DOI: 10.37791/2687-0649-2020-15-6-27-41

Решение задачи препроцессинга календарных данных в ходе реализации технологии Data Mining

Б. В. Окунев[1*]***, А. С. Шурыкин***[1]
[1] *Филиал Национального исследовательского университета «МЭИ» в г. Смоленске, Россия*
** ok-bmv@rambler.ru*

Аннотация. В настоящий момент «грязные» данные, то есть данные низкого качества, становятся одной из главных проблем эффективного решения задач Data Mining. Так как исходные данные аккумулируются из самых разных источников, то вероятность попадания «грязных» данных весьма высока. В связи с этим одной из важнейших задач, которую приходиться решать в ходе реализации Data Mining-процесса, является первоначальная обработка (очистка) данных, то есть препроцессинг. Необходимо заметить, что препроцессинг календарных данных является достаточно трудоемкой процедурой, которая может занимать до половины всего времени реализации технологии Data Mining. Сокращения времени, затрачиваемого на процедуру очистки данных, можно достичь, автоматизировав процесс с помощью специально разработанных инструментов (алгоритмов и программ). При этом следует помнить, что применение вышеуказанных элементов не гарантирует стопроцентную очистку «грязных» данных, а в некоторых случаях даже может приводить к появлению дополнительных ошибок в исходных данных. Авторами разработана модель автоматизированного препроцессинга календарных данных на основе синтаксического анализа и регулярных выражений. Предлагаемый алгоритм отличается гибкой настройкой параметров препроцессинга, достаточно простой реализуемостью и высокой интерпретируемостью результатов, что в свою очередь дает дополнительные возможности при анализе неудачных результатов применения технологии Data Mining. Несмотря на то, что предлагаемый алгоритм не является инструментом очистки абсолютно всех типов «грязных» календарных данных, он успешно функционирует в значительной части реальных практических ситуаций.

Ключевые слова: препроцессинг данных, интеллектуальный анализ данных, регулярные выражения, парсинг данных, метаданные

Для цитирования: *Окунев Б. В., Шурыкин А. С.* Решение задачи препроцессинга календарных данных в ходе реализации технологии Data Mining // Прикладная информатика. 2020. Т. 15. № 6. С. 27–41. DOI: 10.37791/2687-0649-2020-15-6-27-41