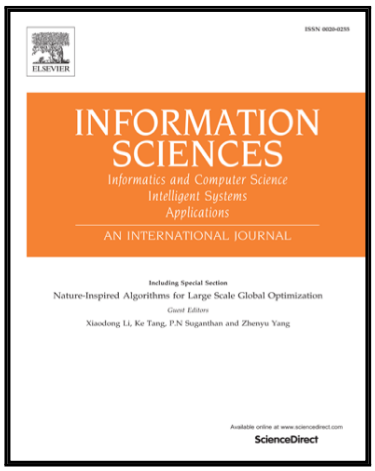
Accepted Manuscript



Column-wise compression of open relational data

Sebastian Wandelt, Xiaoqian Sun, Ulf Leser

PII: S0020-0255(16)31545-6
DOI: 10.1016/j.ins.2018.04.074
Reference: INS 13617

To appear in: *Information Sciences*

Received date: 7 November 2016
Revised date: 18 January 2018
Accepted date: 28 April 2018

Please cite this article as: Sebastian Wandelt, Xiaoqian Sun, Ulf Leser, Column-wise compression of open relational data, *Information Sciences* (2018), doi: 10.1016/j.ins.2018.04.074

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Column-wise compression of open relational data

Sebastian Wandelt[a,b], Xiaoqian Sun[a,b,*], Ulf Leser[c]

[a]*School of Electronic and Information Engineering, Beihang University, Beijing, China*
[b]*Beijing Key Laboratory for Network-based Cooperative ATM, Beijing, China*
[c]*Knowledge Management in Bioinformatics, Humboldt-University of Berlin, Berlin, Germany*

Abstract

The recent growth of open data initiatives has led to a tremendous increase in publicly available data resources. This amount of data, together with a rising interest in people to analyze it, poses severe challenges regarding data storage. Data suppliers often compress their resources with standard compressors. The choice of a compression technique, however, has significant impacts on the compression ratio, the compression speed, and the decompression speed.

In this paper, we provide an empirical analysis on the compression of open data provided in a relational format, such as comma-separated value files. We consider several compression tools and parameter settings. Furthermore, we propose using a novel column-wise compression strategy, where items that have similar properties, are compressed together. We perform a comprehensive analysis on 24 datasets from different domains, such as life sciences, governmental data, finance sector, and public transportation, which cover a wide range of file sizes (from a few MB to several GB). Our results show that the traversal strategy is of paramount importance for achieving high compression ratios; with improvements of up to one order of magnitude. This study further highlights a set of issues for future work on compressing open data.

1. Introduction

The Open Data movement is an emerging force, with the idea that certain data should be freely available for everyone to use [3, 7, 9]. There are multiple initiatives at various organizational levels, such as the Europe 2020 Initiative, which strongly encourage a culture of sharing [14]. Another example is the Open Science Initiative [28], which states that scientific data, whether collected directly as part of an experiment or indirectly as a secondary output of downstream analysis, should be made freely available in electronic form and accessible online [36]. Accordingly, many large datasets were provided for various domains, for instance, health care [1], transportation [31], geographical [18], public sector [19], and education [22]. Clearly, reproducibility and efficiency of scientific processes benefit, the more data is made openly available in a useful manner [25]. Therefore, many researchers have advocated for the reduction of barriers to the availability and reusability of scientific data [32, 30, 27]. Furthermore, a robust citation benefit from open data was found, and that, at least for gene expression microarray data, a substantial fraction of archived datasets are reused, and that the intensity of dataset reuse has been steadily increasing since 2003 [29]. In this light, researchers proposed that data should be considered as first class citizens for scientific sharing and publishing [5].

It has been argued that raw data must be formatted in a standard way, together with appropriate metadata attached [32]. In addition, complexity of open data must be reduced in order to increase attraction [20]. Therefore, open datasets are often released in simple file formats, such as comma-separated value (CSV) files. This file format is chosen since parsing is trivial and many tools support processing of CSV files and importing these files into database systems for post-processing. However, the release of complete raw datasets comes at the price of high storage and transmission costs. Hence, data providers often compress their data simply using the standard compression format zip. The choice of a compressor, however, has a tremendous impact on the result of the compression. A comprehensive analysis of which compressor is used (or should be used) for compression of open data is, to the best of our knowledge, not available.

In this paper, we analyse the compressibility of 24 carefully selected datasets. We report on the effectiveness of different standard compression techniques (zip, bzip2, gzip, LZMA, LZMA2, and PPMd) with different parameter

*Corresponding author: sunxq@buaa.edu.cn

Download English Version:

https://daneshyari.com/en/article/6856323

Download Persian Version:

https://daneshyari.com/article/6856323

Daneshyari.com