c't 4/2002, S. 48: Forschung

Dr. Wolfgang Stieler

Zweckentfremdet

Software-Tools als wissenschaftliche Instrumente

Während die Kompression einer Datei mit einem Zip-Algorithmus sich dazu eignet, Autor und Sprache eines beliebigen Textes zu analysieren, kann man mit dem Netzwerk-Kommando Ping die Lichtgeschwindigkeit bestimmen.

Zwei Physiker an der Universität von Youngstown haben eine elegante Methode beschrieben, um mit Hilfe des Netzwerk-Tools Ping die Ausbreitungsgeschwindigkeit elektromagnetischer Wellen zu messen [1]. Der Versuch (Preprint hier) soll Studenten das Konzept der so genannten ‘stochastischen Resonanz’ vermitteln.

Das Grundprinzip ist simpel: Mit Ping kann man die Laufzeit eines Datenpaketes zwischen zwei vernetzten Rechnern ermitteln. Im Modellfall benutzt man Ethernet-Kabel mit bekannter Länge als Verbindung. Variiert man die Länge des Kabels, kann man so die Geschwindigkeit bestimmen, mit der sich das Signal im Kabel ausbreitet - aus dieser Geschwindigkeit lässt sich relativ leicht die Lichtgeschwindigkeit im Vakuum ableiten. Für gängige Kabellängen ergibt sich allerdings ein Problem: Einige zehn Meter zusätzliches Netzwerkkabel bewirken eine Verlängerung der Signallaufzeit von einigen Nanosekunden; die Messgenauigkeit des Ping-Kommandos liegt dagegen nur im Bereich von Mikrosekunden.

Damit das Experiment trotzdem gelingt, nutzen die Forscher den Effekt der so genannten stochastischen Resonanz aus. Das Phänomen ist seit etwa zehn Jahren bekannt und lässt sich anschaulich am Beispiel einer Murmel in einem Eierkarton erklären: Wenn der Karton behutsam geschwenkt wird, hat die Murmel in der Regel nicht genügend Schwung, um von einer Mulde in eine andere zu rollen. Wenn die Bewegung nun von einem zufälligen Rütteln überlagert wird, dann kann ein zufälliger Impuls im richtigen Moment ausreichen, um die Kugel in eine andere Mulde überwechseln zu lassen. Auf diese Weise können Signale, die eigentlich unterhalb eines gewissen Schwellwertes liegen, durch Rauschen verstärkt werden - wird jedoch das Rauschen zu stark, geht das Signal darin unter [1, 2].

Die Forscher werteten - unter Linux, weil ‘die Autoren mit Windows nicht genügend vertraut waren' - für verschiedene Kabellängen die Laufzeit von je 30 000 Datenpaketen aus. Unabhängig von der verwendeten Verkabelung - Twisted Pair oder Koaxial-Kabel - lässt sich die Lichtgeschwindigkeit mit diesem Experiment auf etwa vier Prozent genau bestimmen.

Analytisches Schrumpfen

Italienische Wissenschaftler haben unterdessen eine Methode demonstriert, um per Software die Sprache eines Textes und mit hoher Wahrscheinlichkeit auch den betreffenden Autor zu identifizieren. Wie Dario Benedetto, Emanuele Caglioti und Vittorio Loreto von der Università degli Studi di Roma 'La Sapienza' in der Fachzeitschrift Physical Review Letters beschreiben [3, 4], 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 - im informationstheoretischen Sinne - des Zeichenstroms. Diese Entropie ist aber proportional zum Informationsgehalt - fügt man also einem Text ein weiteres Stück Text in einer anderen Sprache hinzu, steigt die Entropie an. Der LZ77-Algorithmus benutzt die Eingabe-Zeichenfolge als eine Art Wörterbuch. Wenn eine bereits bekannte Zeichenfolge im Verlauf des Textes wiederholt auftritt, speichert der Algorithmus nicht mehr die Zeichenfolge, sondern nur noch Verschiebungsvektoren, die auf die bekannte Zeichenfolge zeigen.

Die Forscher verglichen nun 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 - dieses Verfahren wird für alle Vergleichstexte durchgeführt. 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 waren. Das Verfahren funktioniert zuverlässig bis zu einer minimalen Länge von nur 20 Zeichen. Die Methode lässt sich - analog zum Sprachentest - auch dazu verwenden, um mit einer Wahrscheinlichkeit von 93 Prozent 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, sondern beispielsweise auch auf DNA- und Protein-Sequenzen.

(wst)

Literatur

[1] P. Jung, P. Hänggi, Physikalische Blätter 11/99, S. 1005

[2] www.umbrars.com/sr/

[3] D. Benedetto, E. Caglioti, V. Loreto, Language Trees u. Zipping, Physical Review Letters No. 88 S. 048702, 28. Januar 2002

[4] Preprint unter xxx.lanl.gov, con-mat/0108530 v2

Kommentare:
URL Korrektur (Andreas Schuster , 10.2.2002 16:16)