

Поиск ассоциативных правил в суперкомпьютерных рейтингах Top500 и Top50

М.Л. Цымблер, П.И. Шумилин

Южно-Уральский государственный университет

В настоящее время суперкомпьютерные рейтинги являются важным средством для отслеживания тенденций развития области высокопроизводительных вычислений [2]. В работе описан поиск ассоциативных правил (устойчивых корреляций между атрибутами) в международном рейтинге Top500 и российском рейтинге Top50 методами интеллектуального анализа данных.

Поиск ассоциативных правил охватывает технические характеристики суперкомпьютера, а также атрибуты организации и страны, в которых он установлен. Для анализа списка Top500 используются следующие атрибуты: количество суперкомпьютеров данной страны в данной редакции списка, показатели индекса человеческого развития (ИЧР) и валового внутреннего продукта на душу населения (ВВП) страны, количество публикаций учеными страны по квартилям Q1-Q4 Scopus. Для списка Top50 в поиск правил вовлечены вычислительная мощность суперкомпьютера, установленного в российской научной организации, показатель валового регионального продукта (ВРП), количество публикаций учеными данной организации по квартилям Q1-Q4 Scopus. С помощью разработанной программной системы сбора и подготовки данных из 26 редакций Top500 образовано 743 объекта, которые представляют собой транзакции, состоящие из значений описанных атрибутов. Для 15 редакций Top50 получено 248 объектов. Поиск правил осуществлялся с использованием алгоритма Apriori [1] в аналитическом пакете KNIME. Было обнаружено более тысячи правил для каждого списка. Несколько десятков правил показывают интересные корреляции.

В результате поиска были найдены следующие интересные правила. Если страна, обладая относительно низким уровнем ИЧР, при этом находится среди лидеров по количеству публикаций в Q1, то она так же имеет относительно большое число суперкомпьютеров, входящих в рейтинг Top500. Для среднего числа публикаций в Q4 характерно обладание страной средним числом публикаций в Q3 и средним количеством суперкомпьютеров, входящих в рейтинг Top500. Если научная организация находится в регионе с лидирующим уровнем ВРП, но имеет относительно низкую вычислительную мощность суперкомпьютера среди остальных систем Top50, то количество публикаций в Q1 учеными из этой аффилиации ниже среднего уровня. Если уровень ВРП региона относительно низок, а количество публикаций в Q1 и Q2 выше среднего уровня, то в организации установлен суперкомпьютер с производительностью выше среднего уровня.

В целом найденные правила подтверждают корреляцию между располагаемым вычислительным ресурсом и научным потенциалом как на уровне стран, так и на уровне российских научных организаций.

Литература

1. Agrawal R., Srikant R. Fast Algorithms for Mining Association Rules // Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, September 1994. Vol. 1215. P. 487–499.
2. Nikitenko D., Zheltkov A. The Top50 list vivification in the evolution of HPC rankings // Parallel Computational Technologies. PCT 2017. Communications in Computer and Information Science. 2017. Vol. 753.