heise online

Meldung vom 28.01.2002 15:04

c't iX Telepolis

[<< Vorige] [Nächste >>]

Zip-Algorithmus identifiziert Autoren

Italienische Wissenschaftler haben eine elegante Methode gefunden, um per Software die Sprache eines Textes und mit hoher Wahrscheinlichkeit auch den betreffenden Autor zu identifizieren. Wie Dariao Benedetto, Emanuele Caglioti und Vittorio Loreto von der Università degli Studi di Roma "La Sapienza" in der Fachzeitschrift *Physical Review Letters* (Physical Review Letters No. 88 S. 048702, 28. Januar 2002) beschreiben (Preprint), lässt sich der Lempel-Ziv-Algorithmus (LZ77), der normalerweise zur Kompression von Dateien verwendet wird, sehr gut für solche Proben einsetzen.

Der Quotient aus der Länge eines komprimierten Textes und der unkomprimierten Länge strebt nämlich gegen die Entropie des Zeichenstroms. Der Algorithmus legt beim Komprimieren eine Art Wörterbuch für bereits bekannte Zeichenfolgen an. Wenn eine solche Zeichenfolge im Verlauf des Textes wiederholt auftritt, so wird nur noch auf den entsprechenden Eintrag im Wörterbuch verwiesen.

Die Forscher verglichen je zehn Textpassagen in zehn verschiedenen europäischen Sprachen miteinander. Sie komprimierten zunächst eine bekannte Zeichenfolge. Danach fügten sie der ursprünglichen Abfolge ein kurzes Fragment des zu prüfenden Vergleichstextes an und komprimierten auch diese Datei. Die Längendifferenz zwischen den beiden komprimierten Zeichenabfolgen liefert nun ein Maß dafür, wie nahe sich die beiden ursprünglichen Zeichenreihen stehen -- eine minimale Längendifferenz trat nur dann auf, wenn die Sprachen der beiden Texte identisch

Top-Meldungen

Informationsfreiheit hat Priorität vor dem Urheberrecht

Öffentliche Petition soll Linux in den Bundestag bringen

Power Mac G4 mit 1 GHz und GeForce-4-Grafik

Tipps für Microsofts Sicherheitskonzept

Aktuelle Meldungen

heise online-Chat: Wie seriös sind Online-Versteigerungen?

SAP will IBM und BEA Konkurrenz machen

Google findet Pop-ups nervig

Ascom-Chef sieht den Telecom-Konzern am Ende der Talsohle

IBM: Linux rechnet sich

Lieferengpass bei Hynix

Offizielles Linux-Release für die Playstation 2

Realnetworks glaubt an den Aufschwung

KPN schluckt E-Plus vollständig

Gebühren für .biz-Registrierung werden zurückgezahlt (Update)

Kartellamt gibt Vorentscheidung zu Libertys Kabelnetzkauf bekannt

Weitere Entlassungen

Suchen nach... Go

7-Tage-News
News-Archiv
News mobil
Newsletter

English Pages

heise jobs

Telefontarife
Internettarife
Provider (Firmen)
Internet-Störungen
Free- & Shareware
Veranstaltungen

Leserforum
Chat-Events

Lexikon Fachwörterbuch Computer

Aktionen
Krypto-Kampagne
Schulen ans Netz
Netz gegen Kinderporno

Abo & Heft
Kontakt, Impressum
Mediainfo

Kiosk Das Archiv via Micropayment

ONLINE-MARKT

waren. Das Verfahren funktioniert zuverlässig bis zu einer minimalen Länge von nur 20 Zeichen. Die Methode lässt sich auch dazu verwenden, um mit einer Wahrscheinlichkeit von 93 Prozent auch den Autor eines Textes zu identifizieren.

In einem dritten Experiment versuchten die Forscher das Verfahren zu nutzen, um Sprachen zu klassifizieren. Dazu verwendeten sie die "Allgemeine Erklärung der Menschenrechte", da deren Formulierung in sehr vielen Sprachen verfügbar ist. Aus ihrer Analyse konstruierten die Forscher einen Stammbaum für 50 Sprachen, der erstaunlich genau dem von Linguisten entwickelten System ähnelt. Die Wissenschaftler sind optimistisch, dass sich das Verfahren nicht nur auf Texte anwenden lässt, sonder auch beispielsweise auch auf DNA- und Protein-Sequenzen. (wst/c't)

Kommentare:
Re: Araber (Richard , 30.1.2002 14:47)
Re: Warum nicht mit bzip2? (der Debil-Fundamentalist , 29.1.2002 18:44)
Re: Warum nicht mit bzip2? (Onkel Wanja , 29.1.2002 15:14)
mehr...