Forschung aktuell | Aus Naturwissenschaft und Technik

13.2.2002

Sprachen erkennen mit Zip

Italienische Forscher nutzen Komprimierung zur Sprachenklassifizierung

Informatik. - Das Kompressionsformat Zip zum Verkleinern von Dateien ist erst recht mit dem Siegeszug des Internets ein vertrauter Begleiter jedes Computer-Nutzers. Italienische Forscher haben nun eine ungewöhnliche neue Anwendung für die nützlichen Helfer gefunden: Mit ihm lässt sich auch die Sprache eines Textes feststellen und in manchen Fällen sogar dessen Autor.

"Normalerweise benutzt man das Zip-Programm, um Datenmengen zu komprimieren: Texte, Bilder, alles, was wir in einen Computer eingeben", sagt der Informatiker Emanuele Caglioti von der Universität Rom. Bis auf ein Fünftel der ursprünglichen Größe lässt sich eine Datei mit Zip verkleinern. Der Zip zugrunde liegende Lempel-Ziv-Algorithmus packt ähnliche Bestandteile einer Datei eng zusammen. Je mehr Ähnlichkeiten es gibt, desto kleiner kann das Zip-Paket schließlich werden. Zusammen mit seinen Kollegen Dario Benedetto und Vittorio Loreto kam Caglioti auf eine neue Anwendung für dieses Verhalten: "Unsere Idee war es, zwei Texte miteinander zu vergleichen. Wir wollten wissen, in welcher Sprache ein Text, in unserem Fall Englisch, verfasst ist. Dazu verkleinerten wir zunächst Zeichenfolgen in Französisch, Englisch und Deutsch und notierten die Komprimierungslängen. Anschließend fügten wir den Zeichenfolgen ein Fragment in englischer Sprache hinzu." Auch die neuen Zeichenfolgen wurden "gezippt". Die rein englischsprachige unter den drei Dateien ist dann am wenigsten gewachsen.

Grund dafür ist der Packalgorithmus von Zip und die gewachsene Häufigkeitsverteilung vom Buchstaben, Silben und Wörtern in unterschiedlichen Sprachen. Das Wortfragment "the" etwa taucht im Englischen sehr häufig, im Deutschen eher selten auf, dort ist aber das "sch" viel verbreiteter. Auf diese Weise lassen sich nicht nur Sprachen, sondern auch Textinhalte definieren. In Rechtstexten etwa kommen bestimmte Begriffe sehr oft vor, die in Liebesromanen meist völlig fehlen. Hat man also einige Referenztexte aus einem Bereich, lässt sich allein über das Komprimieren bestimmen, ob ein zweiter Text ebenfalls dazu gehört. Auch der Stil eines Autors ist durch seine Wortwahl und sprachlichen Vorlieben geprägt, was die italienischen Forscher ebenfalls in einigen Tests erfolgreich zur Autorenbestimmung nutzen konnten. Schließlich lässt sich mit der Zip-Analyse auch das Alter von Texten in gewissem Rahmen feststellen. Denn heute sind andere Begriffe gebräuchlich als in früheren Jahrhunderten - zum Beispiel das Wort "Zip-Datei".

related links

- Übersicht: Sendu A-Z
- Der aktuelle Tag
- Manuskripte der 30 Tage
- Manuskripte seit 2001
- Manuskripte 199
- Computer & Kommunikation
- Wissenschaft im Brennpunkt
- Die Sternzeit
- **Schwerpunkt Stammzellen-Forschung**
- BSE-Sonderseite
- Forschung Aktue es auch als mona "Wissens-Update 5.06 EUR auf CI Hier können Sie bestellen.

[Quelle: Thomas Migge]