Исследование списков TOP500 и Топ-50 методами 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) 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

Статья имеет следующую структуру. В разделе 2 кратко рассматривается история и причины появления суперкомпьютерных рейтингов, текущие результаты их статистических исследований и актуальные направления изучения 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 [9]. Редакции ведутся с 1993 года. В этом рейтинге суперкомпьютеры упорядочиваются по результатам их производительности на тесте Linpack [10].

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 вычисления 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 [11], Graph500 [12], GreenGraph500.

Тем не менее, наибольшую популярность имеет именно TOP500. В связи с этим многие производители суперкомпьютеров проектируют системы, с целью максимизации результатов достигаемых на тесте Linpack. Кроме того, понятие скорости вычислений суперкомпьютера зачастую приравнивается к его эффективности. Однако, существует мнение, что такая оценка некорректна, так как игнорирует другие важные метрики производительности такие как надежность, доступность и удобство использования [8].

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, распределение по области применения суперкомпьютеров, 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. [13].

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 на отрезке $[1 \ldots 10]$, 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 частоту встречаемости 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. Набор B в общем случае может быть произвольным, мы же ограничимся правилами, в консеквенте которых 1-набор. Иначе говоря, в контексте данной работы правила могут быть, например, такие: «если 1 дециль по суммарной производительности систем страны 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 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

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 достижения страны в области высокопроизводительных вычислений и суперкомпьютерных технологий;
2. атрибуты, отражающие экономический уровень развития страны;
3. атрибуты, отражающие научный потенциал страны.

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

В конечном итоге исследуемый объект представляет собой строку, которая содержит информацию о названии страны, присутствующей в редакции рейтинга TOP500 конкретного года, и о характеризующих ее атрибутов для этого же года.

В связи с тем, что новая редакция TOP500 публикуется дважды в год, а данные по количеству научных публикаций, внутреннего валового продукта на душу населения (ВВП)

[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 списка данного года
Num of Systems	TOP500	Общее количество систем страны в редакции списка данного года
RMax	TOP500	Общая Linpack производительность систем страны в редакции списка данного года
HDI	HDR	Величина индекса человеческого развития для данного года
GDPpC	TWB	Величина внутреннего валового продукта на душу населения страны для данного года
Q1-Q4	Scopus	Количество статей, опубликованных авторами данной страны в данном году в источниках, находящихся в данном году в каждом из Q1-Q4 по рейтингу CiteScore
Top10	Scopus	Количество статей, опубликованных авторами данной страны в данном году в 10% лучших источниках по рейтингу CiteScore данного года

и индекса человеческого развития (ИЧР) представляются лишь ежегодно, из 50 существующих на текущий момент редакций списка для исследования использовалась 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 статей различного уровня. Уровень статьи 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, 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 согласно уровню метрики 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 созданное на языке программирования Python приложение, использующее открытый API Scopus. Для этого формировался запрос, содержащий год публикации статьи *PUBYEAR IS <год>*, страну аффилиации хотя бы одного из авторов работы *AFFILCOUNTRY(<страна>)* и список идентификаторов мест публикаций, принадлежащих одному из квартилей *SOURCE-ID(<идентификатор источника>)*. Такие списки создавались на основе файла, содержащего все источники и их квартиль для годов с 2011 до 2017. Для редакций рейтинга TOP500, которые были выпущены с 1993 по 2011 год использовались сведения за 2011 год, в силу отсутствия более актуальных данных для метрики CiteScore.

3.3. Обработка данных

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.

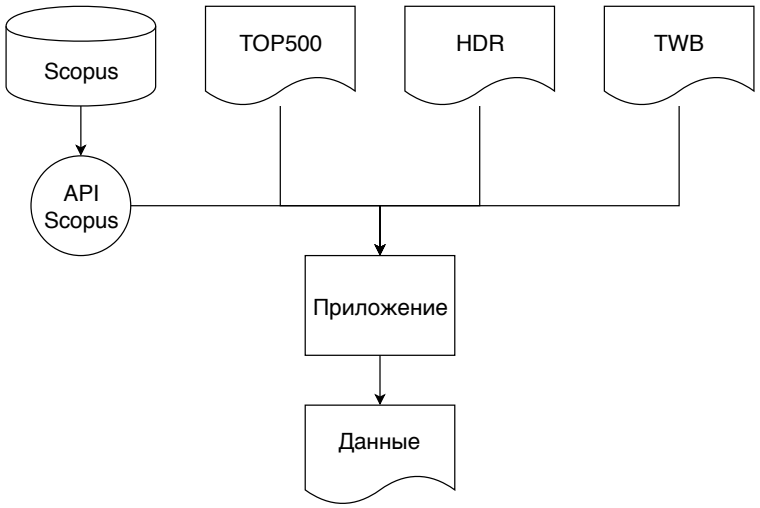

Рис. 1. Схема получения данных

На основе каждой редакции списка TOP500 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 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 [14]. С минимальным уровнем поддержки 1% и минимальным уровнем достоверности 70% было получено 6219 правил.

4. Экспериментальные результаты

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% - 7% является оправданной поддержкой для поиска устойчивых ассоциативных правил.

Как и ожидалось, большая часть полученных ассоциативных правил оказалась логично объяснимыми и интуитивно понятными.

Например, суммарная вычислительная мощность RMax больше всего у тех стран, у которых больше всего суперкомпьютеров в рейтинге, низкий показатель ИЧР связан с низким

уровнем ВВП и дециль Top10 равен децилю Q1, так как научные работы, опубликованные в источниках, входящих в 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 влиянии страны и количественном присутствии в сфере высокопроизводительных вычислений, квартили Q1-Q4 и Top10 – о научном потенциале, HDI и GDPpC – об уровне жизни.

Тогда децили будут обозначать следующие градации указанных характеристик: 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. 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 рейтинга TOP500.

Страны, для которых были обнаружены ассоциативные правила, образовали 7 групп. Они приведены в табл. 2, а примеры правил, в консеквенте которых находится одна страна для каждой из этих групп в табл. 3.

Таблица 2. Группы стран

№	Уровень жизни	Научный потенциал	Страны
1	Высокий	Высокий	Австралия, Германия
2	Средний	Высокий	Великобритания, Канада, Франция
3	Низкий	Высокий	Италия, Китай, Российская Федерация
4	Высокий	Средний	Нидерланды, Швейцария
5	Низкий	Средний	Бразилия, Индия, Польша
6	Высокий	Низкий	Норвегия
7	Низкий	Низкий	ЮАР

Такого рода правила отражают характерные черты страны, проявляющееся через описывающие их признаки, и косвенным образом демонстрируют ее стратегию развития.

Глядя на образованные группы, можно заметить, что высокий уровень научного потенциала достигается странами 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.

4.2. Правила для количества суперкомпьютеров в списках

Рассмотрим следующую табл. 4, где в консеквенте находится атрибут количества систем:

В абсолютном большинстве полученных правил дециль количества систем Num of Systems

Таблица 3. Примеры ассоциативных правил для одной страны из каждой группы

№	Антецедент	Консеквент
1	[HDI: 2, Q1: 2, Q2: 2, Q3: 2]	Country: Germany
2	[HDI: 4, Top10: 2, Q2: 3]	Country: Canada
3	[Num of Systems: 1, HDI: 10, GDPpC: 10, Q1: 1, Q2: 1, Q3: 1, Q4: 1]	Country: China
4	[GDPpC: 1, Q2: 6, Q3: 6, Q4: 6]	Country: Switzerland
5	[HDI: 10, GDPpC: 10, Q1: 4]	Country: India
6	[HDI: 1, GDPpC: 1, Q2: 9, Q4: 9]	Country: Norway
7	[HDI: 10, Q3: 8]	Country: South Africa

Таблица 4. Правила для количества систем

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

и дециль суммарной производительности RMax одинаковы.

То есть явление наличия большого числа относительно слабых суперкомпьютеров, суммарная производительность которых сравнима с той, которой обладают меньшее количество систем, но значительно более производительных, если и имеет место быть, то очень редко. Это свидетельствует о том, что в масштабах скоростей вычислений суперкомпьютеров ситуация превосходства по скорости за счет большего числа менее производительных суперкомпьютеров редко осуществима. Возможно, помимо этого такой подход просто нерелевантен, в силу необходимости нести гораздо большие финансовые затраты для установки суперкомпьютеров, применяющих новые дорогие технологии, позволяющие повысить предел достигаемой мощности. Выгоднее вводить в эксплуатацию несколько систем.

Из правил №1 и №4 видно, что высокий научный потенциал влечет за собой наличие большого числа суперкомпьютеров в рейтинге TOP500. Кроме того исходя из правил №2-№3, №5-№6 страны с различающимся уровнем жизни, но при этом обладающие высоким научным потенциалом, как правило, входят в число лидеров по количеству суперкомпьютерных систем, представленных в рейтинге.

Данное наблюдение позволяет предположить, что научный потенциал страны явным

образом коррелирует со степенью ее развитости в области высокопроизводительных вычислений и существенным образом зависит от нее. То есть преимущество в количестве суперкомпьютеров, а значит и в суммарной производительности, как правило обеспечивает преимущество в научной сфере.

4.3. Правила для экономических атрибутов

Рассмотрим следующую табл. 5, где в консеквенте находятся экономические атрибуты:

Таблица 5. Правила для экономических атрибутов

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

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). Примеры таких типичных правил – №1 и №2. Но были и исключения, которые приведены в таблице с номерами №3-№5.

Особенно интересно правило №3. Оно не соответствует общей тенденции согласованности высокой суммарной мощности суперкомпьютеров и высокого научного потенциала.

Правила №4 и №5 показывают, что для стран с достаточно высоким и средним научным потенциалом часто характерен средний и низкий уровень ВВП на душу населения.

Далее в табл. 6 приведены правила, связанные с квартилями Scopus.

4.4. Правила для квартилей Scopus

Квартили Q1-Q4 и Top10, как правило согласованны и имеют один и тот же, либо соседний дециль. То есть явление, когда страна лидирует в одном квартиле, а в другом занимает низкие позиции, в правилах не отражено.

Приведенные правила вновь подтверждают, что высокий научный потенциал достижим странами с средним или низким уровнем жизни.

5. Заключение

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.

Таблица 6. Правила для квартилей Scopus

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

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. Sharma S., Hsu C.-H., Feng W. Case for a Green500 List // 20th International Parallel and Distributed Processing Symposium, IPDPS -2006

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

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

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

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

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

14. 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