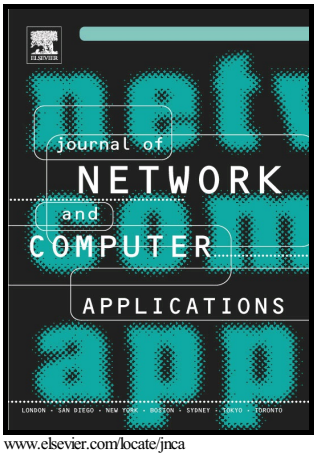
Data popularity measurements in distributed systems: Survey and design directions

C. Hamdeni, T. Hamrouni, F. Ben Charrada



www.elsevier.com/locate/jnca

PII: S1084-8045(16)30120-5
DOI: http://dx.doi.org/10.1016/j.jnca.2016.06.002
Reference: YJNCA1659

To appear in: *Journal of Network and Computer Applications*

Received date: 23 November 2015
Revised date: 20 May 2016
Accepted date: 3 June 2016

Cite this article as: C. Hamdeni, T. Hamrouni and F. Ben Charrada, Data popularity measurements in distributed systems: Survey and design directions *Journal of Network and Computer Applications* http://dx.doi.org/10.1016/j.jnca.2016.06.002

Data popularity measurements in distributed systems: Survey and design directions

C. Hamdeni, T. Hamrouni, and F. Ben Charrada

Computer Science Department, Faculty of Sciences of Tunis, Tunis El Manar University, University Campus, Tunis, Tunisia.
hamdeni.chamseddine@gmail.com, tarek.hamrouni@fst.rnu.tn, f.charrada@gnet.tn

Abstract

Distributed systems continue to be a promising area of research particularly in terms of providing efficient data access and maximum data availability for large-scale applications. For improving performances of distributed systems, several data replication strategies have been proposed to ensure reliability and data transfer speed as well as to offer the possibility to access the data efficiently from multiple locations. Data popularity is one of the most important parameters taken into consideration when designing data replication strategies. It assesses how much the data is requested by the sites of the system. In this paper, the importance of considering the data popularity parameter in replication management is highlighted. Different strategies are then identified and how they rely on the data popularity parameter is illustrated. Different calculation manners of data popularity are hence studied. This allows us to find out which factors are considered in order to assess data popularity. After classifying them into four categories, this work includes a critical discussion about each category. Some important directions for future work are then discussed towards possible solutions for a more effective data popularity assessment.
Keywords Distributed system, replication strategy, data popularity, access pattern, temporal locality.

1 Introduction and motivations

Distributed systems constitute a useful solution to deal with large scale applications that generate huge volumes of data. Replication of data across diverse locations in the system is needed in order to increase data reliability, availability, accessibility, and fault tolerance, while decreasing data access time and network traffic [2, 20]. For this purpose, several data replication strategies have been proposed in many distributed systems. Such systems include mainly data grid [11, 16, 29], cloud storage [30, 38], P2P systems [24, 44, 56] and Content delivery network (CDN) systems [22, 40]. An effective management of data, and more particularly popular ones since highly requested, is then of paramount importance. Indeed, the advantages of relying on popularity prediction is improved decision of what data need increased availability, for improved storage resource utilization, and the possibility to increase data availability preemptively [4].

Data popularity is one of the most important parameters taken into consideration when designing replication strategies. It consists in measuring how much a given piece of data is requested by the system sites. This constitutes key information since it gives an indication of the importance of this data which allows a more intelligent data placement and a large optimization in the storage utilization. Indeed, a reliable localization of the most frequently accessed data w.r.t. their associated numbers of accesses as well as the timing of those