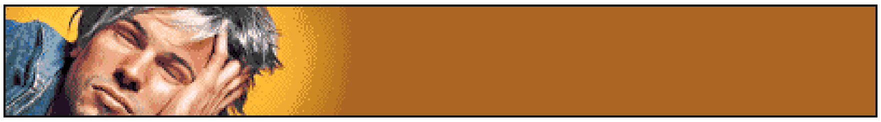

Freitag, 8. Februar 2002

Frontseite
AKTUELL
Kommentar
International
Wirtschaft
Börsen · Märkte
Schweiz
Zürich · Region
Sport
Feuilleton
Vermischtes
Wetter

HINTERGRUND
Dossiers
Netzstoff
Forschung · Technik
Tourismus
Medien · Informatik
Literatur · Kunst
Zeitfragen
Spezialthemen
Buchrezensionen
English Window

NZZ · FINFOX
Ihr Finanzplaner

SERVICE
Veranstaltungen
Restaurantführer
Kreuzworträtsel
Webcam Zürich

NZZ-ARCHIV
NZZ ab 1993
Monatsarchiv
Recherchedienst

NZZ-VERLAG
Kontakte

Kasten: Probe aufs Exempel

Buchstabenstauchung

Datenkompression im Dienst der Sprachwissenschaft

Um Dateien platzsparend zu speichern, verwenden Computerbenützer Kompressionsprogramme, mit denen Dateien auf einen Bruchteil ihrer ursprünglichen Grösse geschrumpft werden können. Aber die populären Kompressionsprogramme - Applikationen wie zum Beispiel Winzip, Pkunzip oder Gnuzip - nützen nicht nur Besitzern von zu kleinen Festplatten; sie können auch der Sprachwissenschaft dienen.

gsz. In der renommierten Fachzeitschrift «Physical Review Letters» erschien Ende Januar eineArbeit, in der zwei Mathematiker und ein Physiker aus Italien eine neuartige Anwendung fürKompressionsprogramme beschrieben: Sie benutzten solche Zipping-Software, um die Autorschaft anonymer Texte zu eruieren und um Verwandtschaftsverhältnisse moderner Sprachfamilien zu ergründen.

Ein Mass für den Informationsgehalt

Grundlegend für das Verständnis der Komprimierung ist der Begriff der Entropie. In der Physik ist Entropie ein Mass für die Unordnung einesSystems (z. B. eines Gases), in der Telekommunikation stellt sie ein Mass für den Informationsgehalt einer Nachricht dar. Die tausendfache Wiederholung des Zeichens Null besitzt zum Beispiel einen sehr kleinen

Suchen

Weitere Artikel
Kultur und Kult der Kleinheit

Der Desktop-PC ist geschrumpft

Unabwendbares Zerwürfnis?

Der steile und steinige Weg zurück in die Zukunft

Wo das Kluge entsteht

Warum nicht einen Teppich finanzieren?

Wenig Innovation

Die dritte Heimat

Ein Rabbiner, der keiner war

Sämtliche Artikel

Herausgegriffen
Netzstoff
Eine Auswahl von Hintergrund-Artikeln zum Thema Internet

Spezialthemen:
Telekommunikation

besitzt zum Beispiel einen sehr kleinen Informationsgehalt und somit eine sehr tiefe Entropie; eine Datei mit dieser Zeichenfolge kann auf die prägnante Formel «1000-mal die Null» komprimiert werden. Andererseits enthält eine zufällige Folge von Nullenund Einsen eine sehr hohe Entropie. Im ungünstigsten Fall kann sie überhaupt nicht komprimiert werden, die einzige Art, eine solche Zufallsfolge zu speichern, ist die Wiederholung jedes Zeichens.

Die relative Entropie gibt an, wie viel Speicherplatz verschwendet wird, wenn eine Zeichenfolge mit einer Methode komprimiert wird, die für eine andere Folge optimiert wurde. Als Beispiel kann das Morsealphabet dienen, das für die englische Sprache optimiert ist. Dem häufigsten Buchstaben in der englischen Sprache, dem «e», wurde die kürzeste Zeichenfolge zugeordnet: ein Punkt. Für selten vorkommende Buchstaben werden längere Zeichenfolgen verwendet, wie zum Beispiel « · » für «q». Aber für andere Sprachen ist das Morsealphabet nicht optimal, denn die Länge der Codes entspricht nicht mehr der Häufigkeit der Buchstaben. Die relative Entropie gibt dann an, wie viele zusätzliche Striche oder Punkte benötigt werden, um zum Beispiel einen italienischen Text mit dem für die englische Sprache optimierten Morsealphabet zu übermitteln.

Fast alle gebräuchlichen Kompressionsprogramme beruhen auf einem Verfahren, das zwei israelische Forscher vor einem Vierteljahrhundert entwickelt haben. Der Computerwissenschafter Abraham Lempel und der Elektroingenieur Jacob Ziv, Professoren am Technion in Haifa, entwarfen Algorithmen, die Dateien sehr effizient zusammenstauchen. Die Methode beruht auf der Tatsache, dass sich bestimmte Zeichenfolgen in einerDatei sehr oft wiederholen. Beim ersten Erscheinen werden sie in einer Art Wörterbuch abgelegt. Bei nochmaligem Auftauchen wird dann einfach auf das erste Vorkommen verwiesen. Der Verweis benötigt weniger Platz als die Wiederholung der ursprünglichen Zeichenfolge. Die Liste der häufigen Zeichenfolgen folgt aber nicht den Vorgaben eines Standardwörterbuchs. Vielmehr lernt der Algorithmus jedes Mal, welche Zeichenfolgen in der vorliegenden Datei häufig vorkommen, und schneidert die Komprimierung auf die spezifische Datei zu. Mit zunehmender Länge sinkt der benötigte Speicherplatz einer mit dem Lempel-Ziv-Algorithmus komprimierten Datei gegen die Entropie des Textes ab.

Stilbrüche

Dario Benedetto, Emanuele Caglioti und Vittorio Loreto von der Università degli Studi di Roma La Sapienza fütterten Kompressionsprogramme mit 90 Texten von 11 verschiedenen italienischen Autoren (Dante Alighieri, Pirandello usw.).* Zuerst komprimierten sie einen Text, dessen Verfasser bekannt war. Dann hängten sie einen kurzenanonymen Text hinten an. Es ging darum, festzustellen, ob die relative Entropie des kombiniertenTextes hoch oder niedrig war. Stammt der angehängte Text von einem anderen Autor, müsste die relative Entropie hoch sein. Der Algorithmus müsste nun die Sprachgewohnheiten des neuen Autors dazulernen und würde mehr Platz benötigen, um die Datei zu speichern. Der angehängte Text durfte aber nicht zu lang sein, denn der Algorithmus lernt sehr schnell die Eigenarten des neuen Autors. Je kürzer also die komprimierte Datei mit den beiden Texten ist, desto wahrscheinlicher ist es, dass der angehängte Text vom gleichen Autor stammt. Das Experiment zeitigte interessante Ergebnisse: In fast 95 Prozent der Fälle konnten die Texte mit Hilfe des Kompressionsprogramms den richtigen Autoren zugeordnet werden.

In ihrem Eifer hatten die Autoren allerdings nicht bemerkt - oder im Literaturverzeichnis zu erwähnen vergessen -, dass sie nicht die ersten Forscher waren, die literarische Texte mittels mathematischer Methoden auf ihre Autorschaft überprüften. So hatte sich der Linguistikprofessor George Zipf schon 1932 an der Harvard-Universität mit Studien zur Wortfrequenz befasst (NZZ vom 21. 11. 01). Dem Schotten George Yule gelang es in dem 1944 erschienenen Werk «The statistical study of literary vocabulary», die Schrift «De imitatione Christi» aus dem 15. Jahrhundert dem Mystiker Thomas a Kempis zuzuordnen. Und schliesslich müssen in diesem Zusammenhang auch die «Federalist Papers» aus dem 18. Jahrhundert erwähnt werden, über deren Autorschaft die amerikanischen Statistiker R. Frederick Mosteller und David L. Wallace 1964 Aufschluss geben konnten.

Evolution der Sprachen

Mit einem zweiten Experiment versuchten Benedetto, Caglioti und Loreto den Verwandtschaftsgrad verschiedener Sprachen herauszufinden. Zwei Sprachen, die aus der gleichen Familie stammen, haben nämlich eine geringere relative Entropie und sollten daher effizienter komprimiert werden können als ein Sprachenpaar, dasnicht miteinander verwandt ist. Insgesamt untersuchten die Forscher 52 europäische Sprachen.Auch hier waren ihre Bemühungen von Erfolg gekrönt. Mit Hilfe des Zip-Programmes konnten sie die Zugehörigkeit der Sprachen zu den jeweils richtigen linguistischen Gruppen feststellen. Zum Beispiel besitzen Rätoromanisch und Italienisch wenig relative Entropie und sind somit verwandt. Dagegen haben Schwedisch und Kroatisch eine hohe relative Entropie und müssen somit verschiedenen Familien angehören. Winzip konnte sogar Maltesisch, Baskisch und Ungarisch als Sprachen identifizieren, die in keine der bekannten Familien hineinpassen. Der Erfolg ihrer Methode liess bei den Forschern die Hoffnung aufkommen, dass die Entropiemessung durch Zipping-Software auch auf andere Datenfolgen wie DNA-Sequenzen oder Aktienkurse anwendbar sein könnte.

* D. Benedetto, E. Caglioti, V. Loreto: Language trees and zipping. Physical Review Letters, Vol. 88, 28. Januar 2002 (link.aps.org/abstract/PRL/v88/e048702).

Probe aufs Exempel

gsz. Die nebenstehend geschilderte Studie inspiriert zu eigenen Tests. Als Grundlage dienten dazu Artikel, die der Schreibende im Januar in der NZZ veröffentlicht hatte, insgesamt 18 Texte mit 14 000 Wörtern und

105 000 Zeichen - die meisten über Begebenheiten in Israel. Nach Entfernung aller Titel, Zwischentitel und Datumszeilen wurden die Texte als Ascii-Dateien gespeichert und in das Kompressionsprogramm Winzip gefüttert. Als der Experimentator das Resultat sah, erlitt er einen Schreck: Die Komprimierung hatte das, was er während eines Monats mit viel Mühe und Schweiss verfasst hatte, um nicht weniger als 66 Prozent zusammengestaucht. Die unausweichliche Schlussfolgerung war, dass bloss ein Drittel des Geschriebenen Gehalt hatte, zwei Drittel stellten pure Entropie dar, waren also redundant.

Der etwas niedergeschlagene Journalist versuchte sich mit dem Argument zu trösten, dass es die versierte Aneinanderreihung der Wörter sei, die Gehalt biete, und nicht der Text an sich. Zum Beweis für diese These wurden die 14 000 Wörter alphabetisch geordnet und neu komprimiert. Siehe da: Die sinnlose Abfolge von Wörtern konnte um über 80 Prozent zusammengeschrumpft werden und bot somit weniger als 20 Prozent Information. (Das ist nicht verwunderlich, denn «Israel» und «israelisch» beispielsweise kamen 231-mal vor, und das Wort «palästinensisch» in allen seinen Abwandlungen insgesamt 195-mal.) Dies bedeutet, dass die richtige Aneinanderreihung, wie sie eben nur ein sachkundiger Journalist vornehmen kann, etwa 15 Prozent mehr Gehalt bietet als ein Wörterbuch. Kein grosser Trost, aber immerhin.

Leider erhielt das eben zurückgewonnene Selbstvertrauen gleich einen neuen Dämpfer. Eine völlig randomisierte Sammlung der 14 000 Wörter konnte nämlich nur um 60 Prozent komprimiert werden, was doch zu beweisen scheint, dass eine zufällige Sammlung der von diesem Korrespondenten geschriebenen Wörter noch mehr Gehalt aufweist als die verfassten Artikel. (Eine zufällige Zeichenfolge hat ja per Definition immer die grösste Entropie.)

Für das eigentliche Experiment braucht es vier Texte: jeweils einen langen (1000 Wörter) und einen kurzen (50 Wörter) von gsz. und von S. B., dem Redaktor dieser Beilage. Die beiden kurzen Texte wurden je einmal an die beiden langen Artikel angehängt, und die vier Dateien wurden komprimiert. Die Ergebnisse entsprachen den Resultaten der italienischen Forscher: Wennder kurze gsz.-Artikel, der ursprünglich 462 Zeichen aufwies, an den langen S.-B.-Artikel angehängt wurde, benötigte Winzip 209 zusätzliche Zeichen. Wenn er aber an den langen gsz.-Artikel angehängt wurde, benötigt die Komprimierung bloss 159 zusätzliche Zeichen. Damit war bewiesen, dass der kurze Text nicht von S. B., sondern von gsz. verfasst worden war. Doch auch hier muss der Israel-Korrespondent frustriert feststellen, dass er in seiner Berichterstattung nicht immer nur Neues bietet. Die grosse Platzersparnis, die die Komprimierung seiner beiden Texte aufweist, deutet darauf hin, dass der kurze Artikel zu einem grossen Teil mit dem langen Artikel redundant ist.

8. Februar 2002

Kontakt
Impressum