Parallel Distributed Text Mining in R

Stefan Theussl\textsuperscript{1}, Ingo Feinerer\textsuperscript{2}, and Kurt Hornik\textsuperscript{1}

\textsuperscript{1} Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, A-1090 Wien, Austria
Stefan.Theussl@wu-wien.ac.at, Kurt.Hornik@wu-wien.ac.at
\textsuperscript{2} Institute of Information Systems, DBAI Group, Technische Universität Wien, A-1040 Wien, Austria
feinerer@dbai.tuwien.ac.at

Summary. During the last decade text mining has become a widely used discipline utilizing statistical and machine learning methods, both in academics and for business intelligence. Recently R \cite{R2008} has gained explicit text mining support via the \texttt{tm} \cite{tm2008} package. This infrastructure provides sophisticated methods for document handling, transformations, filters, and data export (e.g., term-document matrices).

However, the steady growth and availability of large data sets poses new challenges for such a text mining framework: the corpora cannot be efficiently processed on a single computer mainly due to memory restrictions. On the other hand, there is now an increasing number of multicore processors, or even high performance computing environments, i.e., distributed and highly integrated computing clusters.

We propose techniques to take advantage of high performance computing via adding layers to the \texttt{tm} package which provide parallelism and distributed allocation: In detail we identify parts of \texttt{tm} which are sensitive to speed and performance, break these parts up into suitable building blocks for parallel processing and finally encapsulate the emerging parallelism in a functional programming style. A key factor in large scale text mining is the efficient management of data. Therefore, we show how distributed storage can be utilized to facilitate parallel processing of large data sets. This approach offers us a reliable, flexible, and scalable high performance solution for distributed text mining.

Key words: Text mining, Parallelism, Distributed allocation, High performance computing

References