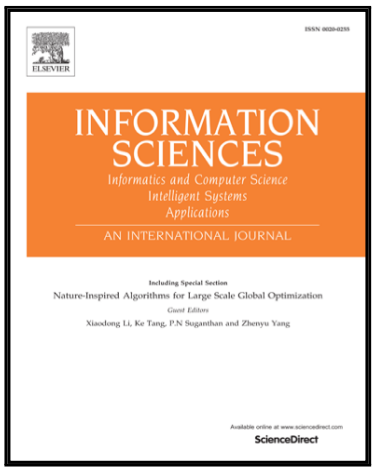

Handling distributed XML queries over large XML data based on MapReduce framework

Hongjie Fan, Zhiyi Ma, Dianhui Wang, Junfei Liu

PII: S0020-0255(18)30284-6
DOI: 10.1016/j.ins.2018.04.028
Reference: INS 13571

To appear in: *Information Sciences*

Received date: 11 September 2017
Revised date: 3 April 2018
Accepted date: 5 April 2018

Please cite this article as: Hongjie Fan, Zhiyi Ma, Dianhui Wang, Junfei Liu, Handling distributed XML queries over large XML data based on MapReduce framework, *Information Sciences* (2018), doi: 10.1016/j.ins.2018.04.028

Handling distributed XML queries over large XML data based on MapReduce framework

Hongjie Fan[a,b], Zhiyi Ma[a,b,*], Dianhui Wang[c], Junfei Liu[d]

[a] *School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China*
[b] *Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education, Beijing, 100871, China*
[c] *Department of Computer Science and Information Technology, La Trobe University, Melbourne, VIC 3086, Australia*
[d] *National Engineering Research Center for Software Engineering, Peking University, Beijing 100871, China*

Abstract

With the increase in available extensible markup language (XML) documents, numerous approaches to querying have been proposed in the literature. XPath queries and Twig pattern queries are the two basic approaches, directly affecting the efficiency of XML operations. Distributive manipulation of massive XML data is challenging. This paper aims to develop an efficient distributed XML query processing method using MapReduce, which simultaneously processes several queries on large volumes of XML data. First, we split up a large-scale XML data file into file-splits and put them in a distributed storage system. Then, we present an efficient algorithm to compute different fragments of the document tree using the MapReduce framework in parallel. In order to efficiently handle a large amount of XML data, we built a partition index and used a random access mechanism for specific queries. The experiment results show that our proposed approach is efficient with good scalability.

Keywords: XML, XPath Query, Twig Query, Hadoop, MapReduce

*Corresponding author
Email address: mazhiyi@pku.edu.cn (Zhiyi Ma)
URL: hjfan@pku.edu.cn (Hongjie Fan), dh.wang@latrobe.edu.au (Dianhui Wang), liujunfei@pku.edu.cn (Junfei Liu)