4.3 Offline im WWW browsen mit `wwwoffle`

Das Internet ist eine schier unerschöpfliche Informationsquelle, die für jeden etwas Interessantes zu bieten hat. Die meisten dieser Informationen sind über das World Wide Web (WWW) mit Hilfe eines Browsers, z. B. dem Netscape Navigator, abrufbar. Voraussetzung für die Nutzung der Informationen im WWW ist natürlich die Anbindung an das Internet, die in der Regel über einen sogenannten ISP (Internet Service Provider) erfolgt. Insbesondere Privatpersonen, aber auch kleinere Unternehmen verfügen in der Regel nicht über eine permanente Verbindung zum Internet. Stattdessen wird die Verbindung je nach Bedarf – manuell oder automatisch – über eine Telefonleitung hergestellt. An Kosten fallen neben dem Grundtarif die Gebühren für die Wählleitung an, die natürlich durch möglichst kurze Online-Zeiten minimiert werden sollen. Dieses Ziel kann durch geeignete Werkzeuge unterstützt werden, die im folgenden beschrieben werden.

Zum Browsen im WWW ist normalerweise eine direkte Verbindung zum Internet erforderlich, um die gewünschten Seiten lesen zu können. Oftmals werden hierbei Seiten aufgrund ihres Informationsgehalts nach dem erstmaligen Ansehen häufig wieder geladen. Besteht keine ständige Verbindung zum ISP, muß die Verbindung jeweils neu aufgebaut bzw. aufrecht erhalten werden, wobei Gebühren anfallen. Denkt man an kleinere Netzwerke, in denen mehrere Benutzer im WWW browsen möchten, fällt dieser Effekt natürlich um so mehr ins Gewicht, da zum einen mehr Seiten gelesen und zum anderen auch gleiche Seiten erneut geholt werden müssen.

Ein Lösung für dieses Problem stellt das Unix-Tool mit dem Namen `wwwoffle` dar, daß im wesentlichen von Andrew M. Bishop entwickelt wurde. Das Akronym `wwwoffle` steht für „WWW offline“, also die Möglichkeit, offline, ohne Verbindung zum Provider, Web-Seiten zu lesen. Natürlich kann man auch mit `wwwoffle` nicht ganz ohne Internet-Provider auskommen. Der große Nutzen dieses Werkzeugs liegt darin, daß man die online angewählten Seiten später im offline-Betrieb in Ruhe lesen kann, ganz so, als wäre man weiterhin mit dem Provider verbunden. Darüber hinaus kann ein `wwwoffle`-Server als Proxy-Server in keineren Netzwerken verwendet werden. Dadurch wird das doppelte Laden identischer Seiten über die kostenpflichtige Leitung zum ISP vermieden. Neben diesen Grundfunktionen bietet `wwwoffle` weitere nützliche Hilfen: So ist es möglich, bestimmte Seiten überwachen zu lassen, die automatisch bei einer Verbindung zum ISP geladen und lokal im Cache abgelegt werden. Auch können ganze Web-Bäume, also Seiten, die sich auf dem Web-Server in unterschiedlich tief geschachtelten Unterverzeichnissen befinden, automatisiert geladen werden. Schließlich läßt sich verhindern, daß bestimmte Seiten im Cache gespeichert werden, was immer dann wichtig ist, wenn solche Seiten sicherheitsrelevante Informationen enthalten (z. B. ein vom Benutzer eingegebenes Passwort). Insgesamt läßt sich mit `wwwoffle` die Online-Zeit sowohl für den Privatmann als auch für

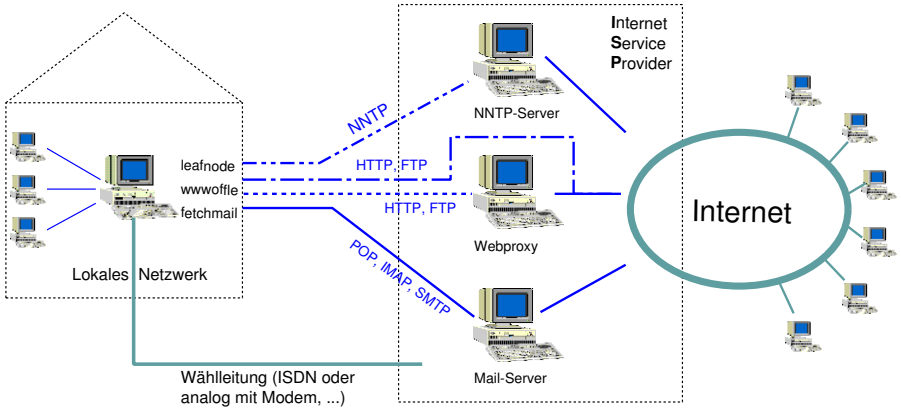

Abbildung 4.9: Schematische Darstellung der Funktion eines ISP

kleinere Netzwerke minimieren, und es entfällt die Notwendigkeit, eine bereits besuchte Seite erneut über das Internet zu laden, da sie jetzt lokal gelesen werden kann.

4.3.1 Arbeitsweise

Das `wwwoffle`-Paket besteht grundsätzlich aus zwei Programmen: `wwwoffled` und `wwwoffle`. `wwwoffled` dient als Proxy, zu dem WWW-Browser eine Verbindung aufbauen können, um Web-Seiten anzufordern, die von `wwwoffled` in einen Cache geschrieben werden. `wwwoffle` ist ein Programm, das der Steuerung und Interaktion von `wwwoffled` dient. Mit Hilfe einer zentralen Konfigurationsdatei (`wwwoffle.conf`, s. u.) ist es möglich, die Funktionsweise des Proxies über zahlreiche Optionen zu steuern.

Wählt ein Benutzer eine bestimmte URL an, kontaktiert der Browser normalerweise direkt den entfernten Rechner, der durch den ersten Teil der URL (dem Namen-Teil, wie z. B. `www.suse.de`) bestimmt ist, und wartet auf die entsprechenden Daten als Antwort. Manche ISP bieten dem Benutzer die Möglichkeit zur Nutzung eines Proxy, der als Cache für eine große Zahl von Web-Seiten dient. Auf diese Weise soll, ähnlich wie bei `wwwoffle`, vermieden werden, daß dieselben Seiten mehrfach über das Internet geholt werden müssen. Unter Umständen muß ein Benutzer sogar einen bestimmten Proxy-Server des ISP einstellen, um überhaupt Seiten des WWW besuchen zu können. Eine schematische Darstellung dieses Aufbaus bietet Abbildung 4.9.

Eine solche Einstellung wird in der Konfiguration des Web-Browsers vorgenommen, bei `netscape` beispielsweise über die Einstellung im Menü *`Edit -> Preferences -> Advanced -> Proxies`*. Hier können die Namen und die Portnummern

der Proxy-Server für die verschiedenen Dienste wie HTTP und FTP konfiguriert werden. Über die entsprechenden Namen und Adressen kann der ISP Auskunft geben, wenn er denn überhaupt einen solchen Proxy unterstüzt bzw. vorschreibt. Zur Vereinfachung dieser Prozedur kann der ISP eine Konfigurationsdatei anbieten, deren URL lediglich im Browser (bei `netscape` das Feld *Automatische Proxy Konfiguration* in dem oben genannten Menü) eingetragen werden muß; alle Proxy-Einstellungen aus den Vorgaben des ISP werden so übernommen.

`wwwoffle` stellt nun lediglich einen (weiteren) Proxy dar, der jedoch im Gegensatz zum evtl. vorhandenen Proxy-Server des ISP auf dem eigenen (bzw. einem der im lokalen Netz liegenden) Rechner arbeitet. Wenn ein Web-Browser eine Verbindung mit einem Web-Server aufbaut, geschieht dies normalerweise auf der Port-Nummer `80`. Eine Port-Nummer kann als Kennummer für einen bestimmten Dienst, wie z. B. auch `wwwoffle`, angesehen werden, unter der genau dieser Dienst auf einem bestimmten Rechner „erreicht“ werden kann. Die Verbindung zwischen Browser und `wwwoffled` erfolgt nicht auf dem Standard-Port 80, sondern in der Default-Konfiguration auf der Port-Nummer 8080. Um `wwwoffle` nutzen zu können, müssen der Rechnernername, auf dem der `wwwoffled`-Prozeß läuft, und die oben angegebene Port-Nummer, auf der `wwwoffled` angesprochen werden kann, in der Proxy-Konfiguration des Web-Browsers eingestellt werden. Alternativ kann, wie bereits gesagt, eine `Proxy.pac`-Datei erstellt werden, deren URL in den Browser eingetragen wird. Dieses Methode wird in Abschnitt 4.3.3.2 auf Seite 303 beschrieben.

Versucht ein Benutzer anschließend, über seinen Web-Browser auf eine Seite im WWW zuzugreifen, wird zunächst eine Verbindung zu dem `wwwoffled`-Prozeß hergestellt. Dieser Prozeß erzeugt einen neuen `wwwoffled`-Prozeß, der die Anfrage bearbeitet. Zunächst sieht dieser Prozeß im lokalen Cache, der normalerweise unter `/var/spool/wwwoffle` liegt, nach, ob die gewünschte Seite bereits angefordert wurde und daher schon vorliegt. Wenn ja, wird die Seite aus dem Cache gelesen und an den Browser geschickt. Liegt die gewünschte Seite jedoch nicht im `wwwoffle`-Cache, hängt es vom aktuellen `wwwoffle`-Arbeitsmodus ab, was weiter geschieht. `wwwoffle` kennt drei verschiedene Arbeitsmodi: den Offline-Modus, der immer dann gewählt werden sollte, wenn zur Zeit keine Verbindung zum Internet besteht, den Online-Modus, falls der Rechner, auf dem `wwwoffled` läuft, zur Zeit über den ISP mit dem Internet verbunden ist, und schließlich den Autodial-Modus, bei dem zunächst immer versucht wird, die geforderte Seite aus dem Cache zu holen, wobei das Netzwerk in diesem Fall als letzte Rettung dient. In diesem Modus geht `wwwoffle` davon aus, daß eine Anfrage an das Netwerk automatisch zu einem Verbindungsaufbau zum Provider führt. Für diesen Modus muß die Verbindung zum ISP also so konfiguriert sein, daß eine Anfrage an eine Adresse außerhalb des lokalen Netzes zu einem automatischen Verbindungsaufbau zum ISP führt. Die Konfiguration des `wwwoffled`-Arbeitsmodus erfolgt entweder über das Programm `wwwoffle` oder

interaktiv mit Hilfe des Web-Browsers und einer zu diesem Zweck von wwwoffle bereitgestellten, lokalen Web-Seite. Mehr zu diesem Thema im folgenden Abschnitt.

Was geschieht nun bei einer Anfrage des Web-Browsers nach einer bestimmten Seite? Die Anforderung einer nicht im Cache liegenden Seite führt im Online-Modus dazu, daß die angeforderte Seite über das Internet geholt, an den Browser geliefert und zusätzlich in den lokalen wwwoffle-Cache geschrieben wird. Tritt beim Laden dieser Seite ein Fehler auf (der entfernte Server ist z. B. nicht erreichbar), meldet wwwoffle diesen Fehler und legt ein Backup einer evtl. existierenden älteren Version dieser Seite in seinem Cache an, so daß später im Offline-Modus immer noch auf die ältere Version der Seite zugegriffen werden kann.

Falls der ISP die Verwendung eines Proxy vorschreibt, also keine direkte Verbindung zu einem Web-Anbieter, sondern lediglich eine Verbindung seinem eigenen Proxy-Server anbietet, der seinerseits die gewünschte Seite anfordert, kann wwwoffle so konfiguriert werden, daß er seine Anfragen ebenfalls an den Proxy des ISP weiterreicht und nicht versucht, eine direkte Verbindung aufzubauen.

Da auch FTP-Verbindungen von wwwoffle verwaltet werden und somit Dateien, die per FTP aus dem Internet mit Hilfe eines Web-Browsers geladen werden, im lokalen Cache landen, ist es wichtig, ausreichend Platz für das Cache-Verzeichnis von wwwoffle bereitzustellen. Falls FTP-Verbindungen vom Web-Browser aus nicht über wwwoffle laufen sollen, muß einfach die Proxy-Konfiguration des Browsers so eingestellt werden, daß für FTP-Verbindungen kein oder evtl. der vom Provider vorgegebene Proxy eingestellt wird. Bei Netscape erfolgt die Einstellung, wie bereits gesagt, im Menü *Edit -> Preferences -> Advanced -> Proxies*. Eine weitere Möglichkeit zu verhindern, daß FTP-Dateien im Cache abgelegt werden, ist der auf Seite 302 beschriebene Konfigurationsabschnitt DontCache der wwwoffle.conf-Datei.

Falls sich wwwoffle im Offline-Modus befindet, wird die Anfrage nach einer Seite registriert, so daß diese Seite automatisch geholt wird, wenn später eine Verbindung zum ISP besteht und wwwoffle in den Online-Modus versetzt wird. Neben der automatischen Registrierung einer Seite bei einem Offline-Zugriff bietet sich die Möglichkeit, wwwoffled so zu konfigurieren, daß der Benutzer in einer von wwwoffle generierten Seite wählen kann, ob er diese Seite zum Download vormerken will oder, ohne Registrierung, einfach eine Fehlermeldung erhält.

Dem Benutzer meldet wwwoffle die Registrierung einer Seite in Form einer dynamisch erzeugten HTML-Seite, in der der Benutzer zum einen über die Registrierung informiert wird und zum anderen die Möglichkeit hat, die gerade vorgenommene Registrierung wieder zu löschen. Abbildung 4.10 auf der nächsten Seite zeigt eine solche Seite. Der Cancel-Link erlaubt die Stornierung von Anfragen, während sich über den Options-Link Einstellungen zur Seitendarstellung vornehmen lassen. Durch die Anwahl dieses Link erscheint eine neue wwwoffle-

Seite, die in Abbildung 4.11 auf der nächsten Seite dargestellt ist. Hier kann der Benutzer beispielsweise festlegen, ob Frames oder Skripten, die zu der Seite gehören, ebenfalls geladen werden sollen. Darüber hinaus kann er bestimmen, ob er nur diese eine Seite, oder aber z. B. auch alle anderen Seiten, die auf dem Web-Server im gleichen Verzeichnis oder in Unterverzeichnissen bis z. B. zur zweiten Ebene stehen, laden lassen möchte.

Über den Link `Monitor` ist es möglich festzulegen, in welchen Abständen die Seite neu aus dem Netz geladen werden soll. Auf diese Weise kann eine Verbindung zum ISP dazu genutzt werden, bestimmte Seiten (z. B. Börsenkurse) automatisch zu aktualisieren, so daß immer eine aktuelle Seite im lokalen `wwwoffle`-Cache verfügbar ist. Die festzulegenden Parameter bestehen aus dem Monat (Jan, Feb, ...), dem Tag des Monats, dem Wochentag und der Stunde, zu der die Seite geholt werden soll. Wird für den Stundenwert keine Angabe gemacht, interpretiert `wwwoffle` dies als Anweisung, diese Seite einmal an dem durch die übri-

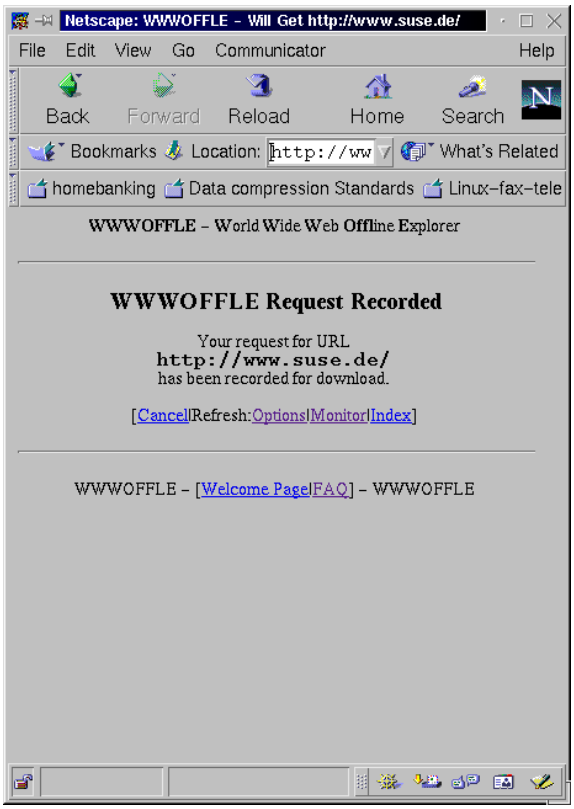

Abbildung 4.10: `wwwoffle`-Ausgabe bei einer Offline-Seitenanfrage

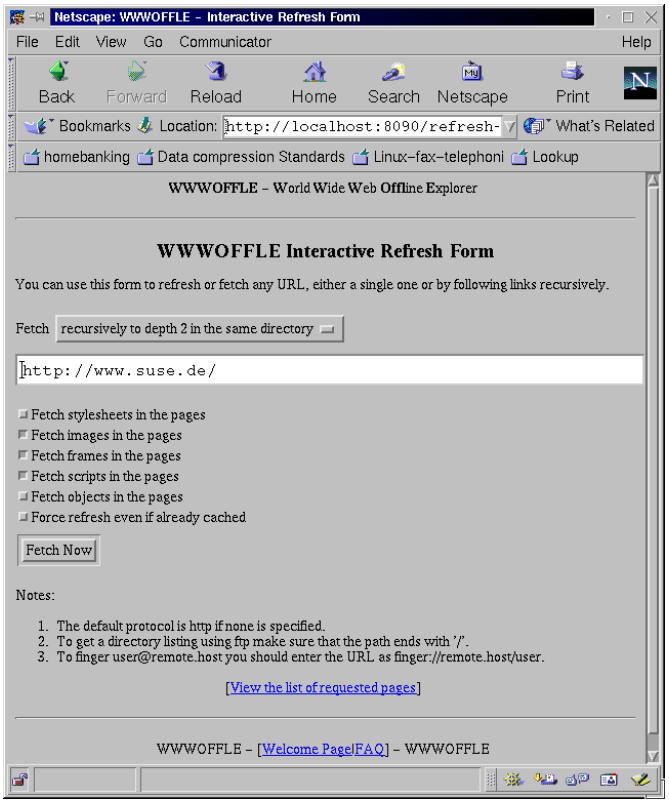

Abbildung 4.11: Download-Optionen in wwwoffle

gen Parameter bestimmten Tag zu holen. Darüber hinaus können auch mehrere Stundenwerte, durch Kommata oder Leerzeichen voneinander getrennt, angegeben werden. Das Feld für den Tag, an dem eine Seite geholt werden soll, darf auch leer gelassen werden. In diesem Fall wird die Seite an allen Tagen, die durch die anderen Angaben bestimmt werden, geholt. Neben der Prüfung einzelner Seiten ist auch ein regelmäßiges rekursives Laden von Seiten möglich. Dazu muß lediglich nach der Festlegung der Optionen für die gewünschte Seite (siehe Bild 4.10 auf der vorherigen Seite) der Fetch-Button gedrückt werden. Anschließend erscheint ein weiterer Dialog, in dem mitgeteilt wird, daß diese Seite(n) zum Download vorgemerkt wurde(n). Wird nun auf derselben Seite der unten stehende Monitor-Link angewählt, können die bereits beschriebenen Monitor-Optionen angegeben werden, um die Häufigkeit des rekursiven Downloads einzustellen.

Um ohne Verbindung zum Internet browsen zu können, muß wwwoffled wissen, ob sich der Rechner, auf dem der Prozeß arbeitet, gerade im Offline- oder im

Online-Modus befindet, also ob gerade eine Verbindung zum Internet über den ISP besteht. Diese Entscheidung kann `wwwoffled` nicht alleine treffen. Stattdessen muß dies dem Prozeß entweder interaktiv über die `wwwoffle`-Kontrollseite oder durch Aufruf des Programms `wwwoffle` mit dem Parameter `-online` bzw. `-offline` mitgeteilt werden. Die Verwendung der `wwwoffle`-Kontrollseite ist nur zu Testzwecken sinnvoll, weitaus besser ist ein vollautomatischer Wechsel. Hierzu muß der Administrator des Rechners dafür Sorge tragen, daß beim Wechsel zum Online-Modus das Kommando `wwwoffle -online`, und bei Beendigung der Verbindung das Kommando `wwwoffle -offline` ausgeführt wird. Diese Notwendigkeit entfällt nur dann, wenn `wwwoffle` sich im Autodial-Modus befindet, d. h., wenn der Versuch eines Verbindungsaufbaus zu einem entfernten Rechner im Internet automatisch zur Herstellung der Verbindung zum ISP führt. In diesem Fall muß `wwwoffle` nur einmalig nach dem Start in den Autodial-Modus versetzt werden, was durch Aufruf des Kommandos `wwwoffle -autodial` geschieht.

Auf Möglichkeiten, die Aufrufe `wwwoffle -online` bzw. `wwwoffle -offline` automatisch beim Verbindungsaufbau ausführen zu lassen, wird im Abschnitt 4.3.3.4 auf Seite 307 eingegangen.

4.3.2 Installation und Grundkonfiguration

`wwwoffle` ist Bestandteil vieler Linux-Distributionen, wie z. B. bei SuSE Linux. Das Programm kann in diesem Fall einfach über das jeweilige Installationswerkzeug installiert werden. Alternativ kann der Quellcode von `wwwoffle` auch von dessen Homepage, `http://www.gedanken.demon.co.uk/wwwoffle/index.html`, frei bezogen werden.

Im Anschluß an die Installation befindet sich – je nach Konfiguration des Pakets bei der Übersetzung des Quelltextes – das Programm `wwwoffle` unter dem Verzeichnis `/usr/bin` bzw. `/usr/local/bin`, das Programm `wwwoffled` unter `/usr/sbin` bzw. `/usr/local/sbin`. Die zentrale Konfigurationsdatei `wwwoffle.conf` liegt normalerweise in `/etc/wwwoffle` oder in `/var/spool/wwwoffle`. In diesem Verzeichnis befindet sich auch der Cache, in dem Web-Seiten von `wwwoffle` abgelegt werden. Um Probleme zu vermeiden, sollte man überprüfen, ob dieses Verzeichnis existiert und der dafür vorgesehenen Benutzerkennung gehört (auf SuSE-Systemen z. B. der Kennung `wwwrun`). Die Benutzerkennung muß mit dem in der Konfigurationsdatei angegebenen Wert für `run-uid`, die Benutzergruppe mit dem für die Variable `run-gid` zugeordneten Wert übereinstimmen. Die Konfigurationsdatei kann entweder mit einem Texteditor angepaßt oder aber später direkt über den Web-Browser editiert werden. In Mehrbenutzerumgebungen, wo es nicht sinnvoll ist, daß jeder Benutzer die Konfigurationsdatei nach seinen Wünschen ändern darf, kann das editieren dieser Datei durch ein Passwort geschützt werden.