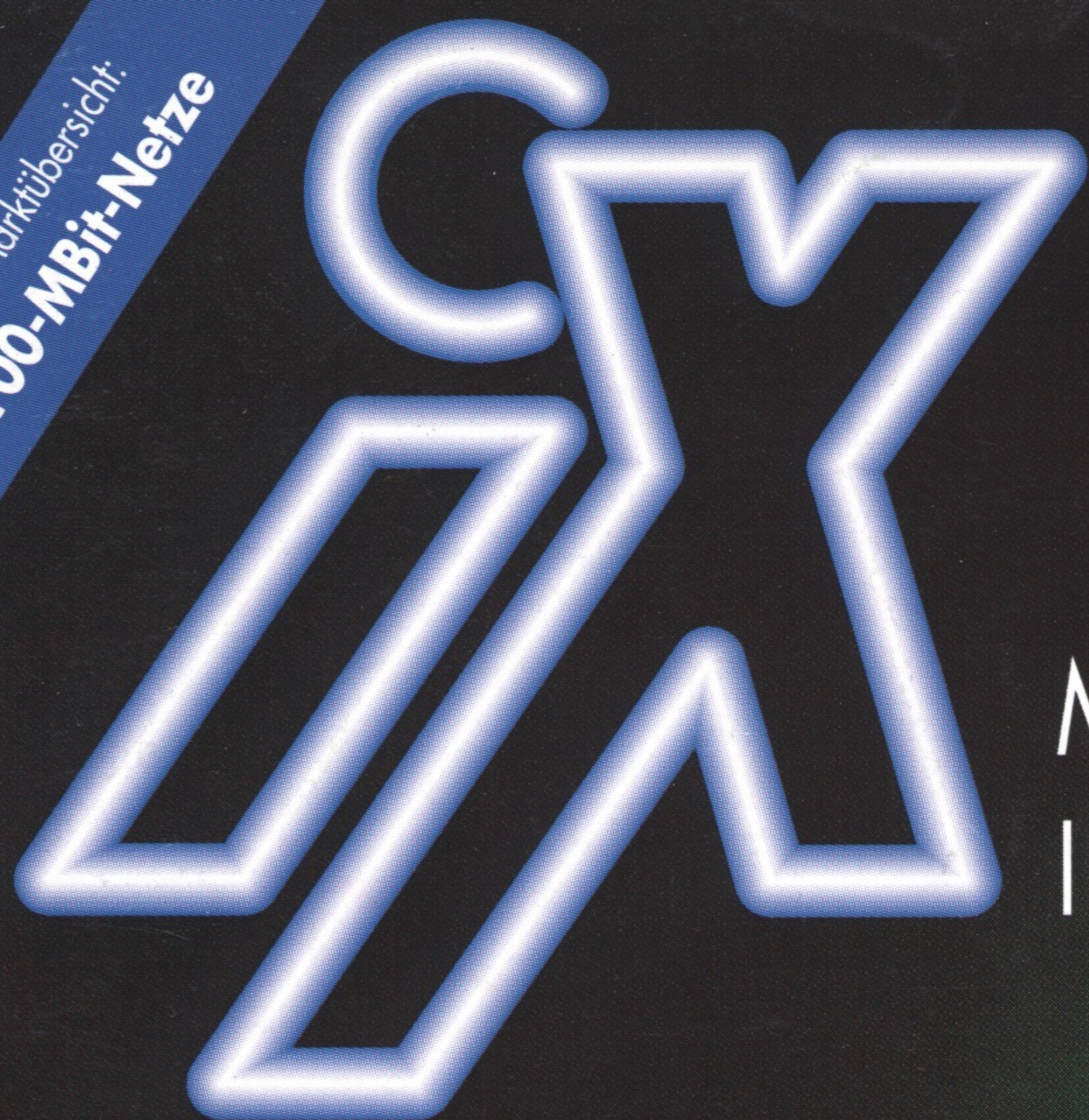


ips, Marktübersicht:
100-MBit-Netze

öS 62,- sfr 7,50 **DM 7,50** H 10554

Juni 1997

6



MAGAZIN FÜR PROFESSIONELLE
INFORMATIONSTECHNIK

Verglichen:

500-MHz-Alphas für NT und Unix

World Wide Web:

**HTML-Alternative
XML**

JDBC/ODBC:

Java-DBMS-Kopplung

C++-Programmierung:

**SGI Developer Magic
Qt für X11 und
Windows**

Software-Installation:

**Pkgadd, RPM,
NT-Setup**

Publishing:

**PostScript 3
Ghostscript 5**

Flach, strahlungsarm, videotauglich:

42"-Farbplasmadisplay



Script durchsucht News-Gruppen
nach Schlagworten

Was gibt's Neues?



Michael Schilli

Einige interessante Newsgroups, darunter auch *comp.lang.perl.misc*, kränken an zuviel Verkehr. Kein berufstätiger Mensch kann bei über 100 neuen Meldungen täglich auf dem laufenden bleiben. Ein Perl-Script schaut automatisch nach, ob es etwas Neues gibt.

Stellt man in einer Newsgroup eine Frage und harrt auf Antworten, gibt es wie fast immer zwei Möglichkeiten: Entweder kontrolliert man täglich die Neuerscheinungen mit einem Newsreader oder nutzt das Script *chknews.pl*, das die Meldungen nach Themen durchforstet und Neuerscheinungen meldet. Das Kommando

```
chknews.pl comp.lang.perl.misc 'my question'
```

gibt alle Artikel der Newsgroup *comp.lang.perl.misc* aus, in deren Subject-Header die Zeichenkette 'my question' auftaucht. *chknews.pl* verträgt sogar (Perl-taugliche) reguläre Ausdrücke. So findet

```
chknews.pl de.comp.lang.perl 'L\S*\s+Wall'
```

in der deutschsprachigen Perl-Newsgroup Artikel, die den Namen des Perl-Autors beispielsweise in den Formen

'L Wall' 'L. Wall' oder 'Larry Wall' im Titel führen.

Objekte für Netzaufgaben aus *libnet*

Für den täglichen Aufruf des Scripts sorgt entweder ein Eintrag in der *crontab* oder aber das entsprechende Kommando wird in das *.profile* der täglich benutzten Shell eingefügt. Im letzteren Fall kommt die Option *-once* zum Einsatz, die das Script dazu bewegt, nur einmal täglich den Newsserver zu befragen und bei allen folgenden Aufrufen ressourcenfreundlich zu terminieren.

Die Variablen *\$newsrsc* und *\$newsrserver* erhalten in den Zeilen 7 und 8 Werte zugewiesen, die den lokalen Gegebenheiten entsprechen müssen. *\$newsrsc* enthält den Namen der Datei, in der sich das Script bereits gemeldete

Veränderungen 'merkt', um den Anwender nicht mit den immer gleichen Ausgaben zu ermüden. Im Listing ist dies *\$HOME/data/news.dat*.

\$newsrserver enthält den Namen des nächsten erreichbaren Newsserver, den meist der Internet-Provider zur Verfügung stellt. Das Script liest ihn – im Beispiel – aus der Environment-Variablen *NNTPSERVER*, die den entsprechenden Wert enthalten muß.

In Zeile 12 kommt der NNTP-User-Agent ins Spiel, ein Teil der von Graham Barr implementierten Bibliothek *libnet*, die Schnittstellen zum Internet anbietet. Sie ist nicht zu verwechseln mit der in der letzten iX vorgestellten *libwww*, die sich vor allem dem WWW widmet [1]. Beide liegen kostenlos auf dem CPAN zur Abholung bereit (siehe 'Was tun?').

Hashes für assoziative Zugriffe

Bei gegebener Kommandozeilenoption *-once* setzen die Zeilen 14 bis 16 die Variable *\$once_a_day*, die im weiteren Verlauf des Scripts eventuell einen vorzeitigen Abbruch auslöst. Sein 'Gedächtnis', das Newsgroup-Namen mit den Nummern bereits aufgetauchter Artikel enthält, realisiert *chknews.pl* in der Form von Keys, die zur Laufzeit im Hash *%subjects_old* liegen. In den Zeilen 18 bis 29 gelangen sie am Scriptbeginn 'zu Fuß' von der Persistenzdatei *\$pfile* nach *%subjects_old*. Daten neu gefundener Artikel hält *chknews.pl* im Hash *%subjects_new*. Sie wandern zum Programmschluß zurück auf die Festplatte. Mit einer Kombination von Newsgroup und Artikelnummer als Key bietet der Hash als Datenstruktur die Möglichkeit, schnell festzustellen, ob eine Meldung schon vorlag, da Hashes

Nachtrag

Das in der letzten iX vorgestellte Script *urlchk.pl* zeigte sich unbeeindruckt, falls sich die Ausgabe eines CGI-Scripts bei konstanter Länge nur dem Inhalt nach veränderte. Ersetzt man die Zeile

```
$len = length($response->content());
```

durch diese

```
$len = unpack('%16C*', $response->content());
```

und vergleicht so statt der Länge die 16-Bit-Checksumme des Dokuments, fallen auch rein inhaltliche Veränderungen auf.

Was tun?

Alle erwähnten Module sind auf im Comprehensive Perl Archive Network (CPAN) erhältlich, in Deutschland beispielsweise hier:

<ftp://ftp.leo.org/pub/comp/programming/languages/perl/CPAN/>

<ftp://ftp.rz.ruhr-uni-bochum.de/pub/CPAN/>

<ftp://ftp.uni-hamburg.de/pub/soft/lang/perl/CPAN/>

<ftp://ftp.uni-erlangen.de/pub/source/Perl/CPAN>

<ftp://ftp.gmd.de/packages/CPAN>

chknews.pl benötigt diese Module (jeweils relativ zum CPAN): *modules/by-module/Net/libnet-1.05.tar.gz* *modules/by-module/IO/IO-1.15.tar.gz*

Unter Windows NT lassen sich die erwähnten Module nicht ohne weiteres installieren, da dort *MakeMaker* und etliche andere, zur Standarddistribution gehörenden Pakete nicht mitgeliefert werden. Der NNTP-Zugriff läßt sich jedoch trotzdem realisieren, indem man *libnet* beispielsweise mit *Winzip* auspackt und das Verzeichnis *Net* nach *perNlib* kopiert. Dann fehlt in *perNlib\Net* nur noch *Config.pm*, das man sich zum Beispiel von einer Linux-Maschine besorgen kann. Es gibt natürlich keine Garantie, daß das eben beschriebene Verfahren mit allen Teilen der *libnet* korrekt funktioniert.

assoziative Zugriffe erlauben. Zusätzlich liegt der aktuelle julianische Tag des Jahres unter dem Schlüssel *_last_check* in der Persistenzdatei, damit das Script erfahren kann, wann es zum letzten Mal lief.

In Zeile 33 wird endlich der Newsagent aktiv. Der Konstruktor des NNTP-Objekts nimmt zunächst Kontakt zum Newsserver auf. Vier Zeilen später startet eine Schleife, die pro Durchgang eine Newsgroup und einen regulären Ausdruck aus den Kommandozeilenparametern extrahiert und bearbeitet. Die *group*-Methode des *newsagent*-Objekts in Zeile 45 liefert mit dem Newsgroup-Namen als Argument die Anzahl der dort verfü-

baren Artikel sowie die Nummer des ersten und des letzten Artikels (als Liste) zurück. Dieses Verfahren funktioniert mit dem vor allem unter Linux verbreiteten NNTP-Server *inn* dann nicht, wenn er auf derselben Maschine wie *chknews.pl* läuft, da er den *GROUP*-Request in diesem Fall nicht versteht. Folgende Zeilen schaffen hier Abhilfe

```
$newsagent->command('mode', 'reader');
$newsagent->response();
```

Sie sind direkt hinter dem Konstruktoraufbau *Net::NNTP->new* in das Script einzufügen.

Da nur die Titel und nicht der gesamte Text der Artikel interessieren,

```
1 #!/usr/bin/perl
2 #####
3 # chknews.pl - Newsgroups nach neuen Titeln durchsuchen
4 #####
5 # Parameter
6
7 my $newsrc = "$ENV{HOME}/data/news.dat";          # "Gedächtnis"
8 my $newsserver = "$ENV{NNTPSERVER}";              # Newsserver
9
10 #####
11
12 use Net::NNTP;                                     # NNTP User Agent
13
14 $once_a_day = 0;                                   # Parameter für "Nur einmal
15                                           # täglich" gesetzt?
16 @ARGV = grep { !(/^~once$/ && ($once_a_day = 1)) } @ARGV;
17
18 if(open(FILE, "<$newsrc")) {                       # Gespeicherte Daten einlesen
19     while(<FILE>) {
20         chop;
21         # Abbrechen, falls -once
22         if($once_a_day && (($last_check) = /^_last_check: (\d+)/)) {
23             exit 1 if $last_check eq (localtime(time))[7];
24         }
25         $subjects_old{$_} = 1;                     # Subject setzen
26     }
27     close(FILE);
28 }
29
30 $| = 1;                                             # Ausgabe nicht puffern
31
32 $newsagent = Net::NNTP->new($newsserver);          # News Agent erzeugen
33
34 # Für alle Newsgroup/Subject-
35 # Paare auf der Kommandozeile
36
```

te/e po/is

das magazin
der netzkultur

film

Specials:

Film & Computer — die Zukunft des Kinos!

Special Effects, virtuelle Schauspieler, interaktiver „Film“ im WWW u.a...

Kunst im Netz — Kunst zwischen Technologie und sozialer Relevanz; Künstler als Programmierer, Programmieren als Kunst; Nachrichten von den Rändern der Netzwelt.

Aktuell:

Staatsanwaltschaft versus Provider: Faustrecht im Netz? Wem gehört das Web? Diskussionen um neue Top Level Domains.

Rubriken:

PopTarts präsentieren das „weibliche Datenset“. Tom Bass berichtet von der inoffiziellen Medienlandschaft in Ex-Jugoslawien. Mark Amerika über Hypertext-Erzählungen sowie neue Texte von Stanislaw Lem, Joel de Rosnay, Ivo Skoric u.a.

Der Zugang zu Telepolis unter
<http://www.heise.de> ist kostenlos -
klicken Sie sich ein!

www.heise.de/tp

Telepolis-Redaktion: Kühbachstr.11 · 81543 München
Tel.: 089/625004-70 · Fax.: 089/625004-66

```

37 while(($newsgroup, $regex) = splice(@ARGV, 0, 2)) {
38
39     usage("Argument Error") unless defined $regex;
40
41     print "Scan [$newsgroup] for [$regex] ... ";
42
43     # Verfügbare Artikelnummern
44     # Newsgroup holen
45     my ($total, $first, $last) = $newsagent->group($newsgroup);
46     my $count = 0;
47     # Treffer für aktuelle Group
48
49     unless(defined $total) {
50         # Fehler?
51         usage("Unknown newsgroup $newsgroup");
52     }
53
54     # Subjects holen
55     $subjects = $newsagent->xhdr('Subject', [$first, $last]);
56
57     # Fehler?
58     die "xhdr failed" unless defined $subjects;
59
60     # über alle gefundenen
61     # Subjects iterieren
62     foreach $mesgno (keys %{$subjects}) {
63         # Match des gefundenen Subjects
64         # auf spezifizierten Regex?
65         if($subjects->{$mesgno} =~ /$regex/i) {
66             $subjects_new{"$newsgroup $mesgno"} = 1; # Treffer speichern
67
68             # Ausgeben, falls noch nicht
69             # bekannt
70             unless($subjects_old{"$newsgroup $mesgno"}) {
71                 print "\n" unless $count++;
72                 print "    $subjects->{$mesgno}\n";
73             }
74         }
75     }
76     print "No News\n" unless $count;
77 }
78 $newsagent->quit();
79 # Newsagent freigeben
80 # Persistente Daten
81 # wieder ablegen
82 open(FILE, ">$newsrsrc") || die "Cannot open file $newsrsrc";
83
84 foreach $i (keys %subjects_new) {
85     print FILE "$i\n";
86 }
87
88 # Heutiges Datum ablegen,
89 # falls -once spezifiziert
90 print FILE "_last_check: ", (localtime(time))[7], "\n" if $once_a_day;
91 close(FILE);
92
93 #####
94 sub usage {
95     # Usage ausgeben
96     #####
97     (my $program = $0) =~ s,.*,,g;
98     print <<EOT;
99     $program: @_
100
101     usage:
102     $program [-once] newsgroup subject [newsgroup2 subject2] ...
103     -once: check only once a day
104 EOT
105
106     exit 1;
107 }

```

Chknews.pl überprüft Newsgroups auf neue Artikel zu bestimmten Themen.

holt die *xhdr*-Methode in Zeile 52 anschließend die *Subject*-Header aller verfügbaren Artikel vom Newsserver. Sie liefert die Referenz auf einen Hash zurück, der als Keys die Artikelnummern und als Werte die -titel enthält. Als Parameter erwartet *xhdr* den Namen des Header-Feldes sowie eine Referenz auf eine Liste mit den Nummern des ersten und des letzten Artikels. *XHDR*-Requests sind zwar nicht Bestandteil von NNTP, die meisten Newsserver dürften sie jedoch korrekt

behandeln. Sollte es Schwierigkeiten geben, könnte eventuell die Methode *head()* aus dem gleichen Modul weiterhelfen.

Suche im Subject mit regulären Ausdrücken

Über alle Hash-Keys iteriert die *for*-Schleife in Zeile 59. Zeile 62 stellt fest, ob der Titel eines Artikels mit dem vorgegebenen regulären Ausdruck

zur Deckung zu bringen ist. Der Schalter *i* stellt dabei die Unterscheidung zwischen Groß- und Kleinschreibung ab. Zeile 64 erzeugt einen Eintrag für den Hash *%subjects_new*, der alle neu gefundenen Titel enthält. Er setzt sich – wie in *%subjects_old* – aus dem Namen der Newsgroup und der laufenden Nummer des Artikels zusammen.

Ab Zeile 68 prüft das Script dann, ob der Artikel schon bekannt ist, indem es im Hash *%subjects_old* nachsieht. Ist er neu, sorgen die beiden nächsten Zeilen für die entsprechende Meldung. Zeile 78 gibt die durch den Newsagent belegten Ressourcen wieder frei.

In den Zeilen 84 bis 91 gelangen die Hash-Einträge von *%subjects_new* in die Persistenzdatei *\$newsrsrc*, um dort bis zum nächsten Scriptaufruf zu überwindern. Zeile 89 schreibt zusätzlich den aktuellen julianischen Tag in die Datei, falls die *-once*-Option aktiv ist. Die *usage*-Funktion in den Zeilen 94 bis 107 gibt lediglich die korrekte Signatur des Scripts aus. Sie tritt in Aktion, wenn ein erforderlicher Parameter nicht vorhanden ist.

Damit die Persistenzdatei nicht endlos wächst, legt *chknews.pl* immer nur die neu gefundenen Titel in *\$newsrsrc* ab. Wer mehrere Newsgroups nach mehreren Mustern durchsuchen will, sollte das besser durch einen Aufruf von *chknews.pl* erledigen, damit das 'Gedächtnis' intakt bleibt. Beispielsweise durchstößt

```
chknews.pl -once de.alt.tv.misc 'ARD|ZDF' \
comp.lang.perl.modules '
```

die Newsgroup *de.alt.tv.misc* nach Titeln, die öffentlich-rechtliche Sender erwähnen und zusätzlich die Newsgroup *comp.lang.perl.modules* nach allen verfügbaren Titeln, die wenigstens ein Zeichen enthalten. (ck)

MICHAEL SCHILLI

ist Softwareentwickler bei Black Sun Interactive, San Francisco.

Literatur

- [1] Michael Schilli; Webseiten kontrollieren; Altbaken oder aktuell?; iX 5/97, S. 166 ff.
- [2] Randal L. Schwartz; Monatliche Kolumnen in 'Web Techniques'; <http://www.teleport.com/~merlyn/WebTechniques>
- [3] Michael Schilli; Effektives Programmieren mit Perl 5; Addison-Wesley 1996; ISBN 3-8273-1095-4