RProtoBuf: Protocol Buffers for R

Romain François

romain@r-enthusiasts.com

Dirk Eddelbuettel

edd@debian.org

Submitted to *useR! 2010*

Abstract

Protocol buffers are a flexible, efficient, automated mechanism for serializing structured data—think XML, but smaller, faster, and simpler. Users define how they want the data to be structured once in a `proto` file and then use special generated source code to easily write and read structured data to and from a variety of data streams and using a variety of officially supported languages— Java, C++, or Python—or third party implementations for languages such as C#, Perl, Ruby, Haskell, and now R via the **RProtoBuf** package.

The **RProtoBuf** package implements R bindings to the C++ protobuf library from Google. It uses features of the protocol buffer library to support creation, manipulation, parsing and serialization of protocol buffers messages. Taking advantage of facilities in the **Rcpp** package, **RProtoBuf** uses S4 classes and external pointers to expose objects that look and feel like standard R lists, yet are managed by the underlying C++ library. These objects also conform to the language-agnostic definition of the message type allowing access to their content from other supported languages.

As the protocal buffers library does not offer any built-in support for networked access to protocol buffer streams, we intend to take advantage on ongoing changes to the R system to expose a native R server. This is work-in-progress, but we hope to be able to report on this at the conference.

Keywords: R, C++, Serialization, Data Interchange, Data Formats