Исследование списков TOP500 и Топ-50 методом интеллектуального анализа данных

М.Л. Цымблер, П.И. Шумилин

Южно-Уральский государственный университет

В статье применяется метод поиска ассоциативных правил для исследования данных о количестве систем и их производительности в редакциях суперкомпьютерных рейтингов TOP500, Топ-50 в их взаимосвязи с некоторыми экономическими (ИЧР и ВВП для TOP500, ВРП для Топ-50) и научными (квартили Scopus) показателями на предмет наличия скрытых закономерностей, которые не находят отражения в результатах, полученных при использовании статистических методов. Приводятся примеры обнаруженных закономерностей и их интерпретация.

Ключевые слова: top500, топ50, суперкомпьютерные рейтинги, ассоциативные правила.

1. Введение

В современном мире критически важна отрасль высокопроизводительных вычислений. Благодаря производительности, которую обеспечивают суперкомпьютерные системы, появляется возможность решать сложные вычислительные задачи в области медицины, моделирования физических процессов, метеорологии, промышленности и многих других научно-технических областях человеческой деятельности. Но производство и эксплуатация суперкомпьютеров требует больших финансовых затрат. В связи с этим, необходимо максимально рационально использовать ресурсы. Этого позволяет добиться осведомленность о влиянии суперкомпьютеров на общее благополучие и конкурентноспособность страны, что позволяет делать выводы о релевантности потенциальных инвестиций и принимать верные стратегические и управленческие решения.

В настоящее время для извлечения полезной информации из большого массива накопленных данных успешно применяются методы data mining. Эта технология обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности [12].

Целью данной работы является поиск таких знаний для данных суперкомпьютерных рейтингов TOP500 и Топ-50.

Статья имеет следующую структуру. В разделе 2 приведен обзор работ по тематике исследования суперкомпьютерных рейтингов. В разделе 3 описывается инструментарий и детали способа проведения исследования. В разделе 4 приводятся основные выводы и трактовки обнаруженных закономерностей.

2. Обзор работ

Для того, чтобы отслеживать динамику, анализировать тенденции и делать прогнозы, существует множество рейтингов суперкомпьютерных систем, ранжирующих их по определенному критерию. Это – удобный способ формализовать и систематизировать информацию для последующего анализа и визуализации. Первый появившийся и самый известный среди них – TOP500 [4]. Редакции ведутся с 1993 года. В этом рейтинге суперкомпьютеры упорядочиваются по результатам их производительности на тесте Linpack [3]. Так как данный тест отражает производительность суперкомпьютеров только с позиции их способности решать системы линейных алгебраических уравнений, то его результаты могут не соответствовать производительности, которую демонстрирует суперкомпьютер, применяющийся

для вычислений реальных задач. Это стало причиной дискуссии [7,9] вокруг объективности и правильности применения данного теста для оценки производительности суперкомпьютеров и привело к созданию альтернативных рейтингов, которые призваны оценить производительность систем на задачах, задействующих другие особенности архитектуры и приближенных к реальным приложениям. Среди них мировые рейтинги Green500 [6], Graph500 [8], GreenGraph500. Тем не менее наибольшую популярность имеет именно TOP500. Кроме того, по его примеру впоследствии начали формироваться региональные суперкомпьютерные рейтинги. Один из них – Топ-50, содержащий информацию о 50 самых производительных систем на территории СНГ.

Так как с конца XX века по настоящее время накоплено большое количество данных о суперкомпьютерных системах на различных тестах производительности, то эти данные используются для поиска закономерностей и анализа тенденций. Методами статистики был исследован характер эволюции суперкомпьютерных систем, ее основные черты. В работе [5] данные редакций размещены в координатах ранга суперкомпьютеров и количества процессоров, что позволяет выявить новые шаблоны. Рассмотрено распределение суперкомпьютеров по рангу в зависимости от архитектуры, также проиллюстрирована связь между рангом суперкомпьютера и производителем систем с массивно-параллельной архитектурой, распределение по области применения суперкомпьютеров, темпы роста их производительности и другие тренды. Однако потенциально опасна ситуация неправильной интерпретации накопленных данных. Например, как показано в работе [11], ведение статистики по доле количества суперкомпьютеров неверна и искажает реальное положение вещей, которое ведет к неправильным выводам и управленческим решениям. Точнее было бы использовать долю общей производительности суперкомпьютеров. Исследованием было подтверждено влияние инвестиций в суперкомпьютерную отрасль на национальную инновационную систему и развитие страны в целом, а также рассматривался вопрос корреляции между вычислительной мощностью и количеством публикаций учеными. [10]. Идею составления рейтинга TOP500 для суперкомпьютеров аналогичным образом воплощают и для мобильных устройств [13]. Интересное наблюдение заключается в том, что эволюция технологического совершенствования мобильных устройств отличается от эволюции суперкомпьютерных систем. В частности, доля производительности на ядро возрастает медленнее, чем доля оперативной памяти на ядро, что является противоположной тенденцией в сравнении с развитием суперкомпьютеров.

Несмотря на важные результаты, полученные с помощью методов статистики, существуют другие подходы для исследования данных, способные расширить знания об исследуемом объекте. В этой работе осуществляется попытка использовать метод интеллектуального анализа данных для обнаружения скрытых закономерностей.

3. Методы исследования

В исследовании применяется один из методов интеллектуального анализа данных – поиск ассоциативных правил. Задача поиска ассоциативных правил состоит в обнаружении таких устойчивых корреляций среди объектов в транзакционной базе данных, что присутствие одного объекта влечет за собой присутствие другого.

Постановка задачи и основные определения содержатся в разделе 3.1. Описание характеристических атрибутов и способов сбора данных приводится в разделе 3.2. Процесс обработки этих данных описывается в разделе 3.3.

3.1. Общий вид задачи поиска ассоциативных правил

Дадим основные определения в соответствии с работой [1] и адаптируем их под рассматриваемую задачу.

Пусть имеется множество литералов $I = \{i_1, \ldots, i_m\}$, называемых *характеристиче-*

скими атрибутами страны. Каждый из характеристических атрибутов может принимать целые значения $[1 \dots 10]$, которые обозначают дециль страны по данному атрибуту.

Набором A назовем конечное подмножество значений характеристических атрибутов из I. Тогда *k-набором* будет являться набор, состоящий ровно из k значений таких атрибутов.

Транзакцией T назовем m-набор, где $m = |I|$. *Транзакционная база* D – множество транзакций T. Таким образом, транзакционная база образована наборами фиксированной длины, которые состоят из значений каждого из характеристических атрибутов. Транзакция T содержит набор A, если $A \subseteq T$.

Поддержкой набора $support(A)$ назовем вероятность присутствия в транзакции всех значений характеристических атрибутов из набора A в транзакциях из D:

$$support\,(A) = P\,(A) \tag{1}$$

Ассоциативным правилом назовем импликацию вида $A \Rightarrow B$ такую, что A – произвольный набор, B – 1-набор, $A \cap B = \varnothing$. Правило может быть прочитано как условие «если A, то B», или «из набора A следует B». A назовем *консеквентом*, который обозначает вывод, а B – *антецедентом* правила, что имеет значение условия. Иначе говоря, в контексте данной работы правила могут быть, например, такие: «если страна имеет 1 дециль по суммарной производительности суперкомпьютеров из редакции рейтинга и значение индекса человеческого развития в 1 дециле, то это США» или «из того, что Китай находится во 2 дециле по количеству публикаций в Q2, следует то, что он так же находится во 2 дециле Q3».

Поддержка правила $support(A \Rightarrow B)$ – доля транзакций, которые содержат одновременно наборы A и B, среди всех транзакций, то есть

$$support\,(A \Rightarrow B) = P\,(A \cup B) \tag{2}$$

Достоверность правила $confidence(A \Rightarrow B)$ — доля транзакций, которые содержат наборы A и B, среди тех, которые содержат A, то есть

$$confidence\,(A \Rightarrow B) = P(B|A) = \frac{support\,(A \cup B)}{support\,(A)} \tag{3}$$

Другими словами, достоверность является условной вероятностью A и B при условии A и демонстрирует, на сколько часто наличие значений характеристических атрибутов из набора A влечет за собой наличие значений характеристических атрибутов из набора B.

Минимальным уровнем поддержки min-support и *минимальным уровнем достоверности min-confidence* называют установленные пороговые значение поддержки и достоверности соответственно. *Устойчивое правило* – ассоциативное правило, поддержка и достоверность которого не меньше соответствующих минимальных значений.

3.2. Сбор данных и характеристические атрибуты

3.2.1. TOP500

Используемые данные находятся в свободном доступе и предоставляются официальными ресурсами в сети Интернет: TOP500[1], Human Development Reports[2] (HDR), The World Bank[3] (TWB) и Scopus[4]. Сведения о выбранных атрибутах и официальных открытых источниках получения данных приведены в табл. 1. Атрибуты разбиты на 3 группы:

[1] http://www.top500.org
[2] http://hdr.undp.org/en/
[3] https://data.worldbank.org/
[4] https://www.scopus.com

1. атрибуты, отражающие степень присутствия страны в области высокопроизводительных вычислений;
2. атрибуты, отражающие экономический уровень развития страны;
3. атрибуты, отражающие научный потенциал страны.

Таблица 1. Обозначения атрибутов, их источников и семантики

Атрибут	**Источник**	**Семантика**
Country	TOP500	Страна, системы которой присутствуют в редакции данного года
Num of Systems	TOP500	Общее количество систем страны в редакции данного года
RMax	TOP500	Общая Linpack производительность систем страны в редакции данного года
HDI	HDR	Значение индекса человеческого развития страны для данного года
GDPpC	TWB	Значение внутреннего валового продукта на душу населения страны для данного года
Q1-Q4	Scopus	Количество научных работ, опубликованных авторами данной страны в данном году в источниках, находящихся в данном году в каждом из Q1-Q4 по рейтингу CiteScore

В конечном итоге исследуемый объект представляет собой строку, которая содержит информацию о названии страны, присутствующей в редакции конкретного года, и о характеризующих ее атрибутов для этого же года.

В связи с тем, что новая редакция публикуется дважды в год, а данные по количеству научных публикаций, внутреннего валового продукта на душу населения (ВВП) и индекса человеческого развития (ИЧР) предоставляются лишь ежегодно, из 50 существующих на текущий момент редакций для исследования использовались только 26 – та половина, которую образуют ноябрьские редакции с 1993 по 2018 год.

Так как атрибут количества опубликованных научных работ за год является обобщенным и не учитывает фактор качества публикаций, для более точной оценки научного потенциала страны вводятся отдельные атрибуты, благодаря которым осуществляется учет количества публикаций различного уровня. Уровень научной работы в исследовании оценивается по рейтингу ее места публикации. Для этого используются квартильная оценка источников базы данных Scopus. Источники включают в себя такие места публикаций, как журналы, книжные серии, материалы конференций и отраслевые издания.

Места публикации разбиты на категории, отражающие их уровень востребованности и авторитет в научном сообществе. Такими категориями являются квартили Q1-Q4, где к Q1 относятся первые 25% источников, в Q2 – следующие 25% и так далее. Квартили назначаются по метрике CiteScore.

Данные о количестве публикаций в квартилях Q1-Q4 получались через разработанную программу, использующую открытый API Scopus. Формировались запросы, содержащие год публикации статьи *PUBYEAR IS <год>*, страну аффилиации хотя бы одного из авторов работы *AFFILCOUNTRY(<страна>)* и список идентификаторов мест публикаций, принадлежащих каждому из квартилей Q1-Q4 *SOURCE-ID(<идентификатор источника>)*. Такие списки создавались на основе файла, содержащего все источники и их квартиль.

Данные для ИЧР и ВВП имеются с 1993 по 2017 год. Данные о квартилях источников – с 2011 по 2017 год. При отсутствии для редакции рейтинга данного года актуальной информации использовались данные наиболее близкие к этому году.

3.2.2. Топ-50

Аналогичная работа была проведена и для редакций Топ-50. Официальные открытые источники данных: Топ-50 [1], ЕМИСС [2], Scopus. Сведения об атрибутах приведены в табл. 2. Группы, на которые разбиваются атрибуты, сохранены и имеют тот же смысл, что и для TOP500, только не для страны, а для научного учреждения Российской Федерации (РФ).

Таблица 2. Обозначения атрибутов, их источников и семантики

Атрибут	**Источник**	**Семантика**
Affiliation	Топ-50	Научное учреждение РФ, системы которого присутствуют в редакции данного года
RMax	Топ-50	Общая Linpack производительность систем научного учреждения РФ в редакции данного года
GRPpC	ЕМИСС	Значение валового регионального продукта на душу населения (ВРП) субъекта РФ для данного года
Q1-Q4	Scopus	Количество научных работ, опубликованных авторами из данного научного учреждения РФ в данном году в источниках, находящихся в данном году в каждом из Q1-Q4 по рейтингу CiteScore

Принимая во внимания те же причины, что и для TOP500, для исследования была выбрана только половина редакций. Она состоит из 1 декабрьской редакции 2004 года и 14 сентябрьских редакций с 2005 по 2018 год.

Данные о количестве публикаций схожим образом были получены с использованием API Scopus. В запросах вместо команды *AFFILCOUNTRY(<страна>)* была использована команда *AF-ID(<идентификатор научного учреждения>)*. Список таких идентификаторов для аффилиаций, присутствующих в редакциях, был получен из информации доступной о каждом научном учреждении, зарегистрированном в Scopus.

Данные для ВРП имеются с 2004 по 2016 год. Данные о квартилях источников – с 2011 по 2017 год. При отсутствии для редакции рейтинга данного года актуальной информации использовались данные наиболее близкие к данному году.

3.3. Обработка данных

3.3.1. TOP500

Для возможности применения алгоритма поиска ассоциативных правил необходимо определенным образом подготовить данные. В связи с этим было разработано приложение на языке программирования Python, которое осуществляет обработку и агрегацию данных. Архитектура решения для экстракции данных представлена на рис. 1.

На основе каждой редакции формировался ряд стран, хотя бы один суперкомпьютер которых представлен в этой редакции, и для каждой страны вычислялось суммарное количество суперкомпьютерных систем, суммарная Linpack производительность всех систем данной страны в данном редакции. Эти данные дополнялись значениями ИЧР и ВВП для

[1] http://top50.supercomputers.ru
[2] https://fedstat.ru

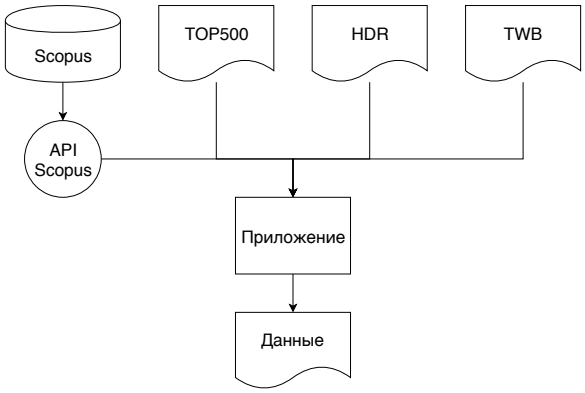

Рис. 1. Схема получения данных

каждой страны по соответствующему году, а также суммарным количеством научных публикаций за этот год в источниках, принадлежащих каждому из квартилей Q1-Q4.

Чтобы учесть временной контекст, заключающийся в различной значимости конкретных значений в зависимости от года, и обеспечить ассоциативным правилам достижение минимального уровня поддержки, абсолютные значения каждого из численных атрибутов были заменены относительными децильными характеристиками по каждой редакции, то есть по каждому году. Первый дециль ставился в соответствие десятой части лучших значений, второй дециль – следующей десятой части и так далее.

Обработанные данные были объединены в итоговый выходной файл, содержащий 716 объектов. Полученный файл использовался для поиска ассоциативных правил, осуществление которого производилось с применением классического алгоритма Apriori [1], реализованным в программной среде для анализа данных KNIME Analytics Platform [2]. С минимальным уровнем поддержки 1% и минимальным уровнем достоверности 50% было получено 2889 правил.

3.3.2. Топ-50

Аналогичным образом осуществлялась обработка данных и для редакций Топ-50. Архитектура решения для экстракции данных осталась такой же, изменились только источники получения данных.

На основе каждой редакции формировался ряд научных учреждений РФ, хотя бы один суперкомпьютер которых представлен в этой редакции, и для каждого такого учреждения вычислялась суммарная Linpack производительность всех систем из данной редакции установленных в этом учреждении. Эти данные дополнялись значениями ВРП для каждого региона по соответствующему году, а также суммарным количеством научных публикаций за этот год в источниках, принадлежащих каждому из квартилей Q1-Q4.

В связи с тем, что разбиение на децили слишком детально и не обеспечивает достаточный уровень поддержки для полученного набора данных, вместо децильной дискретизации по каждому году использовалась квартильная. То есть для каждого числового значения атрибута научных учреждений РФ ставился в соответствие номер квартиля так, что к 1 квартилю относились лучшие 25% значений, ко второму следующие 25% и так далее.

Обработанные данные были объединены в итоговый выходной файл, содержащий 248 объектов. Ассоциативные правила обнаруживаются таким же способом, что и для данных TOP500. С минимальным уровнем поддержки 2% и минимальным уровнем достоверности 50% было получено 1361 правило.

4. Экспериментальные результаты

В разделе приведены некоторые из полученных ассоциативных правил, которые в части консеквента или антецедента содержат атрибут, связанный с суперкомпьютерными рейтингами (Num of Systems или RMax). В разделе 4.1 представлены результаты для суперкомпьютерного рейтинга TOP500, а в разделе 4.2 – для Топ-50.

4.1. Ассоциативные правила для редакций TOP500

В силу способа представления данных в виде разбиения на децили, поддержка каждого из правил не может превосходить 10% от общего числа объектов. Таким образом, диапазон 1% - 5% является оправданной поддержкой для поиска интересных ассоциативных правил. Ассоциативные правила разбиты по логическим группам и размещены в таблицах.

При расшифровке правил атрибуты Num of Systems, RMax свидетельствовали о вычислительном ресурсе страны, квартили Q1-Q4 – о научном потенциале страны, HDI и GDPpC – об уровне жизни и благосостоянии страны. А децили обозначали следующие градации указанных характеристик среди стран, представленных в редакциях: децили 1-3 – относительно высокий (значительный, большой) уровень, лидирующие позиции; децили 4-6 – средний уровень; децили 7-10 – относительно низкий (незначительный, малый) уровень.

В табл. 3 приведены правила, в консеквенте которых находится характеристический атрибут количества суперкомпьютерных систем в редакциях рейтинга.

Таблица 3. Правила для количества систем в редакциях рейтинга

№	Антецедент	Консеквент
1	[HDI: 2, Q4: 1, Q1: 1]	Num of Systems: 1
2	[HDI: 5, Q3: 2]	Num of Systems: 2
3	[HDI: 7, Q1: 2]	Num of Systems: 2
4	[RMax: 10, Q1: 10]	Num of Systems: 9

В абсолютном большинстве полученных ассоциативных правил дециль количества систем Num of Systems и дециль суммарной производительности RMax одинаковы. То есть ситуация обладания страной высокой производительностью за счет малого числа высокотехнологичных и более мощных суперкомпьютеров в устойчивых правилах не отражена.

Для каждого квартиля Q1-Q4 было получено устойчивое правило следующего вида: если лидерство в данном Qx, где x это номер квартиля, то лидерство по количеству суперкомпьютерных систем в редакции. Данное наблюдение позволяет предположить, что научный потенциал страны явным образом коррелирует со степенью ее присутствия в области высокопроизводительных вычислений и зависит от количества суперкомпьютеров, которыми обладает страна.

Видно, что высокий уровень жизни и высокий научный потенциал влекут за собой лидерство по количеству суперкомпьютеров в редакции рейтинга (№1). Вместе с тем обладание большим числом суперкомпьютеров характерно и для стран, имеющих средний и низкий уровень жизни, но при этом высокий научный потенциал (№2-3). Низкий уровень научного потенциала влечет за собой малое количество суперкомпьютерных систем (№4).

Таким образом, страны с высоким научным потенциалом, как правило, обладают высокими вычислительными мощностями. При этом децили, отражающие уровень жизни, могут быть как высокими, так и низкими.

Правила для экономических характеристических атрибутов представлены в табл. 4.

Низкий уровень ВВП влечет за собой низкий уровень ИЧР. Видно, что, обладая средним

Таблица 4. Правила для ВВП и ИЧР

№	Антецедент	Консеквент
1	[Num of Systems: 6, GDPpC: 1]	HDI: 1
2	[Num of Systems: 2, Q1: 2, Q2: 2]	GDPpC: 5
3	[Num of Systems: 4, Q1: 3, Q2: 3]	GDPpC: 7
4	[Num of Systems: 6, GDPpC: 10]	HDI: 10
5	[Num of Systems: 9, GDPpC: 10]	HDI: 10

числом суперкомпьютеров, страна может иметь как высокий, так и низкий уровень жизни (№1, №4). Однако высокий научный потенциал совместно с большим количеством суперкомпьютеров характерен для стран со средним уровнем жизни (№2). Но высокий научный потенциал и среднее число суперкомпьютеров часто встречается и у стран с низким уровнем жизни (№3). Низкий уровень ВВП и малое число суперкомпьютеров ведет к низкому уровню жизни (№5).

Таким образом, количество суперкомпьютеров, которым владеет страна, не определяет ее уровень жизни. Есть примеры как стран с высоким уровнем жизни и с небольшим числом суперкомпьютеров, так и стран с низким уровнем жизни и со средним их количеством.

В табл. 5 приведены правила, в консеквенте которых расположены значения децилей количества научных публикаций в источниках по квартилям Scopus. Квартили Q1-Q4, как

Таблица 5. Правила для квартилей Scopus

№	Антецедент	Консеквент
1	[Num of Systems: 1, HDI: 2]	Q4: 1
2	[Num of Systems: 2, GDPpC: 5]	Q1: 2
3	[Num of Systems: 2, HDI: 7]	Q1: 2
4	[Num of Systems: 9, GDPpC: 10]	Q1: 10

правило согласованны, то есть имеют один и тот же, либо соседний дециль. Это означает, что явление, когда страна лидирует в одном квартиле, а в другом занимает низкие позиции, в устойчивых правилах не отражено.

Высокий научный потенциал следует из владения страной большим количеством суперкомпьютеров и высокого, среднего или низкого уровня жизни (№1-3). Однако при наличии у станы малого числа суперкомпьютеров и низкого уровня жизни для нее характерен низкий научный потенциал (№4).

Видно, что высокий научный потенциал достижим странами с различным уровнем жизни, но при условии обладания страной большим количеством суперкомпьютеров.

4.2. Ассоциативные правила для редакций Топ-50

В силу разбиения данных на квартили максимальная поддержка правила может достигать 25%. Диапазон для поиска интересных правил 2% - 12%. При расшифровке правил

атрибут RMax характеризовал мощность вычислительных ресурсов, имеющихся в научном учреждении, GRPpC – степень экономического развития региона, квартили Q1-Q4 – научный потенциал учреждения. Образованные группы значений квартилей трактовались следующим образом: 1 квартиль соответствует лидерским, высоким показателям, 2 квартиль – повышенным, выдающимся, 3 квартиль – средним, 4 квартиль – базовым, типичным, обычным показателям.

В табл. 6 представлены ассоциативные правила, в консеквенте которых находится квартиль производительности суперкомпьютеров. Научные учреждение с высоким научным по-

Таблица 6. Правила для производительности

№	Антецедент	Консеквент
1	[GRPpC: 1, Q1: 1]	RMax: 1
2	[GRPpC: 4, Q1: 2, Q3: 2]	RMax: 2
3	[GRPpC: 3, Q2: 1, Q3: 1]	RMax: 3
4	[GRPpC: 3, Q2: 4, Q3: 4]	RMax: 3
5	[GRPpC: 4, Q1: 4]	RMax: 4

тенциалом и находящиеся в экономически развитом регионе, как правило, имеют мощный суперкомпьютер (№1). Находясь в экономическом регионе с базовым уровнем развития, но обладая повышенным научным потенциалом, учреждение скорее всего имеет суперкомпьютер повышенной мощности (№2). Однако, находясь в экономических регионах со средним уровнем развития, учреждения как с высоким научным потенциалом, так и с базовым, могут иметь суперкомпьютеры средней мощности (№3-4). Базовый научный потенциал совместно с базовым или средним экономическим развитием региона, как правило, ведет к обладанию научными учреждениями суперкомпьютерами с базовой или средней вычислительной мощностью (№4-5).

В общем случае, чем более экономически развит регион и чем более высок научный потенциал, тем более мощным суперкомпьютером располагает учреждение. Однако существуют примеры достижения повышенного научного потенциала несмотря на расположение в базовом экономическом регионе, а также высокого научного потенциала несмотря на использованием суперкомпьютера средней мощности.

В табл. 7 представлены ассоциативные правила, в консеквенте которых находится квартиль ВРП. Уровень вычислительной мощности и научного потенциала учреждения соответствует уровню экономического развития региона, в котором оно находится (№1, №3, №5). Однако, вместе с тем, научные учреждения с повышенным уровнем научного потенциала, но с суперкомпьютером средней мощности также находится в регионе с высоким уровнем экономического развития (№2). Средний уровень производительности суперкомпьютера наряду с высоким научным потенциалом характерен для научных учреждений, расположенных в регионах со среднем экономическим развитием (№4). Учреждения с высокой мощностью суперкомпьютера и средним научным потенциалом также могут быть расположены в регионе, имеющем базовый уровень экономического развития (№6).

Таким образом, учреждения с высокой производительностью суперкомпьютеров не обязательно находятся в экономически развитых регионах страны. Хотя часто, чем выше научный потенциал учреждения и располагаемая им вычислительная мощность, тем в более экономически развитом регионе оно расположено.

В табл. 8 представлены ассоциативные правила, в консеквенте которых находится номер квартиля Scopus. Научные квартили согласованны. То есть Q1-Q4 принимают одинаковые

Таблица 7. Правила для ВРП

№	Антецедент	Консеквент
1	[RMax: 1, Q1: 1]	GRPpC: 1
2	[RMax: 3, Q3: 2]	GRPpC: 1
3	[RMax: 2, Q1: 2, Q2: 2, Q4: 2]	GRPpC: 2
4	[RMax: 3, Q2: 1]	GRPpC: 3
5	[RMax: 4, Q1: 4]	GRPpC: 4
6	[RMax: 1, Q1: 3]	GRPpC: 4

Таблица 8. Правила для квартилей Scopus

№	Антецедент	Консеквент
1	[RMax: 1, GRPpC: 1]	Q1: 1
2	[RMax: 2, GRPpC: 4, Q2: 2, Q3: 2]	Q4: 2
3	[RMax: 4, GDPpC: 1, Q2: 3]	Q1: 3
4	[RMax: 4, GRPpC: 4, Q2: 4]	Q1: 4

номера квартилей. Если научное учреждение обладает суперкомпьютером с высокой производительностью и находится в регионе с высоким уровнем экономического развития, то это учреждение обладает высоким научным потенциалом (№1). Повышенная мощность суперкомпьютера научного учреждения, находящегося в регионе с базовым уровнем экономического развития влечет за собой повышенный уровень научного потенциала (№2). Однако научный потенциал учреждений, находящихся в экономически высоко развитом регионе и регионе с базовым уровнем экономического развития, имеют примерно одинаковый среднебазовый научный потенциал, в связи с тем, что располагают суперкомпьютерами с базовой вычислительной способностью (№3-4).

Таким образом, влияние факторов обладания научным учреждения вычислительным и экономическим ресурсом оказывает роль на итоговый научный потенциал. Но уровень влияния фактора мощности суперкомпьютера, которым распоряжается научное учреждение, имеет большее значение.

5. Заключение

В работе описан процесс проведенного анализа данных суперкомпьютерных рейтингов TOP500 и Топ-50 с использованием метода поиска ассоциативных правил.

Основные тенденции и закономерности, выявленные в ходе исследования следующие. Высокий уровень жизни не является необходимым условием для достижения страной высокого научного потенциала. Уровень научного потенциала страны, как правило, имеет прямую зависимость от количества суперкомпьютеров страны, входящих в редакции TOP500. Количество суперкомпьютеров страны в редакциях TOP500 не определяет уровень жизни страны.

Уровень производительности суперкомпьютера не определяется однозначно уровнем

экономического развития региона. Но в более развитых регионах, как правило, расположены научные учреждения с более мощными суперкомпьютерами. Высокий научный потенциал не обязательно связан с высокой производительностью суперкомпьютера, но производительность суперкомпьютера, как правило, определяет научный потенциал.

Таким образом, как в масштабах страны, так и для российских научных учреждений, удалось обнаружить корреляцию между располагаемым вычислительным ресурсом и научным потенциалом. Однако как на уровне стран, так и на уровне научных учреждений устойчивых правил, которые бы иллюстрировали существенные различия в количестве публикаций научных работ в источниках среди различных квартилей Scopus, обнаружено не было.

Литература

1. Agrawal R., Srikant R. Fast Algorithms for Mining Association Rules // In Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, September 1994. –Vol. 1215. –P. 487-499.

2. Berthold M.R., Cebron N., Dill F., Gabriel T.R., Kötter T., Meinl T., Ohl P., Thiel K., Wiswedel B. KNIME - the Konstanz information miner: version 2.0 and beyond // –ACM SIGKDD Explorations Newsletter, June 2009. –ACM New York, NY, USA. –Vol. 11. –Issue 1. –P. 26–31. DOI:10.1145/1656274.1656280

3. Dongarra J.J The LINPACK Benchmark: An explanation. // Supercomputing. Lecture Notes in Computer Science, vol 297. Springer, Berlin, Heidelberg, 1988. DOI:10.1007/3-540-18991-2_27

4. Dongarra J.J., Meuer H.W., Strohmaier E. TOP500 Supercomputer Sites. –1999.

5. Feitelson D.G. On the Interpretation of Top500 Data // The International Journal of High Performance Computing Applications. –1 May 1999. –Vol. 13. –Issue 2. –P. 146–153. DOI: 10.1177/109434209901300204

6. Feng W., Scogland T., The Green500 List: Year One // IEEE International Symposium on Parallel & Distributed Processing, -2009, DOI: 10.1109/IPDPS.2009.5160978

7. Kramer W. Top500 Versus Sustained Performance – the Top Problems with the TOP500 List – And What to Do About Them // Parallel Architectures and Compilation Techniques (PACT), – Minneapolis, MN, USA, 19–23 Sept. 2012. 21st International Conference on. IEEE. P. 223-–230.

8. Murphy R.C., Wheeler K.B., Barrett B.W., Ang J.A. Introducing the Graph 500 // Cray User's Group (CUG). –2010. –Vol. 19. –P. 45-74.

9. Strohmaier E., Meuer H.W., Dongarra J., Simon H.D. The TOP500 List and Progress in High-Performance Computing // Computer, Nov. 2015. –Vol. 48. –Issue 11. –IEEE. –P. 42–49. DOI:10.1109/MC.2015.338

10. Zelenkov Y.A., Sharsheeva J.A. Impact of the Investment in Supercomputers on National Innovation System and Country's Development // International Conference on Parallel Computational Technologies. –Springer, Cham, 2017. –C. 42–57.

11. Абрамов С.М. Правда, искажающая истину. Как следует анализировать Top500? // Вестн. ЮУрГУ. Сер. Выч. матем. информ. –2013. –Т. 2. –В. 3. –С. 5–31. DOI:10.14529/cmse130301

12. Дюк В.А., Флегонтов А.В., Фомина И.К. Применение технологий интеллектуального анализа данных в естественнонаучных, технических и гуманитарных областях // Известия Российского государственного педагогического университета им. А.И. Герцена. 2011. №138. С. 77–84.

13. Ячник О.О., Никитенко Д.А., Соболев С.И. Мобильный Linpack: первый опыт введения рейтинга производительности мобильных устройств // Вычислительные методы и программирование: Новые вычислительные технологии. –2018. –Т. 19, № 4. –С. 464–469.