УДК 004

АЛГОРИТМ ФОРМАТНОЙ ПРОВЕРКИ XML-ФАЙЛА В НЕСКОЛЬКИХ ПОТОКАХ

Кравцов М.Д., студент гр. ПИм-151, II-курс
Научный руководитель: С.А. Веревкин, ст. преподаватель
Кузбасский государственный технический университет
имени Т.Ф. Горбачева
г. Кемерово

Перед обработкой любой структурированной информации необходимо проверить ее структуру, чтобы избежать дальнейших ошибок при обработке. Однако это может оказаться нелегкой задачей, если мы имеем большой объём. В этом случае от ЭВМ требуется много ресурсов, в частности, оперативной памяти, но в рамках статьи опустим эти детали и будем считать, что ресурсов используемой ЭВМ достаточно, чтобы проверить файл, но даже в этом случае потребуется длительное время на выполнение процедуры форматной проверки по ряду причин. Одной из главных причин можно выделить тот факт, что процесс проверки занимает длительное время. Это может вызвать, как минимум, неудобства, если требуется уложиться в некоторые временные рамки. Для подобных задач существуют готовые решения, как программные, так и программно-аппаратные, но они могут не подходить, например, в силу своей дороговизны или возможностей интеграции. Именно поэтому может возникнуть вопрос о написании собственного программного решения.

Как уже говорилось выше, процесс валидации XML-файла выполняется в 1 потоке, а значит, не используются все ресурсы процессора. Начнем с того, что нужно разбить этот процесс на несколько потоков – это существенно снизит временные затраты и позволит использовать все ресурсы процессора.

Для начала условимся, что весь объём информации записан в одном XML-файле, а его объём, скажем, равен 10Гб. Не учитывая возможности установленного накопителя, считаем, что этап загрузки файла в оперативную память почти не тормозит весь процесс проверки.

Вторым этапом будет индексация всех узлов XML-данных. На этом этапе неизвестно, имеются ли форматные ошибки, поэтому индексирование будет строго по нахождению имен узлов. По завершению этапа имеется таблица с индексами узлов и их положения в документе.

Следующий этап заключается в условном делении документа, используя таблицу индексов, на блоки, которые далее будут проходить процедуру форматной проверки. Каждый блок представляет собой XML-данные, соответствует заранее определенной схеме и имеет существенно меньший объём информации, по сравнению с исходным документом. Это гарантирует, что проверка этого блока не займет много времени.

Последний этап – форматная проверка. На этом этапе каждый из блоков проходит форматную проверку с заранее подготовленным описанием формата блока используя стандартные средства валидации, доступные во многих языках программирования. Каждый процесс валидации блока выполняется в отдельном потоке, что и позволяет использовать все ресурсы центрального процессора и существенно сократить временные затраты на процесс валидации в целом.