

Inhaltsanalyse von Webseiten: Probleme und Lösungsansätze

Harald Klein¹

Die verschiedenen Formen der Information

Im Internet liegen die Informationen meist auf Webseiten vor, die aus verschiedenen Teilen bestehen. Das sind die reinen Texte, Bilder, Töne und Filme/Animationen. Die Grenzen sind teilweise fließend, da viele Links oft auch als Bilder realisiert sind. Die Navigation wird auch oft Graphiken vorgenommen, die aus nichts weiterem als Text mit farbigem Hintergrund bestehen. Weiter erschwerend kommt hinzu, wenn die Seiten mit Frames realisiert wurden, denn dann weicht die technische Struktur des Angebots von der sichtbaren ab. Bei der Analyse von e-mails, Diskussionforen, Gästebüchern oder Newsgroups kommen nicht textuelle Elemente eher selten vor.

Formulierung von Erkenntnisinteressen

Bevor irgendeine Analyse durchgeführt wird, sollte das Erkenntnisinteresse festgelegt werden, danach richtet sich die Auswahl der zu untersuchenden Angebote. Danach müssen die Date(i)e(n) ausgewählt werden, dies ist ein mehrstufiger Prozess. Generell muss man sich auch entscheiden, ob man nur die Textinformation zur Analyse heranziehen will oder auch die AV-Inhalte.

Vorgehensweise im Überblick

Zuerst müssen die Websites selektiert werden, die relevantes Material enthalten. Von diesen Websites werden im nächsten Selektionsprozess die Dateien ausgewählt, die für die Untersuchung relevant sind. Dazu gibt es Programme, die einem das mühsame Speichern von Hand abnehmen, sogenannte Offline-Reader. Diese Programme kopieren ganze Websites oder Teile davon auf die eigene Festplatte und setzen dabei vorhandene Links so um, dass ein Lesen lokal möglich ist. Bestimmte Dateien können beispielsweise ausgeschlossen werden, das gleiche gilt für das Verfolgen von Links.

1 Social Science Consulting, Königseer Str. 9, 98708 Gehren, Tel/Fax: 036783/80284, email: webmaster@intext.de

Danach hat man die ganzen Dateien auf der Festplatte und muss diese zusammenführen, meist zu einer Datei. Es treten dabei einige Probleme auf, die hier genannt werden sollen:

- Definieren von Textsegmenten bzw. der Analyseeinheit. Je nach Erkenntnisinteresse müssen sinnvolle Einheiten gebildet werden. Bei der Analyse von Webseiten wird das die einzelne Seite sein, bei Kleinanzeigen die Anzeige, bei Universitäten die ganze Website. Möchte man die Daten computerunterstützt analysieren, müssen die Daten auch so umgeformt werden, dass das Programm sie auch verarbeiten kann. Die Texte werden mittels Steuersequenzen in Texteinheiten unterteilt, dabei werden auch die Werte von externen Variablen gesetzt. Das Format dieser Roh-text genannten Texte ist von der Software abhängig und mehr oder weniger benutzerfreundlich. TextQuest erlaubt, dass eine Zeile oder ein Absatz eine Texteinheit ist, so dass HTML-Seiten schnell verarbeitet werden können. Andere Programme brauchen einen Wert für jede externe Variablen in bestimmten Spalten der Zeilen oder am Beginn einer Zeile (aber nicht jeder Zeile), oder als Steuersequenz überall auf der Zeile.
- Definieren von externen Variablen. Sie dienen zur Abgrenzungen der Analyseeinheiten bzw. Textsegmente und enthalten Informationen über diese, z.B. Datum, Grösse, Anzahl der Bilder, Anzahl der Links, mit Frames oder ohne, Sprache der Website usw. Wie schon bei der Definition von Textsegmenten ist auch hier darauf zu achten, was die Software für eine computerunterstützte Analyse erforderlich macht.

Die Realisierung dieser Arbeiten ist sehr zeitintensiv, Programme zur mehr oder weniger automatischen Transformation existieren meines Wissens (noch) nicht.

Vorgehensweise im Detail

Der arbeitsintensivste Teil einer computerunterstützten Analyse von Webseiten ist die Umformung der Seiten, in dem einerseits die Informationen im HTML-Code extrahiert und in externe Variablen umgesetzt werden, andererseits der HTML-Code entfernt wird. Eine lineare Struktur des Textes liegt nicht vor, sie wird aber von den meisten Textanalyseprogrammen gefordert. Bei der Analyse von HTML-Code sind einige andere Entscheidungen zu treffen, sowohl inhaltlicher als auch technischer Natur:

- Zugriffsprobleme: Schon beim Zugriff auf die Webseiten kann es Probleme geben, die entsprechenden Informationen zu speichern:
 - Fehler 404: Seite existiert nicht (mehr). Manche Server laden automatisch die entsprechende Seite nach, wenn diese umgezogen ist. Im kommerziel-

len Bereich muss man damit rechnen, dass die Homepage geladen wird, dann kann man in Endlosschleifen geraten.

- geschützte Seiten: Passwort oder Zugriffssysteme (Age-Check) erlauben keinen Zugriff auf die gewünschte Information. Bei einigen Offline-Readern kann man allerdings Passwort und Userkennungen für die Websites einrichten.
- der Server antwortet nicht. Eine Verbindung ist nicht zustande gekommen, das kann verschiedene technische Ursachen haben. Man kann es später dann noch einmal versuchen, ob ein Laden der Webseiten funktioniert.
- Links: man muss entscheiden, ob man ihnen folgt oder nicht. Wenn man ihnen folgt, kann man in Endlosschleifen geraten oder auf Seiten, die für den Untersuchungsgegenstand irrelevant sind. Links können wie folgt unterschieden werden:
 - innerhalb derselben Seite: dieselbe Information ist mehrfach vorhanden und kann in Endlosschleifen führen. Sie sind mit Sprungmarken operationalisiert (z.B. www.forum.de/forum/#hilfe). Diese Links kann man überlesen.
 - andere Seite: innerhalb der Site oder zu einer anderen Site. Wenn man den Offline-Reader die ganze Website herunterladen lässt, macht das Verfolgen der siteinternen Links wenig Sinn. Bei externen Links muss man entscheiden, ob sie verfolgt werden und wenn ja, bis zu welcher Tiefe.

Die Linkseiten müssen in die Ursprungsseite eingefügt werden.

- AV-Inhalte (Grafik, Audio, Video): oft ist in Grafiken Text enthalten, der von Interesse sein kann, diese Frage stellt sich generell. Software zur Analyse von AV-Inhalten gibt es, so dass dem von dieser Seite her kein Hindernis im Wege steht. Meist wird man aber wegen des hohen Aufwandes die AV-Inhalte von der Analyse ausschliessen.
- Sprache: einige Seiten existieren in verschiedenen Sprachen, so dass derselbe Inhalt mehrfach existiert, und einige Wörter in der einen Sprache haben eine verschiedene - oder sogar konträre Bedeutung - in einer anderen Sprache, z.B. *man* in Englisch und Deutsch, oder *motte, haut, fiel* in Deutsch und Französisch, oder *car, noise, or, for* oder *gave* in Englisch und Französisch.
- die Menge der Daten: Hindernisse sind die maximale Partitionsgrösse von Festplatten sowie der Platzbedarf der Textanalyseprogramme für interne Dateien. TEXTPACK braucht B. relativ viel Platz, während andere Programme wie TextQuest mehr Hauptspeicher (RAM) brauchen. Andere Programme kommen mit grossen Textmengen nicht zurecht. Die Dateien müssen nach ihrer Konvertierung zu einer Eingabedatei zusammenkopiert werden.

An einem Beispiel von Kontaktanzeigen soll gezeigt werden, wie die Konvertierung von Webseiten in ein Rohdatenformat (TextQuest) aussieht.

Beispieldateien: Webseite und Rohtext

<title>Massachusetts dating - Ladies Ads dating</title>

Age: 31

Sex: Female

Sexual Preference: Straight

Location: Webster, MA

Personal Message: I'm 31, dark blonde hair, brown eyes, 5'7", 125 lbs. I'm in great shape because I love to work out. My hobbies include mountain biking, hiking, rollerblading, skiing, sailing, vintage racing. I love to travel. I also love all of my seasons.

E-mail Address: *****

Member Number: 1089851

Member since June 09, 1998

Age: 32

Sex: Female

Sexual Preference: Straight

Location: Boston, Massachusetts

Personal Message: I'm a 32 y/y mulatto woman, seeking a SWM 25-38. A business man, who loves animals, nature, children, being outdoors, cooking, alterna-rock music, togetherness, and having fun. I enjoy cooking, family, nature, outdoors, movies, clubs, dinning (in/out), raves, rollerblading, beaches, cruises, children, spirituality. I'm about 5'5", 170lbs (a bit over weight). Light brown skin. dark eyes/hair (hair is med. length).

E-mail Address: *****

Member Number: 1690124

Member since May 23, 1998

</body>

</html>

\\$1(001-female-male-Mass-June-98-self) I'm 31, dark blonde hair, brown eyes, 5'7", 125 lbs. I'm in great shape because I love to work out. My hobbies include mountain biking, hiking, rollerblading, skiing, sailing, vintage racing. I love to travel. I also love all of my seasons.

\\$1(002-female-male-Mass-May-98-self) I'm a 32 y mulatto woman, \\$7(foreign) seeking a SWM 25-38. A business man, who loves animals, nature, children, being outdoors, cooking, alterna-rock music, togetherness, and having fun. \\$7(self) I enjoy cooking, family, nature, outdoors, movies, clubs, dining (in/out), raves, rollerblading,

beaches, cruises, children, spirituality. I'm about 5'5", 170 lbs (a bit over weight).
Light brown skin, dark eyes/hair (hair is med. length).

Kriterien und Software für das Herunterladen von Webseiten und Websites

Programmname	Web- snake 1.23	Grab- A-Site 3.0.14.	Back Street 1.4	Black Widow 3.63	Win HTTrack 1.24 b	Web Mir- ror 1.33	Web Reaper	Web VCR
Bedienung								
Hilfesystem	x	x	x	x	x	x	x	
Assistent/Wizard	x				x	x		x
Übersichtlichkeit	x			x	x	x	x	x
Scheduler	x					x		
Download								
User und Password	x			x	x	x	x	x
mehrere Websites	x			x	x	x	x	x
mit Bookmarks								x
Suchtiefe	x	x		x		x	x	
Links verfolgen intern	x			x			x	x
Links verfolgen extern	x			x			x	x
Dateinamen kon- vertieren (8.3)		x		x	x			
Dateistruktur behalten	x	x	x	?	x		x	
Filterkriterien								
nach Dateityp	x	x	x	x	x		x	x
mit Suchbegriffen								
Dateigrösse (min/max)	max			x	max		x	
Updatefunktion		x				x	x	
Abbruchkriterien								
Timeout		x			x		x	
Fehlversuche		x			x		x	
Festplatte voll	x		x					
verbrauchte Zeit		x					x	x
Zahl der Dateien	x	x						
Datenmenge	x	x			x		x	x

Stichworte: Inhaltsanalyse, Webseiten, Offline-Reader

Literaturangaben

Bürgi, Dieter (1999). Spiegeln und speichern. Webserver-Inhalte unter Windows auf lokale Medien übertragen. *c't* 19/1999, 242-247.

Franklin, Dave (1999). *Offline Reader*. [WWW document]. URL <http://www.davecentral.com>

Klein, Harald (1999). *TextQuest-Handbuch*. Gehen: Eigenverlag.

Klein, Harald (1999). *Textanalyse Software* [WWW document]. URL <http://www.intext.de/TEXTANAD.HTM>