Исследование списков TOP500 и Топ-50 методом интеллектуального анализа данных

М.Л. Цымблер, П.И. Шумилин

Южно-Уральский государственный университет

В статье исследуются данные редакций суперкомпьютерных рейтингов TOP500, Топ-50 и их взаимосвязь с 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

Итак, целью работы является поиск устойчивых ассоциативных правил, связывающих значения атрибутов, входящих в различные группы.

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 [8]. Редакции ведутся с 1993 года. В этом рейтинге суперкомпьютеры упорядочиваются по результатам их производительности на тесте Linpack [9]. 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 [2, 3] 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 [10], Graph500 [11], 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 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 [4] рассмотрена взаимосвязь позиций суперкомпьютеров в рейтинге TOP500 и их производителей, приведено распределение по области применения суперкомпьютеров и распределение

мощностей в зависимости от позиции в рейтинге. Однако потенциально опасна ситуация неправильной интерпретации накопленных данных. Например, как показано в статье [5], 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. [12]. Идею составления рейтинга TOP500 для суперкомпьютеров аналогичным образом воплощают и для мобильных устройств [1]. 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, такой подход не обладает способностью всестороннее исследовать данные. В этой работе осуществляется попытка использовать методы 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 значений характеристических атрибутов в транзакционной базе данных, что присутствие одного атрибута 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 [7] и адаптируем их под рассматриваемую задачу.

Пусть имеется множество литералов $I = \{i_1, \ldots, i_m\}$, называемых *характеристическими атрибутами* страны. Каждый из характеристических атрибутов может принимать целые значения $[1 \ldots 10]$, 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 появления 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 систем страны из редакции рейтинга и значение индекса человеческо-

го 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 объектов, 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 приведена в табл. 1. Атрибуты разбиты на 3 группы:

1. атрибуты, отражающие степень присутствия страны в области высокопроизводительных вычислений посредством позиции в рейтинге TOP500;
2. атрибуты, отражающие экономический уровень развития страны;
3. атрибуты, отражающие научный потенциал страны.

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) представляются лишь ежегодно, из 50 существующих на текущий момент редакций списка 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

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

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 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
Top10	Scopus	Количество научных работ, опубликованных авторами данной страны в данном году в 10% лучших источниках по рейтингу CiteScore данного года

источников базы данных Scopus. Источники включают в себя такие места публикаций, как журналы, книжные серии, отраслевые публикации и материалы конференций.

Научные журналы 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. Помимо квартилей выделяется еще одна категория Top10 – 10% лучших научных источников публикаций.

Данные о количестве публикаций в квартилях Q1-Q4 и Top10 получались через 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, Top10 *SOURCE-ID(<идентификатор источника>)*. Такие списки создавались на основе файла, содержащего все источники и их квартиль для годов с 2011 до 2017. Для редакций, которые были выпущены с 1993 по 2011 год использовались сведения за 2011 год, в силу отсутствия более актуальных данных для метрики CiteScore.

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).

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

Данные о количестве публикаций схожим образом были получены с использованием API Scopus. В запросах вместо команды *AFFILCOUNTRY(<страна>)* была использована

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

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	TOP500	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
Top10	Scopus	Количество научных работ, опубликованных авторами из данного научного учреждения РФ в данном году в 10% лучших источниках по рейтингу CiteScore данного года

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.

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.

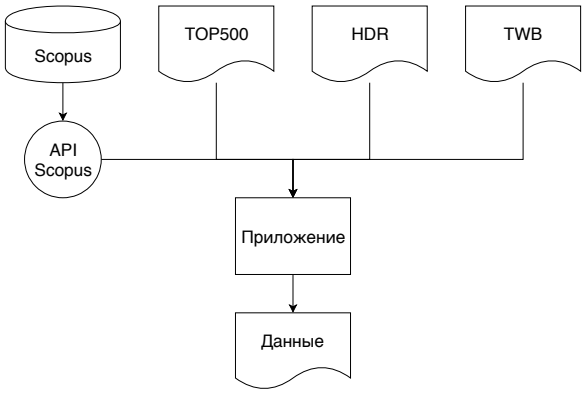

Рис. 1. Схема получения данных

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 каждой страны по соответствующему году, а также суммарным количеством научных публикаций за этот год в источниках, принадлежащих каждому из квартилей Q1-Q4 и Top10.

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.

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 [7], реализованным в программной среде для анализа данных KNIME Analytics Platform [13]. С минимальным уровнем поддержки 1% и минимальным уровнем достоверности 40% было получено 4995 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 и Top10.

В связи с тем, что разбиение на децили слишком детально и не обеспечивает достаточный уровень поддержки для полученного набора данных, дискретизация абсолютных значений численных атрибутов была произведена следующим образом. Если упорядочить объекты по невозрастанию для данного характеристического атрибута, то первым 10% объектам, ставится в соответствие 1 квантиль по данному атрибуту, следующим 20% объектов – 2 квантиль, следующим 20% – 3 квантиль, оставшимся 50% – 4 квантиль.

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 1% и минимальным уровнем достоверности 60% было получено 2668 правил.

4. Экспериментальные результаты

В разделе приведены некоторые из полученных ассоциативных правил. В разделе 4.1 представлены результаты для суперкомпьютерного рейтинга TOP500, а в разделе 4.2 – для Топ-50.

4.1. 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.

Ожидаемо, значительная часть полученных ассоциативных правил оказалась логично объяснимой и интуитивно понятной. Например, страны-лидеры по суммарной вычислительной мощности RMax являются таковыми, так как обладают большим количеством суперкомпьютеров в рейтинге, низкий показатель ИЧР следует из низкого уровня ВВП, и дециль количества публикаций в Top10 совпадает с децилем для Q1, так как научные работы, опубликованные в источниках Top10, входят и в Q1. Вместе с тем удалось обнаружить менее тривиальные и более интересные правила, которые разбиты по логическим группам

и приведены далее.

При расшифровке правил атрибуты Num of Systems, RMax будут свидетельствовать о вычислительном ресурсе страны, квартили Q1-Q4 и Top10 – о научном потенциале страны, HDI и GDPpC – об уровне жизни и благосостоянии страны. Тогда децили будут обозначать 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.

4.1.1. Правила для стран

Следует заметить, что максимальная поддержка набора, в котором фигурирует конкретная страна, может достигать порядка 3,5% при условии присутствия хотя бы одного суперкомпьютера страны в каждой из 28 обработанных редакций.

Страны, для которых были обнаружены ассоциативные правила, образовали 8 групп. Они приведены в табл. 3.

Таблица 3. Группы стран

№	Уровень жизни	Научный потенциал	Страны
1	Высокий	Высокий	США, Германия
2	Средний	Высокий	Великобритания, Канада, Франция, Япония
3	Низкий	Высокий	Италия, Китай, Российская Федерация
4	Высокий	Средний	Австралия, Нидерланды, Швейцария, Швеция
5	Низкий	Средний	Бразилия, Индия, Польша, Республика Корея, Испания
6	Высокий	Низкий	Норвегия, Дания
7	Средний	Низкий	Австрия
8	Низкий	Низкий	ЮАР

Такого рода правила отражают характерные черты стран и косвенным образом демонстрируют их стратегии развития, приоритеты и достижения в данных областях. Глядя на образованные группы, можно заметить, что, например, страны G8 достигают высокого научного потенциала при высоком, среднем и низком уровне жизни. Это свидельствует о том, что уровень жизни не определяет публикационную активность. То есть нельзя сказать, что высокий уровень жизни гарантирует стране высокий научный потенциал, так же как и нельзя сказать, что низкий уровень жизни ведет к низкому научному потенциалу. Но в общем случае полученные результаты подтверждают интуитивные ожидания о степени развитости стран в этих характеристиках. Например те, что высокий уровень жизни свойственен европейским странам, США, Австралии.

4.1.2. Правила для атрибутов высокопроизводительных вычислений

В табл. 4 приведены правила, в консеквенте которых находится характеристический атрибут количества суперкомпьютерных систем в редакциях рейтинга.

В абсолютном большинстве полученных ассоциативных правил дециль количества си-

Таблица 4. Правила для количества систем в редакциях рейтинга

№	Антецедент	Консеквент
1	[HDI: 2, Top10: 1, Q1: 1, Q4: 1]	Num of Systems: 1
2	[Top10: 2, Q1: 2, Q2: 2, Q4: 2]	Num of Systems: 2
3	[HDI: 5, Top10: 2, Q1: 2, Q4: 2]	Num of Systems: 2
4	[HDI: 7, Top10: 3, Q1: 3]	Num of Systems: 3
5	[Q3: 8, Q4: 9]	Num of Systems: 7

стем 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 правилах не отражена.

Из правила №1 видно, что высокий уровень жизни и высокий научный потенциал влекут за собой лидерство по количеству суперкомпьютеров в редакции рейтинга. Вместе с тем из правил №3 и №4 следует, что страны со средним и низким уровнем жизни, но так же имеющие высокие научные показатели, находятся в числе лидеров по количеству суперкомпьютеров. А правила №2 и №5 демонстрируют, что децили научного потенциала определяют соответствующий дециль количества суперкомпьютеров.

Для каждого квартиля Q1-Q4 и Top10 было получено устойчивое правило следующего вида: если лидерство в данном Qx, где x это номер квартиля, или Top10, 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 нее. Таким образом, на основании приведенных правил можно заключить, что страны с высоким научным потенциалом, как правило, обладают высокими вычислительными мощностями. При этом децили, отражающие уровень жизни, могут быть как высокими, так и низкими.

4.1.3. Правила для экономических атрибутов

В основном правила, имеющие на месте консеквента дециль экономического атрибута (HDI или GDPpC) содержат в антецеденте страну и/или экономический атрибут, не находящийся в консеквенте (GDPpC или HDI соответственно). Кроме того, часто в такого рода правилах дециль ВВП совпадает с децилем ИЧР. Но присутствуют и исключения. Некоторые из них представлены в табл. 5.

Из правил №1 и №3 видно, что следствием среднего научного потенциала может являться как высокий, так и низкий уровень жизни. Правила №2 и №4 свидетельствуют о том, что существуют страны с низким уровнем жизни, обладающие высоким и низким научным потенциалом. А правила №5, №6 и №7 показывают, что если страна имеет высокий научный потенциал и большое количество суперкомпьютеров в редакции рейтинга, то она может иметь средний, высокий или низкий уровень жизни. Таким образом, видно, что однозначная закономерность между научным потенциалом и уровнем жизни в стране отсутствует.

4.1.4. Правила для научных атрибутов

В табл. 6 приведены правила, в консеквенте которых расположены значения децилей количества научных публикаций в источниках по квартилям Scopus.

Квартили Q1-Q4 и Top10, как правило согласованны, то есть имеют один и тот же, либо

Таблица 5. Правила для ВВП и ИЧР

№	Антецедент	Консеквент
1	[Q2: 4, Q3: 5]	HDI: 1
2	[Q1: 3, Q2: 3, Q3: 3, Top10: 3, Q4: 3]	HDI: 7
3	[Top10: 6, Q2: 5, Q1: 6]	HDI: 9
4	[Top10: 9, Q3: 8, Q1: 9]	HDI: 10
5	[Q3: 1, Q4: 1, Num of Systems: 1, Q1: 1, Q2: 1, RMax: 1]	GDPpC: 2
6	[Q1: 2, Num of Systems: 2, Q3: 2, RMax: 2]	GDPpC: 5
7	[Q1: 3, Q2: 3, Top10: 3, Num of Systems: 4]	GDPpC: 7

Таблица 6. Правила для квартилей Scopus

№	Антецедент	Консеквент
1	[Q2: 1, Q3: 1, Q4: 1]	Q1: 1
2	[Num of Systems: 1, HDI: 2]	Q4: 1
3	[Num of Systems: 2, RMax: 2, GDPpC: 5]	Q3: 2
4	[Top10: 9, Q1: 9, Q2: 9, Q3: 9]	Q4: 9

соседний дециль. Это означает, что явление, когда страна лидирует в одном квартиле, а в другом занимает низкие позиции, в правилах не отражено.

Правила №2, №3 показывают, что высокий научный потенциал следует из антецедента с большим количеством суперкомпьютеров независимо от уровня жизни в стране. Правила №1, №4 иллюстрируют согласованность значений научных характеристических атрибутов как для высокого, так и для низкого уровня научного потенциала. Приведенные правила вновь подтверждают, что высокий научный потенциал достижим странами со средним или низким уровнем жизни.

4.2. Топ-50

Поддержка каждого из характеристических атрибутов зависит от квантиля, к которому он относится. Для 4 квантиля максимальная поддержка может быть 50%, для 3 и 4 по 20%, а для 1 только 10%. Поэтому диапазон поддержки для поиска интересных правил зависит от квантилей характеристических атрибутов, присутствующих в наборе. Тем не менее установим нижнюю но составляет не менее 2%.

При расшифровке правил атрибут RMax характеризует мощность вычислительных ресурсов, имеющихся в данном научном учреждении, GRPpC – степень экономического развития региона, квартили Q1-Q4, Top10 – научный потенциал учреждения.

Образованные группы значений квантилей трактуются следующим образом: 1 квантиль соответствует лидерским, высоким показателям, 2 квантиль – повышенным, выдающимся, 3 квантиль – средним, 4 квантиль – базовым, типичным показателям.

4.2.1. Правила для научных учреждений

В табл. 7 представлены ассоциативные правила, в консеквенте которых находится научное учреждение.

Максимально возможная поддержка для правил, в которых фигурирует научное учреждение составляет порядка 6%. Поэтому диапазон для поиска интересных правил с таким характеристическим атрибутом установим от 2% до 6%.

Таблица 7. Правила для научных учреждений

№	Антецедент	Консеквент
1	[GRPperCapita: 1, Q3: 1, Linpack: 1, Q4: 1, Q2: 1, Q1: 1, Top10: 1]	МГУ
2	[Q1: 4, Top10: 4, Q4: 4, Linpack: 2, GRPperCapita: 4]	ЮУрГУ
3	[GRPperCapita: 1, Q1: 2, Linpack: 1, Q2: 2, Top10: 2, Q3: 2]	Курчатовский институт
4	[Q1: 3, Linpack: 2, Top10: 3, GRPperCapita: 4]	ННГУ
5	[Linpack: 4, Q3: 1, Q4: 1, Q2: 1, Q1: 1]	СПбГУ
6	[Q4: 2, Q1: 1, Top10: 1]	МФТИ

Данные правила иллюстрируют научный потенциал, мощность вычислительной системы и экономическое положение, присущие университету. Например, из правила №1 видно, что МГУ характеризует лидерство по всем характеристикам. А, например, в правиле №5 СПбГУ, имея суперкомпьютер, по мощности относящийся к базовому децилю, так же находится в лидерах по научному потенциалу.

Однако, очевидно, что выбранных атрибутов недостаточно, чтобы комплексно проанализировать причины и следствия положения университетов, делать выводы об эффективности используемых ресурсов.

4.2.2. Правила для экономического атрибута

В табл. 8 представлены ассоциативные правила, в консеквенте которых находится дециль ВРП.

Таблица 8. Правила для ВРП

№	Антецедент	Консеквент
1	[Linpack: 1, Q2: 1, Q1: 1, Top10: 1]	GRPperCapita: 1
2	[Q1: 2, Q3: 3]	GRPperCapita: 1
3	[Linpack: 4, Q3: 3, Top10: 2]	GRPperCapita: 1
4	[Q2: 4, Q3: 4, Q1: 4, Top10: 4, Q4: 4]	GRPperCapita: 4
5	[Q3: 4, Q1: 4, Top10: 4, Linpack: 2]	GRPperCapita: 4

Из правила №1 можно заключить, что суперкомпьютеры с высокой производительностью установлены в учреждениях с лидирующим научным потенциалом, которые расположены в экономически развитых регионах России. Но научное учреждение так же может

демонстрировать повышенный и средний научный потенциал, находясь в развитом регионе, как видно из правила №2. Правила №4 и №5 свидетельствуют о том, что базовый научный потенциал влечет за собой нахождение в экономически слабо развитом регионе как с наличием суперкомпьютера с повышенным уровнем производительности, так и без него.

4.2.3. Правила для производительности суперкомпьютеров

В табл. 9 представлены ассоциативные правила, в консеквенте которых находится дециль производительности суперкомпьютеров.

Таблица 9. Правила для производительности

№	Антецедент	Консеквент
1	[GRPperCapita: 1, Q2: 1, Q1: 1, Top10: 1]	Linpack: 1
2	[GRPperCapita: 1, Q2: 4, Q3: 4, Q1: 4, Top10: 4, Q4: 4]	Linpack: 4
3	[Q2: 4, Q3: 4, Q1: 4, Top10: 4, Q4: 4]	Linpack: 4
4	[GRPperCapita: 1, Q1: 4, Top10: 4]	Linpack: 4

Научное учреждение с лидирующим научным потенциалом и находящееся в экономически развитом регионе, как правило, имеет мощный суперкомпьютер. В то же время нахождение научного учреждения в регионе с высоким уровнем ВРП и базовым научным потенциалом по всем квартилям Scopus влечет за собой обладание относительно слабым суперкомпьютером. Если научное учреждение имеет низкий научный потенциал, то оно обладает низким вычислительным ресурсом, даже при условии нахождения в экономически развитом регионе.

Таким образом, видно, что нахождение научного учреждения в экономически развитом регионе России не гарантирует обладание производительной вычислительной системой высокого класса.

4.2.4. Правила для научных атрибутов

В табл. 10 представлены ассоциативные правила, в консеквенте которых находится дециль квартиля Scopus.

Таблица 10. Правила для квартилей Scopus

№	Антецедент	Консеквент
1	[Q3: 1, Q4: 1, Q2: 1, Top10: 1]	Q1: 1
2	[Linpack: 1, Q2: 1, Top10: 1]	Q1: 1
3	[GRPperCapita: 1, Linpack: 1, Q2: 1, Top10: 1]	Q1: 1
4	[GRPperCapita: 4]	Q2: 4
4	[Linpack: 4, Q4: 2]	Q1: 2

Из правила №1 видна согласованность научных квартилей. Правило №2 показывает, что если учреждение с высоким научным потенциалом в Q2 и Top10 обладает производительным суперкомпьютером, то оно лидер в Q1. Хотя из правила №4 видно, что продвинутый

научный потенциал возможен и с базовой вычислительной мощностью.

5. Заключение

В работе осуществлен процесс data mining над данными суперкомпьютерных рейтингов TOP500 и Топ-50.

Основные тенденции и закономерности, выявленные в ходе исследования следующие: Высокий уровень жизни в стране не является необходимым условием для достижения высокого научного потенциала. Уровень научного потенциала страны, как правило, согласован с количеством суперкомпьютеров страны, входящих в редакции рейтинга. Многие правила подтверждают зависимость количества.

Итого, можно заключить, что число публикаций явным образом коррелирует с количеством суперкомпьютерных систем. А высокий уровень жизни не является определяющим фактором для формирования научного потенциала страны.

Литература

1. Ячник О.О., Никитенко Д.А., Соболев С.И. Мобильный Linpack: первый опыт введения рейтинга производительности мобильных устройств // Вычислительные методы и программирование. 2018. Т. 19. С. 464-469.

2. 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.

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

4. 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

5. Абрамов С.М. Правда, искажающая истину. Как следует анализировать Top500? // Параллельные вычислительные технологии 2013. –Челябинск. –1-5 апреля 2013

6. 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.

7. 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.

8. Dongarra J.J., Meuer H.W., Strohmaier E. TOP500 Supercomputer Sites // -1999

9. Dongarra J.J The LINPACK Benchmark : An Explanation // LNCS, volume 297 -1988

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

11. Murphy R.C., Wheeler K.B., Barrett B.W., Ang J.A. Introducing the Graph 500 // -2010

12. Zelenkov Y.A., Sharsheeva J.A. Impact of the Investment in Supercomputers on National Innovation System and Country's Development

13. 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