

Praxis der Softwareentwicklung Wintersemester 2013/2014

TrackBack

Application for analysing third-party-content and locating its provider

Handbuch

Team:

Aris Clepe
Michael Herzog
Oliver Löffler
Marco Roehr
Fabian Stolz
Ferdinand Swoboda

Institut für Telematik
Forschungsbereich DSN
Prof. Dr. Hartenstein

Betreuer:

Dipl.-Inform. Till Neudecker, Dr.-Ing. Sebastian Labitzke

Inhaltsverzeichnis

1	Vorwort	3
2	Einleitung	3
2.1	Inhaltliche Einführung	3
2.2	Hinweise zur Anwendung des Handbuchs	4
3	Installation	4
3.1	Entpacken und Verknüpfung erstellen.	4
3.2	Starten des Programms als Administrator	4
3.3	Aktualisieren der Geo-Datenbank	4
4	Benutzeroberfläche	5
5	Anwendungsszenarien	6
5.1	Szenario 1: Der erste Crawl	6
5.2	Szenario 2: Wiederverwendung alter Daten	7
5.3	Szenario 3: Auswertung der analysierten Daten	7
6	Erklärung der Funktionen	8
6.1	Die Menüleiste	8
6.1.1	Importieren	8
6.1.2	Exportieren	8
6.1.3	Beenden	9
6.1.4	Blacklist bearbeiten	9
6.1.5	Sprache ändern	9
6.1.6	Protokoll anzeigen	9
6.1.7	Geo-Datenbank aktualisieren	9
6.1.8	0 Byte ignorieren	10
6.1.9	Hilfe	10
6.1.10	Version	10
6.2	Die Toolbar	11
6.2.1	Start/Stop	11
6.2.2	Kartentab erstellen	11
6.2.3	Graphentab erstellen	11
6.2.4	Diagrammtab erstellen	11
6.3	Anzeige für Statusinformationen	12
6.4	Der Tabbereich	12
6.4.1	Willkommenstab	13
6.4.2	Browsertab	13
6.4.3	Kartentab	14
6.4.4	Graphentab	16
6.4.5	Diagrammtab	16
6.5	Die Crawlkonfiguration	17
6.5.1	Quelle	18
6.5.2	Crawlalgorithmus	19
6.5.3	Abbruchkriterium	19

6.5.4	Filter	20
6.5.5	Fehlerhafte Eingaben	22
7	Behandlung von Problemen	23
7.1	Das Programm startet nicht korrekt oder zeigt nach dem Start nicht das gewünschte Verhalten	23
7.2	Das Protokoll öffnet eine Nachricht, dass er nicht gelöscht werden konnte	23
7.3	Das Programm hängt mitten im Crawlvorgang komplett	23
8	Glossar	24
8.1	Cluster	24
8.2	Crawl/Crawlvorgang	24
8.3	FreeGeoIP	24
8.4	Initiale URLs	24
8.5	Kante	24
8.6	Knoten	24
8.7	Marker	24
8.8	Regex	24
8.9	TPC/Third-Party-Content	25

1 Vorwort

Dieses Produkt richtet sich an Nutzer, die daran interessiert sind, welche Daten und Verknüpfungen beim Besuch einer Webseite auftreten. Es ist im Besonderen für Forschungseinrichtungen gedacht, die sich mit Web Analytics beschäftigen. Es lässt sich aber auch in Büroumgebungen oder Privathaushalten einsetzen. Das Produkt soll dem Nutzer die Möglichkeiten geben den Third-Party-Content anzuzeigen, der bei dem Besuch einer Webseite bzw. einer Liste bestimmter Webseiten auftritt. Ihm stehen dazu verschiedene Visualisierungs- und Analysemöglichkeiten zur Verfügung. Durch eine intuitiv bedienbare Oberfläche ist das Programm für jeden Benutzer bedienbar. Es werden daher keine Vorkenntnisse vorausgesetzt.

2 Einleitung

2.1 Inhaltliche Einführung

Bei der Anwendung “TrackBack – Application for analysing third-party-content and locating its provider“ handelt es sich um eine eigenständige Softwareapplikation zur automatisierten Analyse von Drittanbieter-Webinhalten und Lokalisierung der Provider dieser Drittanbieter auf einer Weltkarte.

TrackBack wurde mit dem Ziel der wissenschaftlichen Nutzung und der Nutzung durch interessierte Anwender in einer Microsoft Windows – Systemumgebung entwickelt.

Um der Masse der Webinhalte auf der ganzen Welt Herr zu werden, bedient sich TrackBack der Technik des Webcrawlers. Webcrawler sind autonome Computerprogramme, die das World Wide Web durchforsten und dabei alle Webinhalte, mit denen sie in Berührung kommen, analysieren.

Über Hyperlinks gelangt ein Webcrawler von anfänglich nur wenigen URLs zu weiteren Webdokumenten, die in einer durch spezielle Algorithmen zu bestimmenden Reihenfolge besucht werden. Durch in diesen Dokumenten neu gefundenen URLs ist es theoretisch möglich, alle verlinkten und öffentlich zugänglichen Seiten des Internets zu erreichen.

Neben verschiedenen Eingabemöglichkeiten und Ausgabedarstellungen kann der Nutzer ebenso den Crawlvorgang durch Angabe von Abbruchbedingungen, Einschränkungen und der Wahl eines Crawlmodells beeinflussen.

Eine Webseite, beziehungsweise eine Liste von Webseiten, werden anhand deren Quelltexte auf Verweise untersucht, die zu Drittanbietern führen.

Sind die Drittanbieter identifiziert, werden über deren IP- Adressen deren geographische Standorte ausgemacht. Auf einer Weltkarte wird schließlich visualisiert, wo sich jene befinden.

Ebenfalls können bei Auswertung mehrerer Seiten die Häufigkeit auftretender Anbieter anhand der Färbung der Herkunftsländer erkannt werden.

Zusammen mit den von TrackBack vorgenommenen Messungen des Datenvolumens der Third-Party-Content(TPC)-Provider werden diese Informationen in für den Nutzer anschaulichen Statistiken aufbereitet.

2.2 Hinweise zur Anwendung des Handbuchs

Das Handbuch gibt Auskunft über die Benutzung des Programms.

Es enthält ausführliche, bebilderte Anleitungen der einzelnen Funktionen und beschreibt auch erste Anwendungsszenarien.

Mit Hilfe des Handbuchs soll es dem neuen Nutzer möglich sein das Produkt vollständig bedienen zu können bzw. die Bedienung zu erlernen.

Fortgeschrittene Nutzer können das Handbuch als Referenz und Erinnerungshilfe nutzen.

3 Installation

Der Installationsvorgang benötigt folgende drei Schritte:

3.1 Entpacken und Verknüpfung erstellen.

Das Programm wird in einem Archiv mit dem Namen „TrackBack.zip“ geliefert. Zum entpacken verwenden Sie ein geeignetes Programm, z.B. 7-Zip (bei Windows 7 normalerweise vorinstalliert).

Entpacken Sie das Programm in das von ihm vorgesehene Verzeichnis. Beispiel mit 7-Zip:

- Erstellen Sie einen Ordner im Verzeichnis ihrer Wahl, z.B. C:\Users\„Benutzername“\TrackBack.
- Klicken Sie mit der rechten Maustaste auf das Archiv und wählen Sie '7-Zip -> Datei entpacken...'.
- Geben Sie den Pfad des eben erstellten Ordners an und klicken Sie auf Ok.
- Das Archiv ist nun entpackt.

Um das Programm nun auszuführen, müssen Sie die Datei „TrackBack.exe“ ausführen.

Diese können Sie in ihrem entpackten Verzeichnis unter dem relativen Pfad „TrackBack\bin\Release\TrackBack.exe“ finden.

Für einen schnelleren Zugriff, erstellen Sie über das Rechtsklickkontextmenü eine Verknüpfung von „TrackBack.exe“ und verschieben Sie diese an den Ort, an dem Sie darauf zugreifen wollen.

3.2 Starten des Programms als Administrator

Vor dem ersten normalen Start des Programms muss es einmalig als Administrator ausgeführt werden, um das Protokoll einzurichten. Das Protokoll bezieht seine Daten aus dem Windows Event Log. Dieser benötigt spezielle Berechtigungen. Nachdem das Programm einmal als Administrator gestartet wurde kann es zukünftig ohne diese Berechtigung ausgeführt werden. Bis auf das Löschen des Protokolls (siehe [Kapitel 7.2](#)) lassen sich danach alle Funktionen ohne Administratorrechte ausführen.

3.3 Aktualisieren der Geo-Datenbank

Nachdem Sie das Programm erfolgreich gestartet haben, sollten Sie die Geo-Datenbank aktualisieren. Wählen Sie dazu unter 'Optionen -> Geo-Datenbank aktualisieren'. Weitere Informationen dazu finden Sie im [Kapitel 6.1.7](#).

4 Benutzeroberfläche

Die Oberfläche ist, wie in Windows üblich, durch Fenster realisiert. Es gibt ein Hauptfenster (Abbildung 1), das den Großteil der Steuerelemente und Funktionen bereitstellt. Dieses Fenster ist in 4 Bereiche unterteilt:

- Die Menüleiste enthält Optionen und sekundäre Funktionen, wie etwa Spracheinstellung oder den Import/Export.
- Die Toolbar stellt die Hauptfunktionen Start, Stop und die Anzeigemöglichkeiten der analysierten Daten.
- Rechts neben der Toolbar befindet sich eine Anzeige für Statusinformationen.
- Der untere große Teil ist ein Behälter für Tabs. Dieser dient zur Anzeige und Analyse der gesammelten Daten.

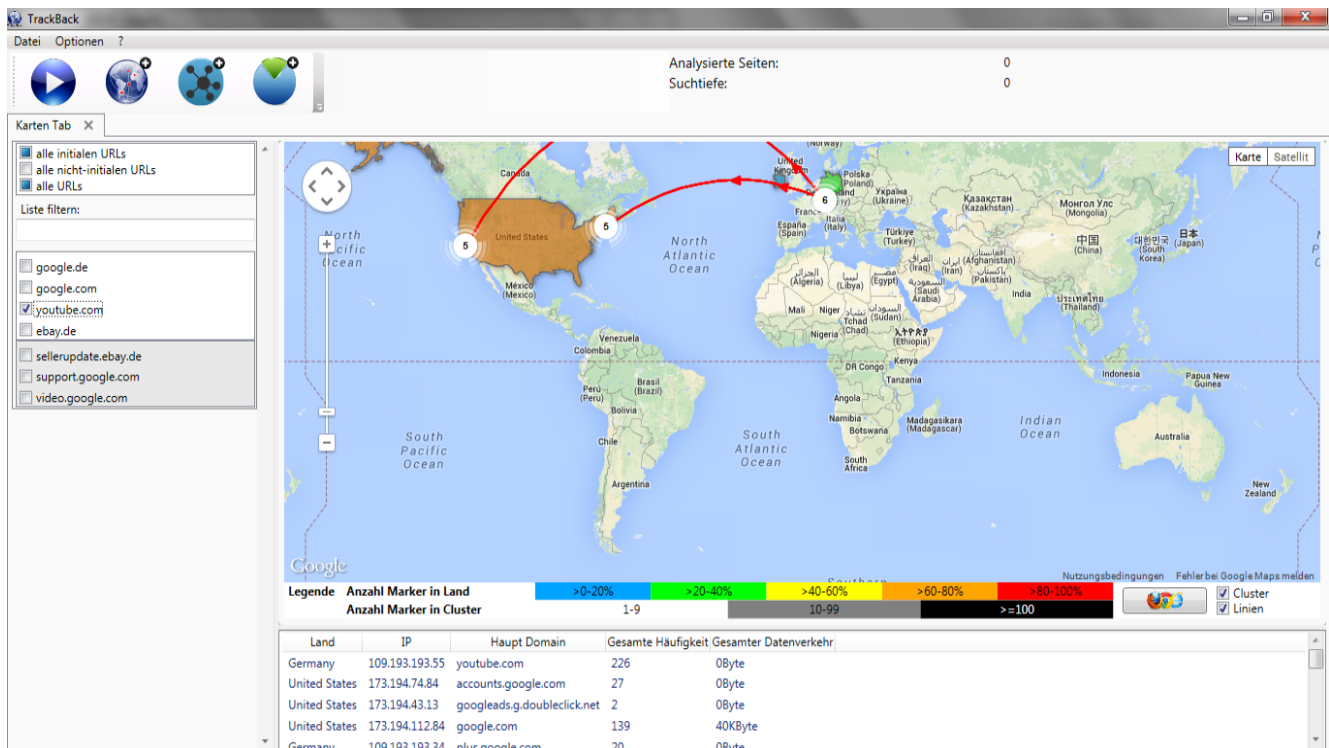


Abbildung 1: Das Hauptfenster

Der Tabbehälter ist je nach angezeigtem Inhalt noch einmal unterteilt. Die Bereiche innerhalb eines Tabs lassen sich wie in Abbildung 2 dargestellt an den Rahmen verschieben. Weitere Informationen zur Benutzeroberfläche befinden sich in Kapitel 6 - Erklärung der Funktionen.

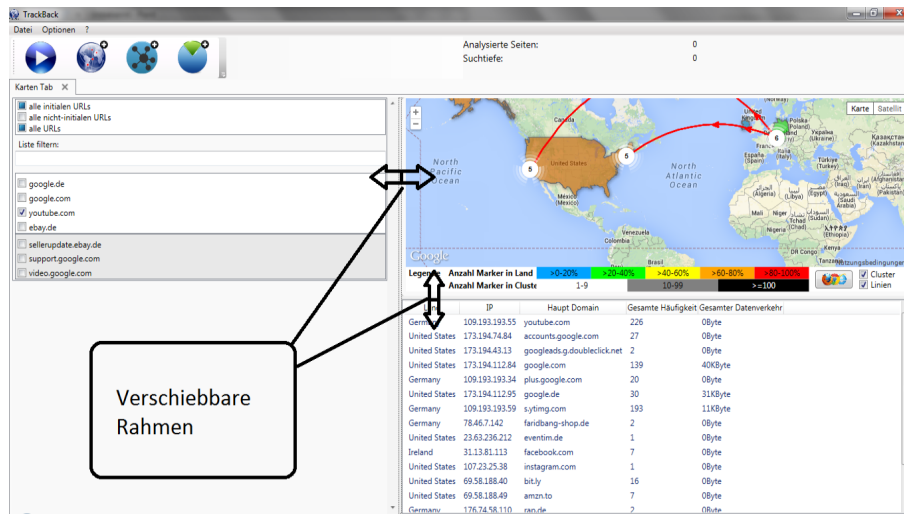


Abbildung 2: Verschiebbare Rahmen innerhalb eines Tabs

5 Anwendungsszenarien

Dieses Kapitel stellt erste Anwendungsszenarien vor und gibt ein paar Abläufe an, die einen schnellen Einstieg und sofortige Ergebnisse/Erfolgslebnisse liefern.

Hinweis: Die hier angegebenen Abläufe sind auf das Wesentliche beschränkt. Detaillierte Informationen finden Sie in [Kapitel 6 - Erklärung der Funktionen](#).

5.1 Szenario 1: Der erste Crawl

Dieses Szenario gibt den Ablauf eines neuen Crawl an.

- Starten Sie das Programm.
- Drücken Sie 'Optionen -> Geo-Datenbank aktualisieren' und warten Sie bis die Meldung, dass die Datenbank aktualisiert wurde erscheint. Dies kann einige Minuten dauern und ist nur beim allerersten Crawlvorgang zwingend nötig.
- Drücken Sie auf den Startknopf (der erste Knopf in der Toolbar) und das Crawlkonfigurationsfenster öffnet sich.
- Wählen Sie eine Quelle, aus der Sie ihre URLs beziehen wollen. (z.B: Alexa)
- Wählen Sie die zusätzlichen Bedingung der Quelle. (z.B: Seitenzahl 5, Land - DE Germany)
- Wählen Sie den Crawlalgorithmus. (z.B: RandomWalk)

- Wählen Sie ein Abbruchkriterium. (z.B: Seitenzahl 10)
- Wählen Sie ihre gewünschten Filter (z.B: Robots.txt Filter)
- Drücken Sie im Konfigurationsfenster auf 'Ok'.
- Warten Sie bis der Crawl abgeschlossen ist. Sie können den Fortschritt in der Informationsanzeige beobachten.
- Nun können Sie ihre Ergebnisse betrachten (weiter Szenario 3) oder diese über 'Datei -> Exportieren' sichern und beliebig importieren (Szenario 2).

5.2 Szenario 2: Wiederverwendung alter Daten

In dieses Szenario werden alte gesicherte Daten importiert, um sie zu betrachten.

- Starten Sie das Programm.
- Drücken Sie in der Menüleiste 'Datei -> Importieren'.
- Suchen Sie im angezeigten Dialog die Datei, in der Sie ihre Daten gesichert haben und klicken Sie auf 'Öffnen'.
- Nun können Sie ihre Ergebnisse betrachten (weiter Szenario 3).

5.3 Szenario 3: Auswertung der analysierten Daten

Dieses Nutzungsszenario beschreibt die Betrachtung und Auswertung der analysierten Daten. Aufgrund der vielfältigen Analyseoptionen und unterschiedlicher Auswertungsziele, kann dieses Szenario nur bedingt die gewünschten Ergebnisse liefern und lediglich als Beispiel dienen.

- Führen Sie einen neuen Crawl durch oder Importieren Sie alte Daten.
- Drücken Sie auf einen der Tabknöpfe in der Toolbar. (z.B: Kartentab - 2 Knopf von Links)
- Machen Sie in der linken Liste einen Haken bei allen URLs, die Sie betrachten wollen. (z.B: alle URLs)
- Nun werden die gewünschten Ergebnisse grafisch (z.B auf der Karte) angezeigt.
- Je nach gewähltem Tab können Sie hier nun mit der grafischen Darstellung arbeiten.(z.B: auf eine Markierung auf der Karte klicken um Informationen darüber zu bekommen)

6 Erklärung der Funktionen

6.1 Die Menüleiste

Die Menüleiste enthält Optionen und sekundäre Funktionen (Abbildung 3). Sie befindet sich wie üblich unter Windows am oberen Rand des Programmfensters. Nachfolgend sind alle Funktionen in diesem Menü aufgelistet und erklärt.

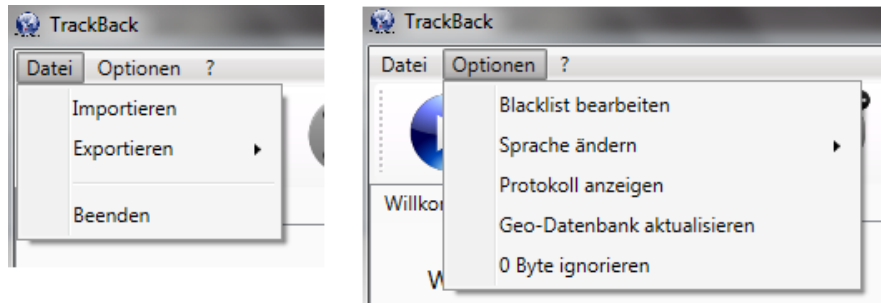


Abbildung 3: Die Menüleiste und ihre Funktionen

6.1.1 Importieren

Die Importortfunktion finden Sie unter 'Datei -> Importieren'. Mit ihr können Sie Daten aus vorherigen Crawlvorgängen, die Sie vorher über die Exportfunktion gesichert haben, laden und sie (erneut) auswerten. Drücken Sie 'Importieren', dann öffnet sich ein Dialogfenster, in dem Sie den Speicherort der gesicherten Ergebnisse angeben können.

! Hinweis: Das Programm speichert die Daten in einer txt-Datei. Das manuelle Ändern der gesicherten Datei oder das Laden einer beliebigen txt-Datei wird nicht die erwarteten Ergebnisse erzeugen. Fehlerhafte oder falsche Dateien können zu Funktionsstörungen führen.

6.1.2 Exportieren

Die Exportfunktion finden Sie unter 'Datei -> Exportieren'. Mit ihr können Sie ihre Ergebnisse sichern, um sie später mit Importieren erneut zu verwenden. Außerdem lässt sich die Statistik eines Tabs exportieren. Drücken Sie 'Exportieren', dann öffnet sich ein Dialogfenster, in dem Sie den Speicherort angeben können, an dem ihre Daten in einer txt-Datei gespeichert werden.

Exportieren besitzt zwei Auswahlmöglichkeiten, durch die angegeben werden, was man exakt exportieren möchte.

Datenstruktur: Hier werden die gesamten gesammelten Daten exportiert, um sie erneut laden zu können.

Statistik: Hier wird die Statistik (Tabelle im unteren Teil des Tabs) des momentan aktiven/sichtbaren Tabs exportiert. Damit lassen sich verschiedene Auswahlen der URLs (Liste auf der linken Seite des Tabs) speichern und außerhalb des Programms vergleichen/betrachten.

6.1.3 Beenden

Diese Funktion beendet das Programm. Sie ist äquivalent zu dem in Windows üblichen Kreuz in der oberen rechten Ecke des Fenster.

6.1.4 Blacklist bearbeiten

Durch das Drücken von 'Optionen -> Blacklist bearbeiten' wird ein Texteditor geöffnet. In diesen ist die Datei mit dem Namen custom geöffnet. Hier können Sie nun URLs eingeben, die Sie auf ihre persönliche Sperrliste haben möchten. Aktivieren Sie nun in der Crawlkonfiguration die Option 'Blacklist Filter' können Sie die Liste 'Custom' angeben, die ihre ungewünschten URLs enthält. Weitere Informationen dazu finden Sie im [Kapitel 6.5.4](#)

6.1.5 Sprache ändern

Mit diesem Menüpunkt lässt sich die Sprache des Programms ändern. Wählen Sie im zugehörigen Untermenü ihre gewünschte Sprache aus und die Anzeige wird sich anpassen.

Hinweis: Bereits geöffnete Tabs passen weder ihren Namen, noch ihren Inhalt an die Sprache an. Für einen Tab in der neuen Sprache muss ein neuer Tab geöffnet werden.

6.1.6 Protokoll anzeigen

Hiermit wird ein Protokollfenster geöffnet. In diesem werden unter anderem auftretende Fehler protokolliert. Wird zum Beispiel eine Seite beim Crawlen übersprungen, weil sie nicht geladen werden kann, dann wird das im Protokoll festgehalten.

6.1.7 Geo-Datenbank aktualisieren

Zur Lokalisierung der Provider verwendet das Programm eine Datenbank von FreeGeoIP. Da der Onlinezugriff begrenzt ist wird hier die Möglichkeit des Anbieters genutzt, die gesamte Datenbank lokal zu betreiben. Die Datenbank wird etwa einmal im Monat aktualisiert. Um immer die aktuellen Daten zu nutzen, verwenden Sie 'Optionen -> Geo-Datenbank' aktualisieren. Dieser Vorgang kann einige Minuten dauern.

! Hinweis: Bevor Sie den ersten Crawlvorgang starten, müssen Sie diese Funktion nutzen. Andernfalls wird es zu einem Fehler führen, da das Programm die Provider nicht lokalisieren kann.

6.1.8 0 Byte ignorieren

Diese Funktion sorgt dafür, dass Einträge mit 0 Byte innerhalb der Daten verschwinden. Standardmäßig werden beim Crawlen alle Links erfasst, egal ob von ihnen Inhalt geladen wird oder nicht. (Abbildung 4) Durch 'Optionen -> 0 Byte ignorieren' werden nur noch die Links angezeigt, die auch wirklich Inhalt laden oder in der ursprünglichen Liste der zu crawlenden URLs enthalten sind. (Abbildung 5)

Land	IP	Haupt Domain	Gesamte Häufigkeit	Gesamter Datenverkehr
Germany	109.193.193.55	youtube.com	226	0Byte
United States	173.194.74.84	accounts.google.com	27	0Byte
United States	173.194.43.13	googleads.g.doubleclick.net	2	0Byte
United States	173.194.112.84	google.com	139	40KByte
Germany	109.193.193.34	plus.google.com	20	0Byte
United States	173.194.112.95	google.de	30	31KByte
Germany	109.193.193.59	s.ytimg.com	193	11KByte
Germany	78.46.7.142	faridbang-shop.de	2	0Byte
United States	23.63.236.212	eventim.de	1	0Byte
Ireland	31.13.81.113	facebook.com	7	0Byte
United States	107.23.25.38	instagram.com	1	0Byte

Abbildung 4: Eine Statistik mit 0 Byte Einträgen

Land	IP	Haupt Domain	Gesamte Häufigkeit	Gesamter Datenverkehr
Germany	109.193.193.55	youtube.com	226	0Byte
United States	173.194.112.84	google.com	139	40KByte
United States	173.194.112.95	google.de	30	31KByte
Germany	109.193.193.59	s.ytimg.com	193	11KByte

Abbildung 5: Eine Statistik ohne 0 Byte Einträge

6.1.9 Hilfe

Mit '? -> Hilfe' rufen Sie dieses Handbuch auf. Dadurch können Sie auch aus dem Programm heraus Informationen über das Programm und seine Funktionen erhalten oder diese Nachschlagen.

6.1.10 Version

Diese Funktion ruft ein Fenster auf, das unter anderem Informationen über die Version und die Autoren des Programms enthält.

6.2 Die Toolbar

Die Toolbar (Abbildung 6) enthält die wichtigsten Funktionen des Programms. Über sie wird ein Crawlvorgang gestartet und die verschiedenen Tabs zur Auswertung erzeugt.



Abbildung 6: Die Toolbar

6.2.1 Start/Stop

Der Startknopf (Abbildung 6 - (1)) öffnet das Crawlkonfigurationsfenster, mit dem ein neuer Crawlvorgang gestartet werden kann. Ist dieser Vorgang am laufen, dann ändert sich das Startsymbol in ein Stoppsymbol, mit dem Sie den aktuellen Crawlvorgang abbrechen können. Dies ist vor allem dann nötig, wenn Sie keine Abbruchbedingungen (siehe [Kapitel 6.5](#)) angegeben haben, damit Sie die gesammelten Daten auch auswerten zu können.

6.2.2 Kartentab erstellen

Der Knopf für einen Kartentab (Abbildung 6 - (2)) erstellt einen neuen Tab, der eine Karte zur Betrachtung der Daten verwendet. Weitere Informationen in [Kapitel 6.4.3](#).

6.2.3 Graphentab erstellen

Der Knopf für einen Graphentab (Abbildung 6 - (3)) erstellt einen neuen Tab, der einen Graphen zur Betrachtung der Daten verwendet. Weitere Informationen in [Kapitel 6.4.4](#).

6.2.4 Diagrammtab erstellen

Der Knopf für einen Diagrammtab (Abbildung 6 - (4)) erstellt einen neuen Tab, der Diagramme zur Betrachtung der Daten verwendet. Weitere Informationen in [Kapitel 6.4.5](#).

6.3 Anzeige für Statusinformationen

Diese Anzeige gibt Statusinformationen während, aber auch vor des Crawlvorgangs. Sie befindet sich rechts neben der Toolbar im oberen Bereich des Fensters.


Zum Beispiel wird hier angezeigt, was gerade passiert, nachdem man auf 'Optionen ->Geo-Datenbank aktualisieren' gedrückt hat. (Abbildung 7)



Analysierte Seiten:
Suchtiefe:
Lade herunter: <http://dev.maxmind.com/static/csv/codes/maxmind/region.csv>

Abbildung 7: Statusinformationen über die Aktualisierung Geo-Datenbank

Während eines Crawlvorgangs bekommt man je nach Einstellung Informationen über den aktuellen Status des Crawls.



Analysierte Seiten: 5 / 15 (33%)
Suchtiefe: 2

Abbildung 8: Crawlvorgang: Randomwalk mit Abbruch nach 15 gecrawlten URLs

6.4 Der Tabbereich

Dieser Bereich enthält die Hauptanzeige als Tabs. Es gibt zwei Sorten von Tabs:

- Interaktive Tabs (Karte, Graph, Diagramm) sind zur Auswertung der gesammelten Daten.
- Nichtinteraktive Tabs (Willkommen, Browser) dienen zur Anzeige bestimmter Informationen

Einem interaktiven Tab kann man mit einem Rechtsklick einen neuen Namen geben. So lässt er sich von anderen seiner Art unterscheiden und kann seinen angezeigten Inhalt beschreiben.

Ein intaktiver Tab ist in drei Teile aufgeteilt (siehe Abbildung 9) :

- Eine Liste mit URLs, die zur Auswahl der zu betrachtenden URLs dient.
- Eine Tabelle, die Statistik der ausgewählten URLs anzeigt.
- Ein Anzeigebereich für eine Karte, einen Graphen oder Diagramme.

In der Liste der URLs kann man sich die herausuchen und markieren, die man im Anzeigebereich und der Statistik betrachten möchte. Diese Markierungen sind für jeden Tab eigenständig. So lassen sich z.B. mehrere Kartentabs mit verschiedenen Markierungen öffnen. Zwischen diesen Tabs könnte man nun Vergleiche anstellen.

Die Liste enthält, neben der Möglichkeit in den URLs zu suchen, eine Aufteilung in initiale URLs (weiß unterlegt) und URLs, die während des Crawlvorgangs dazu gekommen sind(grau unterlegt). Aus diesen Bereichen lassen sich auch jeweils alle URLs durch 'alle initialen URLs'/ 'alle nicht initialen URLs' gleichzeitig markieren. Auch beiden Gruppen also alle URLs lassen sich sofort markieren.

Desweiteren lässt sich die Statistik unten mit der spezifischen Markierung über 'Datei -> Exportieren -> Statistik' exportieren, wenn der jeweilige Tab aktiviert/sichtbar ist.

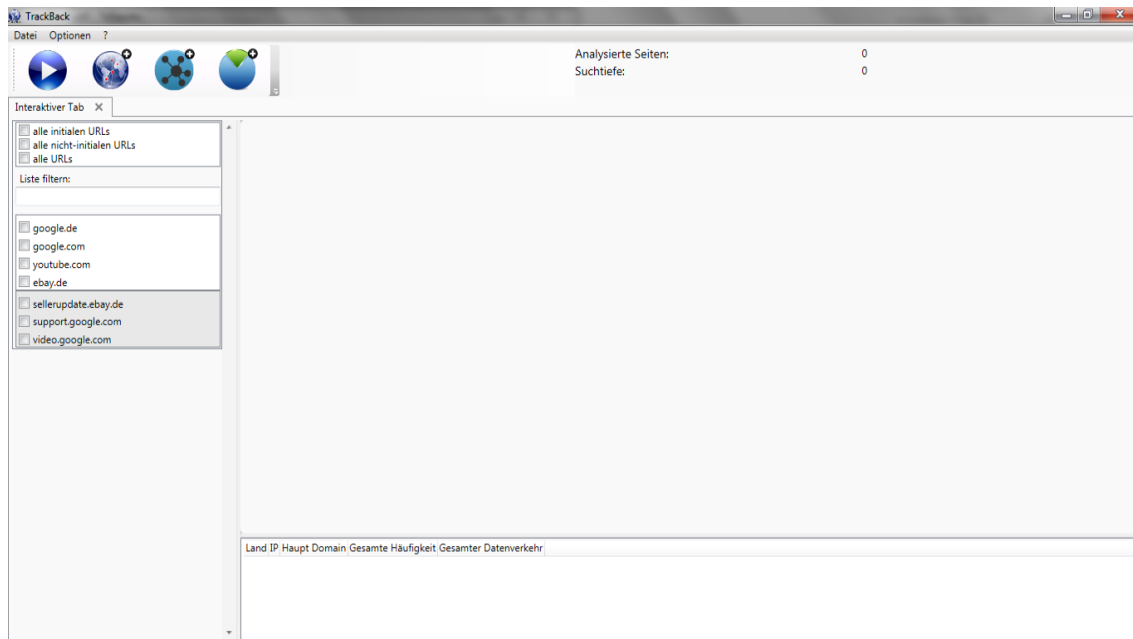


Abbildung 9: Die Bereiche eines interaktiven Tabs

Hinweis: Haben zwei URLs die selbe IP-Adresse als Provider werden sie in der Liste verschmolzen und als eine URL angezeigt. Welche URLs verschmolzen wurden, kann man sich in einen Kartentab genauer ansehen, wenn man auf die Standorte klickt.

6.4.1 Willkommenstab

Dieser Tab wird beim Start des Programms angezeigt. Er enthält einen kurzen Text, der als Starthilfe oder Schnelleinstieg dienen soll. Beim Start eines Crawlvorgangs wird dieser Tab automatisch geschlossen.

6.4.2 Browsertab

Dieser Tab wird während eines Crawlvorgangs angezeigt. Er enthält ein nicht interaktives Browserfenster, das die aktuelle Aktion des Crawlers visualisieren soll. Nach Beendigung des Crawlvorgangs wird dieser Tab automatisch geschlossen.

6.4.3 Kartentab

Im Kartentab werden die analysierten Information auf einer Weltkarte dargestellt. Einzelne URLs bzw. die verschmolzenen URLs mit gleicher IP-Adresse werden hier als ein Knoten(Marker) dargestellt. Befindet sich auf einer URL ein Link zu einer anderen, dann wird dies mit einer gerichteten Kante zwischen den Knoten gekennzeichnet. Diese Kanten lassen sich bei Bedarf mit 'Linien' ein-/ausblenden.

Liegen Knoten geographisch zu nah beieinander, z.B. weil der Zoomfaktor zu klein ist, dann werden sie in einem Cluster zusammengefasst. Die Cluster können ebenfalls bei Bedarf über 'Cluster' aktiviert/deaktiviert werden.

Die Länder werden in dem Prozentanteil der in ihnen vorkommenden Knoten markiert. Die Farben der Länder und der Cluster lassen sich in der Legende ablesen. Abbildung 10 zeigt die beschriebenen Eigenschaften eines Kartentab.

Der Knopf mit den drei Symbolen zeigt den beschriebenen Inhalt in ihrem Standartbrowser an. Bei großen Datenmengen empfiehlt es sich die Karte in diesem externen Browser zu betrachten, da der interne gewissen Einschränkungen unterliegt und teilweise starke Leistungseinbrüche zeigt.

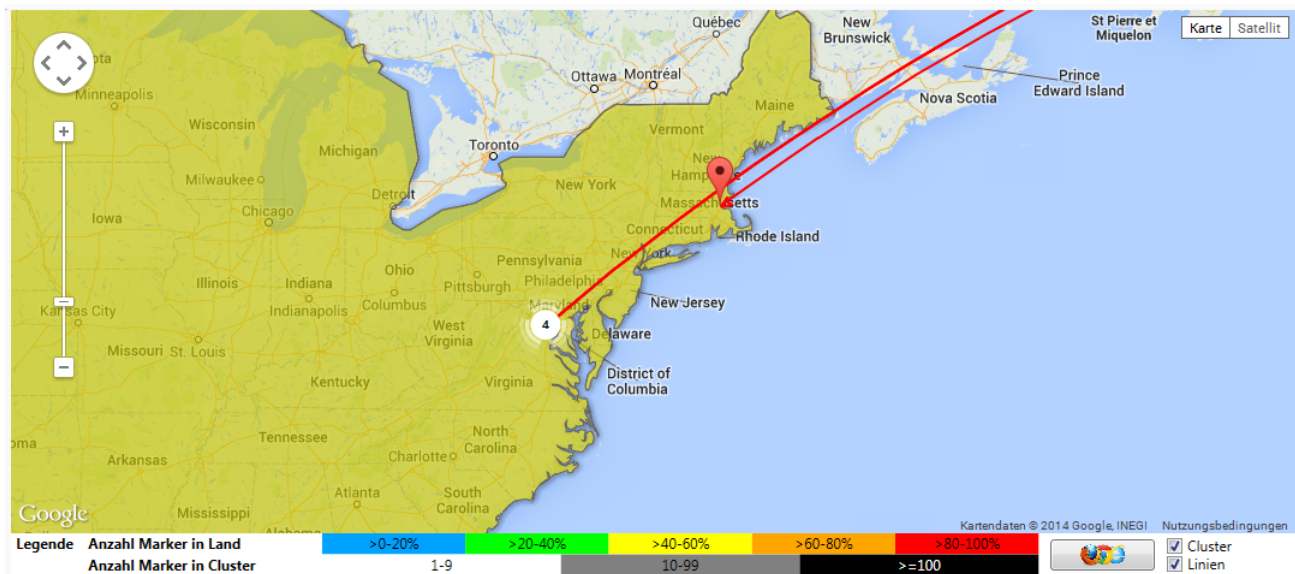


Abbildung 10: Der Inhalt eines Kartentabs

Vergrößert man nun die Karte oder klickt direkt auf einen Cluster, dann wird dieses aufgelöst und es werden ein oder mehr Marker angezeigt. Klickt man nun auf einen dieser Marker, gibt es 2 Möglichkeiten:

- Der Marker besteht aus mehreren Knoten, die den gleichen Standort besitzen. In diesem Fall wird der Marker spiralförmig (Siehe Abbildung 11) in mehrere aufgeteilt. Klicken Sie hier auf einen Marker passiert das gleiche wie in Möglichkeit 2.
- Der Marker ist ein einzelner Knoten und es öffnet sich ein Fenster, das Informationen anzeigt (Abbildung 12). Diese Informationen beinhalten z.B. die IP-Adresse, geographische Daten aber auch unter 'Urls' eine Liste mit allen URLs, die in diesem Knoten miteinander verschmolzen sind.

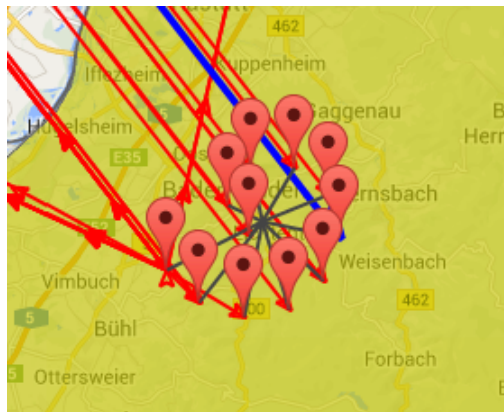


Abbildung 11: Verteilung der Knoten bei gleichem Standort

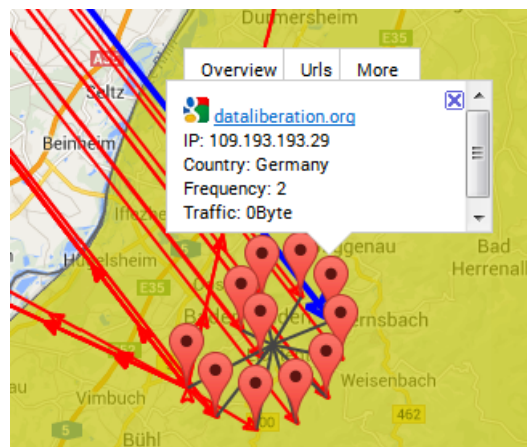


Abbildung 12: Informationen eines Knotens

6.4.4 Graphentab

In einem Graphentab werden die analysierten Daten als Graph dargestellt. Wie im Kartentab gibt es hier Knoten und die gerichteten Verknüpfungen zwischen ihnen. Die Knoten lassen sie beliebig anordnen, indem man sie mit gedrückter linker Maustaste verschiebt. Fährt man über einen Knoten drüber, werden alle ausgehenden Kanten und die dazugehörigen Nachbarknoten blau markiert. Alle eingehenden Kanten werden rot markiert. (Abbildung 13)

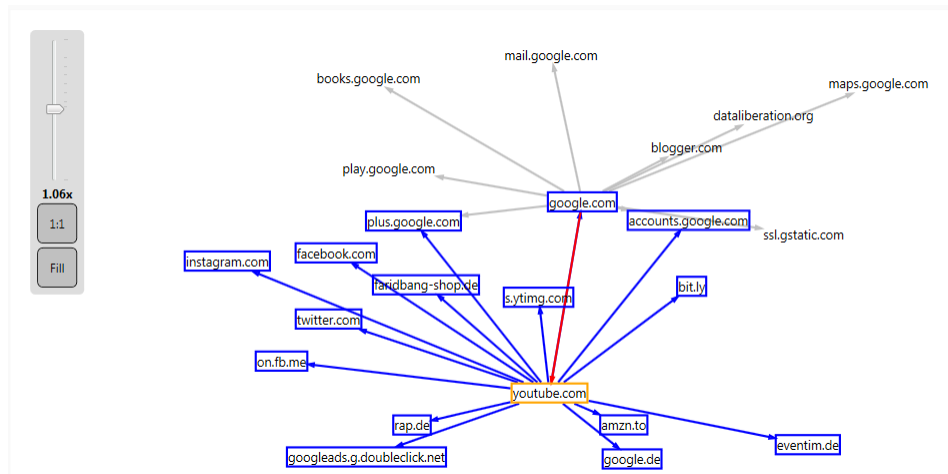


Abbildung 13: Beispiel für einen Graphen

6.4.5 Diagrammtab

Im Diagrammtab können Sie sich die analysierten Daten in verschiedenen Diagrammtypen anzeigen lassen. Sie können alle Typen sowohl mit der Häufigkeit als auch mit dem Datenvolumen betrachten. Zusätzlich lassen sich relative Werte anzeigen.

Über 'Weitere Diagramme zeigen' lassen sich bis zu 4 Diagramme nebeneinander anzeigen, wobei die Aufteilung sich mittels der Trennstriche skalieren lässt. Dies empfiehlt sich für einen direkten Vergleich bei gleich ausgewählten Daten. Für den Vergleich zweier unterschiedlicher Auswahlen der linken Liste müssen sie einen zusätzlichen Diagrammtab öffnen.

6.5 Die Crawlkonfiguration

Die Crawlkonfiguration ist das Menü, in dem alle Einstellungen für einen Crawlvorgang gewählt werden. Es erscheint nachdem man im Hauptfenster auf den Startknopf gedrückt hat.

Sie ist in 4 Bereiche aufgeteilt, die sich nach Bedarf ein-/ausblenden lassen. (Abbildung 14)

- Mit der Quelle werden die initialen URLs gewählt, die (zuerst) gecrawlt werden sollen.
- Der Crawlalgorithmus legt fest in welcher Reihenfolge die URLs nach der initialen Liste gewählt werden.
- Abbruchbedingungen können gewählt werden, um den Crawlvorgang nach einem bestimmten Kriterium automatisch zu unterbrechen.
- Filter können die URLs, die betrachtet werden sollen einschränken.

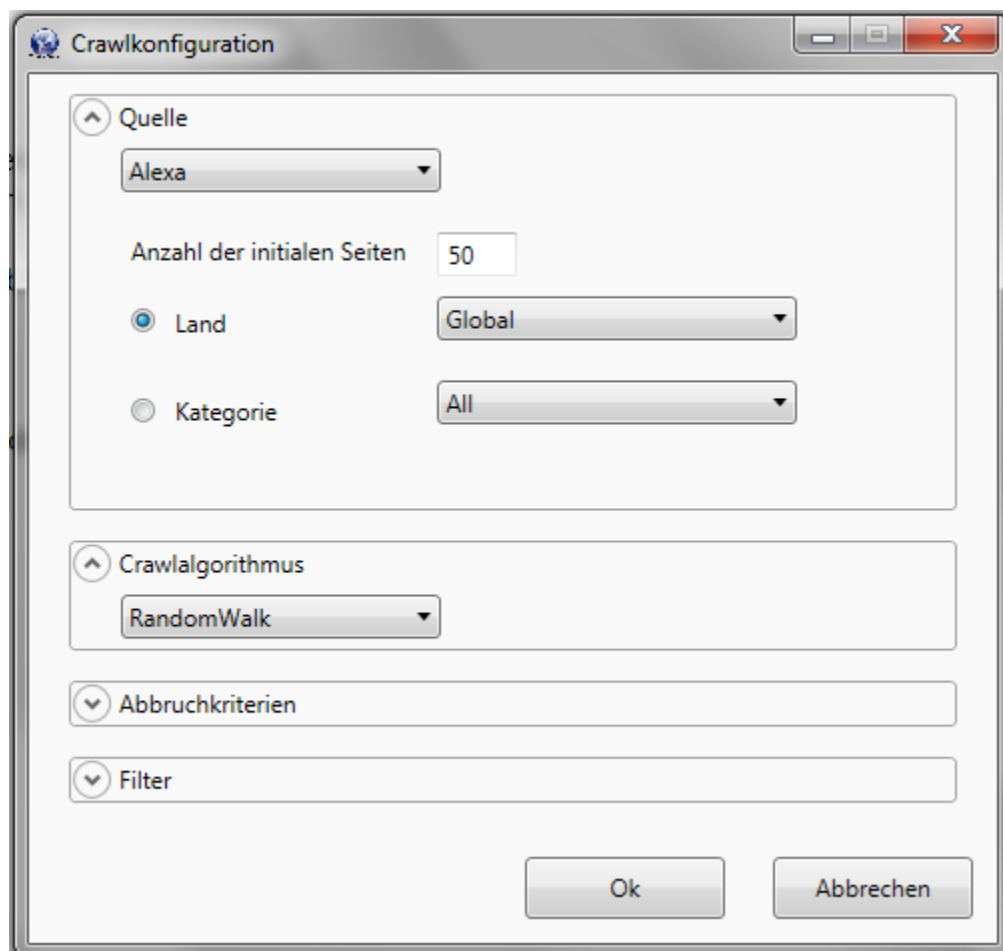


Abbildung 14: Die Crawlkonfiguration nach dem Öffnen

6.5.1 Quelle

Mit der Quelle werden die initialen URLs gewählt, die (zuerst) gecrawlt werden sollen. Hier haben Sie drei verschiedene Möglichkeiten URLs auszuwählen:

Alexa

Alexa.com ist eine Webseite, die die meistbesuchtesten Internetseiten auflistet. Es gibt verschiedene Listen, die entweder nach Land oder nach Kategorie gewählt werden können. Diese Listen können Sie hier auswählen und gleichzeitig einstellen, wie viele (max. 500) der URLs der Liste Sie crawlen wollen. (Abbildung 15)

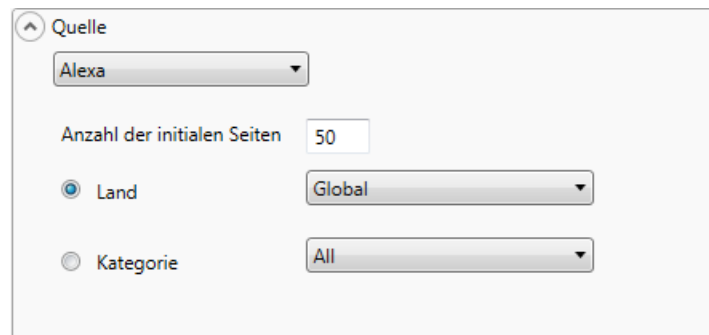


Abbildung 15: Auswahloptionen für die Quelle Alexa.com

URLs

Mit 'URLs' können Sie ihre eigenen URLs eintragen. Dazu haben Sie hier ein Textfeld zur Verfügung, in dem Sie ihre URLs eintragen können. (Abbildung 16)

Der Knopf 'URL Vervollständigen' setzt vor jede Zeile ein `http://`, um die notwendige Form aller angegebenen Adressen zu gewährleisten.

! Wichtig: Pro Zeile darf maximal eine URL stehen.

! Wichtig: Die eingetragenen URLs müssen einem bestimmtem Format entsprechen. Die eingegebene URL sollte mit `http://`, `https://` oder `ftp://` beginnen. Sollte das nicht der Fall sein, wird die Eingabe als falsch gewertet und Sie können nicht fortfahren. Drücken Sie 'URL Vervollständigen' um diese Bedingung zu erfüllen. Ist die Eingabe immer noch falsch(siehe [6.5.5 Fehlerhafte Eingaben](#)), stimmt in mindestens einer Zeile das Format nicht mit dem Soll überein. Womöglich steht ein Leerzeichen innerhalb einer Zeile. Für weitere Informationen erkundigen Sie sich über Uniform Resource Identifier(URI) mit dessen Hilfe das Programm die URLs auf Korrektheit überprüft.

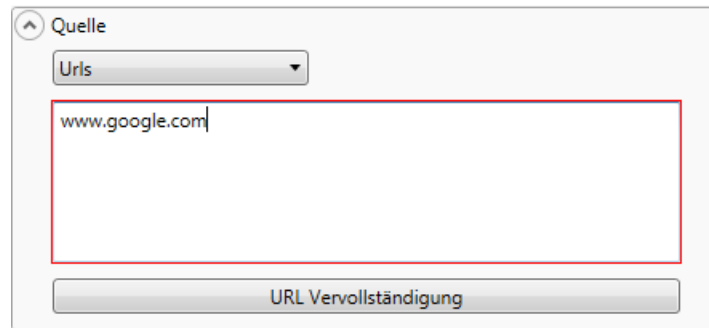


Abbildung 16: Auswahloptionen für die Quelle URLs

Browserhistory

Neben Alexa und eigenen URLs stehen Ihnen noch drei weitere Quellen zur Verfügung. Bei diesen handelt es sich um die Browserhistory des Internet Explorers, von Chrom und von Firefox. Wählen Sie einen aus und geben Sie die Anzahl der URLs an, die Sie aus der History Ihres Browsers crawlen wollen.

Hinweis: Damit Sie die History Ihres Chrombrowsers nutzen können, müssen Sie zuvor alle offenen Instanzen von Chrom schließen. Diese Meldung erscheint auch bei der Auswahl von Chrom. Sie können diese getrost ignorieren, falls Sie Chrom bereits geschlossen haben.

6.5.2 Crawlalgorithmus

Der Crawlalgorithmus legt fest in welcher Reihenfolge die URLs nach der initialen Liste gewählt werden. Es stehen drei Algorithmen zur Verfügung:

- Randomwalk wählt einen zufälligen Link aus der gesamten Liste aus.
- BreadthFirstSearch(Breitensuche) wählt zuerst alle Links, die in der initialen Liste neu gefunden wurden(Suchtiefe 1). Danach wählt er alle Links aus Suchtiefe 2 usw.
- DepthFirstSearch(Tiefensuche) wählt immer direkt die neu hinzugekommenen Links. Gibt es keine neuen Links geht er eine Ebene(entspricht Suchtiefe) nach oben.

6.5.3 Abbruchkriterium

Abbruchbedingungen können gewählt werden, um den Crawlvorgang nach einem bestimmten Kriterium automatisch zu unterbrechen. Werden keine Bedingungen gewählt, läuft der Crawlvorgang solange, bis man ihn manuell mit dem Stopknopf unterbricht. Hier können Sie wählen zwischen:

- Maximale Seitenanzahl: Der Crawlvorgang bricht ab, nachdem die angegebene Anzahl an gecrawlten Seiten erreicht wurde.
- Suchtiefe: Der Crawlvorgang bricht ab, wenn die angegebene Zahl erreicht wurde (Breitensuche) oder geht keine Ebene tiefer als die angegebene Zahl (Tiefensuche). Beachten Sie, dass Randomwalk diese Bedingung nicht hat, da hier aus allen Ebenen per Zufall gewählt wird.

6.5.4 Filter

Filter können die initialen und die gecrawlten URLs, die betrachtet werden sollen einschränken. Sie können so viele Filter wählen wie Sie wollen, indem Sie einen Haken bei dem entsprechenden Filter setzen.

Hinweis: Das Verwenden von Filtern kann die Geschwindigkeit des Crawlvorgangs verringern. Insbesondere der Robots.txt Filter und der Blacklist Filter (mit langen Listen wie Top500Adult) können sich hier bemerkbar machen.

Es stehen die folgenden vier Filter zur Verfügung:

Regex Filter

Mit diesem Filter können Sie die zu crawlenden URLs auf einen bestimmten Typ eingrenzen. Dazu müssen Sie angeben, welche Form die noch zu crawlenden URLs haben sollen. Das tun Sie, indem Sie einen regulären Ausdruck (genauer gesagt ein Regex in C#) in das dafür vorgesehene Textfeld eintragen. In der Abbildung 17 sehen Sie den Ausdruck `.*google\.com.*` in das Textfeld. Das bedeutet nun, dass nur die URLs in Betracht gezogen werden, die das Teilwort 'google.com' enthalten.

Bei der Bildung dieser regulären Ausdrücke sind einige Dinge zu beachten:

- Der Punkt '.' steht hier für ein beliebiges Zeichen.
- Der Stern '*' wiederholt ein Zeichen/eine Zeichenfolge beliebig oft.
- '.' steht in diesem Zusammenhang also für eine beliebige Zeichenfolge.
- Um den Punkt innerhalb einer Zeichenfolge zu benutzen, verwenden Sie '\.'
- Daraus lassen sich nun simple URLs Muster bilden, die sie als regulären Ausdruck nutzen können.
- Für alle URLs mit deutscher Domain müssten sie nun `.*\.de.*` eingeben.

Hinweis: Die Anwendung des Regex Filters verzeiht keine Fehler. Ein Ausdruck, der der Form eines Regex entspricht aber nicht genaustens angegeben wird, führt unweigerlich zu einem falschen Ergebnis. Bspw. wird der reguläre Ausdruck `www.google.com` dazu führen, dass der Crawlvorgang direkt abbricht, wenn in der Liste der URLs nur `http://www.google.com` ist. Ob der angegebene Ausdruck zu diesem falschen Ergebnis führt oder nicht hängt stark von der gegebenen Liste ab. Sollten Sie sich also nicht sicher sein, ob ihr Ausdruck ein fehlerfreies Ergebnis liefert, dann verzichten Sie besser auf den Regex Filter. Für weitere Informationen zu dem Thema sollten Sie sich mit den Stichworten `Regex in C#` beschäftigen.

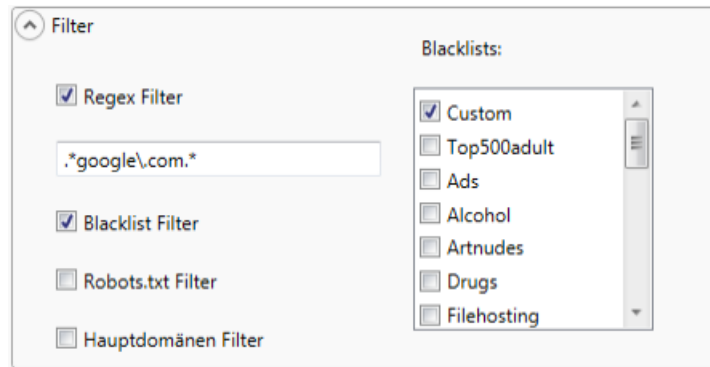


Abbildung 17: Auswahl der Filter

Blacklist Filter

Beim Blacklist Filter können Sie mehrere Blacklists auswählen, die dann mit ihrer Liste der URLs abgeglichen werden und Übereinstimmungen im Crawlvorgang ausnehmen. Hierbei stehen Ihnen einige vordefinierte Listen zur Verfügung (z.B. Porn oder Drogen). Auch die in [Kapitel 6.1.4](#) angesprochene eigene Blacklist kann hier ausgewählt werden. Abbildung 17 zeigt die eigene Blacklist in der Auswahl.

Hinweis: Sämtliche Filter sind nicht standardmäßig aktiviert. Insbesondere sollten Sie Filter wie Porn oder Top500Adult überprüfen, wenn Sie pornografische Inhalte meiden möchten.

Robot.txt Filter

Nach der Übereinkunft des Robots-Exclusion-Standard-Protokolls liest ein Webcrawler (Robot) beim Auffinden einer Webseite zuerst die Datei robots.txt im Stammverzeichnis einer Domain. In dieser Datei kann festgelegt werden, ob und wie die Webseite von einem Webcrawler besucht werden darf. Website-Betreiber haben so die Möglichkeit, ausgesuchte Bereiche ihrer Webpräsenz für (bestimmte) Suchmaschinen zu sperren.

Der Robots.txt Filter kümmert sich genau um diese Übereinkunft und entfernt alle Webseiten aus der zu crawlenden Liste der URLs (auch neu Hinzugekommene), die in der robot.txt festgelegt haben, dass sie nicht gecrawlt werden möchten.

Außerdem merkt das Programm sich, welche der Seiten er nicht crawlen darf, um so die sehr leistungshungrige und regelmäßige Überprüfung der robots.txt so kurz wie möglich zu machen.

Hinweis: Es wird empfohlen den Robot.txt Filter aus rechtlichen Gründen bei jedem Crawlvorgang zu verwenden.

Hauptdomänen Filter Da viele Webseiten Links auf ihre eigene Seite haben, kann es passieren, dass eine Crawlvorgang sich sehr lange nur innerhalb einer Hauptdomäne und ihren Unterdomänen bewegt. Hier ist es praktisch den Crawlvorgang auf die Hauptdomäne zu beschränken. Dazu existiert der Hauptdomänen Filter. Er kürzt neue URLs auf ihre Hauptdomäne und fügt sie der Liste nur hinzu, wenn sie noch nicht schon existieren.

6.5.5 Fehlerhafte Eingaben

Die Crawlkonfiguration besitzt einige Textfelder, die das Überprüfen ihrer Eingaben erfordern. Dabei werden falsche oder nicht dem richtigen Format entsprechende Eingaben direkt durch einen roten Rahmen um das Textfeld gekennzeichnet. Gleichzeitig wird der 'Ok' Knopf deaktiviert, um Programmfehler durch falsche Werte zu vermeiden. Jedes Textfeld besitzt einen separaten Tooltip, der ihnen helfen soll die richtige Eingabe zu tätigen. Dabei kann sich der Tooltip eines Textfeldes je nach Eingabe unterscheiden. Fahren Sie über das Textfeld, um den aktuell passenden Tipp zu bekommen. Zusätzliche Tipps können Sie in den jeweiligen Kapiteln hier im Handbuch finden. Abbildung 19 zeigt ein Beispiel des beschriebenen.

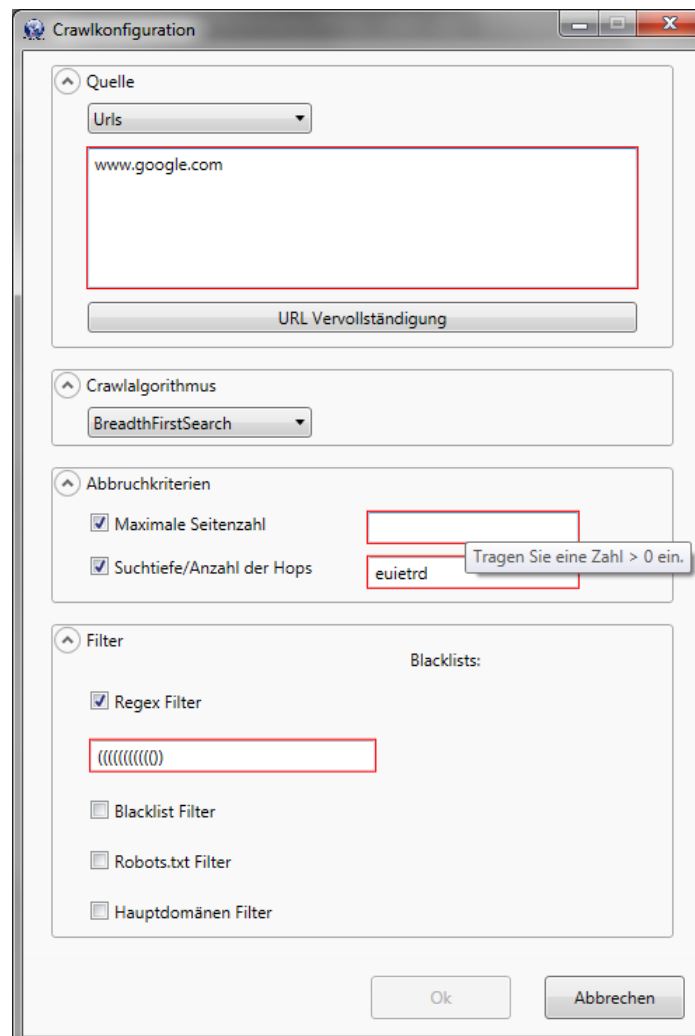


Abbildung 18: Fehler werden direkt angezeigt und Tooltips geben nützliche Hinweise

7 Behandlung von Problemen

7.1 Das Programm startet nicht korrekt oder zeigt nach dem Start nicht das gewünschte Verhalten

Bitte stellen Sie sicher, dass Sie das Programm korrekt installiert haben. Schauen Sie dazu in [Kapitel 3 - Installation](#) nach, welche Schritte zur richtigen Einrichtung nötig sind.

Achten sie insbesondere darauf, dass Sie das Programm vor der ersten Nutzung einmal als Administrator ausführen, um das Protokoll richtig einzurichten.

Folgende Fehlermeldung beim Start des Programms kann auf dieses Problem hinweisen.

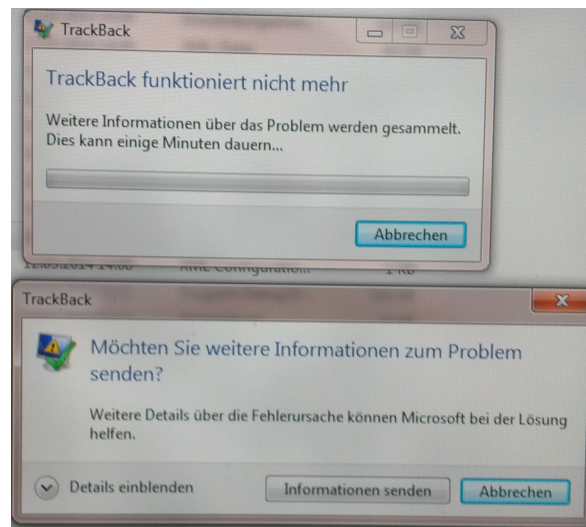


Abbildung 19: Fehler beim Starten des Programms

7.2 Das Protokoll öffnet eine Nachricht, dass er nicht gelöscht werden konnte

Das Protokoll bezieht seine Daten aus dem Windows Event Log. Dieser benötigt spezielle Berechtigungen. Aus diesem Grund ist es nötig für jeden Löschvorgang des Protokolls das Programm als Administrator auszuführen. Auch dann wenn man es zur Installation bereits einmal als Administrator ausgeführt hat.

7.3 Das Programm hängt mitten im Crawlvorgang komplett

Bisher ist dieses Problem unter folgenden Bedingungen eingetreten:

<http://forum.chip.de> war in der Liste der zu crawlenden URLs. Bisher ist unklar wieso das Programm beim Crawlen dieser Webseite einfriert.

Zu diesem Zeitpunkt lässt es sich nur noch über den Taskmanager schließen.

Es wird empfohlen alle betroffenen URL in die custom Blacklist einzutragen und diese beim Crawlen mit anzugeben, wenn man nicht ausschließen kann, dass sie in der Liste landet.

8 Glossar

Hier werden Definitionen gegeben, die im Zusammenhang mit dem Programm oder dem Handbuch neu entstanden sind oder die eine bestimmte/andere Definition besitzen. Allgemeine Fachbegriffe werden hier nicht erklärt. Begriffe die mit dem Programm in Verbindung stehen, hier aber nicht aufgelistet sind, z.B. Funktionen, werden im jeweiligen Kapitel erklärt.

8.1 Cluster

Cluster dienen der Übersichtlichkeit auf der Karte. In einem Cluster werden in einem bestimmten Radius und je nach Zoomfaktor mehrere Marker zusammengefasst.

8.2 Crawl/Crawlvorgang

Ein Crawlvorgang ist das Sammeln der URLs/ihrer Links. Er wird gestartet, nachdem die Einstellungen in der Crawlkonfiguration mit 'Ok' bestätigt wird und beendet, indem man den Stopknopf drückt bzw. die Abbruchkriterien eingetreten sind.

8.3 FreeGeoIP

Ein Anbieter zur Lokalisierung von IP-Adressen. Es gibt eine Onlineabfrage und die Möglichkeit die Datenbank lokal zu betreiben. Das Programm nutzt Letzteres.

8.4 Initiale URLs

Die initialen URLs sind die URLs der Liste, die zu Beginn eines Crawlvorgangs ausgesucht wurde. Z.B. die AlexaTop100.

8.5 Kante

Eine Kante ist eine Verbindung zwischen Knoten. Ihre Bedeutung ist hier, dass in einer URL in einem Knoten ein Link zu einer URL eines anderen Knoten ist.

8.6 Knoten

Ein Knoten ist ein Teil des Graphen, der die Datenstruktur bildet. In ihm befinden sich mehrere Informationen wie IP-Adresse, Geodaten und die verschmolzenen URLs. Er wird in der Karte durch einen Marker dargestellt und im Graphentab durch den Namen seiner Hauptdomain.

8.7 Marker

Ein Marker ist eine Markierung auf der Karte. Sie entspricht dem üblichen Aussehen von Markierungen von GoogleMaps. Ein Marker stellt entweder mehrere Marker oder einen Knoten/eine URL im Datenbestand dar.

8.8 Regex

Ein Regex (für Regular Expression) ist die Umsetzung des Prinzips der regulären Ausdrücke in C#.

8.9 TPC/Third-Party-Content

Webinhalt von Drittanbietern. Z.B. Links auf einer Webseite, die auf eine komplett andere Webseite verweisen.