

weemo

Perl als Backend-System für eine Web-Suchmaschine

Deutscher Perl Workshop 2007
München, 21.02.2007

Stefan Fischerländer



Neomo - ist das nicht der Fisch?

- Neomo ist eine Web-Suchmaschine
 - ca. 200 Mio. deutschsprachige Dokumente indexiert
 - Geschäftsmodell: Erstellung und Betrieb von Enterprise-Search-Lösungen mit Fokus Spezialwebsuchen
 - Referenzen: DeTeMedien, GelbeSeiten Marketing, VDI Verlag
- Betreiber: Neomo GmbH & Co. KG
- Sitz: Köln
- Gesellschafter: 5 Telefonbuchverlage + Stefan Fischerländer

Neomo - Making Of ...

- Das Backend
 - Crawler, Indexer, Link-Analyse
 - besteht ausschließlich aus Perl
 - etwa 10.000 Zeilen Code - inklusive Kommentare und Dokumentation
- Das Frontend
 - zuständig für die Beantwortung von Suchanfragen
 - aus Performancegründen in C/C++
 - aktuell: Wir portieren immer mehr Funktionalität nach Perl und/oder Python

Gründe für Perl im Backend

- Rapid Application Development
- Reguläre Ausdrücke als Sprachelement
- Perl ist fehlertolerant - Daten auf Webseiten sind pures Chaos
- Informatik-Professor: „Mit Perl kann man keine Suchmaschine bauen!“ - Der entscheidende Ansporn.

Die Performance-Frage

- Backend-Prozesse sind nicht extrem performance-kritisch: „Perl wartet auf die Festplatte genau so schnell wie C“
- Suchmaschine relativ leicht parallelisierbar.
- Schnelleres Coden bedeutet: Mehr Zeit für die Optimierung der Algorithmen.
- Perl = C für ganz Faule:
 - Perl verbindet die einzelnen Komponenten und erledigt I/O.
 - Die eigentlichen Aufgaben übernehmen RegExps, BerkeleyDB, Lynx und sort - allesamt hoch optimierte und getestete C-Programme.
 - Vergesst OOP - das ist die wahre Wiederverwendung von Code!

Die Zuverlässigkeit

- Backend-Architektur:
 - Cluster aus etwa 30 x86-Servern
 - OS: Debian GNU/Linux, Perl: 5.8.4
- bearbeitete Datenmenge:
 - etwa 5 bis 10 Terabyte pro Monat
- Ausfälle:
 - Festplatten und Controller
 - in zwei Jahren ein einziges Problem mit Perl - und das entpuppte sich als peinlicher Programmierfehler

Die dunklen Seiten

- Auf Uni-verwöhnte Einsteiger wirkt Perl abschreckend:
 - Code sieht „schlimm“ aus
 - OOP ist gewöhnungsbedürftig
 - implizites Verhalten verwirrt
 - CPAN - mächtig und anarchisch („Welcher der 37 XML-Parser ist denn nun der richtige für meinen Zweck?“)
 - Diskussionen um Perl 6
- Das typische Perl-Programm entsteht durch evolutionäre Vorgänge aus einem Einzeiler - das (ver-)führt zu schlechtem Programmdesign und schwer lesbarem Code.
- Überlegung: Umstieg auf Python

Und ein letzter Tipp ...

Life's too short for static
typing.

Danke für Ihre Aufmerksamkeit!

heomo