


Web Loupe



Benutzerdokumentation

Version 0.5

Februar 2005

*Ein Crawler
zur Analyse und Visualisierung
von Website-Strukturen*

Ansprechpartner

Katja Langholz & Matthias Kahlau

Email:

kaimka@users.sourceforge.net
webspirit@users.sourceforge.net

Webseiten:

<http://www.webloupe.de.vu>
<http://www.sourceforge.net/projects/webloupe/>

Inhaltsverzeichnis

1 Einleitung.....	4
1.1 Was ist WebLoupe.....	4
1.2 Warum WebLoupe.....	4
1.3 Benutzerdokumentation.....	5
2 Die Software „WebLoupe“.....	5
2.1 Open-Source-Lizenz.....	5
2.2 Softwarefunktionen.....	5
2.3 Softwarekomponenten.....	6
2.3.1 Die Benutzeroberfläche.....	7
2.3.2 WebLoupe multi-threaded Crawler.....	8
2.3.3 Visualisierungstool Touchgraph.....	8
2.3.4 Visualisierungstool HyperGraph.....	9
2.4 Abgrenzung zu anderen Crawlern.....	10
3 Installation.....	10
3.1 Technische Voraussetzungen.....	10
3.2 Download von WebLoupe.....	10
3.3 Installieren und Starten von WebLoupe.....	11
4 Benutzung von WebLoupe.....	12
4.1 Vorbereitung.....	12
4.2 Einstellungen und Meldungen.....	12
4.2.1 Crawler Einstellungen vor dem Start.....	12
4.2.2 Crawler Starten und Beenden.....	15
4.2.3 Statuszeile.....	16
4.3 TreeView/Statistics.....	17
4.4 Touchgraph.....	19
4.5 HyperGraph.....	22
5 Ausblick.....	24
6 Ansprechpartner.....	25
7 Quellenangaben.....	26

1 Einleitung

1.1 Was ist WebLoupe

WebLoupe ist eine Software zur Analyse und Visualisierung von Webseiten und ihren Verknüpfungen. Sie basiert auf einer Web Crawler-Technologie (auch Spider oder Robot genannt), d.h. Webseiten werden nach Links durchsucht und diese weiterverfolgt, um weitere Webseiten einzubeziehen. Dabei können Daten der einzelnen Webseiten gesammelt werden, wie Titel, Dateigröße, enthaltene Bilder etc. Was genau ein Crawler leisten kann, hängt von der Charakteristik der einzelnen Lösungen ab. WebLoupe ist nicht der einzige Web Crawler, den es gibt, soll aber eine echte Alternative zu anderen Produkten darstellen.

Die Entwicklung von WebLoupe hat im November 2004 im Rahmen eines Studienprojektes der Informatik an der Fachhochschule Kaiserslautern, Standort Zweibrücken begonnen. Es sollte eine Software zum Thema Informationsarchitektur entwickelt werden. Dabei mussten Risiken wie Visualisierung und Performanz der Anwendung, sowie die relativ knappe Entwicklungszeit zum Abgabetermin im Februar 2005 berücksichtigt werden. Die genaue Ausprägung der Software wurde aber vom Dozenten nicht festgelegt, sondern konnte nach eigenen Vorstellungen entwickelt werden. Vorgabe war jedoch die Entwicklung unter einer Open-Source Lizenz. Das bedeutet, dass WebLoupe nach Abgabe der ersten Version von den ursprünglichen Entwicklern und von Interessenten weiter entwickelt werden kann, z.B. zur Implementation zusätzlicher Features.

Die Implementierung ist mit der Programmiersprache Java von SUN Microsystems in der aktuellen Version 5.0 erfolgt. Diese Version der Java 2 Standard Edition (J2SE) bietet mehr Möglichkeiten bei der Umsetzung im Vergleich zu Vorgängerverionen.

Der Vertrieb des Produkts findet über das Internet statt. Dazu wird es auf der größten Plattform des Internets für open-source Softwareentwicklung – SourceForge.net – veröffentlicht. Zusätzlich wird das Produkt auf einer eigenen Produktwebseite (URL: <http://www.webloupe.de.vu>) zur Verfügung gestellt, von der ebenfalls der Download und die Kontaktaufnahme zu den Entwicklern möglich ist.

1.2 Warum WebLoupe

Im Internet herrscht seit geraumer Zeit ein regelrechtes Chaos, durch die Flut und Heterogenität von Informationen. Durch die Fülle von Informationen, die leicht von jedem im Internet publiziert werden können, wird es schwierig, das Gesuchte zu finden.

Die Komplexität und Informationsdichte des Internets, welches in den letzten Jahren zu einem Massenmedium geworden ist, hat die Ausrichtung des Produkts motiviert. Durch die schlechte Benutzbarkeit vieler Webseiten ist es für viele Internetbenutzer schwierig, den Überblick zu behalten und Informationen zu finden. Daher wird mit WebLoupe eine Software bereitgestellt, welche zur Analyse, Visualisierung und Exploration von lokalen oder öffentlich zugänglichen Webseiten verwendet werden kann. WebLoupe kann die Informationsrecherche, aber auch die Administration von Webseiten durch die Analyse von den Inhalten und der Erreichbarkeit von Webseiten, und die anschließende Auflistung der Ergebnisse.

1.3 Benutzerdokumentation

Diese Dokumentation wurde mit dem Ziel geschrieben, Benutzer der Software bei der Verwendung zu unterstützen und grundlegende Funktionalitäten zu erklären. Sie erhalten nähere Informationen über die Funktionen von WebLoupe (Kap. 2), und erhalten Hinweise zur Installation (Kap. 3) und Benutzung (Kap. 4).

2 Die Software „WebLoupe“

2.1 Open-Source-Lizenz

WebLoupe ist ein Programm, das bei SourceForge.net, dem größten open-source Anbieter im Internet, zum freien Download zur Verfügung gestellt wird. Das Programm kann nicht nur einfach benutzt werden, sondern der Programmcode ist durch die open-source Lizenz auch für jeden zugänglich, kann erweitert und ins eigene open-source Programm eingebaut werden [1].

WebLoupe wird unter der Open-Source Lizenz GNU General Public License (GPL) entwickelt. Mehr Informationen hierzu unter <http://www.opensource.org>. Die Lizenz finden sie ebenfalls auf der Produktwebseite von WebLoupe unter <http://www.webloupe.de.vu>, und ist auch im Programmpaket von WebLoupe enthalten.

2.2 Softwarefunktionen

Die Anwendung WebLoupe dient zur Analyse und Visualisierung der Informationsarchitektur von Webseiten. Dabei werden öffentliche Internetseiten oder lokal gespeicherte Seiten heruntergeladen, analysiert und kann mittels einer grafischen Benutzeroberfläche als Baumstruktur, tabellarisch und in Form interaktiver Graphen dargestellt werden.

WebLoupe hat eine anwenderfreundliche Benutzeroberfläche, die folgende Funktionen und Einstellungen bereithält:

- Linkverfolgung von Webseiten
- Online/Offline Suche
- Überprüfung der Gültigkeit von Links (Broken Links)
- Statusanzeige des Crawlvorgangs einzelner Seiten
- Repräsentation von Webseiten mittels Baumstruktur, Tabellenübersicht und interaktivem Visualisierungstool:
 - in der Baumstruktur werden die Seiten mit Titel und URL angegeben; mit Doppelklick auf einen Knoten kann die Seite im Standardbrowser geöffnet werden. Broken Links werden in roter Schrift hervorgehoben.
 - die Tabellenübersicht zeigt optional Titel, URL, HTTP Response Code, Status, Mime-Typ, Anzahl gefundener Zeichen und Bilder pro Seite, Dateigröße einer Seite und evt. gefundene Keywords an
 - das Visualisierungstool (TouchGraph) repräsentiert die Seite mittels Knoten und Kanten und enthält weitere Funktionen zum Zoomen, Rotieren, Verändern und Navigieren des Graphen
- Die Visualisierung (TouchGraph) kann als Bild gespeichert werden
- Ausserdem ist die Möglichkeit vorhanden, eine GraphXML Datei zu generieren und diese mit

dem HyperGraph Java Applet zur interaktiven Visualisierung zu verwenden

- Statuszeile, die angibt, wie viele URLs gefunden wurden, wie viele Seiten nicht erreichbar sind, von wie vielen Seiten Ressourcen erfolgreich heruntergeladen wurden und wie viele Schlüsselwörter gefunden wurden. Dabei wird der Crawlingprozess in der Statuszeile dynamisch angezeigt

Das Verhalten des Crawlers und die Anzeige von Ergebnissen können über die Benutzeroberfläche konfiguriert werden:

- maximale Anzahl der Seiten
- maximale Crawlingtiefe
- Einschränkung des Suchbereiches (alles, nur innerhalb der Startseite, .com, .de, .edu, .net, .org)
- Auswahl des Mime-Typs (text/html, alle)
- Angabe der maximalen Größe von Dateien (in Byte), die heruntergeladen und durchsucht werden sollen
- Angabe des maximalen Connection Timeouts (Dauer des Versuchs, die Verbindung zu einer URL aufzubauen; danach wird der Versuch abgebrochen)
- ein oder mehrere Keywords, nach denen auf den Seiten gesucht werden soll, können angegeben werden
- Statistics - Informationen über Seiten, die der tabellarischen Übersicht der Ergebnisse aufgelistet werden sollen:
 - Titel
 - URL
 - HTTP Response Code
 - Status
 - Mime Typ
 - Anzahl der Zeichen pro Seite
 - Anzahl der Bilder pro Seite
 - Dateigröße

In den nachfolgenden Kapiteln werden die Funktionen und Einstellungsmöglichkeiten von WebLoupe näher erläutert.

2.3 Softwarekomponenten

WebLoupe besteht aus einem multi-threaded Crawler, der Webseiten herunterlädt und ihre Inhalte analysiert, um weiterführende Links und andere Eigenschaften herauszufinden.

WebLoupe bietet eine grafische Benutzeroberfläche für die Konfiguration des Crawlers und zur Anzeige der Ergebnisse. Die Ergebnisse eines Crawlvorgangs werden in einer navigierbaren Baumstruktur und in einer Tabelle dargestellt.

Ausserdem besteht WebLoupe aus weiteren Visualisierungstools. Mit dem integrierten TouchGraph, können die Verknüpfungen von Webseiten visuell und interaktiv dargestellt werden. Ein Exportmodul dient zur Generierung von GraphXML Dateien, die mit dem HyperGraph Applet angezeigt werden können, das im Umfang von WebLoupe enthalten ist.

2.3.1 Die Benutzeroberfläche

Die Benutzeroberfläche stellt für den User die Interaktionsmöglichkeit mit dem Programm dar. Implementiert wurde sie mit Java Swing. Nähere technische Informationen erhalten Sie hierzu in der Entwicklerdokumentation.

Es können verschiedene Einstellungen vorgenommen werden, um den Crawler für die Suche zu konfigurieren (siehe Kapitel 4.2). Wird der Crawler gestartet, wird der Crawling-Vorgang schon während des Crawlens dynamisch in einer tabellarischen Übersicht und als navigierbare Baumansicht dargestellt (siehe Kapitel 4.3).

Ist dieser Vorgang abgeschlossen, kann die graphische Visualisierung der Webseite (TouchGraph) betrachtet werden, und durch die erzeugte Struktur navigiert werden. Nähere Informationen hierzu in Kapitel 4.4.

Anschließend ist es möglich, die graphische Visualisierung als Bild lokal abzuspeichern und eine GraphXML Datei zu generieren, die unter Verwendung eines mitgelieferten Applets zur interaktiven Visualisierung verwendet werden kann (HyperGraph).

Mit dem Applet kann das Ergebnis einer Suche entweder einfach lokal im Browser, aber auch auf einer Webseite angezeigt werden, wenn man das Applet in seine Webseite einbaut und die generierte Datei sowie weitere erforderliche Dateien, die im Umfang von WebLoupe enthalten sind, auf dem Server mit der Webseite bereitstellt. Nähere Informationen hierzu in Kapitel 4.5.

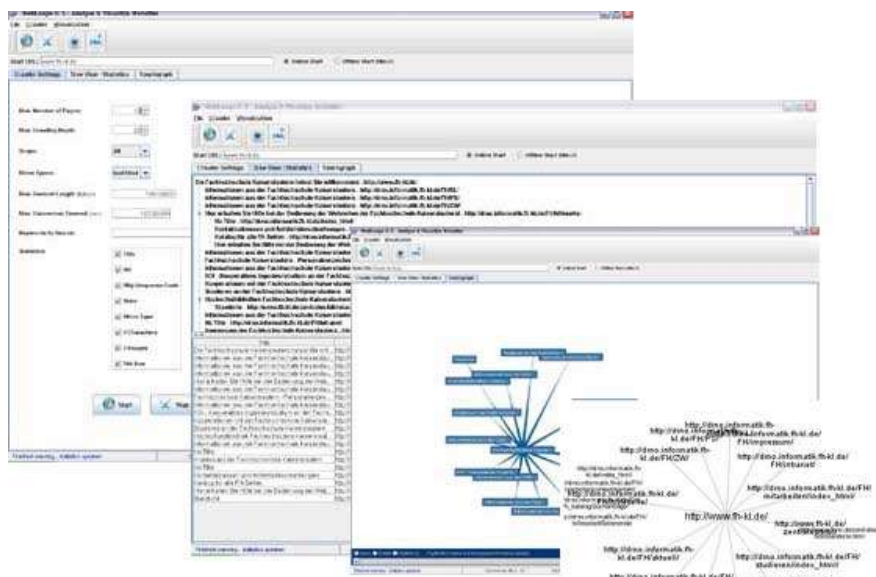


Abbildung 1 Benutzeroberfläche von WebLoupe 0.5

2.3.2 WebLoupe multi-threaded Crawler

Der Algorithmus für das Crawlen von Webseiten wurde als Breitensuche [2] implementiert. Breitensuche ist ein Fachbegriff der Informatik, welcher ein Verfahren zum Durchsuchen bzw. Durchlaufen der Knoten eines Graphen bezeichnet. Die Struktur, die Webseiten durch ihre Verknüpfungen bilden, kann man sich als eine Baumstruktur bzw. als Graph mit Knoten und Kanten vorstellen.

Eine Webseite enthält meist eingebettete URLs, die auf weitere Seiten verweisen. Geht man von der Startseite aus, wären das die Kinder (Unterknoten) in der ersten Ebene. Diese Kinder haben wiederum eingebettete Links, die auf weitere Seiten verweisen. Das sind die Kinder der zweiten Ebene, u.s.w.. Bei einer Breitensuche werden erst alle Webseiten einer Ebene gesammelt, dann die der nächsten Ebene u.s.w..

Der Crawler ist multi-threaded. Das bedeutet, dass das Programm mehrere Arbeiten gleichzeitig verrichten kann. Muss der Crawler beispielsweise auf den Download einer URL warten, weil der Server nicht sofort reagiert, kann indessen eine andere URL heruntergeladen werden. Das erhöht die Geschwindigkeit des Crawlvorgangs. Ein weiterer Vorteil ist, dass eine Anwendung, bei der mehrere Threads verwendet werden, auch mehr Rechenzeit vom Betriebssystem zugeteilt bekommt, wenn parallel noch andere Anwendungen laufen.

Robots.txt werden von WebLoupe in dieser Version 0.5 noch nicht unterstützt. Diese Dateien können von Administratoren zu einer Webseite auf dem Server abgelegt werden, um dort festzuhalten, welche Verzeichnisse von Crawlern, oder auch Robots, durchsucht werden sollen. Die Fähigkeit, robots.txt Dateien zu interpretieren, soll WebLoupe in einer nachfolgenden Version erhalten.

Nähere Informationen zur Konfiguration des Crawlers finden sie in Kapitel 4.2.

Technische Informationen zur Implementation des WebLoupe multi-threaded Crawler erhalten sie in der Entwicklerdokumentation auf der Webseite des Projekts unter folgender Adresse:
<http://www.webloupe.de.vu>

Die Entwicklerdokumentation ist ausserdem im Programmpaket von WebLoupe enthalten, das den Programmcode enthält (WebLoupe-0.5-src.zip). Es kann ebenfalls unter ober genannter Adresse heruntergeladen werden.

2.3.3 Visualisierungstool Touchgraph

Zur graphischen Visualisierung wurde das open-source Tool „Touchgraph“ als Baustein von WebLoupe eingebaut und an die Erfordernisse von WebLoupe angepasst.

TouchGraph ist ein Programm, das zur Visualisierung von Graphen, also Gebilden aus Knoten und Kanten dient. Ein Knoten kann beliebig viele Kinder (Unterknoten) haben. In den Knoten werden die Titel der Webseiten angezeigt. Die Verknüpfungen von Webseiten werden über die Kanten dargestellt.

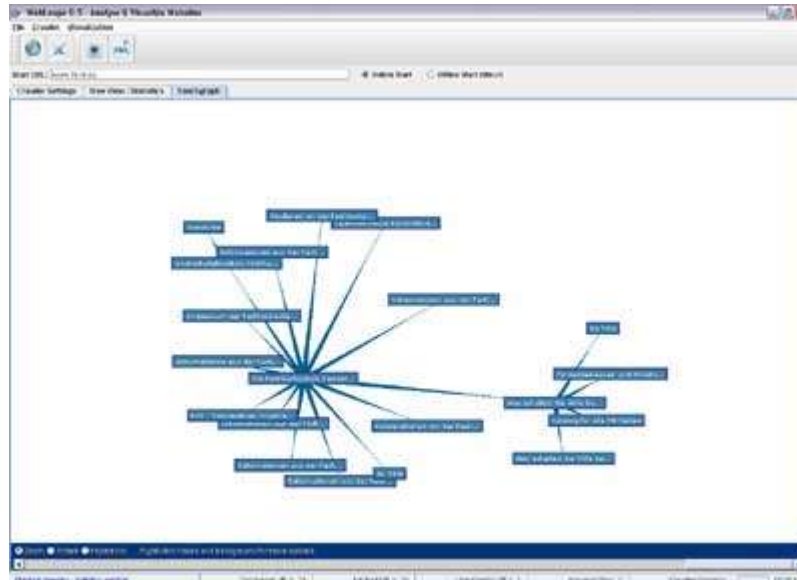


Abbildung 2 Touchgraph

Die Visualisierung ist interaktiv, d.h. Knoten können ein- und ausgeblendet und verschoben werden, der Graph kann heran- oder herausgezoomt und rotiert werden. Diese Funktionen können mit der Maustaste, im Kontextmenü und in der Bedienleiste unterhalb des TouchGraph ausgewählt und eingestellt werden. Nähere Informationen zur Benutzung erhalten sie in im Kapitel 4.4.

Wollen sie mehr über TouchGraph selbst erfahren, können sie deses Projektwebseite unter folgender Adresse besuchen:

<http://www.touchgraph.com>

2.3.4 Visualisierungstool HyperGraph

HyperGraph ist ein open-source Java Programm mit Funktionen zur interaktiven Visualisierung von Graphen, also Knoten, die über Kanten verbunden sind. Ein Knoten kann auch hier beliebig viele Kinder (Unterknoten) haben. Zur Beschreibung der Baumstruktur wird von HyperGraph das GraphXML Format verwendet. GraphXML ist eine Beschreibungssprache für Graphen im XML Format, die Attribute von Knoten und deren Verknüpfungen ausdrücken kann.

In WebLoupe wurde eine Exportfunktion implementiert, welche anhand der erstellten Baumstruktur des Crawlers eine GraphXML Datei generieren kann.

Diese Datei kann vom HyperGraph Applet eingelesen werden. Die graphische Visualisierung der Webseite kann so in einem Applet betrachtet werden, das beispielsweise in eine Webseite eingebaut werden kann. In den Knoten werden die URLs von Webseiten angezeigt. Mit einem Mausklick auf den Knoten kann die Webseite im Standardbrowser geöffnet werden. Genauere Informationen zur Generierung der GraphXML Datei und Verwendung des Applets erhalten Sie im Kapitel 4.5.

Weitere Informationen zu HyperGraph selbst und Beispiele finden sie auf dessen Projektwebseite:

<http://hypergraph.sourceforge.net/>

2.4 Abgrenzung zu anderen Crawlern

WebLoupe ist ein eigenständiges und komplettes Programm zum Crawlen von Webseiten, zur Analyse von Inhalten, und zur Visualisierung der Ergebnisse. Die meisten im Netz kostenlos angebotenen (open-source) Crawler stellen nur einen Programmbaustein dar, der zur weiteren Implementierung für eigentliche Crawlerprogramme von Entwicklern verwendet werden kann, und für sich allein keine Option zur Benutzung durch Nicht-Programmierer darstellt. Der multi-threaded Crawler von WebLoupe basiert aber nicht auf einem bereits vorhanden Crawler, sondern wurde selbst entwickelt.

Ebenso differenziert sich WebLoupe durch seine vielfältige Repräsentation von Webseiten. Webseiten werden tabellarisch, und hierarchisch mittels Baumstruktur in verschiedenen Arten dargestellt, und können interaktiv navigiert und angepasst werden.

In der Tabelle werden vielfältige Informationen über die Erreichbarkeit und Inhalte gefundener Webseiten angezeigt. Sogar eine Keywordsuche ist möglich.

Mit der Möglichkeit, das Ergebnis eines Suchvorgangs als Bild in Form eines TouchGraph abzuspeichern, oder als GraphXML zu exportieren, und mit dem mitgelieferten Applet auf Webseiten anzuzeigen, ergeben sich völlig neue Möglichkeiten. Programme, die ähnliches wie WebLoupe leisten können, sind unseres Wissens nicht kostenlos zu bekommen.

3 Installation

3.1 Technische Voraussetzungen

Voraussetzung für die Ausführung von WebLoupe ist eine passende Laufzeitumgebung, das Java Runtime Environment (JRE) der Java 2 Standard Edition (J2SE). Für WebLoupe wird mindestens die aktuelle Version 5.0 (auch 1.5 genannt) benötigt.

Das JRE 5.0 kann kostenlos unter folgender Adresse für alle gängigen Betriebssysteme heruntergeladen werden (ca. 15 MB): <http://java.sun.com/j2se/1.5.0/download.jsp>

Für WebLoupe wird auch ein Paket angeboten, welches bereits ein gebündeltes Java Runtime Environment für Microsoft Windows Betriebssysteme enthält. (WebLoupe-0.5-exeJre.zip). Wird dieses Paket verwendet, wird keine gesonderte Laufzeitumgebung benötigt. Nähere Informationen hierzu in Kapitel 3.2.

3.2 Download von WebLoupe

WebLoupe wird einmal in Form einer ausführbaren JAR Datei (.jar) angeboten, für die ein JRE, mindestens in der Version 5.0, auf dem Betriebssystem installiert sein muss. Sie kann auf allen gängigen Betriebssystemen, für die ein JRE zur Verfügung steht, ausgeführt werden.

WebLoupe gibt es aber auch in Form von .exe Dateien zur Benutzung auf einem Microsoft Windows Betriebssystem. Eine Windows .exe kann wie die JAR Datei einfach per Doppelklick ausgeführt werden, wird aber durch ein WebLoupe Icon dargestellt. Ausserdem gibt es sie in 2 Varianten, einmal ohne und einmal mit gebündelter JRE.

Die Variante ohne JRE (ca. 0.7 MB, mit Benutzerdokumentation) sucht nach Doppelklick nach der benötigten JRE. Ist nicht mindestens ein JRE 5.0 auf dem Zielsystem vorhanden, wird der Anwender über einen Dialog benachrichtigt, und kann die JRE nach entsprechender Auswahl direkt von oben genannter Webseite des Herstellers kostenlos herunterladen.

Die zweite Variante ist zwar entsprechend größer (ca. 20.7 MB), dafür wird aber keine vorherige Installation des JRE benötigt. Sie ist im Programmpaket von WebLoupe enthalten, d.h. WebLoupe kann nach Doppelklick auf die .exe sofort gestartet werden, unabhängig von dem Vorhandensein einer JRE.

Die mit WebLoupe gebündelte JRE ist unsichtbar für das Betriebssystem, da sie nicht installiert wird. Sie wird daher nicht automatisch von anderen Anwendungen verwendet.

Ingesamt stehen folgende Pakete zum Download bereit:

- WebLoupe-0.5-jar.zip (die jar-Datei und die Benutzerdokumentation, ca. 0,7 MB)
- WebLoupe-0.5-exe.zip (die exe-Datei und die Benutzerdokumentation, ca. 0,7 MB)
- WebLoupe-0.5-exeJre.zip (die exe-Datei gebündelt mit dem JRE, und die Benutzerdokumentation, ca. 20.7 MB)
- WebLoupe-0.5-src.zip (der Programmcode für Entwickler, die Benutzer- und Entwickler dokumentation, ca. 1,66 MB)

Sie können unter folgenden Adressen kostenlos heruntergeladen werden:

1. <http://www.webloupe.de.vu>
2. <https://sourceforge.net/projects/webloupe/>

WebLoupe untersteht der open-source Lizenz Gnu General Public License (GPL). Die Lizenz ist in den Downloadpaketen von WebLoupe enthalten, kann aber auch unter der ersten der oben angegebenen Adressen eingesehen und heruntergeladen werden.

3.3 Installieren und Starten von WebLoupe

- Download von
 - *WebLoupe-0.5-jar.zip* (plattformunabhängig)
 - oder *WebLoupe-05-exe.zip* (Microsoft Windows)
 - oder *WebLoupe-05-exeJre.zip* (Microsoft Windows)
- Entpacken der Datei in ein Verzeichnis. Das .zip Format ist ein gängiges Archivierungsformat zur Reduzierung der Größe von Dateien, und kann unter Microsoft Windows beispielsweise einfach entpackt werden, indem die Datei mit der rechten Maustaste ausgewählt wird, um das Kontextmenü zu öffnen, und dort die Entpacken Funktion des Archivierungsprogramms auswählt, z.B. „Extract to WebLoupe-0.5-exe/“.
- Zum Starten der Anwendung Doppelklicken der .jar oder .exe Datei im WebLoupe-Verzeichnis. (Beachten sie, dass zur Benutzung der JAR-Datei ein Java Runtime Environment auf dem Betriebssystem installiert sein muss, mindestens in der Version 5.0, siehe Kapitel 3.2.)
- Zu den .jar oder .exe Dateien können auch Desktop Verknüpfungen erstellt werden. Unter Microsoft Windows müssen sie dazu mit der rechten Maustaste die ausführbare Datei auswählen, um das Kontextmenü zu öffnen, und dort „Verknüpfung erstellen“ auswählen. Dann wird ein

Icon auf dem Desktop erstellt, das sie zum Starten von WebLoupe doppelklicken können.

4 Benutzung von WebLoupe

4.1 Vorbereitung

1. Stellen Sie eine aktive Internetverbindung her (Offline und Online Suche; bei der Offline Suche kann zwar von einer lokal gespeicherten Webseite ausgegangen werden, zur Verfolgung von Verknüpfungen zu Internetseiten durch den Crawler muss aber natürlich trotzdem eine aktive Internetverbindung hergestellt sein.)
2. Starten Sie WebLoupe, um die Benutzeroberfläche zu öffnen

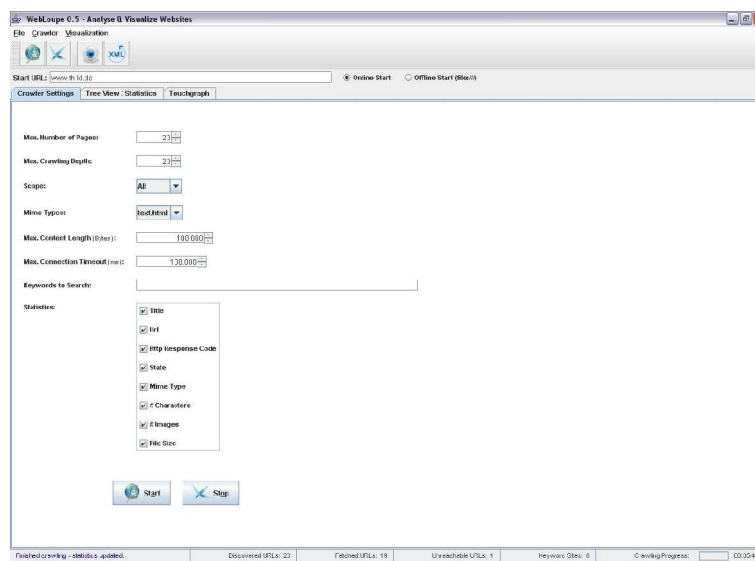


Abbildung 3 Benutzeroberfläche – Crawler Settings

4.2 Einstellungen und Meldungen

4.2.1 Crawler Einstellungen vor dem Start

- Geben sie die **Start URL** in das dafür vorgesehene Textfeld ein. Das ist die URL, von der der Crawler mit der Suche beginnen soll. Es können nur URLs angegeben werden, die das „http“ oder das „file“ Protokoll verwenden. Andere Protokolle werden momentan nicht von WebLoupe unterstützt.
- Wählen Sie ob, sie eine Online oder Offline Suche vornehmen möchten. Die Offline Suche ist dafür vorgesehen, dass eine lokal gespeicherte Webseiten-Datei als Ausgangspunkt der Suche angegeben werden kann. Für die Offline Suche müssen sie im Feld der StartURL das Dateiprotokoll mit dem absoluten Pfad zu der Datei in der URL spezifizieren.

- Bsp.: file:///C:/homepage.html

Für die Online Suche ist es bei URLs, die mit „www“ oder „www2“ beginnen, nicht erforderlich, das Protokoll „http://“ voranzustellen. Ansonsten schon.

Die folgende Abbildung zeigt rot hervorgehoben das Textfeld zur Eingabe der Start URL.



Abbildung 4 Eingabe der URL

- Bevor sie den Crawler starten, können sie weitere **Einstellungen** für den Crawler vornehmen. Ein Bildschirmfoto der Benutzeroberfläche für die Einstellungen sehen sie im Anschluss an die folgenden Erläuterungen.

➤ *Max. Number of Pages:*

Hier kann die maximale Seitenanzahl, die durchsucht werden soll, angegeben werden. Erreicht der Crawler diese Anzahl von Seiten, bricht der Crawlvorgang ab.

➤ *Max. Crawling Depth:*

Hier kann die maximale Crawlingtiefe angegeben werden. Der Crawler wird somit auf eine bestimmte hierarchische Suchtiefe beschränkt. Alle Webseiten, die sich von der Startseite ausgehend in einer tieferen Ebene befinden, werden nicht vom Crawler berücksichtigt.

➤ *Scope:*

Scope bezeichnet eine Bereichseinschränkung. Bestimmte Domains können gefiltert werden, so dass der Crawler nur nach ganz bestimmten Seiten im Netz sucht. Folgende Bereiche stehen zur Auswahl:

- *all* – keine Einschränkung
- *Startseite* – nur innerhalb der Startseite, keine externen Urls
- *.com* – nur Seiten in der .com Domain
- *.de* – nur Seiten in der .de Domain
- *.edu* – nur Seiten in der .edu Domain
- *.net* – nur Seiten in der .net Domain
- *.org* – nur Seiten in der .org Domain

➤ *Mime Type:*

Hier ist eine Einschränkung des Mime-Typs der Dateien möglich, die vom Crawler berücksichtigt werden. Mime-Typen werden zur Kennzeichnung des Inhalts und des Formats von Dateien verwendet. Webseiten, die einem anderen Mime-Typ als dem hier eingestellten angehören, werden nicht nach Links zur Weiterverfolgung durchsucht. Zur Auswahl stehen *all*

für alle und *text/html*. Weitere Informationen zu Mime-Typen finden sie im Internet unter [5].

➤ *Max. Content Length:*

Die maximale Größe der vom Crawler berücksichtigten Dateien kann hier angegeben werden. Alle Dateien, die größer sind, werden nicht nach Links zur Weiterverfolgung durchsucht. Die max. Größe wird in Byte angegeben. 1024 Byte entsprechen einem Kilobyte. 1024 Kilobyte wiederum einem Megabyte. Durch die Angabe in Byte ist eine sehr feine Justierung möglich.

➤ *Max. Connection Timeout:*

Die Dauer des Versuchs, eine Verbindung zur einer Seite herzustellen, kann hier eingeschränkt werden. Braucht der Crawler länger, um eine Verbindung aufzubauen, beispielsweise weil der Server überlastet ist, wird der Verbindungsaufbau unterbrochen und die Seite nicht zum Analysieren heruntergeladen. Dadurch können Blockierungen verhindert werden. Die Zeit wird in Millisekunden angegeben. 1000 Millisekunden entsprechen einer Sekunde. Die Angabe von 0 bedeutet, dass es keinen Connection Timeout gibt!

➤ *Keywords to search:*

Hier können Schlüsselwörter angegeben werden, nach denen auf den Seiten gesucht werden soll. Mehrere Schlüsselwörter werden einfach durch Leerzeichen oder Kommas getrennt. Es wird nicht zwischen Groß- und Kleinschreibung unterschieden.

Die Anzahl von Seiten, auf denen mind. ein Schlüsselwort gefunden wurde, wird dann in der Statuszeile während des Crawlvorgangs dynamisch angezeigt. Gefundene Keywords pro Seite werden nach dem Crawlvorgang in der tabellarischen Übersicht angezeigt.

➤ *Statistics:*

Hier kann bestimmt werden, welche Informationen in der Tabelle, die in der TreeView/Statistics Ansicht nach Ende des Crawlvorgangs angezeigt wird, angezeigt werden sollen. Nähere Informationen hierzu erhalten sie im Kapitel 4.3 zu TreeView/Statistics.

Die folgende Abbildung zeigt ein Bildschirmfoto der Ansicht „Crawler Settings“ von WebLoupe mit den Einstellungsmöglichkeiten.

Max. Number of Pages:	<input type="text" value="20"/>
Max. Crawling Depth:	<input type="text" value="20"/>
Scope:	<input type="button" value="All"/>
Mime Types:	<input type="button" value="text/html"/>
Max. Content Length (Bytes):	<input type="text" value="100.000"/>
Max. Connection Timeout (ms):	<input type="text" value="100.000"/>
Keywords to Search:	<input type="text"/>
Statistics:	<input checked="" type="checkbox"/> Title <input checked="" type="checkbox"/> Url <input checked="" type="checkbox"/> Http Response Code <input checked="" type="checkbox"/> State <input checked="" type="checkbox"/> Mime Type <input checked="" type="checkbox"/> # Characters <input checked="" type="checkbox"/> # Images <input checked="" type="checkbox"/> File Size

Abbildung 5 WebLoupe - Crawler Settings (Einstellungen)

4.2.2 Crawler Starten und Beenden

Um den Crawler zu starten bzw. zu beenden sind verschiedene Möglichkeiten vorgegeben.

➤ *mittels Menü:*

- Crawler>Start: Starten des Crawlingvorgangs
- Crawler>Stop: Beenden des Crawlingvorgangs



Abbildung 6 Menü

➤ *mittels Buttons (rot hervorgehoben):*

- links ist der Startbutton
- rechts davon der Stopbutton



Abbildung 7Buttons

➤ *mit den Shortcut Tasten, die in der Abbildung 6 zum Menü ersichtlich sind:*

- Strg-A zum Starten des Crawlers (unter Windows)
- Strg-O zum Stoppen des Crawlers (unter Windows)

➤ *mit den Mnemonic Tasten, die ebenfalls in der Abbildung 7 zum Menü ersichtlich sind:*

- Alt-c zum Öffnen des Crawler Menüs (nicht notwendig zum Betätigen der Funktion)
- Alt-a zum Starten des Crawlers
- Alt-o zum Stoppen des Crawlers

Der Crawler stoppt automatisch, wenn keine weiteren Seiten mehr gefunden wurden, entweder aufgrund der Sucheinschränkungen, oder aufgrund dessen, dass einfach keine weiteren Seiten in den bereits durchsuchten Seiten entdeckt wurden, oder heruntergeladen werden konnten.

4.2.3 Statuszeile

In der Statuszeile, die sich ganz unten horizontal über die Breite des Anwendungsfensters erstreckt, werden dynamisch während des Crawlvorgangs folgende Meldungen angezeigt:

- Meldungen über den Status und Tätigkeiten des Crawlers, Fehlermeldungen
- Discovered URLs (Anzahl der gefundenen URLs)
- Fetched URLs (Anzahl der erfolgreich heruntergeladenen URLs)
- Unreachable URLs (Anzahl der nicht erreichbaren URLs, z.B. Broken Links oder aufgrund der Sucheinschränkungen ausgeschlossene Urls)
- Keyword Sites (Anzahl der Seiten, auf denen mind. ein Schlüsselwort gefunden wurde)
- Anzeige des Crawlvorgangs durch einen Fortschrittsbalken
- Anzeige der Zeit, die seit dem Start des Crawlers vergangen ist

Finished crawling - statistics updated.	Discovered URLs: 20	Fetched URLs: 19	Unreachable URLs: 1	Keyword Sites: 0	Crawling Progress: <input type="text"/>
---	---------------------	------------------	---------------------	------------------	---

Abbildung 8 Statuszeile

4.3 TreeView/Statistics

Wird der Crawlvorgang gestartet, öffnet sich automatisch die Ansicht mit der Baumstruktur (TreeView) und der Tabelle (Statistics).

Der TreeView zeigt schon während des Crawlens die aktuellen Ergebnisse des Crawlvorgangs an. In der Baumstruktur können Sie die hierarchische Verlinkung der einzelnen Seiten entnehmen, die pro Knoten des Baumes mit Titel und URL angegeben werden. Nicht gefundene Seiten werden in roter Schrift hervorgehoben, so dass „Broken Links“ schneller lokalisiert werden können.

Beim einfachen Anklicken eines Knotens im TreeView mit der linken Maustaste, wird auch in der darunterliegenden Tabelle die entsprechende Zeile mit der URL angezeigt. Mit einem Doppelklick auf einen Knoten des Baumes öffnet sich der Standardbrowser mit der entsprechenden URL.

Im TreeView können Ebenen einfach aus- und eingeklappt werden, indem man einmal auf die „Handles“ auf der linken Seite klickt, die neben Knoten, die Unterknoten haben, dargestellt werden.

Dies kann auch durch einen dreifachen Klick auf einen Knoten mit der linken Maustaste erreicht werden.

Die Tabelle unterhalb des TreeView wird erst nach dem Ende des Crawlvorgangs angezeigt. Dort erhalten Sie pro Zeile Informationen über die einzelnen URLs, die im Laufe des Crawlvorgangs gefunden wurden. Dort können folgende Informationen abgelesen werden:

- Titel
- URL
- HTTP Response Code – die Serverantwort in Form eines Codes [6]:
 - Aussagen über die Erreichbarkeit einer Seite werden gemacht. z.B. 200 = Ok, Seite gefunden; 404 = Seite nicht gefunden – klassischer Broken Link!
- State (Status) – Meldungen des Crawlers zum Status einer URL, der sich im Laufe des Crawlvorgangs normalerweise ändert:
 - *Discovered* – URL gefunden (beim Analysieren einer Ressource)
 - *Fetched* – Ressource der URL erfolgreich heruntergeladen. Das ist notwendig zum Analysieren einer Seite; die Seiten wird aber nicht gespeichert, sondern nach dem Analysieren wieder verworfen.
 - *Parsed* – Ressource erfolgreich geparkt (Inhalte analysiert)
 - *FetchError* – Ressource konnte nicht erfolgreich heruntergeladen werden -> Broken Link; dies ist der Fall, wenn der Http Response Code kleiner als 200 oder größer als 302 ist
 - *MalformedUrl* – URL Syntax falsch (Broken Link)
 - *ParseError* – Fehler beim Parsen der Ressource aufgetreten
 - *ContentLengthOversized* – die Ressource wurde nicht geparkt, und wenn der Server die

Information zur Größe vor dem Download im Response Header herausgibt, auch nicht heruntergeladen, weil sie größer als die max. Dateigröße ist, die in den Einstellungen angegeben wurde

- *MimeTypeExcluded* – die Ressource wurde nicht heruntergeladen, weil ihr Mime-Typ durch die zugehörige Benutzereinstellung ausgeschlossen wurde
- *ContentLengthOversized_MimeTypeExcluded* – Kombination der beiden einzelnen Statuse
- *Timeout* – die max. Zeit zum Versuch des Verbindungsaufbaus zu dieser URL wurde überschritten; die Ressource wurde nicht heruntergeladen, und demnach auch nicht nach weiteren Links durchsucht

- Mime Type – der Mime-Typ [5] der Ressource; „Unknown“ wenn nicht bekannt
- Anzahl gefundener Zeichen pro Seite
- Anzahl gefundener Bilder pro Seite
- File Size (Dateigröße)
- Keywords – gefundene Keywords auf dieser Seite

Welche Spalten in der Tabelle aufgelistet werden, ist abhängig von den Benutzereinstellungen. In der „Crawler Settings“ Ansicht kann unter Statistics angegeben werden, welche Informationen in der Tabelle angezeigt werden sollen.

Die nachfolgende Abbildung zeigt ein Bildschirmfoto der TreeView/Statistics Ansicht. Oben ist die hierarchische Baumstruktur zu erkennen, darunter die Tabelle mit den Einzelergebnissen. Zwischen dem TreeView und der Tabelle ist ein Trennbalken, der zu Vergrößerung/Verkleinerung der angrenzenden Ansichten mit der Maus nach oben oder nach unten gezogen werden kann.

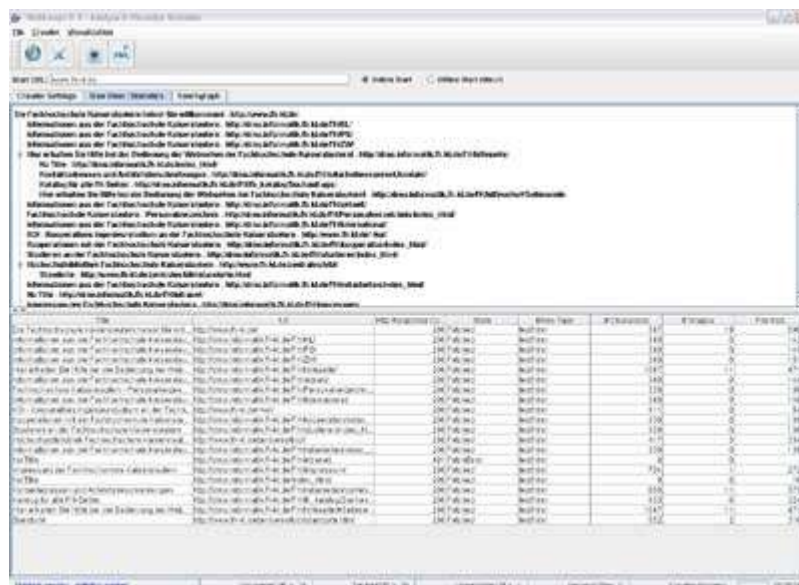


Abbildung 9 TreeView/Statistics

4.4 Touchgraph

Der TouchGraph kann nach Ende des Crawlvorgangs in der „TouchGraph“ Ansicht betrachtet werden. Nach der Auswahl des entsprechenden Tabs (neben dem TreeView/Statistics Tab), dauert es kurz, bis er dargestellt wird.

Der TouchGraph bietet eine effiziente, da raumsparende Möglichkeit, die Visualisierung der Seitenstruktur mittels Knoten und Kanten darzustellen. In den Knoten stehen die Titel der Seiten. Damit der Graph nicht zu unübersichtlich wird, wurde allerdings die Länge der angezeigten Titel auf max. 30 Zeichen begrenzt. Die Verknüpfung zu anderen Seiten wird über Kanten zwischen den Knoten dargestellt.

Zur weiteren Erläuterung eine kleine Einführung in die Terminologie und Eigenschaften unseres Graphen:

- Ein Knoten kann beliebig viele Kinder haben
- Ein Knoten hat höchstens einen Elterknoten
- jeder Knoten hat einen Elternknoten, bis auf den Wurzelknoten
- der Wurzelknoten ist der Knoten, der die Startseite repräsentiert

Diese Struktur entspricht auch der Struktur des Baumes im TreeView.

Interaktion

Zu Beginn werden alle Knoten des Graphen auf einmal dargestellt. Bei einer größeren Anzahl von Knoten wird das schnell unübersichtlich. Die Anzahl der dargestellten Seiten kann man mit einem einfachen Linksklick auf einen Knoten einschränken. Dann wird der ausgewählte Knoten in die Mitte der Ansicht verschoben, seine Kinder und der Elternknoten werden um ihn herum platziert, und alle anderen Knoten werden ausgeblendet.

Kinder, die zu weiteren Knoten in einer anderen Ebene verweisen, haben rechts oben im Kästchen des Knotens eine rote Markierung mit einer Zahl, welche die Anzahl dieser dort ausgeblendeten, direkt benachbarten Knoten anzeigt. Klickt man mit einem einfachen Linksklick auf einen solchen Knoten, wird dieser Knoten in die Mitte der Ansicht verschoben, und seine zuvor ausgeblendeten Kinder und der Vater des Knotens werden wieder um ihn herum angezeigt.

Unterhalb des Graphen sind weitere Einstellungen möglich. Mit der vorausgewählten Zoomfunktion kann der Graph über einen horizontalen Schieberegler vergrößert oder verkleinert werden. Damit ändert sich aber nicht die Größe der Knoten, sondern die Kantenlänge zwischen den Knoten. Das ist nützlich, um viele aufeinander fallende Knoten auseinander zu schieben.

Es kann auch eine Rotationsfunktion ausgewählt werden. Damit kann der Graph mit dem Schieberegler gedreht werden.

Mit der dritten dort auswählbaren Funktion kann man die Stärke des hyperbolischen Effekts mit dem Schieberegler beeinflussen. Er bewirkt, dass der Graph in der Mitte weiter aufgefächert dargestellt wird, als am Rand der Ansicht. Weiter aufgefächert bedeutet einfach, dass, wie beim Vergrößern, die Kantenlänge zwischen den Knoten erhöht wird.

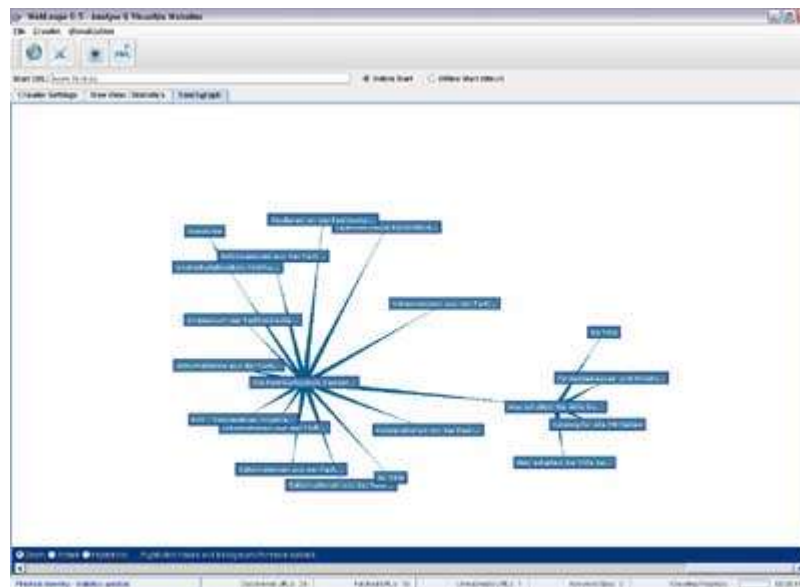


Abbildung 10 Touchgraph

Funktionen der Kontextmenüs:

Mit einem Rechtsklick auf die einzelnen Knoten öffnet sich das Kontextmenü eines Knotens. Dort sind folgende Aktionen möglich:

- *Expand Node:*
Die nicht sichtbaren Nachbarn eines Knotens werden eingeblendet, ohne dass die anderen, bereits sichtbaren Knoten, ausgeblendet werden, was bei einem Linksklick auf einen Knoten passiert.
- *Collapse Node:*
Die Kinder des Knotens werden ausgeblendet.
- *Hide Node:*
Der ausgewählte Knoten wird unwiederruflich ausgeblendet, kann also nicht mehr eingeblendet werden!
- *Center Node:*
Der ausgewählte Knoten wird in die Mitte der Ansicht verschoben.

Mit einem Rechtsklick auf den ausgewählten Hintergrund wird folgende Funktion bereitgestellt:

- *Toggle Controls:*
Die Toolbar mit den Funktionen Zoom, Rotate, Hyperbolic und dem Schieberegler wird aus- oder eingeblendet.

Bisher ist es im Gegensatz zum TreeView nicht möglich, im TouchGraph durch Doppelklick auf einen Knoten den Standardbrowser mit der zum Knoten gehörenden URL zu öffnen. Das soll aber in einer nachfolgenden Version von WebLoupe möglich sein.

Bildschirmfoto:

Bildschirmfotos (Screenshots) des aktuellen TouchGraph können als PNG-Datei lokal auf dem Rechner gespeichert werden. Dies ist aber erst nach Ende des Crawlvorgangs möglich, sobald der TouchGraph dargestellt wird.

Um das Bildschirmfoto anzufertigen, können sie entweder im Hauptmenü unter Visualization die Option Screenshot auswählen, oder das entsprechende Icon (siehe Abbildung 11) in der Toolbar betätigen, oder eine Tastenkombination auswählen, die sie dem Menü entnehmen können.



Abbildung 11

Screenshot
Icon

Nachdem sie die Funktion betätigt haben, öffnet sich ein Dialogfenster zum Speichern des Bilds. Dort können sie auswählen, wo sie das Bild speichern wollen.

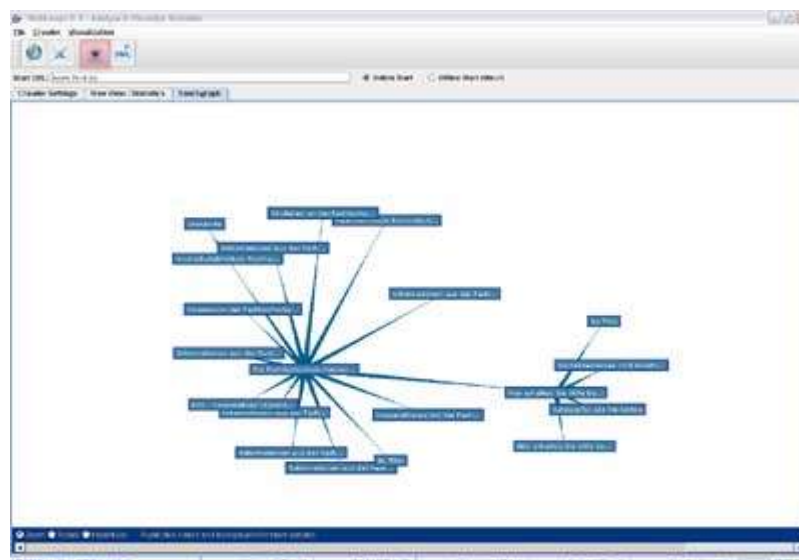


Abbildung 12 Speicherung des Graphen als png-Datei möglich

Nähere Informationen zu TouchGraph selbst und Beispiele finden sie auf dessen Projektwebseite:
<http://www.touchgraph.com/>

4.5 HyperGraph

HyperGraph ist ein Programm zur Visualisierung von Graphen, also Knoten, die über Kanten verbunden sind. Der Unterschied zu TouchGraph besteht in der Darstellung des Graphen bzw. im Visualisierungsalgorithmus. HyperGraph verwendet hyperbolische Geometrie, speziell hyperbolische Bäume, um einen Linseneffekt zu erzeugen. HyperGraph ist ebenfalls interaktiv, hier wird allerdings nach Auswahl eines Knoten der Standardbrowser mit der URL geöffnet.

Die Visualisierung mit HyperGraph ist nicht direkt in WebLoupe integriert. In WebLoupe kann aber eine GraphXML Datei des Crawlergebnisses generiert werden, welche vom HyperGraph Applet eingelesen und angezeigt werden kann. Das HyperGraph Applet ist im Umfang von WebLoupe enthalten, und zwar im Ordner „HyperGraph“ des WebLoupe Pakets.

Neben dem HyperGraph Applet und der generierten GraphXML Datei sind noch zwei weitere Dateien notwendig, um das Applet verwenden zu können: die HTML Datei, mit der das Applet im Browser angezeigt wird, und die GraphXML.dtd, die das Format der GraphXML beschreibt. Diese Dateien sind ebenfalls im Ordner „HyperGraph“ enthalten.

Voraussetzung zur Anzeige des Applets ist ein Browser mit korrekt konfiguriertem Java-Plugin. Das Java Runtime Environment, die sie möglicherweise im Umfang von WebLoupe erhalten haben (WebLoupe-0.5-exeJre.zip), wird dazu nicht verwendet.

Im Folgenden eine kurze Beschreibung zur Vorgehensweise:

1. GraphXML Datei generieren, nachdem der Crawlvorgang beendet ist.

Dazu können sie das zugehörige Icon, das hier unterhalb abgebildet ist, in der Toolbar betätigen. Alternativ können sie auch „HyperGraph Export“ aus dem Visualization Menü des Hauptmenüs auswählen, oder den entsprechenden Shortcut (unter Windows Str-E) oder den Mnemonic (unter Windows Alt-E) anwenden.



Abbildung 13

GraphXML
Export Icon

Nach Auswahl der Funktion öffnet sich ein Dialogfenster, in dem sie den Namen und den Speicherort der generierten GraphXML Datei angeben können.

2. Um das Applet starten zu können, müssen sie benötigten Dateien aus dem Ordner HyperGraph in das selbe Verzeichnis kopieren, wie die exportierte Datei:

- GraphXML.dtd (immer benötigt)
- hyperapplet.jar (immer benötigt)
- HyperGraph.html (eine Beispiel HTML Datei zum Laden des Applets im Browser)



Abbildung 14 Ordner

3. Vor dem Starten müssen sie noch die HTML Datei anpassen. Dazu müssen sie die HyperGraph.html Datei mit einem Texteditor (z.B. Windows Notepad) öffnen, und den Namen der exportierten Datei in das „value“ Feld des „file“ Parameters eintragen, das in der folgenden Abbildung hervorgehoben dargestellt ist. Dies ist nötig, um dem Applet mitzuteilen, welche Datei es laden soll.

```

HyperGraph.html - Editor
Datei Bearbeiten Format Ansicht ?
<html>
<head>
<title>HyperGraph generated with webLoupe 0.5 (http://www.webloupe.de.vu)</title>
</head>
<body>
<p>
  <applet
    code="hypergraph.applications.hexplorer.HEXplorerApplet"
    align="baseline"
    archive ="hyperapplet.jar"
    width="400"
    height="400">!-- adapt layout parameters to your needs -->
    <!-- change the file name to the name of the GraphXML file you generated with webLoupe. -->
    <param name="file" value="HyperGraph-webLoupe-FH.xml">
  </applet>
</p>
</body>
</html>

```

Abbildung 15 HyperGraph.html

- Unterstützung von robots.txt, damit der WebLoupe Crawler nur die Verzeichnisse durchsucht, die von Webmastern dafür vorgesehen sind

6 Ansprechpartner

Katja Langholz & Matthias Kahlau

Email:

webspirit@users.sourceforge.net (Programmierung)

kaimka@users.sourceforge.net (Marketing, Design)

Webseiten:

www.webloupe.de.vu

www.sourceforge.net/projects/webloupe/

7 Quellenangaben

[1] Wikipedia: *Open Source*

http://de.wikipedia.org/wiki/Open_Source

zuletzt gesehen: 18.01.2005

[2] Formelsammlung.de: *Breitensuche*

<http://www.formel-sammlung.de/ld-Breitensuche-296.html>

zuletzt gesehen: 16.02.2005

[3] Codeworx.org Tutorials: *Multithreading*

http://www.codeworx.org/cpp_tuts_1_5.php

zuletzt gesehen: 16.02.2005

[4] Ranking abc: *Crawler mit Robots.txt ansteuern*

<http://www.ranking-abc.de/crawler-steuern.html>

zuletzt gesehen: 18.01.05

[5] SelfHTML Dokumentation: *Diverse technische Ergänzungen / Mime-Typen*

<http://de.selfhtml.org/diverses/mimetypen.htm>

zuletzt gesehen am 16.02.2005

[6] Homepage des W 3 Consortium: *Http Status Codes*

<http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html>

zuletzt gesehen am 16.02.2005