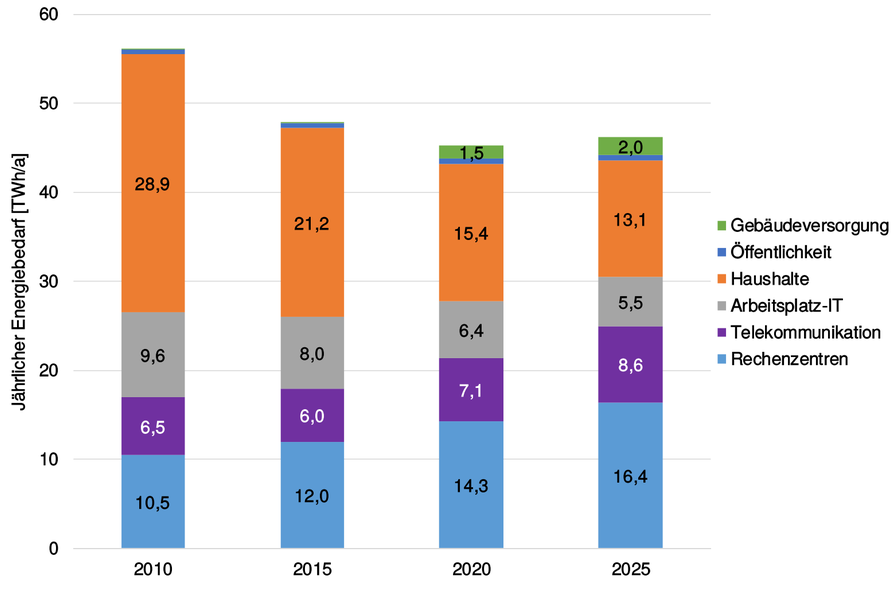
Burrows-Wheeler-Transform Tunneling

Uwe Baier | Institut für theoretische Informatik | Universität Ulm

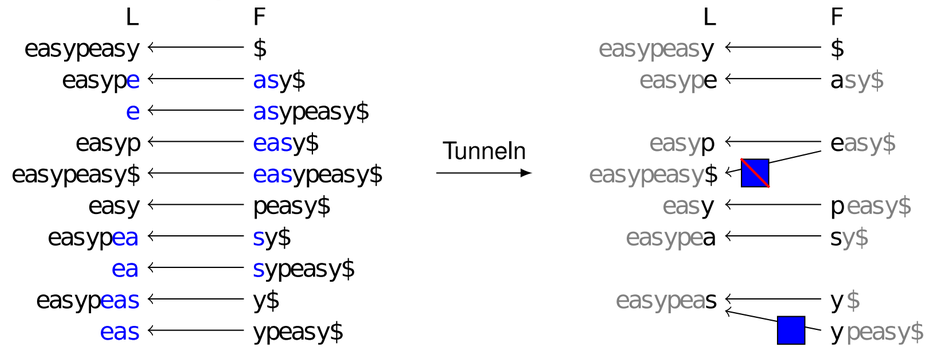
Motivation

- ca. 9.3 % des deutschen Stromverbrauchs im IKT – Sektor (Informations- und Kommunikationstechnik)
- Stromverbrauch von Rechenzentren und Kommunikation steigend (DNA-Analysen sind sehr rechenintensiv)
- Verbesserungsansatz: Datenkompression, Verarbeitung von komprimierten Daten (keine Dekodierung notwendig)



BWT Tunneling

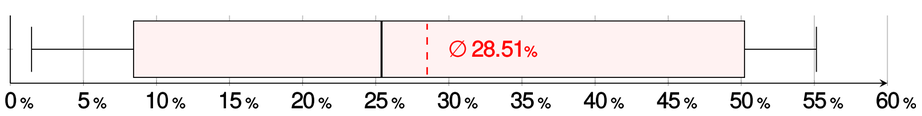
- BWT: Reversible Texttransformation basierend auf Vorgängerbuchstaben von sortierten Suffixen
- Einsatz in Datenkompression und Sequenzanalyse
- Im Rahmen der Dissertation: Entwicklung von Tunneling
- Tunneling: Identische Präfixe von aufeinanderfolgenden sortierten Suffixen können verschmolzen werden
- Tunneling erhält Eigenschaften einer BWT, reduziert dabei aber die Länge einer BWT deutlich



Herkömmliche BWT (links, L-Spalte) und getunnelte BWT (rechts) für den Beispielstring easypeasy$.

Einsatz in der Datenkompression

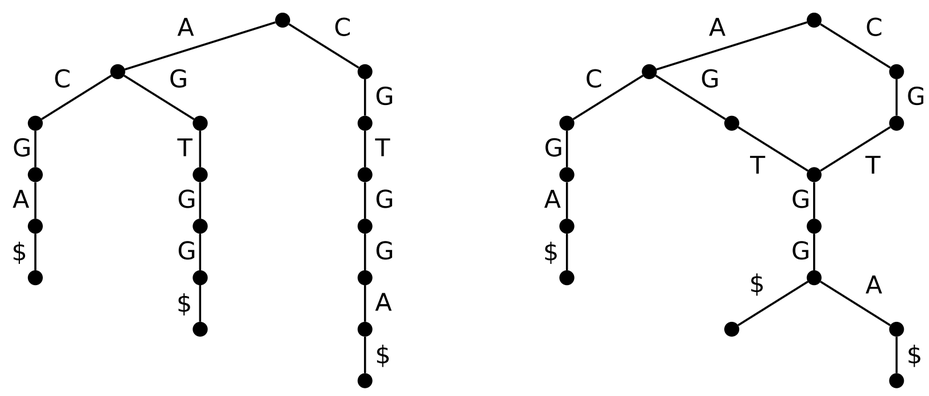
- Erweiterung bekannter BWT-Komprimierer mit Tunneling
- Kompressionsvergleich auf 41 Daten aus verschiedensten frei zugänglichen Textsammlungen (u.a. deutsche und englische Texte, Sourcecode, XML, DNA-Sequenzen, ...)
- Kodierungsverkleinerung vor allem bei großen Dateien (› 50 MB), im Vergleich mit normaler BWT-Kompression um durchschnittlich 29 % und bis zu 55 %
- Kompression wettbewerbsfähig zu den neuesten anderen Kompressionstechniken → effizienter Austausch zwischen BWTs auf verschiedenen Rechnern



Boxplot zur Reduzierung der Kodierungsgröße durch Tunneling bei Dateien › 50MB und einem der besten öffentlich zugänglichen BWT-Komprimierer.

Einsatz in der Sequenzanalyse

- Anwendung auf beliebte Datenstrukturen aus der Sequenzanalyse: DeBruijn-Graphen und Tries (baumartig aufgebaute Wörterbücher)
- Erste Tests: Senkung der Datenstrukturgröße um durchschnittlich 80 % bei repetitiven Eingaben



Herkömmlicher Trie (links) und getunnelter Trie (rechts) für die Strings ACGA$, AGTGG$ und CGTGGA$.

Kontakt: uwe.baier@uni-ulm.de | https://www.uni-ulm.de/in/theo/m/msc-uwe-baier/