УДК 004.4

ВЛИЯНИЕ СПОСОБА ПАРСИНГА XML-ДОКУМЕНТОВ НА ВРЕМЯ ОБРАБОТКИ ИНФОРМАЦИИ

© Кочетков В.С., Логанова Л.В.

e-mail: kcht2312@gmail.com

Самарский национальный исследовательский университет имени академика С.П. Королёва, г. Самара, Российская Федерация

В настоящее время XML является самым привлекательным форматом информационного обмена. Особенно важной представляется его возможность, которая позволяет отделять форматирование от информационного наполнения, а также экспортировать данные и манипулировать ими.

Основными методами работы с XML являются[1]:

- Simple XML
- DOM
- SAX
- StAX

Принимая во внимание, что XML-документы могут содержать избыточные данные, которые не используются для последующего их анализа и, выбирая те или иные методы работы с XML, можно снизить время обработки документов.

В качестве контрольной выборки были использованы открытые данные сервера госзакупок zakupki.gov.ru, доступные на соответствующем FTP, определен набор извлекаемой информации.

Работа включает в себя реализацию различных способов парсинга XML-документов и исследование времени обработки тем или иным методом контрольной выборки документов для выработки рекомендаций относительно использования технологии извлечения данных.

В отличии от DOM-парсера, SAX-парсер осуществляет связь с приложением посредством функций обратного вызова, по этой причине на время обработки файлов влияет количество извлекаемых тегов. Производя контрольные замеры времени исполнения методов SAX-парсера, была выявлена зависимость времени выполнения обработки данных от количества извлекаемых тегов и количества обрабатываемых файлов (документов), представленная в таблице.

Таблица. Время выполнения обработки данных

Число извлекаемых тегов	Количество документов	Время выполнения преобразования, мс					
		1	2	3	4	5	среднее
4	1000	9374	8545	7500	8850	8726	8599
8	1000	12011	12697	12045	11486	12619	12171
16	1000	11853	14653	12235	13596	13277	13122

Полученные результаты позволяют сделать вывод, что при увеличении количества извлекаемых тегов, и соответствующих условий, прописанных в классе парсера, незначительно увеличивается время выполнения их обработки (до 30% при увеличении количества извлекаемых тегов в 2 раза).

Полученные результаты позволяют судить о том, что в рамках работы с XML-документами, содержащими избыточные данных, оправдано использование SAX-парсера, поскольку этот метод не требует загрузки файла как целого объекта, что оптимизирует расход памяти, уменьшает время выполнения обработки данных. В случае увеличения количества извлекаемой информации время обработки изменяется незначительно.

В дальнейшем над извлекаемыми данными может быть выполнена дополнительная обработка, связанная с классификацией данных.

Библиографический список

1. Дэвид Хантер, Джефф Рафтер. XML. Базовый курс, 4-е издание, Вильямс, 2018. – 1344 с.
2. Бретт Маклахлин. Java и XML, 2-е издание, Символ-Плюс, 2016. – 544 страницы.
3. Выгрузка условий конкурсов с zakupki.gov.ru // [habrahabr.ru] –2015 – URL: https://habr.com/post/253201/
4. Интеграция XML данных – другой путь // [habrahabr.ru] –2017 – URL: https://habr.com/ru/post/325186/