Distributed Text Mining with tm

Stefan Theussl[1,*], **Ingo Feinerer**[2], **Kurt Hornik**[1]

1. Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, A-1090 Wien, Austria
2. Institute of Information Systems, DBAI Group, Technische Universität Wien, A-1040 Wien, Austria

* Contact author: Stefan.Theussl@wu-wien.ac.at

Keywords: High performance computing, Text mining, MapReduce, Distributed file system

Text mining is a widely used technique utilizing statistical and machine learning methods to extract patterns or knowledge from large unstructured text data sets. Recently R has gained explicit text mining support via the **tm** [2, 3] package. This infrastructure provides sophisticated methods for document handling, transformations, filters, and data export (e.g., term-document matrices).

However, the availability of very large and always growing text corpora poses new challenges for efficient handling of these data sets mainly due to architectural performance limits of single processor environments and memory restrictions. On the other hand we observe an increasing availability of multicore architectures even in commodity computers and high performance computing environments, i.e., distributed and highly integrated computing clusters.

In this context, we propose to make use of a technique called MapReduce [1] which is widely used in high performance computing because of its functional programming nature. Existing building blocks in **tm** allow for adding new layers to support this kind of parallelism and distributed allocation. In particular we identify compute-intensive parts of **tm**, break these parts up into suitable entities 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 and very large data sets. This approach offers us a reliable, flexible, and scalable high performance computing solution for distributed text mining.

References

[1] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. In *OSDI'04, 6th Symposium on Operating Systems Design and Implementation*, pages 137–150, 2004.

[2] Ingo Feinerer. *tm: Text Mining Package*, 2009. R package version 0.3-3.

[3] Ingo Feinerer, Kurt Hornik, and David Meyer. Text mining infrastructure in R. *Journal of Statistical Software*, 25(5):1–54, March 2008.