социально-экономической географии, Пермский государственный национальный исследовательский университет;
Россия, 614990, Пермь, ул. Букирева, 15

Department of Socio-Economic Geography, Perm State University;
15, Bukireva st., Perm, 614990, Russia
e-mail: l.chekmeneva@mail.ru

Просьба ссылаться на эту статью в русскоязычных источниках следующим образом:
Балина Т.А., Николаев Р.С., Осоргин К.С., Пономарева З.С., Столбов В.А., Чекменева Л.Ю. Эволюция научных подходов к районированию Пермского края: теоретические и методологические аспекты // Географический вестник = Geographical bulletin. 2021. №3(58). С. 45–62. doi:10.17072/2079-7877-2021-3-45-62.
Please cite this article in English as:
Balina, T.A., Nikolaev, R.S., Osorgin, K.S., Ponomareva, Z.V., Stolbov, V.A., Chekmeneva, L.Yu. (2021). Evolution of scientific approaches to the zoning of Perm Krai: theoretical and methodological aspects. *Geographical bulletin*. No. 3(58). Pp. 45–62. doi: 10.17072/2079-7877-2021-3-45-62.

УДК 911.3 DOI: 10.17072/2079-7877-2021-3-62-73

ЦИФРОВОЕ РАЗВИТИЕ СИБИРСКОГО ФЕДЕРАЛЬНОГО ОКРУГА: КЛАСТЕРИЗАЦИЯ РЕГИОНОВ В ОБЛАКЕ ТЕГОВ

Виктор Иванович Блануца
ORCID: http://orcid.org/0000-0003-3958-216X
e-mail: blanutsa@list.ru
Институт географии им. В.Б. Сочавы СО РАН, г. Иркутск, Россия

Свободный доступ к потоку региональных новостей и онлайн-генераторам облака тегов открывает перед социально-экономической географией новые возможности по обработке «больших данных» и выявлению географических закономерностей. Целью исследования стала идентификация современных приоритетов цифрового развития десяти сибирских регионов с помощью их кластеризации в облаке тегов из потока официальных региональных новостей. Сформулированы исходная и альтернативная гипотезы исследования. Установлены двенадцать тегов (ярлыков, ключевых слов), отражающих приоритеты цифрового развития. На основе потока новостей (текстов) за первые пять месяцев 2021 г. от региональных министерств цифрового развития созданы облака из наиболее встречающихся тегов. Заданы пять полос частоты встречаемости тегов. Предложены мера расстояния между регионами в облаке тегов и алгоритм объединения регионов в кластеры. В результате проведенного исследования установлено, что в Сибири нет регионов с одинаковыми приоритетами цифрового развития, а имеющиеся различия позволяют объединить все регионы в два кластера. На этом основании исходная гипотеза о единообразии приоритетов цифрового развития всех регионов была отклонена. Перечислены десять особенностей цифрового развития Сибири. Представлены направления дальнейших исследований по данной проблематике.

Ключевые слова: социально-экономическая география, большие данные, поток новостей, цифровое развитие, облако тегов, кластерный анализ, Сибирский федеральный округ.

DIGITAL DEVELOPMENT OF THE SIBERIAN FEDERAL DISTRICT: CLUSTERING OF REGIONS IN A TAG CLOUD

Victor I. Blanutsa
ORCID: http://orcid.org/0000-0003-3958-216X
e-mail: blanutsa@list.ru
V.B. Sochava Institute of Geography, Siberian Branch of the RAS, Irkutsk, Russia

Free access to the flow of regional news and online generators of tag clouds opens up new opportunities for human geography with regard to processing big data and detecting geographical patterns. The aim of the study was to identify the

current digital development priorities of ten Siberian regions by clustering them in a tag cloud from the stream of official regional news. There were identified twelve tags (labels, keywords) that reflect the priorities of digital development. Based on the flow of news (texts) from the regional ministries of digital development for the first five months of 2021, clouds of the most common tags were created. Five bands of tag frequency were set. A measure of the distance between the regions in the tag cloud and an algorithm for grouping regions into clusters were proposed. It has been found that there are no regions in Siberia with the same priorities for digital development; the existing differences allow us to group all the regions into two clusters. On this basis, the initial hypothesis about the uniformity of digital development priorities in all the regions has been rejected. Ten features of the digital development of Siberia are listed. The directions of further research on this issue are presented.

K e y w o r d s : human geography, big data, news flow, digital development, tag cloud, cluster analysis, Siberian Federal District.

Введение

Вхождение в эпоху «больших данных» ставит перед социально-экономической географией проблему осмысления и пространственного количественного анализа (в режиме реального времени) непрерывного потока разнообразной информации, включая неструктурированные качественные сведения [2; 9; 11–13]. Первой попыткой осмысления новых источников территориально-распределенных качественных данных можно считать исследование «Горожане как датчики», заложившее основы анализа «добровольной географической информации» [10]. Однако географические данные можно извлекать не только из социальных сетей и отчетов операторов сотовой связи, но и из потоков новостей в виде текстов. Основы количественного измерения параметров текста заложены в контент-анализе [1; 7; 8; 17]. Дальнейшее развитие оценки частоты встречаемости слов в тексте привело к созданию онлайн-генераторов облака тегов (от англ. tag – ярлык, метка, ключевое слово). К настоящему времени наиболее известными в мире сетевыми ресурсами построения облака тегов являются www.wordclouds.com, www.tagxedo.com, www.wordart.com, www.imagechef.com, www.worditout.com, www.wordcloud.pro и www.wordshift.org, а в нашей стране – www.облакослов.рф. Под облаком тегов понимается визуальное представление списка ярлыков, в котором размер шрифта каждого ярлыка пропорционален частоте его встречаемости в тексте. Обычно облако тегов используется для визуализации особенностей стиля выступления политиков, официальных документов, научных публикаций и других текстов, прямо не связанных с географическими данными.

Относительно недавно появились исследования по созданию облаков тегов, связанных с конкретными географическими местами. Анализ одной российской (www.elibrary.ru) и семи международных (www.link.springer.com, www.onlinelibrary.wiley.com, www.sciencedirect.com, www.login.webofknowledge.com, www.scopus.com, www.journals.sagepub.com, www.ideas.repec.org) библиографических баз данных позволил вывить 12 журнальных статей по этой проблематике, первая из которых была опубликована в 2011 г., а остальные – в 2015–2020 гг. В первой работе изучалась частота встречаемости фамилий в территориальных ячейках Великобритании (кроме ячеек Северной Ирландии), кластеризация которых позволила выявить 20 «фамильных регионов» с разными облаками наиболее встречающихся фамилий (тегов) [16]. Среди остальных исследований можно отметить разделение территории города Шанхая на «подпространства», расположенные между линиями метро и имеющие специфические облака тегов из онлайн-текстов географического описания «точек интересов» [14], оценку демографических аспектов по округам США на основе геотегов из Twitter [5], анализ пространственных взаимодействий между туристами и местными жителями в десяти городах США [15], рейтинг туристических дестинаций в американских городах с помощью кластеризации географических мест в облаке тегов [19], изучение поведения людей в различных частях города по геотегам социальных сетей [20] и выявление восприятия городской идентичности посредством кластеризации облака тегов, полученных из электронных средств массовой информации

в городах США [6]. Среди работ, появившихся после 2020 г., целесообразно отметить англо-итальянское исследование по преобразованию фотографий восьми островов (Канарские острова и 4 острова в Средиземном море), взятых из Instagram, в облака тегов с заданием «индекса расстояния в изображениях» между островами [4].

Обобщение существующего опыта позволяет прийти к выводу, что под кластеризацией территорий (регионов) понимается их объединение в априори не заданное количество кластеров, внутри которых достигается максимальное сходство регионов по заданным параметрам (признакам), а между кластерами – максимальное различие регионов. Если в качестве параметров выступают теги, то такая группировка называется кластеризацией регионов в облаке тегов.

В отечественных географических исследованиях облака тегов, привязанные к конкретным территориям, не анализировались (по данным www.elibrary.ru). Поэтому в рамках цикла работ по географическому анализу цифрового развития сибирских регионов была поставлена цель идентификации различия современных (2021 г.) приоритетов цифровизации десяти регионов Сибирского федерального округа с помощью кластеризации территорий в облаке тегов из потока официальных региональных новостей. Данная цель трансформировалась в две гипотезы исследования – исходную и альтернативную. Первая из них заключалась в том, что отсутствуют различия между региональными приоритетами цифрового развития. Основанием для этого могут служить два априорных предположения: региональные министерства цифрового развития выстраивают свою деятельность в соответствии с распоряжениями, поступающими из Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации, и в силу этого приоритеты развития всех регионов должны быть одинаковыми (все территории входят в один кластер); анализируемые регионы составляют единое целое – Сибирский федеральный округ, в котором приоритеты не могут быть различными. Тогда альтернативная гипотеза заключается в наличии региональных различий, приводящих к формированию нескольких кластеров (групп регионов со сходными приоритетами цифрового развития).

Методы исследования

В структуре правительства каждого сибирского региона имеется организация, отвечающая за цифровое развитие. В трех регионах такая организация называется «Министерство цифрового развития и связи» (Алтайский край, Кемеровская и Новосибирская области), в двух – «Министерство цифрового развития» (Республика Алтай и Красноярский край), а в оставшихся пяти регионах называется по-разному (Министерство информации и связи Республики Тыва, Государственный комитет цифрового развития и связи Республики Хакасия, Министерство экономического развития Иркутской области, Министерство промышленности, связи, цифрового и научно-технического развития Омской области и Департамент цифровой трансформации Администрации Томской области). Данные организации проводят различные мероприятия, способствующие цифровому развитию региона. Информация об этом размещается в разделе «Новости» на официальных сайтах (www.digital.altayreg.ru, www.mis.rtyva.ru, www.r-19.ru/authorities/executive-authorities/state-committee-of-digital-development/, www.digital.alregn.ru, www.it.krskstate.ru, www.irkobl.ru/sites/economy/?type=special, www.digital42.ru, www.digit.nso.ru, www.mps.omskportal.ru/oiv/mps/, www.digital.tomsk.gov.ru). Поток новостей (текстов) с перечисленных сайтов составил исходную базу данных для выявления региональных приоритетов цифрового развития.

Анализ сайта Минцифры России (www.digital.gov.ru) показал, что федеральное министерство переходит от реализации государственной программы Российской Федерации «Информационное общество (2011–2020 гг)» к выполнению национальной программы

«Цифровая экономика Российской Федерации». При этом в текущей деятельности (2021 г.) министерство нацелено на цифровизацию всех видов человеческой деятельности в России, включая достижение цифровой зрелости отраслей экономики и социальной сферы, а также повышения цифровой грамотности россиян. Это приведет к цифровой трансформации не только отраслей, но и регионов, что потребует развития информационной инфраструктуры для устранения цифрового неравенства, обеспечения информационной безопасности для информационных систем, создания новых информационных технологий и распространения услуг цифрового государственного управления. При таком понимании цифрового развития в потоке региональных новостей целесообразно выделять 12 тегов (перечислены в алфавитном порядке; каждый тег не имеет пробелов между словами): ИнформационнаяБезопасность, ИнформационнаяИнфраструктура, ИнформационноеОбщество, ИнформационнаяСистема, ИнформационнаяТехнология, ЦифроваяГрамотность, ЦифроваяЗрелость, ЦифроваяТрансформация, ЦифроваяЭкономика, Цифровизация, ЦифровоеГосударственноеУправление и ЦифровоеНеравенство. Некоторые из перечисленных тегов могли принимать иной вид (например, ЦифроваяТехнология вместо ИнформационнаяТехнология), но в редакторе новостей приводились к принятому виду. Следует также отметить, что сроки рассмотрения федерального проекта «Цифровой регион» сдвинулись на третий квартал 2021 г. (по сведениямwww.tadviser.ru/index.php/Статья:Федеральный_проект_Цифровой_регион), что исключило тег ЦифровойРегион из текущего анализа.

Облако тегов создается с помощью онлайн-генератора (основные из них перечислены во введении), в настройках которого можно выбрать геометрическую форму облака (квадрат, круг, овал и др.), палитру тегов, их взаимное расположение (горизонтальное или с различными углами наклона), количество визуализируемых тегов, минимальную частоту их встречаемости и «стоп-слова» (вспомогательные слова, не отображаемые в облаке). В онлайн-генератор вводятся текст или ссылка на сетевой ресурс с необходимыми текстами, после чего нажимается кнопка «Сгенерировать облако». Результатом генерации являются облако и список тегов с указанием частоты встречаемости. До или после создания облака (в зависимости от особенностей онлайн-генератора) производится редактирование (исходного текста или итогового облака). Оно призвано объединить устойчивые словосочетания в теги и представить их в именительном падеже единственного числа (например, «цифровую технологию», «информационной технологии» и «цифровые технологии» объединяют в тег ИнформационнаяТехнология, а «Иркутской области» и «Иркутской областью» – в ИркутскаяОбласть).

Поскольку на сайтах министерств цифрового развития разных регионов размещаются неодинаковые объемы текста, то при выявлении приоритетов целесообразно отказаться от абсолютных значений частоты встречаемости анализируемых тегов. Международный опыт создания облаков тегов, привязанных к конкретным территориям, показывает, что наиболее значимо попадание тега в первую сотню, а далее частота встречаемости тегов имеет наименьшие значения и в силу этого нивелируются различия между тегами. При этом в разных исследованиях подчеркивается особая значимость первых 10 [18], 15 [16], 20 [4], 25 [5] или 50 [18] наиболее встречающихся тегов. Поэтому имеет смысл разделить первую сотню наиболее встречающихся тегов на четыре квартиля, а все остальные теги (с порядковыми номерами 101 и более) признать равнозначными. Тогда в потоке новостей от каждого региона будем отслеживать попадание 12 анализируемых тегов цифрового развития в пять полос частоты встречаемости с порядковыми номерами тегов 1–25, 26–50, 51–75, 76–100, 101 и более. Теги, попавшие в двух сравниваемых регионах в одну и ту же полосу значений в своем региональном облаке, будут считаться равнозначными, а теги из разных полос – неравнозначными с приоритетностью тега из полосы с более высокой частотой

встречаемости. При такой оценке расстояние d(X, Y) между регионами X и Y будет равно сумме различий между частотными полосами 12 анализируемых тегов. Например, в двух сравниваемых регионах 10 тегов относятся к последней – пятой – полосе, а два тега – к первой (в одном регионе) и третьей (в другом регионе) полосам, что приводит к межрегиональному расстоянию d(X, Y) = 4 $(10 \times (5-5) + 3 - 1 + 3 - 1)$. Здесь максимально возможное расстояние достигается в том случае, когда все теги одного региона попадают в первую полосу, а другого региона – в пятую полосу $(12 \times (5 - 1) = 48)$.

При анализе облаков тегов обычно используется кластерный анализ [6; 16; 18]. Возможны разные способы кластеризации регионов [2]. С учетом специфики расстояния d(X, Y), предлагается следующий алгоритм объединения (группировки) регионов в кластеры: задается группировочный шаг $(\Delta d = d(X,Y)_{max} \div 10)$; на первом шаге объединяются регионы X и Y в том случае, если d(X, Y)≤ Δd, а к ним присоединяется любой регион Z при $d(X,Z) \leq \Delta d$ и $d(Y,Z) \leq \Delta d$; такие операции повторяются до тех пор, пока на первом шаге не будут объединены регионы, расстояние между которыми не превышает Δd; на втором и последующих шагах эти операции повторяются для kΔd, где k – порядковый номер группировочного шага. В данном алгоритме на десятом шаге все регионы объединяются в один кластер. Для визуализации последовательности группировки регионов в кластеры обычно используется дендрограмма (древовидная диаграмма), на которой в нашем случае может быть представлено максимум 10 вариантов кластеризации. Среди них необходимо выбрать наиболее оптимальный. Здесь можно использовать теоретико-графовый подход к выбору наилучшего варианта, которому соответствует наиболее сложный ярус леса (множества древовидных графов) или шаг группировки [2, с. 93–96]. Если наибольшая сложность достигается на последнем шаге (образуется один кластер), то подтверждается исходная гипотеза исследования; если это происходит на любом другом шаге (несколько кластеров), то исходная гипотеза отклоняется и принимается альтернативная гипотеза.

Результаты и обсуждение

Для создания облака тегов по каждому сибирскому региону использовался поток новостей от регионального министерства цифрового развития за первые пять месяцев (январь–май) 2021 г. Из потока были исключены все даты, цифры, числительные, частицы, местоимения и вспомогательные слова без содержательной нагрузки. Тексты были загружены в два онлайн-генератора (www.wordclouds.com и www.облакослов.рф), которые одинаковым образом распределили 12 анализируемых тегов по 5 полосам частоты встречаемости тегов в каждом региональном потоке новостей. К примеру, по Новосибирской области получилась следующая последовательность тегов цифрового развития (знаками «=» и «>» отмечено вхождение тегов в одну или разные полосы): ЦифроваяТрансформация = ЦифроваяЭкономика > Цифровизация > ИнформационнаяТехнология = ЦифроваяЗрелость > остальные равнозначные теги (полосы 1, 2, 3 и 5 разделяются знаком «>»). Результаты по всем сибирским регионам сведены в табл. 1. На основе этих данных можно сделать следующие выводы о приоритетах цифрового развития Сибири: (а) вне основного внимания (в пятой полосе) во всех регионах оказались информационное общество и цифровое государственное управление; (б) наибольший интерес (попадание в первую полосу) представляло развитие цифровой экономики, информационных технологий и цифровой трансформации; (в) отсутствовали регионы, уделяющие основное внимание (первые четыре полосы) всем анализируемым тегам; (г) в двух регионах только к одному тегу проявили интерес (вне пятой полосы); (д) отсутствовало максимальное внимание к рассматриваемой проблематике (первая полоса) в двух регионах.

По первым ста тегам региональных облаков видно (рис. 1), что наиболее часто (максимальный размер шрифта в каждом облаке) в потоке новостей цифрового развития упоминались названия регионов (в 7 случаях), а также «МФЦ» (многофункциональный центр; Томская область), «края» (Красноярский край) и «связи» (Республика Тыва). Если эти данные сопоставить с приоритетностью 12 анализируемых тегов, то можно прийти к следующим региональным выводам: (е) в Республике Алтай основное внимание уделялось не цифровому развитию (только тег ИнформационаяСистема в третьей полосе), а развитию «связи» (максимальная частота встречаемости в потоке новостей); (ж) в Томской области при повышенном внимании к МФЦ в первую полосу попали и информационные технологии; (з) в Омской области половина тегов цифрового развития вошла в первые четыре полосы, но они реже использовались в новостях, чем «развития», «связи», «данных» и «трека»; (и) в Красноярском крае теги цифрового развития находились в подчиненном состоянии относительно общего развития края; (к) в Республике Хакасия чаще упоминали правительство и документы, чем цифровое развитие, в котором акцент сделан на СЭД (система электронного документооборота).

Таблица 1

Полосы частоты встречаемости (от 1 до 5) тегов в потоке новостей (январь – май 2021 г.) на официальных сайтах министерств цифрового развития в регионах Сибирского федерального округа (составлено автором)
Frequency bands (from 1 to 5) of tags in the news stream (January – May 2021) on the official websites of the digital development ministries in the regions of the Siberian Federal District (compiled by the author)

Теги	*Частота встречаемости*									
	Регион									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
ИнформационнаяБезопасность	5	5	5	5	3	2	3	5	3	5
ИнформационнаяИнфраструктура	5	5	4	5	5	4	5	5	5	5
ИнформационноеОбщество	5	5	5	5	5	5	5	5	5	5
ИнформационнаяСистема	3	5	2	5	5	5	5	5	5	5
ИнформационнаяТехнология	5	4	3	2	1	1	1	3	1	1
ЦифроваяГрамотность	5	5	5	5	5	5	4	5	5	5
ЦифроваяЗрелость	5	5	5	5	5	5	5	3	5	5
ЦифроваяТрансформация	5	2	5	1	1	2	1	1	2	5
ЦифроваяЭкономика	5	1	3	1	1	1	1	1	4	5
Цифровизация	5	5	5	3	2	5	2	2	2	5
ЦифровоеГосударственноеУправление	5	5	5	5	5	5	5	5	5	5
ЦифровоеНеравенство	5	5	5	5	5	5	5	5	4	5

Примечание. Регионы: 1 – Республика Алтай, 2 – Республика Тыва, 3 – Республика Хакасия, 4 – Алтайский край, 5 – Красноярский край, 6 – Иркутская область, 7 – Кемеровская область – Кузбасс, 8 – Новосибирская область, 9 – Омская область, 10 – Томская область.

Note. Regions: 1 – Republic of Altai, 2 – Republic of Tyva, 3 – Republic of Khakassia, 4 – Altai Territory, 5 – Krasnoyarsk Territory, 6 – Irkutsk Region, 7 – Kemerovo Region – Kuzbass, 8 – Novosibirsk Region, 9 – Omsk Region, 10 – Tomsk Region.

По приоритетам цифрового развития видно (см. табл. 1), что наименьшее различие между регионами Сибири отмечалось между Красноярским краем и Кемеровской областью (отличие на одну полосу по тегу ЦифроваяГрамотность, что привело к $d(X, Y) = 1$), а наибольшее – между Республикой Алтай и Кемеровской областью ($d(X, Y) = 20$). Для удобства интерпретации абсолютные значения расстояния между регионами по 12 тегам цифрового развития были переведены в относительные величины (путем деления на максимальное расстояние) и сведены в единую матрицу (рис. 2), которая использовалась для кластеризации регионов и, соответственно, проверки исходной гипотезы исследования.

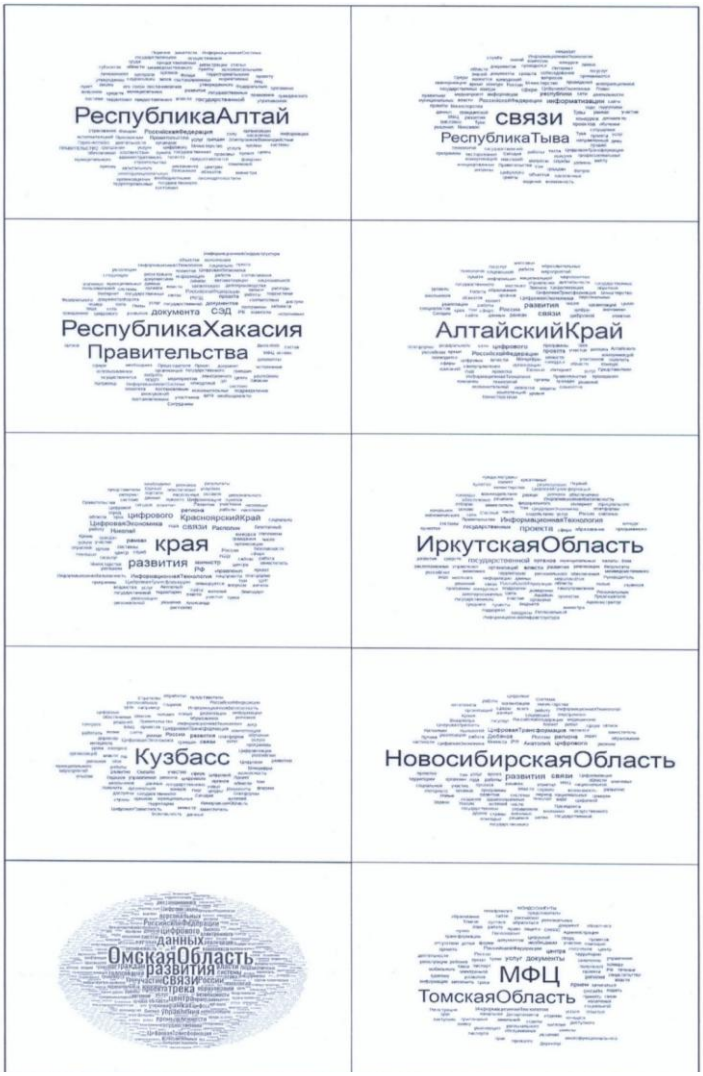

Рис. 1. Облака тегов сибирских регионов, полученные из потока новостей региональных министерств цифрового развития (январь – май 2021 г.) с помощью онлайн-генераторов www.wordclouds.com (для Омской области) и www.облакослов.рф (для остальных регионов)

Fig. 1. Tag clouds of Siberian regions obtained from the news stream of the regional ministries of digital development (January – May 2021) using online generators www.wordclouds.com (for the Omsk region) and www.облакослов.рф (for the other regions)

При величине группировочного шага $\Delta d = 0{,}10$ кластеризация регионов началась с объединения Красноярского края и Кемеровской области ($d(X, Y) = 0{,}05$), после чего Алтайский край объединился с Новосибирской областью ($d(X, Y) = 0{,}20$), а Республика Алтай – с Республикой Хакасия ($d(X, Y) = 0{,}30$). Последовательность группировки сибирских регионов в кластеры показана на дендрограмме (рис. 2). Из 10 шагов на 7 появлялись новые варианты кластеризации. Если провести расчеты сложности графов, получаемых путем разрезания дендрограммы горизонтальной линией на каждом шаге (методика приведена в [2]), то наиболее оптимальным вариантом будет выделение двух кластеров на шестом шаге (расстояние между регионами внутри кластеров составляет не более 0,60; сложность оценивается в 0,354). Сравнение этого варианта с объединением всех регионов в один кластер (0,328) дает основание ($0{,}354 > 0{,}328$) отклонить исходную гипотезу исследования. Таким образом, установлено, что в Сибирском федеральном округе между регионами наблюдались (январь–май 2021 г.) различия в приоритетах цифрового развития, подтверждающие альтернативную гипотезу исследования.

Таблица 2

Симметричная матрица расстояний (в относительных единицах) между сибирскими регионами по двенадцати тегам цифрового развития из потока новостей от региональных министерств за первые пять месяцев 2021 г. (составлено автором)
Symmetric matrix of distances (in relative units) between Siberian regions for twelve tags of digital development from the regional ministries' news stream for the first five months of 2021 (compiled by the author)

Матрица расстояний										
Регион	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	–	0,50	0,30	0,75	0,95	0,85	1,00	0,85	0,80	0,30
2	0,50	–	0,50	0,25	0,45	0,35	0,50	0,35	0,60	0,50
3	0,30	0,50	–	0,65	0,85	0,65	0,90	0,75	0,80	0,40
4	0,75	0,25	0,65	–	0,20	0,40	0,25	0,20	0,45	0,60
5	0,95	0,45	0,85	0,20	–	0,30	0,05	0,30	0,25	0,65
6	0,85	0,35	0,65	0,40	0,30	–	0,35	0,60	0,45	0,55
7	1,00	0,50	0,90	0,25	0,05	0,35	–	0,35	0,30	0,70
8	0,85	0,35	0,75	0,20	0,30	0,60	0,35	–	0,55	0,75
9	0,80	0,60	0,80	0,45	0,25	0,45	0,30	0,55	–	0,50
10	0,30	0,50	0,40	0,60	0,65	0,55	0,70	0,75	0,50	–

Примечание. Номера регионов приведены по табл. 1.
Note. The region numbers correspond to those specified in Table 1.

В первый кластер вошли семь регионов, для которых весьма важно цифровое развитие. Если номера частотных полос всех тегов для этих регионов суммировать, то получится следующая последовательность предпочтений в первом кластере: ЦифроваяТрансформация = ЦифроваяЭкономика > ИнформационнаяТехнология > Цифровизация > ИнформационнаяБезопасность > ЦифроваяЗрелость > ИнформационнаяИнфраструктура = ЦифроваяГрамотность = ЦифровоеНеравенство > ИнформационноеОбщество = ИнформационнаяСистема = ЦифровоеГосударственноеУправление. Для второго кластера, объединяющего Томскую область, Республики Алтай и Хакасия, характерны иные приоритеты: ИнформационнаяТехнология > ИнформационнаяСистема > ЦифроваяЭкономика > ИнформационнаяИнфраструктура >, остальные равнозначные теги. Для детализации региональных различий первый кластер может быть разделен на два подкластера (см. рис. 2): в один из них войдут Красноярский край и Кемеровская область с присоединившими к ним Омской и Иркутской областями, а в другой – Алтайский край и Новосибирская область с подключением Республики Тыва.

Обсуждение полученных результатов целесообразно проводить путем сравнения с ранее проведенными исследованиями по рассматриваемой проблематике. Однако приоритеты цифрового развития сибирских регионов ранее не выявлялись. Поэтому остается только косвенное сравнение полученной кластеризации с основными социально-экономическими показателями анализируемых регионов. Прямое сопоставление цифровых приоритетов с данными официальной статистики исключено по причине использования разных единиц измерения. В такой ситуации с некоторой условностью допустимо только сопоставление дендрограмм, полученных для одних и тех же регионов по единому алгоритму, но при разных мерах различия. Как пример, сравним наш результат с кластеризацией сибирских регионов по численности населения и валовому региональному продукту (ВРП) как основным социально-экономическим показателям. Официальные данные по этим показателям приведены в последнем статистическом сборнике «Регионы России» [3].

Максимальное различие по людности сибирских регионов составило примерно 2646 тыс. человек (между Красноярским краем и Республикой Алтай). Тогда по предлагаемой методике величина группировочного шага равна 264,6 тыс. человек,

а объединение всех регионов в один кластер осуществляется за 10 шагов. Последовательность группировки регионов представлена на дендрограмме (рис. 3). Наиболее оптимальный вариант (сложность графов составляет 0,276) достигается на четвертом шаге (расстояние между регионами внутри кластеров не превышает 0,40), в результате чего получаются два кластера. Их сравнение с кластеризацией регионов по приоритетам цифрового развития (см. рис. 2) позволяет утверждать, что результаты проведенного исследования на уровне кластеров почти совпадают с группировкой сибирских регионов по численности населения. Единственное различие связано с Республикой Тыва, которая на сравниваемых дендрограммах относится к разным кластерам. Однако сопоставление порядка объединения в кластеры указывает на существенные различия. Например, Алтайский край и Иркутская область по людности объединяются на первом шаге, а по цифровым приоритетам – на шестом шаге; Омская область по численности населения сначала присоединяется к Алтайскому краю и Иркутской области, а по цифровизации – к Красноярскому краю и Кемеровской области. Поэтому схождение регионов Сибири в кластеры по цифровым приоритетам только частично соответствует группировке этих же регионов по людности.

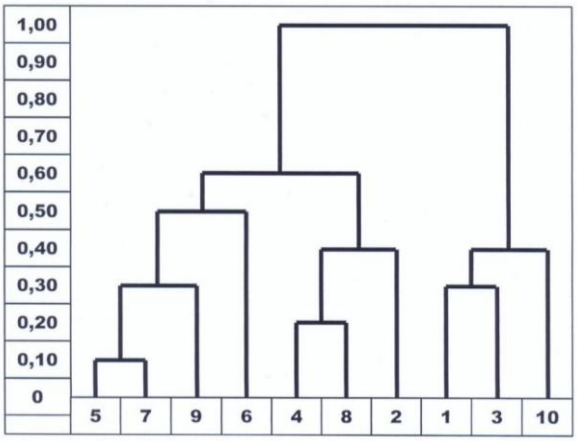

Рис. 2. Дендрограмма объединения сибирских регионов в кластеры по мере увеличения межрегионального расстояния в облаке тегов цифрового развития от 0 до 1,00 (составлено автором)
Регионы: 1 – Республика Алтай, 2 – Республика Тыва, 3 – Республика Хакасия, 4 – Алтайский край, 5 – Красноярский край, 6 – Иркутская область, 7 – Кемеровская область – Кузбасс, 8 – Новосибирская область, 9 – Омская область, 10 – Томская область
Fig. 2. Dendrogram of the clustering of Siberian regions as the interregional distance in the digital development tag cloud increases from 0 to 1.00 (compiled by the author)
Regions: 1 – Republic of Altai, 2 – Republic of Tyva, 3 – Republic of Khakassia, 4 – Altai Territory, 5 – Krasnoyarsk Territory, 6 – Irkutsk Region, 7 – Kemerovo Region – Kuzbass, 8 – Novosibirsk Region, 9 – Omsk Region, 10 – Tomsk Region.

Проделав аналогичные операции по ВРП, был получен иной результат (рис. 4). В этом случае оптимальный вариант группировки регионов (сложность 0,345) образовался на первом шаге (четыре кластера при внутрикластерном расстоянии между регионами до 0,10). Отличие от кластеризации сибирских регионов в облаке тегов цифрового развития (рис. 2) настолько велико, что позволило сделать вывод об отсутствии зависимости определения цифровых приоритетов от размера валового продукта. Сравнение с другими социально-экономическими показателями на данном этапе не проводилось. Поэтому в первом приближении можно утверждать, что последовательность объединения регионов в кластеры по приоритетам цифрового развития отличается от порядка группировки этих же регионов по основным социально-экономическим показателям, а по составу получаемых кластеров имеется только некоторое сходство с объединением регионов по численности населения.

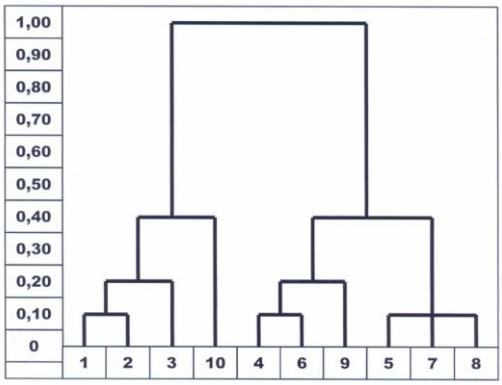

Рис. 3. Дендрограмма объединения сибирских регионов (номера приведены по рис. 2) в кластеры по численности населения (составлено автором по данным из [3])
Fig. 3 Dendrogram of the clustering of Siberian regions by population (numbers are given in Fig. 2) (compiled by the author according to data from [3])

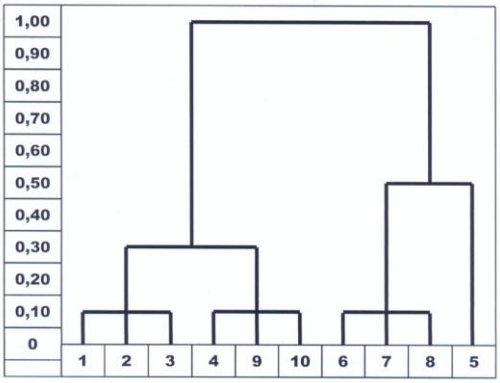

Рис. 4. Дендрограмма объединения сибирских регионов (номера приведены по рис. 2) в кластеры по валовому региональному продукту (составлено автором по данным из [3])
Fig. 4. Dendrogram of the clustering of Siberian regions (numbers are given in Fig. 2) by gross regional product (compiled by the author according to data from [3])

Заключение

Появление «больших данных» в виде потока новостей (текстов) от различных веб-сайтов и наличие свободного доступа к онлайн-генераторам облаков тегов открывают перед социально-экономической географией возможности получения нового знания. Ранее, например, для определения приоритетности региональных проблем использовались экспертные оценки и косвенные индексы (см. проблемное районирование [2, с. 98–104], которые были слишком субъективными. На сегодняшний день при переходе от отдельного субъективного текста к большим массивам текстов происходит нивелирование разнонаправленных оценок и выделение инвариантной структуры, отражающей действительность по закону больших чисел. Новые инструменты были апробированы на примере выявления приоритетов цифрового развития сибирских регионов. В результате проведенного исследования установлено, что в Сибири нет регионов с одинаковыми приоритетами ($d(X, Y) = 0$), а существующие различия в цифровом развитии позволяют объединить регионы в два кластера ($d(X, Y) \leq 0{,}60$).

Дальнейшие исследования могут быть связаны с масштабированием первого опыта оценки цифрового развития регионов через облако тегов (например, для всех 85 регионов России), расширением источников информации (не только сайты министерств, но и электронные средства массовой информации в регионах), увеличением количества анализируемых тегов, выявлением пространственно-временных изменений (к примеру, перераспределение регионов между кластерами по годам), созданием новых мер расстояния (сходства) между регионами в облаке тегов, использованием других методов кластерного анализа, привлечением искусственного интеллекта (глубокое машинное обучение) для автоматического распознавания тегов в виде длинных нечетких словосочетаний, разработкой способов верификации полученных результатов, встраиванием предлагаемой методики в более общие системы методов социально-экономической географии (районирования, географической экспертизы и др.) и формированием регулярно обновляемого рейтинга российских регионов по приоритетности цифрового развития.

Особые перспективы последующих исследований могут быть связаны с планируемым принятием федерального проекта «Цифровой регион». С географических позиций целесообразно зафиксировать ход реализации этого проекта в российских регионах в контексте других приоритетов цифрового развития, что можно сделать через анализ динамики изменения положения тега ЦифровойРегион в региональных облаках тегов. Это позволит не только дополнительно идентифицировать эволюцию кластеров

через сходство (различие) траекторий цифрового развития регионов, но и выявить «регионы-аналоги», которые придерживаются одних и тех же приоритетов развития на протяжении всего анализируемого периода времени.

Благодарности. *Исследование выполнено за счет средств государственного задания (№ регистрации темы АААА-А17-117041910166-3).*

Acknowledgements. *The study was funded as part of a state assignment, topic No. AAAA-A17-117041910166-3.*

Библиографический список

1. *Блануца В.И.* Социально-экономическое районирование как система смыслов: контент-анализ постсоветских публикаций // Географический вестник. 2017. № 4. С. 39–50. doi: 10.17072/2079-7877-2017-4-39-50.

2. *Блануца В.И.* Социально-экономическое районирование в эпоху больших данных. М.: ИНФРА-М, 2018. 194 с.

3. Регионы России. Социально-экономические показатели 2020 // Федеральная служба государственной статистики. URL: https://rosstat.gov.ru/folder/210/document/13204 (дата обращения: 24.05.2021).

4. *Arabadzhyan A., Figini P., Vici L.* Measuring destination image: A novel approach based on visual data mining. A methodological proposal and an application to European islands // Journal of Destination, Marketing & Management. 2021. Vol. 20. e100611. doi: 10.1016/j.dmm.2021.100611.

5. *Bokányi E., Kondor D., Dobos L., Sebõk T., Stéger J., Csabai I., Vattay G.* Race, religion and the city: Twitter word frequency patterns reveal dominant demographic dimensions in the United States // Palgrave Communications. 2016. Vol. 2. e16010. doi: 10.1057/palcomms.2016.10.

6. *De Oliveira Capela F., Ramirez-Marquez J.E.* Detecting urban identity perception via newspaper topic modeling // Cities. 2019. Vol. 93. P. 72–83. doi: 10.1016/j.cities.2019.04.009.

7. *Drisko J.M., Maschi T.* Content Analysis. Oxford: Oxford University Press, 2016. 191 pp.

8. *Feldman R., Sanger J.* The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. New York: Cambridge University Press, 2006. 422 pp.

9. *Ferreira D., Vale M.* Geography in the big data age: An overview of the historical resonance of current debates // Geographical Review. 2020. Vol. 110. e1832424. doi: 10.1080/00167428.2020.1832424.

10. *Goodchild M.F.* Citizens as sensors: The world of volunteered geography // GeoJournal. 2007. Vol. 69. P. 211–221. doi: 10.1007/s10708-007-9111-y.

11. *Graham M., Shelton T.* Geography and the future of big data, big data and the future of geography // Dialogues in Human Geography. 2013. Vol. 3. No. 3. P. 255–261. doi: 10.1177/2043820613513121.

12. *Kitchin R.* Big data and human geography: Opportunities, challenges and risks // Dialogues in Human Geography. 2013. Vol. 3. No. 3. P. 262–267. doi: 10.1177/2043820613513388.

13. *Kwan M.-P.* Algorithmic geographies: Big Data, algorithmic uncertainty, and the production of geographic knowledge // Annals of the American Association of Geographers. 2016. Vol. 106. No. 2. P. 274–282. doi: 10.1080/00045608.2015.1117937.

14. *Li C., Dong X., Yuan X.* Metro-Wordle: An interactive visualization for urban text distributions based on Wordle // Visual Informatics. 2018. Vol. 2. No. 1. P. 50–59. doi: 10.1016/j.visinf.2018.04.006.

15. *Li D., Zhou X., Wang M.* Analyzing and visualizing the spatial interactions between tourists and locals: A Flickr study in ten US cities // Cities. 2018. Vol. 74. P. 249–258. doi: 10.1016/j.cities.2017.12.012.

16. *Longley P.A., Cheshire J.A., Mateos P.* Creating a regional geography of Britain through the spatial analysis of surnames // Geoforum. 2011. Vol. 42. No. 4. P. 506–516. doi: 0.1016/j.geoforum.2011.02.001.

17. *Mayring P.* Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution. Klagenfurt: SSOAR, 2014. 143 pp.

18. *Spyrou E., Korakakis M., Charalampidis V., Psallas A., Mylonas P.* A geo-clustering approach for the detection of areas-of-interest and their underlying semantics // Algorithms. 2017. Vol. 10. No. 1. e35. doi: 10.3390/a10010035.

19. *Zhou X., Xu C., Kimmons B.* Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform // Computers, Environment and Urban Systems. 2015. Vol. 54. P. 144–153. doi: 10.1016/j.compenvurbsys.2015.07.006.

20. *Zhou Z., Zhang X., Guo X., Liu Y.* Visual abstraction and exploration of large-scale geographical social media data // Neurocomputing. 2020. Vol. 376. P. 244–255. doi: 10.1016/j.neucom.2019.10.072.

References

1. Blanutsa, V.I. (2017), “Social'no-ekonomicheskoe rajonirovanie kak sistema smyslov: kontent-analiz postsovetskih publikacij” [Socio-economic regionalization as a system of meanings: Content analysis of post-Soviet publications], *Geograficheskij vestnik*, no. 4, pp. 39–50. doi: 10.17072/2079-7877-2017-4-39-50.

2. Blanutsa, V.I. (2018), “*Social'no-ekonomicheskoe rajonirovanie v epohu bol'shih dannyh*” [Socio-Economic Regionalization in the Era of Big Data], INFRA-M, Moscow.

3. “Regiony Rossii. Social'no-ekonomicheskie pokazateli 2020” [Regions of Russia: Socio-economic indicators 2020], *Federal'naya sluzhba gosudarstvennoj statistiki*. Available at: https://rosstat.gov.ru/folder/210/document/13204 (accessed 24 May 2021).

4. Arabadzhyan, A., Figini, P., Vici, L. (2021), “Measuring destination image: A novel approach based on visual data mining. A methodological proposal and an application to European islands”, *Journal of Destination, Marketing & Management*, vol. 20, e100611, doi: 10.1016/j.dmm.2021.100611.

5. Bokányi, E., Kondor, D., Dobos, L., Sebõk, T., Stéger, J., Csabai, I., Vattay, G. (2016), “Race, religion and the city: Twitter word frequency patterns reveal dominant demographic dimensions in the United States”, *Palgrave Communications*, vol. 2, e16010. doi: 10.1057/palcomms.2016.10.

6. De Oliveira Capela, F., Ramirez-Marquez, J.E. (2019), “Detecting urban identity perception via newspaper topic modeling”, *Cities*, vol. 93, pp. 72–83. doi: 10.1016/j.cities.2019.04.009.

7. Drisko, J.M., Maschi, T. (2016), “*Content Analysis*”, Oxford University Press, Oxford.

8. Feldman, R., Sanger, J. (2006), “*The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*”, Cambridge University Press, New York.

9. Ferreira, D., Vale, M. (2020), “Geography in the big data age: An overview of the historical resonance of current debates”, *Geographical Review*, vol. 110, e1832424. doi: 10.1080/00167428.2020.1832424.

10. Goodchild, M.F. (2007), “Citizens as sensors: The world of volunteered geography”, *GeoJournal*, vol. 69, pp. 211–221. doi: 10.1007/s10708-007-9111-y.

11. Graham, M., Shelton, T. (2013), “Geography and the future of big data, big data and the future of geography”, *Dialogues in Human Geography*, vol. 3, no. 3, pp. 255–261. doi: 10.1177/2043820613513121.

12. Kitchin, R. (2013), “Big data and human geography: Opportunities, challenges and risks”, *Dialogues in Human Geography*, vol. 3, no. 3, pp. 262–267. doi: 10.1177/2043820613513388.

13. Kwan, M.-P. (2016), “Algorithmic geographies: Big Data, algorithmic uncertainty, and the production of geographic knowledge”, *Annals of the American Association of Geographers*, vol. 106, no. 2, pp. 274–282. doi:10.1080/00045608.2015.1117937.

14. Li, C., Dong, X., Yuan, X. (2018), “Metro-Wordle: An interactive visualization for urban text distributions based on Wordle”, *Visual Informatics*, vol. 2, no. 1, pp. 50–59. doi: 10.1016/j.visinf.2018.04.006.

15. Li, D., Zhou, X., Wang, M. (2018), “Analyzing and visualizing the spatial interactions between tourists and locals: A Flickr study in ten US cities”, *Cities*, vol. 74, pp. 249–258. doi: 10.1016/j.cities.2017.12.012.

16. Longley, P.A., Cheshire, J.A., Mateos, P. (2011), “Creating a regional geography of Britain through the spatial analysis of surnames”, *Geoforum*, vol. 42, no. 4, pp. 506–516. doi: 0.1016/j.geoforum.2011.02.001.

17. Mayring, P. (2014), “*Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution*”, SSOAR, Klagenfurt.

18. Spyrou, E., Korakakis, M., Charalampidis, V., Psallas, A., Mylonas, P. (2017), “A geo-clustering approach for the detection of areas-of-interest and their underlying semantics”, *Algorithms*, vol. 10, no. 1, e35. doi: 10.3390/a10010035.

19. Zhou, X., Xu, C., Kimmons, B. (2015), “Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform”, *Computers, Environment and Urban Systems*, vol. 54, pp. 144–153. doi: 10.1016/j.compenvurbsys.2015.07.006.

20. Zhou, Z., Zhang, X., Guo, X., Liu, Y. (2020), “Visual abstraction and exploration of large-scale geographical social media data”, *Neurocomputing*, vol. 376, pp. 244–255. doi: 10.1016/j.neucom.2019.10.072.

Поступила в редакцию: 12.06.2021

Сведения об авторе

Виктор Иванович Блануца
доктор географических наук, ведущий научный сотрудник, Институт географии им. В.Б. Сочавы СО РАН;
Россия, 664033, Иркутск, ул. Улан-Баторская, 1

About the author

Victor I. Blanutsa
Doctor of Geographical Sciences, Leading Researcher, V.B. Sochava Institute of Geography, Siberian Branch of the Russian Academy of Sciences;
1, Ulan-Batorskaya st., Irkutsk, 664033, Russia

e-mail: blanutsa@list.ru

Просьба ссылаться на эту статью в русскоязычных источниках следующим образом:
Блануца В.И. Цифровое развитие сибирских регионов: кластеризация в облаке тегов // Географический вестник = Geographical bulletin. 2021. №3(58). С. 62–73. doi: 10.17072/2079-7877-2021-3-62-73.

Please cite this article in English as:
Blanutsa, V.I. (2021). Digital development of the Siberian Federal District: clustering of regions in a tag cloud. *Geographical bulletin*. No. 3(58). Pp. 62–73. doi: 10.17072/2079-7877-2021-3-62-73.