислить показатели этого процесса. При оценке релевантности той или иной ссылки, влияющей на ее позицию в выдаче поисковика, учитываются индекс цитирования публикации и ее автора, а также известность интернет-источника или того издания, где появилась статья.

1.5. Orbit

Стоит также упомянуть систему интеллектуального анализа патентов Orbit [7], которая может быть полезна при наукометрическом анализе научных направлений. Система предоставляет поиск, который производится по патентам более 90 международных патентных организаций, и различные аналитические инструменты для работы с патентами. Найденные патенты могут быть сохранены в личные файлы, к которым можно применить средства статистического анализа и визуализации. При этом можно создавать неограниченное число файлов с неограниченным числом патентов в них. Удобной функцией является мониторинг всех изменений и оповещение при появлении новых патентов, которые могут быть интересны пользователю.

В ходе анализа происходит выявление цитирований патентов, формирование семейств близких патентов, определяются правопреем-

ники, есть возможность проведения классификации патентов. Результаты представляются в виде различных графиков и отчетов, которые могут быть переданы коллегам внутри системы для совместной работы. Такой анализ, по словам разработчиков, идеально подходит для оценки новых технологий, их конкурентоспособности и потенциала лицензирования.

1.6. Инструменты построения карты науки

Рассмотренные выше системы могут быть использованы для более глубокого анализа состояния дел в научной сфере. Так существуют системы, позволяющие строить и анализировать карту науки, среди них: Bibexcel, CiteSpace II, CoPalRed, IN-SPIRE, Leydesdorff's Software, Network Workbench Tool, Sci2 Tool, VantagePoint и VOSViewer. Перечисленные системы подробно рассматриваются в работе [8]. Основные функции систем заключаются в построении библиометрических цитатных сетей разных видов, кластеризации множества публикаций на меньшие группы, именование полученных групп, анализ наукометрических показателей отдельных публикаций, временной анализ различных показателей. Многие системы используют библиометрические данные для кластеризации, однако есть и исключения, так, например, система IN-SPIRE строит кластера, в которые попадают документы, близкие по содержанию. IN-SPIRE имеет свой текстовый анализатор, с помощью которого документы представляются в виде векторов взвешенных терминов, которые используются для определения близости документов. Все системы имеют средства визуализации, которые позволяют наглядно рассмотреть, как развиваются и взаимодействуют отдельные научные направления, отображают изменения различных наукометрических показателей.

1.7. Science Index (РИНЦ)

Электронная библиотека eLIBRARY.RU [9] позволяет проводить поиск научных журналов, публикаций, авторов и организаций. Поиск производится с помощью российского индекса научного цитирования (РИНЦ), указателя, реализованного в виде базы данных, содержащего информацию о публикациях российских учёных в российских и зарубежных научных изданиях. Это специализированный информационный продукт, в котором собирается и обрабатывается полная библиографическая информация о журнальных статьях, аннотации и списки цитируемой литературы. Такая база позволяет находить как публикации, цитируемые в отдельно взятой статье, так и публикации, цитирующие эту статью.

В системе реализована возможность находить публикации близкие по тематике и схожие по тексту, перемещаться между статьями в пределах выбранного журнала, просматривать все публикации выбранного автора. Поиск можно проводить с учетом морфологии. Найденные публикации можно добавлять в собственные подборки публикаций, после чего можно проводить поиск внутри подборки и проводить анализ различных показателей (различные индексы цитирования, распределение статей по ключевым словам, по тематическим рубрикам и др.). Статьи также распределены по типам (статьи в журналах, диссертации, книги, отчеты), что позволяет проводить исследования по разным видам источников. Можно получить публикационную активность отдельного журнала или автора, однако нельзя посмотреть динамику развития отдельного научного направления, что является одним из недостатков электронной библиотеки. Другим существенным недостатком является отсутствие в базе данных публикаций из зарубежных систем научного цитирования, например, таких как Scopus и Web of Science.

1.8. Exactus Expert

Информационно-аналитическая система Exactus Expert, разрабатываемая в Институте системного анализа Российской академии наук, предназначена для поддержки научно-исследовательской деятельности, поиска точек роста науки и представляет собой комплекс инструментов и сервисов для работы с научными текстами [10]. Индекс системы включает статьи журналов из перечня ВАК, статьи зарубежных научных журналов, материалы российских и зарубежных конференций, авторефераты диссертаций, российские и зарубежные патенты.

Exactus Expert позволяет выявлять научные коллективы, занимающиеся исследованиями в интересующей предметной области, выявлять актуальные научные направления и оценивать их перспективность. Все задачи решаются на основе глубокого лингвистического анализа полных текстов научных публикаций. Для вы-

явления научных направлений используются методы кластеризации, классификации и другие методы интеллектуального анализа текстов. Описание научных направлений представляется в виде ранжированного набора наиболее характерных слов и словосочетаний, позволяющих установить тему направления. Также выполняется анализ динамики различных показателей для направлений и коллективов.

2. Методы определения перспективных направлений исследований

Опишем некоторые методы определения перспективных направлений исследований, которые используют данные, предоставляемые рассмотренными в первом разделе системами.

В работе [11] предлагается метод выявления развивающихся направлений путем кластеризации и анализа цитатных сетей. Первый шаг (1) заключается в сборе данных из различных коллекций научных публикаций по интересующей теме, которая задается с помощью запроса. В качестве такой коллекции, например, может быть выбрана цитатная база данных Web of Science. На втором шаге (2) выполняется построение сетей цитирований для каждого года с накоплением, т.е. построенная для произвольного года сеть содержит в себе сеть предыдущего года. Третий шаг (3) состоит в выделении наибольшей связной компоненты графа. Публикации, не цитируемые или не цитирующие другие работы построенной сети, не учитываются при анализе научного направления. После выявления наибольшей связной компоненты, на следующем шаге (4), сеть делится на кластера с помощью топологического метода кластеризации, который позволяет выявлять группы публикаций с высокой плотностью связей. На пятом шаге (5) выполняется визуализация полученных цитатных карт с использованием открытого продукта Large Graph Layout (LGL), затем (6) определяются позиции и роли статей внутри кластеров и, наконец, на шаге (7) выявляются темы построенных кластеров, характеризующие научные направления. Полученные данные затем используются для выявления развивающихся направлений. Схема метода представлена на Рис. 1.

Рассмотрим подробнее заключительную часть этого метода. На шаге (6) устанавливается роль каждой публикации в кластере, которая определяется следующими параметрами: степенью связанности узла внутри кластера z и коэффициентом участия P. Степень связанности узла показывает, насколько он «хорошо соединен» с остальными узлами кластера, коэффициент участия позволяет измерить, насколько «хорошо распределены» связи узла среди других кластеров. Формулы для вычисления этих параметров впервые приведены в работе [12]. Согласно значению первого параметра z узлы делятся на центральные и периферийные. Второй параметр P делит узлы на местные (провинциальные) – с существенным преобладанием внутрикластерных связей, связующие – с большим числом связей с узлами других кластеров и «родственные» - с преобладанием межкластерных связей. В качестве развивающихся направлений выбираются кластера, обладающие следующими свойствами: 1) центральные узлы имеют большое значение z и малое значение P; 2) центральные узлы опубликованы в последнее время; и 3) тема кластера отличается от тем других кластеров. Тема кластера представляется ключевыми словами и

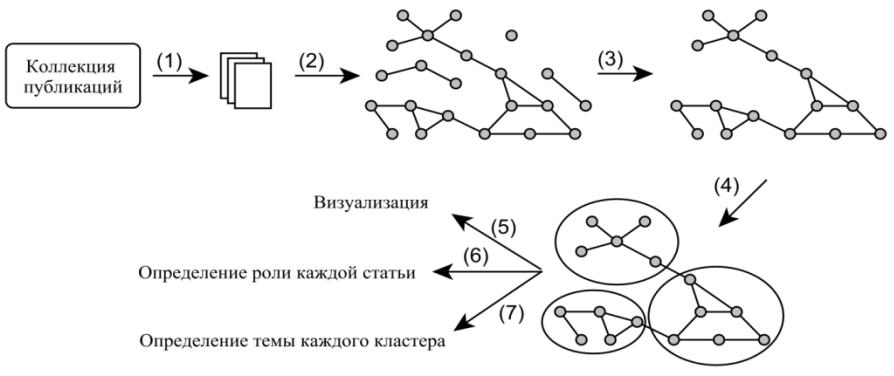

Рис.1. Схема метода, предложенного Shibata et al

словосочетаниями кластера, которые выделяются из текстов публикаций автоматически с помощью методов обработки естественного языка и ранжируются с использованием меры *tf-idf*. Для сравнения тем кластеров в качестве термов, характеризующих тему кластера, выбираются топ-10 ключевых слов.

Дополнительным шагом рассматриваемого метода является выявление в развивающемся кластере статей, которые в будущем должны набрать большое число цитирований и поспособствовать возникновению инноваций. В статье [11] показано, что с ожидаемым ростом цитирований коррелирует такой параметр как *центрированность* статьи (betweenness centrality). Он показывает, какой процент путей между различными узлами сети проходит через рассматриваемый узел, что можно интерпретировать как влияние узла на распространение информации через сеть. Публикация с большим значением центрированности соединяет ранее не связанные публикации, и поэтому, по предположению авторов метода, может стать источником инноваций в отдаленном будущем.

Другой метод, предложенный в работе [13], помещает в один кластер совместно процитированные статьи. Каждый кластер именуется словосочетанием, наиболее часто встречающимся в заголовках цитирующих статей, которое выбирается с помощью метода логарифмического отношения правдоподобия (log-likelihood ratio method (LLR)) [13]. Для анализа полученных кластеров на всем множестве публикаций выявляются следующие типы публикаций: с большим значением центрированности, наиболее цитируемые, публикации с наиболее резким и длительным увеличением цитируемости (всплеск цитируемости). На основании того какие типы публикаций находятся в кластере делается вывод о его характере. Также вычисляются значения *модульности* сети по годам – степень, с которой возможно распределить статьи по группам так, чтобы узлы внутри группы были связаны крепче, чем узлы между группами. Авторы метода полагают, что уменьшение значения модульности в рассматриваемом году может говорить о появлении новаторских, революционных публикаций, для выявления которых необходимо проводить дополнительное исследование.

Рассмотрим еще один метод, предложенный авторами данной статьи, применяемый в системе Exactus Expert [10]. В отличие от рассмотренных методов научные направления выявляются посредством построения кластеров близких по содержанию публикаций на основе анализа полных текстов [14]. Для оценки перспективности научных направлений используется несколько критериев: наличие научных коллективов, работающих в анализируемом направлении и имеющих положительную динамику публикационной активности; уровень качества публикаций коллектива, динамика появления научных результатов, представленных авторами коллектива в публикациях. Коллективы выявляются системой автоматически на основе соавторства и социтирования. Оценка качества публикаций и извлечение научных результатов из текстов также выполняется автоматически с помощью методов интеллектуального анализа текстов.

Алгоритм оценки перспективности научного направления в системе Exactus Expert состоит из нескольких шагов:

1. Сбор документов P научного направления.

2. Извлечение коллективов F анализируемого направления.

3. Для каждого коллектива из F вычисляется обобщенная оценка коллектива, учитывающая критерии наличия коллективов, публикационной активности и уровня качества

$$D_c(C) = \frac{|P(C)|\,Def(P(C))}{|A(C)|},$$

где $P(C)$ – множество публикаций коллектива C, $A(C)$ – множество авторов коллектива C, $Def(P)$ – средний уровень качества публикаций коллектива, способ его вычисления описан в работе [15].

4. Вычисляется возраст самого позднего документа в направлении $Novelty(P)$.

5. Из текстов публикаций автоматически извлекаются научные результаты и вычисляются показатели D_{rp} – оценка динамики практических результатов направления, D_{rt} – оценка динамики теоретических результатов направления.

6. Вычисляется оценка перспективности направления E:

$$E = \frac{\sum_{C \in F} D_c(C)}{|F|\,Novelty(P)} D_{rp} D_{rt}.$$

3. Эксперименты по определению перспективных направлений в области «регенеративная медицина»

Рассмотрим далее, какую информацию можно получить с использованием каждого из приведенных методов в области «регенеративная медицина». Эта обширная научная область включает в себя большое количество различных направлений, для нее характерна высокая скорость появления новых публикаций, поэтому она сложна для экспертного оценивания и представляет интерес для автоматического анализа.

Первый метод применялся авторами работы [11] на данных ресурсов Science Citation Index (SCI) и the Social Sciences Citation Index (SSCI), доступных в системе Web of Science [5]. Сбор публикаций проводился по запросам “regenerative medicine*”, “ES cell*”, “embryonic stem cell*”, “embryo-derived stem cell*”, “ips cell*”, “pluripotent stem cell*”, “adult stem cell*” or “somatic stem cell*”. В результате получены 17824 статьи, опубликованные до конца 2008 года. После построения наибольшей связной компоненты статьи были разделены на кластера с помощью топологического метода кластеризации. Анализ построенных кластеров позволил выделить 5 кластеров, соответствующих развивающимся направлениям исследований. Первый кластер, включающий 1916 статей со средним возрастом 1,8 лет, появился в 2004 году и относился к теме применения эмбриональных стволовых клеток (ЭСК) к клеткам человека. Этот кластер затем породил два других, которые образовались в 2007 году и относились к исследованиям эмбриональных стволовых клеток и взрослых и соматических стволовых клеток соответственно. Эволюционировав, данные кластера совместно образовали в 2008 году другие два кластера. Один из них снова относился к эмбриональным стволовым клеткам, другой – к взрослым и соматическим стволовым клеткам.

Эти результаты оказались правдоподобными и ожидаемыми, но не имеющими существенного значения, поэтому авторы метода провели исследование последнего кластера, для которого снова выполнили шаги (4)-(7), описанные в предыдущем разделе. В результате выделено три больших кластера. Тема первого кластера относилась к общей ДНК, тема второго связана с именами генов Nanog и Oct4, играющих существенную роль для плюрипотентности, тема третьего кластера относилась к индуцированным плюрипотентным стволовым клеткам (ИПСК). Такие результаты оказались полезными для обзора предметной области и могут быть использованы для выбора дальнейшего направления исследований и выявления новых перспектив. Построенные для кластеров ключевые слова представлены в Табл. 1.

Рассмотрим результаты применения **второго метода** к теме “регенеративная медицина”, представленные авторами метода в [13]. Анализируемая коллекция собрана из публикаций, найденных по запросу “regenerative medicine” в системе Web of Science (3875 статей), и расширена цитирующими статьями, которые предположительно должны быть так же релевантны теме. Всего собрано 35963 публикации с 2000 по 2011 год, которые включали 28252 оригинальные научные статьи и 7711 обзорных статей. На основе этих данных построена сеть совместно процитированных публикаций и выполнена кластеризация. В результате выделено 8 кластеров, содержащих в среднем по 60 публикаций.

Табл. 1. Основные направления, выявленные с помощью первого метода

№	Ключевые слова направления	Кол-во статей	Средн. дата
1	Genes, methylation, dna, dna methylation, cell, expression, genome, protein, development, mice, role, x chromosomes, es cells, loci, levels, mechanism, differentiation, modification, stem cells, regulation, histone	813	2005
2	Oct, cell, expression, genes, stem cells, es cells, sox, differentiation, self, germ cells, protein, mice, renewal, development, role, es, stem, factors, promoters, seminoma, marker	751	2006
3	Cell, stem cells, somatic cells, oocytes, embryos, oct, mice, es cells, expression, stem, development, potential, transfer, genes, nucleus, differentiation, state, blastocysts, es, sox, factors	685	2007

Для выявления развивающихся направлений авторы метода рассмотрели наиболее поздний кластер (средний год опубликования – 2008). По заголовкам статей автоматически определена тема кластера: “Induced pluripotent stem cell” (индуцированные плюрипотентные стволовые клетки). Кластер имеет большое число статей с всплеском цитируемости, выделена одна публикация с большим значением центрированности (Takahashi K, 2006, Cell, V126, P663). Такая статья, по словам авторов метода, является ориентиром, который может задать направление дальнейшего развития регенеративной медицины. Три статьи кластера являются одними из наиболее цитируемых среди статей всех кластеров (1273, 1583 и 1841 цитирование). Наименьшее значение модульности сети пришлось на 2009 год, которому соответствуют статьи анализируемого кластера, что означает наличие в кластере революционных статей. Вручную проанализированы опубликованные в 2009 и 2010 годах статьи с высоким всплеском цитируемости и среди них выявлены статьи, которые поднимают вопросы молекулярной и функциональной эквивалентности ИПСК и ЭСК человека, и статьи, сфокусированные на улучшении методов перепрограммирования соматических клеток человека для восстановления плюрипотентного состояния. Авторами метода сделано предположение, что эти статьи и отражают новые образовавшиеся тренды.

Рассмотрим третий метод, предложенный авторами данной статьи и реализованный в системе Exactus Expert. Метод отрабатывал на основе публикаций журналов “Stem cells” и “Journal of Tissue Engineering and Regenerative Medicine”, содержащих в открытом доступе около 2500 публикаций с 1996 по 2014 годы. На множестве этих публикаций проведена кластеризация, в результате которой выявлены научные направления, наиболее крупные из которых представлены в Табл. 2.

На основании анализа полных текстов статей, вошедших в кластера, эксперты предметной области присвоили направлениям названия и подтвердили, что ключевые слова, представленные в Табл. 2, соответствуют темам кластеров. Основные выделенные направления относятся к следующим темам:

- стволовые клетки периферической крови;
- стволовые клетки роговичного эпителия;
- восстановление хрящевой ткани;
- миелоидный лейкоз;
- дендритные клетки;
- лечение диабетов;
- нейрональная дифференцировка;
- дофаминергические нейроны.

Табл. 2. Основные направления, выявленные с помощью системы Exactus Expert

№	Ключевые слова направления	Кол-во статей	Средн. дата
1	antigen histocompatibility, apheresis cell, blood allogeneic, cell elusive, cell lak, cell malignant, characteristic engraftment, chemopriming, colony recombinant, disorder hematologic, diversity cellular, donor granulocyte, donor normal	28	1999
2	biopsy limbal, cell basal, cell corneal, cell limbal, cornea central, cornea develop, cornea human, cornea peripheral, deficiency cell, disease ocular, epithelium corneal, fibroblast limbal, limbus corneal	21	2009
3	cartilage articular, cartilage host, cartilage hyaline, cartilage uninjured, collagen type, defect articular, formation cartilage, osteoarthritis, regeneration cartilage, surface articular, tissue repair	20	2010
4	activity tyrosine, cell bcr-abl-positive, chromosome, crisis blast, crisis myeloid, inhibitor tyrosine, leukemia chronic, mortality transplant-related, response cytogenetic, therapy imatinib	19	2004
5	activity stimulatory, antigen soluble, cell bulk, cell dendritic, cell epidermal, cell immunostimulatory, cell langerhans, colony dc, dc human, dc immature, dc mature, immunostimulatory, molecule class, organ lymphoid, population dc, reaction leukocyte	18	2002
6	animal diabetic, c-peptide human, cell insulin-positive, diabetes autoimmune, diabetes, hyperglycemia, insulin-producing, islet pancreatic, level blood, level glucose, pancreas adult, regeneration cell, secretion insulin	18	2008
7	cell flat, cell rosette, cord ventral, development neural, differentiation motoneuron, differentiation neuroectodermal, gene neural, induction neural, morphogens, motoneurons spinal	18	2009
8	cell floor, cell plate, dopamine, dopaminergic, fate dopaminergic, fate midbrain, fate ventral, hydroxylase tyrosine, identity midbrain, induction floor, induction plate, midline ventral, neuron dopamine, phenotype forebrain, neuron human, rat parkinsonian, shh recombinant, shh exogenous, tissue mesencephalic	15	2009

Построенные направления сопоставимы с направлениями, выделенными с помощью первого метода, что говорит о близости результатов, получаемых с помощью ссылочной кластеризации публикаций и кластеризации по полным текстам. Однако третий метод позволяет проводить кластеризацию без использования цитатных баз и предварительного отбора статей по ключевым словам. Также преимуществом этого метода является возможность анализа полных текстов, что позволяет группировать содержательно близкие тексты, не обязательно цитирующие друг друга. При этом в кластеры не попадают лишние статьи, не относящиеся к направлению, как это происходит в первых двух методах.

Заключение

В статье рассмотрены различные системы наукометрического анализа и исследованы методы выявления перспективных направлений исследований на основании кластеризации. Рассмотренные методы позволили выделить различные направления, однако ключевые слова, построенные с их помощью, не всегда удачно характеризовали тематику направлений. В связи с этим требуется более глубокий анализ публикаций, который должен зависеть от типа содержимого, поскольку значимая информация находится в разных частях текстов разного типа. Так, например, в научных статьях ключевые слова следует выделять в аннотациях, в разделе с описанием методов, но пропускать раздел, содержащий обзор текущего состояния дел, который может включать информацию, слабо связанную с темой статьи.

Специалистам исследуемой предметной области (в данном случае области регенеративной медицины) было бы также интересно получить распределение статей по заранее заданным темам, которые принято выделять в качестве основных направлений этой предметной области. Представленные в статье методы не позволяют в полной мере с использованием кластеризации выполнять такое распределение, так как не все выявляемые кластера совпадают с существующими направлениями. Поэтому дальнейшим развитием автоматического определения перспективных направлений может стать применение методов классификации, с помощью которых станет возможным выполнять детальный анализ предметной области по заданным специалистами темам. Также перспективным является развитие методов, основанных на взаимодействии ссылочной кластеризации и кластеризации по полным текстам.

Литература

1. ScienceResearch [Электронный ресурс] URL: http://www.scienceresearch.com (дата обращения 10.06.2014).
2. Scopus // [Электронный ресурс] URL: http://www.info.sciverse.com/scopus/ (дата обращения 10.06.2014).
3. Scival // [Электронный ресурс] URL: http://www.elsevier.com/online-tools/research-intelligence/products-and-services/scival (дата обращения 10.06.2014).
4. Illumin8 // [Электронный ресурс] URL: http://www.elsevier.com/online-tools/illumin8 (дата обращения 10.06.2014).
5. Web of Science Core Collection // [Электронный ресурс] URL: http://thomsonreuters.com/web-of-science-core-collection/ (дата обращения 10.06.2014).
6. Академия Google // [Электронный ресурс] URL: http://scholar.google.ru/ (дата обращения 10.06.2014).
7. Orbit // [Электронный ресурс] URL: http://www.orbit.com/ (дата обращения 10.06.2014).
8. M.J. Cobo, A.G. López-Herrera, E. Herrera-Viedma, and F. Herrera. Science Mapping Software Tools: Review, Analysis, and Cooperative Study Among Tools, Journal of the American Society for Information Science and Technology, 62(7):1382–1402, 2011.
9. Научная электронная библиотека eLIBRARY.RU // [Электронный ресурс] URL: http://elibrary.ru/defaultx.asp (дата обращения 10.06.2014).
10. Тихомиров И.А., Смирнов И.В., Соченков И.В., Девяткин Д.А., Шелманов А.О., Зубарев Д.В., Швец А.В., Лешкин А.В., Суворов Р.Е. Exactus Expert: Поисково-аналитическая система поддержки научно-технической деятельности // Труды тринадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2012. Белгород: БГТУ, 2012. Т. 4. С. 100-108.
11. Shibata N, Kajikawa Y, Takeda Y, et al. Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. Technol Forecasting Soc Change 2011; 78:274-82.
12. R. Guimera, L.A.N. Amaral, Functional cartography of complex metabolic networks, Nature 433 (2005) 895–900.
13. Chen C, Hu Z, Liu S, Tseng H. Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace, Expert Opin. Biol. Ther. (2012) 12(5):593-608.
14. Девяткин Д.А., Суворов Р.Е., Соченков И.В. Распределенная тематическая кластеризация текстовых документов в системе Exactus Expert // Труды пятой международной конференции «Системный анализ и информационные технологии» (САИТ-2013). Красноярск, 2013. Т. 1, С 200-207.
15. Швец А.В. Экспериментальный метод автоматического определения уровня качества научных публикаций // Труды пятой международной конференции «Системный анализ и информационные технологии» (САИТ-2013). Красноярск, 2013. Т. 1. С. 304-312.

Швец Александр Валерьевич. Инженер-исследователь Института системного анализа Российской академии наук. Окончил Сибирский федеральный университет в 2011 году. Автор 15 научных работ. Область научных интересов: компьютерная лингвистика, математическое моделирование, методы оптимизации, искусственный интеллект.
E-mail: shvets@isa.ru

Девяткин Дмитрий Алексеевич. Инженер Института системного анализа Российской академии наук. Окончил Рыбинскую государственную авиационную технологическую академию в 2011 году. Автор 6 научных работ. Область научных интересов: адаптивные методы дистанционного обучения, классификация и кластеризация текстов, интеллектуальный веб-краулинг, искусственный интеллект. E-mail: devyatkin@isa.ru

Смирнов Иван Валентинович. Старший научный сотрудник Института системного анализа Российской академии наук. Окончил Российский университет дружбы народов в 2003 году. Кандидат физико-математических наук. Автор 34 научных работ. Область научных интересов: искусственный интеллект, компьютерная лингвистика, машинное обучение. E-mail: ivs@isa.ru

Тихомиров Илья Александрович. Ведущий научный сотрудник Института системного анализа Российской академии наук. Окончил Рыбинскую государственную авиационную технологическую академию в 2002 году. Кандидат технических наук. Автор 51 научной работы. Область научных интересов: искусственный интеллект, компьютерная лингвистика, поисковые системы, информационная безопасность, интернет-системы. E-mail: tih@isa.ru

Попов Константин Васильевич. Старший научный сотрудник Института молекулярной биологии им. В.А. Энгельгардта Российской академии наук. Окончил Московский физико-технический институт в 1998 году. Кандидат биологических наук. Автор 40 научных работ. Область научных интересов: биология клетки, регенеративная медицина.
E-mail: Konstantin.v.popov@gmail.com

Ярыгин Константин Никитич. Заведующий лабораторией клеточной биологии Института биомедицинской химии им. В.Н. Ореховича Российской академии медицинских наук. Доктор биологических наук, профессор, член-корреспондент РАН. Автор более 150 научных работ. Область научных интересов: регенеративная медицина, персонализированная медицина, клеточная терапия, клеточная биология, стволовые клетки, трансдифференцировка, индуцированные плюрипотентные клетки. E-mail: kyarygin@ibmc.msk.ru