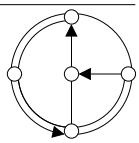
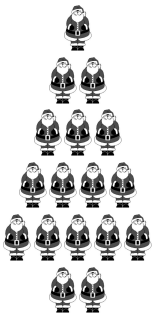
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



WS 2003/04 Prof. R. Wattenhofer / Keno Albrecht

Vernetzte Systeme
Xmas-Übung



Ausgabe: **17. Dezember 2003**
Abgabe: **26. Januar 2004**

1 Web-Crawler

Damit Ihnen über Weihnachten nicht langweilig wird, haben wir uns eine kleine Aufgabe für Sie einfallen lassen. Die Bearbeitung ist freiwillig und nicht prüfungsrelevant, bietet Ihnen (auch gerne in Gruppen) aber die Möglichkeit, Ihre Programmierfertigkeiten zu vertiefen, Interessantes über das WWW herauszufinden – und hoffentlich auch jede Menge Spass. Zusätzlich werden wir als Belohnung die „besten“ Abgaben mit einem kleinen Überraschungspreis prämieren.

Die Idee dieser Aufgabe ist es, einen Web-Crawler zu schreiben, der interessante oder lustige Informationen über das World-Wide-Web liefert. Ein Web-Crawler ist ein Programm, das ausgehend von einer Webseite[1] den dort aufgeführten Links folgt. Dabei kann Ihr Web-Crawler eine Breiten- oder Tiefensuche oder auch irgendeine andere Art der „Fortbewegung“ durchführen.

Wir möchten Ihnen gar nicht allzu viel vorgeben, sondern eher Ihre Kreativität wecken. Sicherlich kommen Sie auf deutlich interessantere Ideen als die folgenden:

- Wie vielen Links müssen Sie ausgehend von einer zufälligen Startseite durchschnittlich folgen, um a) auf der Webseite der ETH zu landen? b) auf einer Webseite zu landen, die das Wort „Sex“ (Computer, TV, ...) beinhaltet? c) wieder auf Ihrer Startseite zu landen?
- Wie gross ist die durchschnittliche/kleinste/grösste Webseite (mit/ohne Bildern etc.)? Wie gross ist das durchschnittliche/kleinste/grösste Bild? Welches ist das häufigste Dateiformat ausser .html (.pdf, .doc, .gif, .png, ...)?
- Wie viele (interne/externe) Links befinden sich durchschnittlich auf einer Webseite?

Es sind beliebige andere internet-orientierte „Statistiken“ denkbar. Wenn Sie interessante Details auch ohne einen Web-Crawler herausfinden können, dann tun Sie das – Sie haben freie Hand!

Zur Vereinfachung Ihrer Aufgabe ist es Ihnen von unserer Seite aus auch erlaubt, vorhandenen Code (aus dem Internet) zu verwenden. Suchen Sie zum Beispiel nach „java web crawler“ bei Google – „WebSPHINX“ sah ganz interessant aus.

Reichen Sie bitte bis zum Abgabetermin Ihren vollständigen Quellcode ein, und präsentieren Sie Ihre Ergebnisse in angemessener Weise.

[1] Eine geeignete Startseite können Sie zum Beispiel von Hand eingeben oder mittels einer Suche nach einem zufälligen Suchwort bei Google finden.