УДК 32.813

АНАЛИЗ ТЕКСТОВ И АВТОМАТИЧЕСКОЕ ФОРМИРОВАНИЕ КЛАСТЕРОВ ПО ТЕМАТИЧЕСКОЙ НАПРАВЛЕННОСТИ

Аскеров Э.З.[1]

[1]*ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации», Москва, e-mail: rigorokhova@fa.ru*

Аннотация

В статье представлены результаты решения проблемы автоматического формирования кластеров тематической направленности. Анализ изменений тематических кластеров выполнен на материале научных статей молодых ученых, это связано с тем. что статьи ученых, признанных в научном мире имеют рейтинг, в то время, когда статьи молодых ученых такого рейтинга не имеют. Реализован подход, основанный на анализе полных текстов научных статей, методы основываются на анализе совместного пересечения словосочетаний/слов. Исследование связей тематической кластеризации проведено с использованием программного продукта VosViewer. Продукт позволяет проводить анализ цитирования, выявлять ключевые темы и паттерны, прослеживать связи между научными областями и учеными. Он способствует обнаружению структурных характеристик сетевых данных, таких как группировка, центральность и степень связанности. Кроме того, VOSviewer дает возможность проводить анализ социальных сетей и исследовать взаимосвязи между участниками с помощью различных метрик и методов визуализации. Описана кластеризация словосочетаний на основе алгоритма модуляции. Для выявления особенностей и тенденций в различных временных промежутках и для визуализации и сравнения данных использована стратегическая диаграмма. Изучение динамики формирования тематических кластеров осуществляется путем анализа изменений и развития внутри кластера, для этого применяются ориентированные графы, которые учитывают направление связей и позволяют анализировать и следить за значимыми изменениями, происходящими внутри кластера. В статье рассмотрен анализ текстов докладов за десятилетний период, который показывает характеристики тенденций и изменений в кластерах тематической направленности, такие как обнаружение формирования новых кластеров, рост или сокращение существующих кластеров, изменение связей и взаимодействий между тематическими группами и другие изменения, которые отражают эволюцию и развитие в обсуждаемых тематических направлениях.

Ключевые слова: анализ текстов, тематическая направленность, кластеры, графы, предметная область, программа анализа текстов.

TEXT ANALYSIS AND AUTOMATIC FORMATION OF CLUSTERS BY THEMATIC FOCUS

Askerov E.[1]

[1]*Financial University under the Government of the Russian Federation, Moscow, e-mail: rigorokhova@fa.ru*

Abstract

The article presents the results of solving the problem of automatic cluster formation of a thematic focus. The analysis of changes in thematic clusters is based on the material of scientific articles by young scientists, this is due to the fact. that articles by scientists recognized in the scientific world have a rating, at a time when articles by young scientists do not have such a rating. An approach based on the analysis of full texts of scientific articles is implemented, the methods are based on the analysis of the joint intersection of phrases/words. The study of the links of thematic clustering was carried out using the VosViewer software product. The product allows you to analyze citations, identify key topics and patterns, and trace connections between scientific fields and scientists. It helps to detect the structural characteristics of network data, such as grouping, centrality, and degree of connectivity. In addition, VOSviewer makes it possible to analyze social networks and explore the relationships between participants using various metrics and visualization methods. Clustering of phrases based on the modulation algorithm is described. A strategic chart is used to identify features and trends in different time periods and to visualize and compare data. The study of the dynamics of the formation of thematic clusters is carried out by analyzing changes and development within the cluster. For this, oriented graphs are used that take into account the direction of connections and allow you to analyze and monitor significant changes occurring within the cluster. The article examines the analysis of the texts of reports for a ten-year period, which shows the characteristics of trends and changes in clusters of a thematic focus, such as the detection of the formation of new clusters, the growth

or reduction of existing clusters, changes in relations and interactions between thematic groups and other changes that reflect the evolution and development in the discussed thematic areas.
Keywords: text analysis, thematic focus, clusters, graphs, subject area, text analysis program

Введение

Информационная система, созданная для регулярного автоматического обновления информации, проведения статистического анализа исследовательской деятельности признана наиболее частым способом представления структуры предметной области. К построению карт применимо два основных подхода [1]. Первый демонстрирует структуру: цитирования двумя авторами одной статьи. Второй показывает совместное пересечение терминов в текстах [2]. При этом, оба подхода схожи. Программное построение карт формируется путем выбора текстов из статей по ключевым запросам к базам данных. Предметная область представлена тематическими кластерами, имеющими внутрикластерные и межкластерные связи.

В случае применения совместного цитирования нескольких связанных текстов, выявление одинаковых частей происходит по авторским ссылкам. Тексты, в которых используются частым образом похожие термины, тексты, тематики, имеют схожее строение. Для выявления структуры предметной области используют кластеризацию терминов из Интернета. Статьи являются текстами, построенные на цитировании, в них находятся термины, которые в свою очередь характеризуют кластеры. Ключевые слова, слова, словосочетания, небольшие тексты являютя терминами.

В 2010 - 2020 гг. было проведено исследование, в котором наблюдалась динамика и тематическая структура технических направлений. Изменение кластеров тематической направленности означает изменение группировки объектов, которые связаны с определенной тематикой или концептуальным направлением. Особенностью проводимого анализа является то, что алгоритм реализовывался на текстах научных статей технической направленности, представленных молодыми учеными на конференции.

Получено, что статьи ученых, признанных в научном мире имеют рейтинг, в то время, когда статьи молодых ученых такого рейтинга не имеют. Тем не менее, факт прохождения отбора трудов для представления на конференции демонстрирует их значимость. Также можно утверждать, что при оценке кластерных изменений предметной области по рассматриваемым материалам риск ошибок выше, в сравнении со статьями, включенных в рейтинг [3,4]. Таким образом, своевременное выявление ошибок в формировании кластеров представляет для целей изучения и практического применения большой интерес. Исследование применения различных видов моделей проведено в целом ряде статей [5,6,7].

Цели исследования

Разработка алгоритма формирования кластеров по выбранным темам в смежных периодах времени и выявление возможности применения на практике разработанного алгоритма.

Материал и методы исследования

В статье представлены результаты анализа изменений тематических кластеров, полученных на материале научных статей молодых ученых по технической направленности, опубликованных на конференциях с 2010 по 2020 годы.

Методы основываются на анализе совместного пересечения словосочетаний/слов в текстах научных статей. [8] Анализ текстов проводился с помощью программного обеспечения VOSViewer.

Выборка терминов, группировка множества объектов проводилась с помощью применением программы VOSViewer (Просмотрщик) [2].

Изменения в кластерах на протяжении всего времени их существования происходят на уровне всех анализируемых смежных сегментов. Создание нейронной сети происходит через i и j. Мера ассоциации вычисляется по формуле 1.

$$s_{ij} = \frac{2ma_{ij}}{k_i k_j} \qquad (1),$$

где a_{ij} – вес связи между терминами i и j, равный количеству исследуемых статей, схожих между собой, k_i и k_j – суммарный вес связей для i и j терминов, m – показатель суммарного веса общих связей в нейронной сети [9].

Подход к анализу является методом, разработанным для исследования и визуализации библиографических данных с помощью программного инструмента VOSviewer, который способен создавать наглядные карты и визуализации. Использование VOSviewer и его основных компонентов (функция «модульности» и параметр «разрешение») позволяют проводить более глубокий и структурированный анализ библиографических данных. Это способствует выявлению тематических групп, группированию ключевых запросов и получению наглядных результатов визуализации, что помогает исследователям лучше понять структуру и содержание в изучаемых библиографических данных. Ключевые запросы, сгруппированные по определенным характеристикам или общим элементам, формируют определенные темы технической направленности. То есть, с помощью VOSviewer и его подхода к анализу, можно выделить важные слова или фразы, которые появляются в связанных тематиках или областях, и наглядно представить их визуально с помощью создания карт в программе.

Результаты исследования и их обсуждение

Стратегические диаграммы, основанные на наборе текстов, относящихся к одному и тому же периоду времени, являются инструментом для оценки развития темы и перспектив развития. Разграничение осуществляется на основе различных групп и характеристик, которые отражены на осях центральности (c) и показателях плотности (d). Эти показатели

рассчитываются с использованием индекса эквивалентности и связей между терминами в конкретном кластере.

Включение каждого кластера в разные квадранты стратегической диаграммы представлено следующим образом:

– 1-й квадрант отражает основную тему, связанную с второстепенными темами. Значения центральности и плотности в кластере при этом являются положительными: $c > 0$ и $d > 0$;

– 2 квадрант включает узкоспециализированную тематику, где центральность низкая или отрицательная, а плотность остается высокой: $c \leq 0$ и $d > 0$;

– 3 квадрант содержит тексты с незначительной направленностью, где как центральность, так и плотность, являются положительными или равными нулю: $c \geq 0$ и $d \geq 0$;

– 4 квадрант отражает потенциально перспективные темы для будущего развития, где центральность высока, а плотность низка или отрицательна: $c > 0$ и $d \leq 0$.

Начало координат устанавливается в точке пересечения медиан значений центральности $\{c_i\}$ и плотности $\{d_i\}$, где i представляет численность кластеров. Это формирует базовую точку для построения стратегической диаграммы.

Отслеживание динамики временных изменений в кластере предполагает использование графов, у которых ребрам присвоено направление. Эти графы отражают компоненты функционирования кластеров на различных временных периодах.

Обозначим $ct = \left(c_i^t\right)$ - количество кластеров в научных трудах молодых ученых, связанных с техническим направлением на конференции, где t – период времени, $i = 1$, k_t – количество кластеров.

Для каждого кластера присутствует метка, указывающая на его содержание, относящееся к определенной тематике. Метка подразумевает, что внутри кластера имеется 2-3 термина, которые связаны с наибольшим количеством ссылок.

В момент перехода между временными интервалами, например от отрезка времени t к $t+1$ каждая группа значений c_i^t преобразовывается в множество $\left(c_j^{t+1}\right), j = 1, \ldots, K_{t+1}$. Переходя от временного периода $t - 1$ к t (по убыванию) каждый кластер c^t_I видоизменяется из упорядоченного множества $\left(c_j^{t-1}\right), j = 1, \ldots, K_{t-1}$. В период осуществления перехода к времени $t+1$ или $t-1$ в рассматриваются и учитываются изменения кластеров, в соответствии с cd – типом развития, cf – типом формирования, Δ – разнице объемов общих составляющих кластеров:

$$sf: \langle c_i^t \cap c_1^{t+1} \rangle / \langle c_i^t \rangle - \langle c_i^t \cap c_n^{t+1} \rangle / \langle c_i^t \rangle = \Delta_{1,n};$$

$$sd: \left\langle c_i^t \cap c_1^{t-1} \right\rangle / \left\langle c_i^t \right\rangle - \left\langle c_i^t \cap c_n^{t-1} \right\rangle / \left\langle c_i^t \right\rangle = \Delta_{1,n}$$

где $\cap$ – пересечение множеств.

1 тип c_i^t видоизменился в c_i^{t+1}, так $\Delta_{1,2} > 3$;

2 тип c_i^t преобразовался в c_i^{t+1} и c_i^{t-1}, так $\Delta_{1,2} \leq 3$ и $\Delta_{1,3} > p$;

3 число кластеров, в которые преобразовался кластер c_i^t, превышает два $j > 2$, $\Delta_{1,j} \leq p$;

4 тип c_i^t отсутствует во множестве $\left(c_j^{t+1}\right)$ значительной частью элементов $\Delta_{0,1} > p$, $c_0^t = c_i^t / c^{t+1}$.

1 тип c_i^t сформировался из $\left(c_j^{t+1}\right)$, $\Delta_{1,2} > p$;

2 тип c_i^t сформировался из $\left(c_j^{t+1}\right)$ и c_{2j}^{t+1}, $\Delta_{1,2} \leq p$ и $\Delta_{1,3} > p$;

3тип. Количество терминов, из которых сформировался кластер c_t, превышает значение 2 (объединение множеств) > 2, $\Delta_{1,j} \leq p$;

4 тип c_i^t сформировался из терминов, не имеющих отношение к множеству $\left(c_j^{t+1}\right)$ ($\Delta_{0.2} > p$, $c_0^t = \left(c_j^t / c_j^{t-1}\right)$), где p является порогом.

cd = 1 и *cf = 1* относятся к темам, сохранившим обновление, *cd = 2* и *cf = 2*.

cd = 3 и *cd = 4* относятся к прекращению тем.

cf = 3 и *cf = 4* относятся к формированию новых тем, опубликованных в более ранних промежутках времени, до тех, которые ранее нигде не публиковались.

Выполнение условий отбора гарантирует правильное формирование кластеров.

Апробация кластеризации терминов проведена с помощью использования программы VOSViewer.

Заданы параметры:

1. К выбору объектов применен параметр «ключевого слова». По Стоп-словарю, отфильтрованы наиболее встречающиеся термины.
2. Присоединялись термины, повторяющиеся в более чем 10 научных статьях.
3. Количество объединений нескольких однородных элементов определялось выбором значения параметра «разрешение». В исследовании, значение равно 1.10.

Таблица 1 – Числовые параметры характеристики текстов

	$t = 1$ 2010	$t = 2$ 2012	$t = 3$ 2014	$t = 4$ 2016	$t = 5$ 2018	$t = 6$ 2020	Показатели общего периода (T)
m_t	951	1496	1259	1133	1571	1080	7490
M_t	27465	1041884	852446	676408	1463022	801662	5363977
n_t	819	1693	139	1131	2530	1322	3908

Результаты анализа зависят от выбранных параметров для изучения частоты совместного употребления терминов. Анализируются высокочастотные термины, связанные с различными трактовками по теме, основанными на обширных концепциях технической направленности. В исследовании использовались научные тексты по техническим темам за период с 2010 по 2020 годы. Обозначим:

T - количество конференций,

t - количество рассматриваемых периодов,

m_t - количество рассмотренных докладов,

M_t - объем текстов/докладов в словесных формулировках,

n_t - количество терминов по годам.

При применении математических и статистических методов анализа книг, периодических изданий и других публикаций, включая характеристики, подвергшиеся анализу, отмечаются пиковые значения количества публикаций [10]. За периоды $t = 2$ и $t = 5$ наблюдается значительный рост значений.

С целью проведения анализа для разных временных промежутков используются стратегические диаграммы, представляющие собой координатную плоскость, разделенную на четыре квадранта. Будем использовать значимость в упорядоченном виде по убыванию в качестве единицы измерения. На вертикальной оси d в диаграмме отображается значимость плотности, а на горизонтальной оси c - центральности кластеров. Каждый термин представлен в виде окружности, где непрерывная линия границы показывает зависимость от количества научных работ по техническим темам, включенных в выбранный кластер, в то время как прерывистая линия указывает на выборку меньшего числа текстов.

Например, кластер "Фотокатализ" переместился из 2-го квадранта в 1-й, но доля выборки технических текстов показала незначительное изменение. Это указывает на то, что связь между кластером и другими темами укрепилась. Кластеры "Кислотность, цеолит" и "Катализ" преобразовались и переместились в 4-й квадрант. Возможности развития нового кластера увеличились.

Кластер «Арретирование» преобразовался в «Освобождение», и теперь он содержит небольшую долю текстов, связанную как с внутренними, так и с внешними терминами относительно тематических кластеров. Кластер «Масло, Вязкость» появился в 2020 году, но не имел значительной активности, в то время как кластер «ZSM, СОЭ» не проявился в изученных научных текстах. Кластер «Метан» остался стабильным и обновился до «Метан, водород», находясь в первом квадранте.

При формировании объединения однородных элементов на соседних временных периодах (T) чаще всего происходят трансформации типов $cd=1$ и $cf=1$. Изменения в технических

тематических текстах *cd=2* и *cf=2* происходят в 12% случаев, удаление в 14%, а возникновение в 36% случаев. Наибольшее количество обновлений отмечается в периоде, где происходит увеличение числа опубликованных научных статей: *t=2* и *t=5*. 5 и 6 кластеры считаются новыми, так как они содержат более 89% новых терминов.

Математические схемы развития и формирования кластеров представляют собой стратегии или модели, которые применяются для анализа и визуализации развития и формирования кластеров. В этих схемах объекты, имеющие парные связи, представлены в виде узлов, а взаимосвязи между ними - стрелками. Тип развития определяет внешний вид и характеристики стрелок и узлов в схеме. Обычно, тип развития отображается через толщину стрелок и форму узлов в соответствии с физическими или статистическими свойствами объектов кластера. Математические схемы развития и формирования кластеров, объекты которых имеют парные связи, строятся в виде ориентированных графов. Тип развития определяет толщина стрелок и форма узлов.

Не имеющие входящих стрелок вершины демонстрируют новый сформированный кластер в конкретном промежутке времени. Важно в процессе анализа выявить не только близкие кластеры, но и удаленные кластеры.

Разработанная система реализации типов развития и формирования кластера (*cd* и *cf*) за десятилетний период 2010 - 2020 г. показала:

- сохранение темы (*cf =1 cd=1*) в 24 % случаев из 100;
- в 23% случаев была использована схема *cf =f4 cd = 1*, которая предусматривает создание новой темы с сохранением предыдущей;
- 15% от общего числа случаев характеризуется кратковременным существованием темы *cf = 4 cd = 4* (15%). Он характеризует случаи кратковременного существования темы. Этот тип был наиболее активен в период с 2012 по 2018 год. Некоторые из 16 типов развития кластера (6) проявили себя незначительно *cf =f1 cd =4* и *cf =f3 cd = 4*.

Технические тексты, сохраняющие единую тему в отдельные периоды времени, реже прекращают существование, в сравнении с кластерами, созданными из нескольких терминов. Им свойственно с течением времени трансформироваться.

Заключение

Проведен анализ полных текстов на основе построенных на шести коллекциях научных статей технической направленности, представленных молодыми учеными на конференциях 2010-2020 гг. Для построения наглядных карт и визуализации библиографических данных использовался программный инструмент VOSViewer. На уровне формирования кластеров в их поведении и дальнейшем развитии для каждого периода времени выявлены закономерности, построены графы формирования терминов для каждого рассматриваемого

временного периода, проведен анализ стратегических диаграмм начального и завершающего периодов. Дальнейшее развитие основано на наблюдении и анализе поведения кластеров, сформированных на узких специфических терминах с низкой частотой ключевых запросов. Кластерный анализ рейтинговых публикаций в тематических журналах, вызывает особый интерес и может быть использован в дальнейших исследованиях для выявления изменений в направлениях публикуемых научных исследований.

Список литературы

1. Callon, M., Courtial, J.P., Laville, F. Co-word analysis as a tool for describing the network of interaction between basic and technological research: the case of polymer chemistry // Scientometrics. 1991. N 22, P. 155– 205.
2. Кандилас, В., Апхэм, С. П., Унгар, Л. Х. Анализ сообществ знаний с использованием кластеров переднего плана и фона. [Электронный ресурс]. http://citeseerx.ist.psu.edu/ viewdoc/download?doi=10.1.1.146.3141&rep= rep1&type=pdf (дата обращения: 02.02.2024).
3. Ван Экк, Н.Дж., Уолтман, Л. Визуализация библиометрических сетей [Электронный ресурс]. https://link.springer.com/chapter/10.1007 (дата обращения: 02.02.2024).
4. Betancourt, C., Chen, W.-H. Deep reinforcement learning for portfolio management of markets with a dynamic number of assets. // Expert Systems with Applications, Volume 164, February 2021, 114002 - URL: .https://doi.org/10.1016/j.eswa.2020.114002.
5. Cobo, M.J., Lypez-Herrera, A.G., Herrera-Viedma, E., Herrera, F. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field [Электронный ресурс]. https://www.sciencedirect.com/science/article/pii/ S1751157710000891 (дата обращения: 02.02.2024).
6. Нгуен, Т.Т. и др. Многоцелевая система обучения с глубоким подкреплением // Инженерные приложения искусственного интеллекта, Том 96, ноябрь 2020, 103915. [Электронный ресурс]. - https://doi.org/10.1016/j.engappai.2020.103915(дата обращения: 02.02.2024).
7. Пальмов, С.В., Артюшкина, Е.С. Глубокое обучение: определение и отличительные особенности. //Форум молодых ученых. 2020. № 3 (43). С. 311-316.
8. Чистова, Е.В., Шелманов, А.О., Смирнов И.В. Применение глубокого обучения к моделированию диалога на естественном языке. // Труды Института системного анализа Российской академии наук. 2019. Т. 69. № 1. С. 105-115.
9. Потемкин, А.В. Обработка разнородной информации с помощью глубокого обучения нейронных сетей. // Мягкие измерения и вычисления. 2019. № 9 (22). С. 44-48.
10. Малый, Н. Отслеживание и прогнозирование областей роста в науке [Электронный ресурс]. http://www.scimaps.org/exhibit/docs/small.pdf (дата обращения: 02.02.2024).