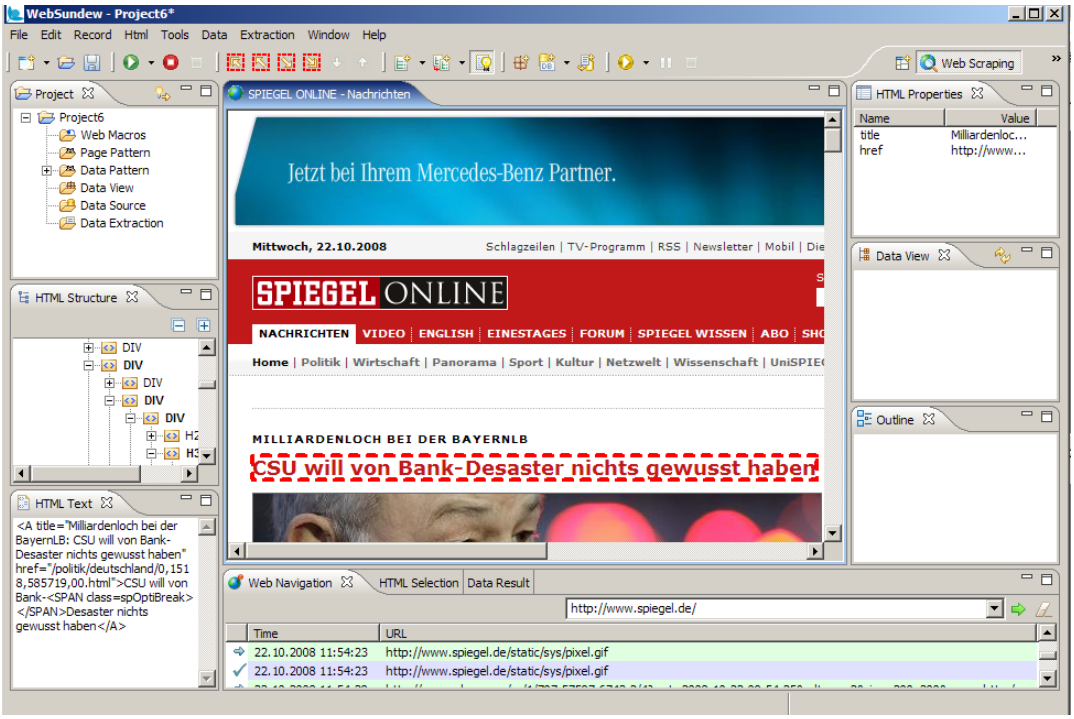
IWI Discussionpaper
#43 (08. Juni 2010)[1]
ISSN: 1612-3646

Analyse der Potenziale betrieblicher Anwendungen des Web Content Mining

Naum Neuhaus[2], **Karsten Sohns**[3] und **Michael H. Breitner**[4]



[1] Kopien oder eine PDF-Datei sind auf Anfrage erhältlich: Institut für Wirtschaftsinformatik, Leibniz Universität Hannover, Königsworther Platz 1, 30167 Hannover (www.iwi.uni-hannover.de).

[2] Candidatus Diplom-Ökonom, Institut für Wirtschaftsinformatik der Leibniz Universität Hannover (info@airbrushwear.com).

[3] Diplom-Ökonom, wissenschaftlicher Mitarbeiter und Doktorand (sohns@iwi.uni-hannover.de).

[4] Professor für Wirtschaftsinformatik und Betriebswirtschaftslehre und Direktor des Instituts für Wirtschaftsinformatik der Leibniz Universität Hannover (breitner@iwi.uni-hannover.de).

Inhaltsverzeichnis

Abkürzungsverzeichnis

Abb.	Abbildung
Abs.	Absatz
AGB	Allgemeine Geschäftsbedingungen
Art.	Artikel
ASP	Active Server Pages
BDSG	Bundesdatenschutzgesetz
bspw.	Beispielsweise
bzw.	Beziehungsweise
ca.	Circa
CAPTCHA	Completely Automated Public Turing Test To Tell Computers and Humans Apart
CERN	Conseil Européen pour la Recherche Nucléaire
CIA	Central Intelligence Agency
CMS	Content Management System
Corp.	Corporation
CSO	Corporate Security Officer
CSV	Character Separated Values
DM	Deutsche Mark
EG	Europäische Gemeinschaft
EM	Expectation Maximization
et al.	Et alli
etc.	et cetera
FSB	Federalnaja Sluschba Besopasnosti
FTD	Financial Times Deutschland
GbR	Gesellschaft bürgerlichen Rechts
GG	Grundgesetz
ggf.	Gegebenenfalls
GmbH	Gesellschaft mit beschränkter Haftung
HTML	Hypertext Markup Language
HTTPS	HyperText Transfer Protocol Secure
i.d.R.	In der Regel
i.V.m.	In Verbindung mit
IBM	International Business Machines Corporation
ID	Identifikationsnummer
IP	Internet Protocol
IT	Informationstechnologie
Kfz	Kraftfahrzeug
KNN	Künstliche neuronale Netze
Ldt.	Limited
LLC	Limited Liability Company
MB	Megabyte
Mio.	Million
Mrd.	Milliarde
NLP	Natural Language Processing
NP	Nichtdeterministisch Polynomiell
OCR	Optical Character Recognition
OHG	Offene Handelsgesellschaft
PAS	Professionelle Autosuche
PC	Personal Computer
PDF	Portable Document Format
PHP	Personal Home Page Tools
S.	Seite

SMS	Short Massage Service
SPD	Sozialdemokratische Partei Deutschlands
SQL	Structured Query Language
STC	Suffix Tree Clusterung
SZ	Süddeutsche Zeitung
TFIDF	Term Frequency - Inverse Document Frequency
TSV	Tab-Separated Values
UrhG	Urhebergesetzbuch
URL	Uniform Resource Locator
US	United States
USA	United States of America
vgl.	Vergleiche
WCE	Web Content Extractor
WCM	Web Content Mining
WIE	Web Info Extractor
WWW	World Wide Web
XML	Extensible Markup Language
z. B.	Zum Beispiel

1 Einleitung

Seit das World Wide Web (WWW) 1989 im CERN von Berners-Lee entwickelt und der Öffentlichkeit zugänglich gemacht wurde, ist es rasant und unaufhörlich gewachsen. Heute ist das Internet mit etwa einer Billion Seiten die größte Informationsdatenbank der Welt (vgl. hierzu und zum Folgenden Alpert, Hajaj, 2008). Die Datenflut der im Internet veröffentlichter Inhalte reißt nicht ab, denn jeden Tag kommen etwa eine Milliarde weiterer Seiten hinzu. Dies lässt Vermuten, dass sich in diesem Informationsdschungel wahre Schätze verbergen. Aus betriebswirtschaftlicher Sicht lassen sich aus dem Internet wertvolle Informationen über die Konkurrenz, Kunden, Preise, Trends, Produkte und vielem mehr extrahieren und verwerten. Nicht nur Unternehmen der Informationswirtschaft haben bereits erkannt, dass ein Wissensvorsprung in Form eines schnellen und flexiblen Zugriffs auf relevante Internetinhalte zu einem erheblichen Wettbewerbsvorteil verhelfen kann. Für Unternehmen bietet das Internet die Möglichkeit, Informationen finden und sammeln. Trotz technischen Fortschritts erfolgt die Suche, Interpretation und Integration von Internetdokumenten noch überwiegend manuell mit Hilfe gängiger Suchmaschinen. Nicht zuletzt zeugt der Erfolg von Google Inc. von der Bedeutung und Attraktivität des Geschäftes mit Online-Suchdiensten.

Da die manuelle Extraktion relevanter Inhalte aus dem exponentiell wachsenden WWW immer aufwendiger und damit teurer wird, rücken moderne, automatisierte Methoden der Inhaltsextraktion in den Mittelpunkt der wissenschaftlichen und betrieblichen Betrachtung. Von zentraler Bedeutung sind hierbei insbesondere Anwendungen zur maschinellen Erfassung, Klassifizierung und Integration von Webdokumenten.

Gegenstand vorliegender Untersuchung sind Methoden und Verfahren des Web Content Mining sowie seiner Potenziale im Hinblick auf betriebliche Anwendungen. Diese junge Forschungsdisziplin beschäftigt sich mit der automatischen Generierung relevanten Wissens aus den Ressourcen des weltweiten Internets. Das oberste Ziel von Web Content Mining (WCM) ist die maschinelle Extraktion aktueller und nützlicher Inhalte aus verfügbaren Webseiten und -dokumenten sowie eine strukturierte Präsentation der gewonnenen Erkenntnisse. Im Gegensatz zur Analyse von strukturierten Tabellen oder Archiven stellt die fehlende Einheitlichkeit von Webseiten die größte Herausforderung für WCM-Anwendungen dar. Dennoch sollen moderne WCM-Programme schon jetzt in der Lage sein, einen großen Teil der manuellen Internetrecherche abzulösen und zu automatisieren.

Der Markt für WCM-Software ist sehr jung und in der praktischen Anwendung wissenschaftlich weitestgehend unerforscht. Ebenso neu ist die Betrachtung von Geschäftsprozessen und –Konzepten, die mit Hilfe von WCM optimiert bzw. vollzogen werden könnten. Das Ziel der vorliegenden Untersuchung ist die Erforschung unterschiedlicher Anwendungsgebiete von WCM. Vor dem Hintergrund dieser Erkenntnisse werden daraufhin acht WCM-Programme auf ihre Leistungs- und Anwendungsfähigkeit getestet und mit einander verglichen.

Im Rahmen vorliegender Diplomarbeit sollen Potenziale betrieblicher Anwendungen des WCM analysiert und dargestellt werden. In diesem Zusammenhang werden im Verlauf der Untersuchung folgende Fragen beantwortet.

- Inwieweit ist WCM in der Lage das betriebliche Informationsmanagement zu entlasten oder zu optimieren?
- Welche neuen Geschäftskonzepte lassen sich mit WCM künftig verwirklichen?
- Welche Anspruchsgruppen profitieren von WCM-Anwendungen?
- Welche Verfahren und Methoden nutzen Entwickler von WCM-Programmen?
- Welche konkreten Anwendungen bietet der Markt für WCM-Software?
- Wie weit sind moderne WCM-Programme fortentwickelt?
- Welche Funktionsweise steckt hinter den jeweiligen WCM-Programmen?
- Wohin geht der Trend von WCM-Anwendungen?

Der Verlauf der Untersuchung ist in drei Kapitel unterteilt. Im Grundlagenteil erfolgt eine theoretische Auseinandersetzung mit WCM. Darin werden grundlegende Begriffe und Aspekte von WCM definitorisch erläutert. Daraufhin werden einige Anwendungsfelder des WCM untersucht, um ein Verständnis für die konkrete Funktionsweise solcher Anwendungen zu erhalten. Mit diesen beiden Kapiteln (2 und 3) wird die theoretische Grundlegung des vorliegenden Forschungsgebiets abgeschlossen.

Im vierten Kapitel werden schließlich WCM-Programme im praktischen Einsatz getestet. Hierfür sollen diese Programme im Rahmen zuvor festgelegter Suchszenarien erlernt und angewendet werden. Die Suchszenarien sollen wiederum aus den im vierten Kapitel entdeckten Anwendungsgebieten abgeleitet werden. Es wird jeweils ein gründlicher Einblick in die Funktionsweisen, Stärken und Schwächen der einzelnen Programme erlaubt. An dieser Stelle

wird sich zeigen, welche entwickelten Anwendungsideen bereits jetzt ausführbar sind und welches WCM-Programm für welche Anforderungen in Frage kommt.

2 Grundlagen

Um künftigen Missverständnissen im Hinblick auf differierende Definitionen vorzubeugen, werden in diesem Kapitel die Schlüsselbegriffe vorliegender Arbeit grundlegend erläutert. Besondere Aufmerksamkeit erfährt in diesem Zusammenhang das Web Mining (Abs. 2.1) als Oberbegriff für einzelne Web Mining-Disziplinen, die sich im Zeitablauf daraus gebildet haben. Abschnitt 2.2 ist dem Web Content Mining gewidmet.

2.1 Web Mining

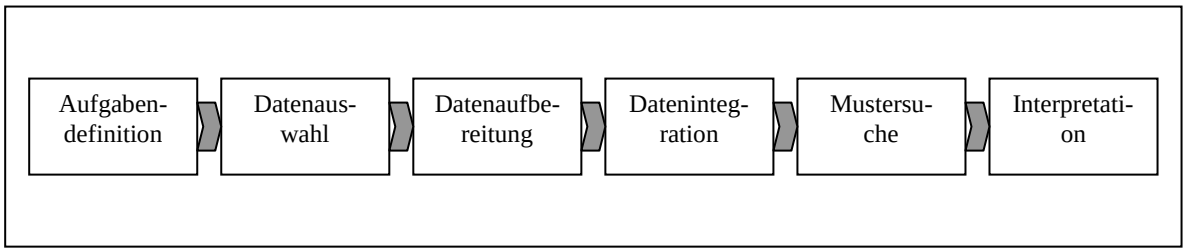
Unter Data Mining werden nach Kollmann unterschiedliche Techniken verstanden, die geeignet sind, "besondere Datenkonstellationen mit möglicher Ursachenklärung aufgrund der Erkennung von Mustern in den Daten einer umfangreichen Datenbank herauszufiltern" (Kollmann 2007, S. 200). Eine geeignete englischsprachige Definition bieten David Hand et al.: "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data novel ways that are both understandable and useful to the data owner" (Hand et al. 2001, S. 1). Hand et al. präzisieren im Gegensatz zu Kollmann, dass die gewonnenen Ergebnisse nicht einfach nur herausgefiltert werden. Der Prozess ist mit der Erkennung von Datenmustern noch nicht zwangsläufig beendet. Darauf folgt eine bedarfsorientierte Darstellung sowie unter Umständen eine Interpretation der Ergebnisse auf Grundlage einer zuvor formulierten Fragestellung (vgl. Gentsch 2002, S.282).

Seit der Umfang elektronischer Daten exponentiell zunimmt, während die Kosten der Speicherung sinken, gehört Data Mining zu den aktivsten Forschungsfeldern der IT-Branche (vgl. Velasquez/Palade 2008, S. 32). Innerhalb von Unternehmen sammeln sich im Zuge der Kommunikation und Organisation erhebliche Mengen von Daten an, hinter denen nützliches Wissen über Kunden, Trends, Werbemaßnahmen, etc. vermutet wird. Da setzt das Data Mining an und findet Anwendung bei großen, strukturierten Beständen numerischer, ordinal- oder nominalskalierter Daten mit interessanten aber verborgenen Informationen (vgl. Kollmann 2006, S. 346). Statistische und mathematische Verfahren des Data Mining dienen dem Management, um aus den internen und externen Datenbanken relevante Informationen und wertvolles Wissen zu generieren (Kollmann 2007, S. 200).

Mit über einer Billion Seiten stellt das Internet heute die mit Abstand größte Datenbank der Welt dar (vgl. Markov/Larose 2007, S. 4). Es gibt mittlerweile kaum Informationen, die nicht irgendwo in den Tiefen des WWW auf ihren Interessenten warten. Das Wachstum der im Internet veröffentlichter Inhalte sprengt alle Grenzen und stellt sogar den Branchenriesen Google Inc. vor große Herausforderungen (vgl. Alpert/Hajaj 2008, S. 1). Das größte Problem des Internets ist die Kehrseite seines größten Erfolges - das Überangebot an Informationen (vgl. hierzu und zum Folgenden Hornig et al. 2008, S. 82). Suchmaschinen liefern zu jedem Suchbegriff Unmengen an Treffern, die hierarchisch nach Ihrer Beliebtheit im Netz sortiert sind. Ein Überangebot an Treffern verspricht jedoch noch lange kein erwünschtes Ergebnis. Vielmehr beeinträchtigt der so genannte "Information Overload" die Entscheidungsfähigkeit der Nutzer und erhöht somit die Kosten der Internetrecherche eines Unternehmens (vgl. Stegbauer 2001, S. 171-173). Die vorliegende Arbeit befasst sich mit der Frage, inwieweit Methoden des Data Mining dazu beitragen könnten, wertvolle Informationen und anschließend relevantes Wissen aus dem WWW zu extrahieren. Der Begriff Web Mining bezeichnet Anwendungen von Data Mining im Zusammenhang mit elektronischen Daten, die durch die Nutzung des Internets generiert werden (vgl. Alpar/Niedereichholz 2000, S. 19). Etzioni liefert hierzu folgende Definition: "Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services" (Etzioni 1996, S. 65).

Abbildung 1 zeigt in Anlehnung an Hippner et al. die typischen Schritte eines Web Mining-Prozesses. Auf Basis der Aufgabenstellung erfolgt zunächst eine Auswahl relevanter Daten. Der darauf folgende Schritt der Datenbereinigung ist von besonders hoher Bedeutung und nimmt auch die meiste Zeit in Anspruch (vgl. Zaiane 2002, S. 27).

Abbildung 1: Einzelne Schritte des Web Mining-Prozesses



Quelle: In Anlehnung an Hippner et al. 2002, S. 8

Wenn es sich um Usage-Daten handelt, dann erfolgt an dieser Stelle auch die Identifikation von Nutzern und Sitzungen (vgl. hierzu und zum Folgenden Hippner). Soweit die Daten aus unterschiedlichen Quellen stammen, ist als nächstes die Integration der verschiedenen Datenquellen erforderlich. Anschließend werden die aufbereiteten Daten mit Hilfe von Data Mining Verfahren auf Muster untersucht. Zuletzt erfolgt eine bedarfsorientierte Interpretation und Umsetzung der gewonnenen Erkenntnisse.

Die Generierung wertvollen Wissens aus dem Internet ist eine große Herausforderung. Das WWW enthält überwiegend unstrukturierte Daten, die weitaus komplexer und dynamischer sind als eine strukturierte betriebliche Datenbank (vgl. Akerkar/Lingras 2008, S. 6). Im Gegensatz zum klassischen Data Mining, welches Daten in Tabellenform benötigt, kommen hier beispielsweise Algorithmen des Text Mining zum Einsatz, da ein Großteil des Internets nur in schwach- oder unstrukturierter Form vorliegt.

Aufgrund der wachsenden Zahl im Internet verfügbarer textbasierter Inhalte und der hohen Nachfrage nach automatischen Methoden zur Aufbereitung und Interpretation der Datenflut, gehört Text Mining zu einem sehr aktiven Forschungsgebiet (vgl. Deinhard/Oswald 2008). Eine sinnvolle Klassifizierung weiterer Internetinhalte liefern Srivastava et al (vgl. Srivastava et al. 2000, S. 12-23).

Tabelle 1: Klassifikation von Internetdaten

Typ	Art der Daten	Beispiele
Content (Inhalt)	Auf Webseiten enthaltene und angebotene Daten und Informationen.	Text-, Bild-, Audio-, Videodateien. Hyperlinks, Metadaten
Structure (Struktur)	Daten, die Rückschlüsse auf die Struktur und Organisation des Inhalts zulassen.	Zusammenstellung von HTML- oder XML-tags, Hyperlinks
Usage (Nutzungsdaten)	Sekundärdaten, die bei der Nutzung des Internets anfallen.	web-server access logs, proxy server logs, Nutzer-Profile, Cookies
User Profile (Benutzerdaten)	Personifizierte Information über bestimmte Internetnutzer.	Daten aus sozialen Netzwerken, Blogs, etc. Konsumgewohnheiten, Interessen, Demografie.

Quelle: In Anlehnung an Srivastava et al. 2000, S. 12-23

Tabelle 1 zeigt eine mögliche Unterteilung von Internetdaten in vier grundlegende Klassen. Während Usage- und User Profile Daten insbesondere für Betreiber von Internetseiten interessant sind, steht für Internetsuchdienste die Struktur von Webseiten im Mittelpunkt der Betrachtung. Der mit Abstand größte und am wenigsten strukturierte Datentyp ist der Inhalt. Die Gewinnung von Erkenntnissen aus überwiegend textbasierten Internetinhalten ist Gegenstand vorliegender Arbeit und bezieht sich auf Webressourcen wie z. B. Handelsbörsen, Webshops, Foren, Communities oder Blogs.

2.2 Web Content Mining

In diesem Abschnitt erfolgt eine definitorische Auseinandersetzung mit dem Begriff WCM und seinen verwandten, erklärungsbedürftigen Techniken.

Vergleichbar mit anderen jungen Forschungsrichtungen ist, das Begriffsverständnis für WCM ist uneinheitlich. Während sich beispielsweise Kollmanns Definition auf die Strukturierung und Systematisierung von Web-Dokumenten-Inhalten beschränkt (Kollmann 2007, S. 201), schreiben andere Autoren in diesem Zusammenhang von "find and retrieve new resources from the web", "extract the actual information from web pages" (Kernahan/Capretz 2006, S. 84) oder gar "machine learning" (Marcov/Larose 2007, S. 59). Tabelle 2 zeigt unterschiedliche Definition von WCM, die sich zum Teil auf inhaltlicher Ebene unterscheiden.

Tabelle 2: Sammlung unterschiedlicher Definitionen von Web Content Mining

Web Content Mining Definitionen	**Quellen**
Kollmann definiert Web Content Mining als eine „Anwendung von Data Mining-Verfahren, die das Ziel hat die Informationssuche im Internet, durch Strukturierung und Systematisierung von Dokumenten-Inhalten zu vereinfachen. Die Dokumente müssen beim anschließenden Informationsscreening leichter und besser erkannt werden können.“	Kollmann 2007, S. 201
Markov und Larose schreiben in diesem Zusammenhang von: "Machine learning and data mining approaches organize the Web by content and thus respond directly to the major challenge of *turning web data into web knowledge*"	Marcov/Larose 2007, S. 59
"Web Content Mining befasst sich mit der Erkennung und Extraktion von Mustern in Web-Dokumenten, die aus Texten, Hyperlinks, Grafiken, Audio- und Videostreams bestehen können.	Deinhard/Oswald 2008, S. 1
The major goals of Web content mining are "find and retrieve new resources from the web", categorize and cluster the resources that have already been found" and "extract the actual information from a web page.	Kernahan/Carpez 2006, S. 84
"Web-content mining techniques are used to discover useful information from content on the web."	Akerkar/Lingras 2008, S. 263
"The objective of mining the web is to find useful information from the web documents." It deals with "mining of document contents and improving the content search with tools such as search engines."	Velasquez/Palade 2005, S. 33
"Beim Web Content Mining (WCM) geht es um die Extraktion von Wissen und Informationen aus Dokumenten, Bildern und anderen Inhaltsformen, aus denen Webseiten bestehen. Dabei kommen insbesondere Verfahren des Text Mining zum Einsatz."	Linder 2005, S. 65
"Das Web Content Mining stellt Methoden und Verfahren bereit, mit deren Hilfe Informationen und damit neues Wissen aus dieser Datenflut automatisch extrahiert werden können.	Dehmer 2006, S. 18
Web Content Mining befasst sich mit "der inhaltlichen Analyse von Webseiten. Hierbei bedient es sich oft der Methoden des Text-Mining oder des Multimedia Data Mining, um in den Dokumenten Muster zu finden bzw. die Dokumente zu klassifizieren und zu gruppieren."	Schildhauer 2003, S. 333

Quelle: Eigene Erstellung

Einige der Autoren in Tabelle 2 verstehen das WCM als einen Vorgang der umgangssprachlich als "Lernen" bezeichnet wird. In erster Reihe lernen hierbei jedoch Maschinen bzw. Computer und erst darauf der menschliche Nutzer. Der Vorgang des maschinellen Lernens aus elektronischen Datenbeständen des WWW erfolgt mit Hilfe von Data Mining-Verfahren. Die Inhalte von Seiten werden ermittelt oder die Seiten thematisch gruppiert (vgl. Salmen 2004, S. 91). Hierbei spielt die Analyse von Freitexten eine große Bedeutung. Das oberste Ziel von WCM ist demnach die automatische Extraktion aktuellen und nützlichen Wissens aus verfügbaren Webseiten und -dokumenten.

Für die vorliegende Arbeit wurde auf Grundlage der aus Tabelle 2 gewonnenen Aspekte eine eigene Definition formuliert:

Web Content Mining ist ein Verfahren, welches sich der Methoden des Data Mining bedient, um aus Webseiten und -dokumenten automatisch aktuelles und nützliches Wissen zu extrahieren und in einer bedarfsgerechten Art zu präsentieren.

Der Prozess der Datenauswertung und -darstellung wird auch oft als Knowledge Discovery-Prozess bezeichnet (vgl. Preißner 2001, S. 185).

3 Betriebswirtschaftliche Anwendungsgebiete des Web Content Mining

3.1 Generierung von Kunden- und Produktinformationen

Das Internet kann als eine nahezu unendliche Menge potenziell entscheidungsrelevanter Webressourcen betrachtet werden. Zahlreiche Anspruchsgruppen wie Webseitenbetreiber, Unternehmen, Marktforschungsinstitute oder Regierungsorganisationen interessieren sich für das aggregierte Wissen, das hinter den WWW-Daten verborgen ist. Gegenstand vorliegenden Abschnitts ist die Extraktion entscheidungsrelevanter Informationen mit Hilfe von WCM-Methoden. Dabei liegt der Fokus auf Informationen, die geeignet sind, einem Unternehmen Wettbewerbsvorteile zu verschaffen.

Klassische Methoden der Markt- und Meinungsforschung beschränken sich in der Regel auf Befragungen, Beobachtungen und Interviews. Die Verbreitung des Internets hat aber auch in diesem Gewerbe zu weit reichenden Veränderungen geführt, womit unterschiedliche Formen der Online-Marktforschung immer stärkeren Einzug in die Marktforschungspraxis halten (vgl. Hoffmann 2006, S. 134). Grundsätzlich lassen sich die Untersuchungsmethoden in Sekundär- oder Primärforschungsarten unterteilen. Zur Gruppe der primären Online-Marktforschung gehören u. a. Online-Fragebogenuntersuchungen, -Interviews und Experimente. Diese Gruppe zeichnet sich dadurch aus, dass die zugrunde liegenden Daten speziell für die Forschungsfrage erhoben werden. Daten der Sekundärforschung jedoch werden zu einem früheren Zeitpunkt und zu einem anderen Zweck gesammelt. Allerdings schließen sich die Methoden der Primär- und Sekundärforschung nicht gegenseitig aus, vielmehr ergänzen sie einander (vgl. Jung 2006, S. 597). Zu den Methoden der Sekundärforschung gehören manuelle Datenbankrecherchen sowie neuerdings das WCM (vgl. Freyer 2006, S. 226).

Soziale Netzwerke, oft auch Online Communities oder Social Network Services genannt, bieten ihren Teilnehmern eine öffentliche Plattform, in der sie Kontakte knüpfen, Freundschaften aufbauen und sich den anderen Mitgliedern präsentieren können (vgl. Ziser 2007, S. 18).Das besondere Merkmal sozialer Netzwerke ist, dass der Betreiber lediglich eine funktionsfähige Plattform zur Verfügung stellt und aufrechterhält. Sämtliche Inhalte werden daraufhin ausschließlich von den Nutzern produziert. Die ersten Dienste solcher Art (Myspace, Facebook) stammen aus den USA und sind nicht älter als vier Jahre. Dennoch gehören laut einer aktuellen Universal McCann-Studie derzeit 57 Prozent aller weltweiten Internetnutzer zumindest einer Community an (vgl. hierzu und zum Folgenden Smith 2008, S. 3). In Deutschland sind es 40 Prozent, von denen wiederum 41 Prozent mindestens einmal die Woche das Netzwerk besuchen. Heute stehen dem Internet-User unzählige soziale Netzwerke zu Auswahl, die zwar unterschiedliche Kundengruppen ansprechen (StudiVZ→ Studenten, S chülerVZ → Schüler, MeinVerein → Vereinsmitglieder, Myspace → Musiker, Entertainer) sich aber dennoch des gleichen Prinzips bedienen.

Zusammenfassend kann gesagt werden, dass der Zugriff auf so umfassende Personendaten noch nie so einfach war wie jetzt. Die enthaltenen Inhalte und Strukturen erlauben den Einblick in die Privatsphäre der Nutzer und geben damit Aufschluss über seine Interessen, Meinungen, Wünsche, Hobbys, Freunde, Aufenthaltsorte, etc.

Einen beeindruckenden Vorgeschmack auf die aggregierten Inhalte von StudiVZ bietet Hagen Fritsch, dem es Dezember 2006 gelang 1.074.574 Profile auszulesen. Auf http://studivz.irgendwo.org präsentiert Fritsch seine Untersuchungsergebnisse. Dort erwähnt er, dass ihm der Zugang zu den Daten mittels simpler Crawling-Programme sehr einfach und problemlos gelang. Mittlerweile sei das „Abgrasen“ von StudiVZ-Profilen aufgrund eingeführter Schutzmechanismen erheblich erschwert worden (vgl. Fritsch, 2006, S. 5). Derartige Schutzmechanismen und die praktische Anwendung von WCM-Software auf StudiVZ werden am Ende dieses Kapitels noch genauer betrachtet.

Der Umfang, indem das Web Content Mining auf soziale Netzwerke angewendet wird, hängt davon ab, wie die erhaltenen Informationen gewinnbringend eingesetzt werden können. Aus der Datenflut wertvolles Wissen zu generie-

ren, ist insbesondere für Betreiber von Online Communities sehr wichtig. Weitere Interessenten sind Unternehmen, die auf Werbung angewiesen sind sowie staatliche Organisationen, wie z. B. Geheimdienste.

Betreiber von Online Communities: Mit der Mitgliederanzahl und der Datenflut steigt für Seitenbetreiber auch die Notwendigkeit der Kontrolle. Hinweisen auf Gewaltverherrlichung, Mobbing, Diskriminierung, Verleumdung, Spam, etc. muss umgehend nachgegangen werden, was sich bei großen Netzwerken wie z. B. Myspace (200 Mio. Mitglieder) StudiVZ (7,5 Mio. Mitglieder) als technisch und organisatorisch schwierig herausstellt (vgl. Mühlenbeck/Skibicki 2008, S. 19). Verfahren des WCM könnten z. B. zur automatischen Identifikation solcher Äußerungen verwendet werden. Fragwürdige oder verdächtige Daten könnten mit Hilfe von Text Mining erkannt und daraufhin mit Klassifikationsverfahren in Gefahrenklassen eingeteilt werden. Den jeweiligen Klassen kann eine adäquate Reaktion nachgeschaltet werden, wonach entschieden wird, ob der Nutzer verwarnt oder ggf. aus der Gemeinschaft ausgeschlossen wird.

Das klassische Erlösmodell, welches hauptsächlich auf der Einblendung von Werbebannern basiert, scheint sich heute nicht mehr auszuzahlen. Seit dem Verkauf von Myspace an die News Corp. setzen die Betreiber von Myspace auf personalisierte Werbemaßnahmen (vgl. Meusers 2007, S. 1). Zur Segmentierung und Klassifikation ihrer und der Mitglieder konkurrierender Plattformen eignen sich Verfahren und Methoden des WCM. Sie dienen der Analyse und Typisierung von Profilen anhand aller verfügbaren Informationen. Dabei ist zu beachten, dass vor der Verwendung persönlicher Daten stets die Einwilligung dieser Person einzuholen ist. Im Folgenden werden mögliche Erlösmodelle für unterschiedliche Anspruchsgruppen kurz vorgestellt.

Unternehmen: Personalisierte Werbemaßnahmen sind für Unternehmen zwar sehr interessant, erfordern jedoch zunächst eine saubere Segmentierung und Klassifikation der Plattformmitglieder durch den Betreiber. Erst wenn entsprechende Strukturen und die Datenbasis vorhanden sind, kann das Unternehmen seine Werbeaktivitäten vorbereiten und den Werbeplatz kaufen. Interessant sind solche Angebote nicht nur für große, namhafte Unternehmen, denen eine individuelle Ansprache des Kunden wichtig ist, sondern auch für kleine Firmen, die mit einem geringen Budget gezielt eine kleine, aber konkrete Zielgruppe ansprechen wollen.

Namhafte Unternehmen sind auf ein ehrliches Feedback bezüglich ihrer Produkte, ihrer Werbung oder Reputation angewiesen. Das WCM ist in der Lage die Datenflut innerhalb sozialer Netzwerke nach bestimmten Äußerungen durchzusuchen und diese zu typisieren, um daraus eine bedarfsgerechte Interpretation abzuleiten. Auf diese Weise können neue Trends identifiziert werden, die zur Optimierung von Produkten und Werbeaktivitäten erforderlich sind.

Unter Zuhilfenahme von Web Mining lassen sich Menschen identifizieren, die mit ihren Aussagen und Empfehlungen eine Vielzahl weiterer Konsumenten erreichen und beeinflussen können. Durch ihre Autorität, Fachwissen oder sonstigen Einfluss auf ihren Freundeskreis kommt den so genannten Meinungsführern im Marketing eine große Bedeutung zu (vgl. Langner 2007, S. 81). Aus der graphischen Darstellung von Freundschaftsverbindungen sowie der Aktivität der Nutzer (z. B. Wer schreibt viele Kommentare? Wer gründet Gruppen? Wer wirbt neue Mitglieder? Wer hat viele Freunde?) lassen sich demnach besonders sendungsorientierte Typen bestimmen. Insbesondere für moderne Werbeformen, wie z. B. das virale Marketing, bei dem gezielt Mundpropaganda ausgelöst wird, kann sich eine Typisierung der Mitglieder als sehr nützlich erweisen (vgl. Langner 2007, S. 27).

Geheimdienste und sonstige Regierungsorganisationen: Derart detaillierte und umfangreiche Datenbanken mit persönlichen Daten, Interessen, Aufenthaltsorten und Freundschaftskreisen stand bis vor zehn Jahren noch nicht mal den internationalen Geheimdiensten zur Verfügung. Auf der einen Seite häufen sich Fälle, bei denen mit Hilfe sozialer Netzwerke Kriminelle ausfindig gemacht werden konnten. Auf der anderen Seite entsteht der Verdacht, dass Geheimdienste wie CIA oder FSB (Russland) diese relativ einfach verfügbaren Informationen für eigene Zwecke missbrauchen (vgl. Ackerman 2006, S. 3; Markoff 2006, S. 1). Jon Callas, CSO eines der führenden Entwicklers von Sicherheitssoftware sagte dazu im Interview: „I am continually shocked and appalled at the details people voluntarily post online about themselves. … social networking websites such as Myspace and Friendster are a snoop's dream" (Marks 2006, S. 1). Das russische soziale Netzwerk odnoklassniki.ru, mit derzeit über 22 Mio. Mitgliedern wurde dieses Jahr mit dem Verdacht auf Bekanntgabe von geheimen militärischen Information konfrontiert (vgl. hierzu und zum Folgenden Cnews 2008, S. 1). Nicht nur gewöhnliche Soldaten, sondern auch Mitglieder von Spezialeinheiten organisierten sich über mehrere Jahre unter Angabe von Aufenthaltsorten, -zeiten und Fotos in Gruppen. Schließlich konnte ein großer Teil des russischen Militärs mit dessen Einsatzorten, Truppenstärken sowie der gesamten Hierarchie in der Online Community abgebildet werden. Es ist daher kaum denkbar, dass derart wichtige und gleichzeitig frei verfügbare Daten keinen Interessenten in Geheimdienst- bzw. Regierungskreisen finden.

3.2 Generierung von Wissensvorteilen auf Handelsmärkten

Auf keinem Markt lässt sich aus einem Informationsvorsprung so viel Profit und aus Desinformation so viel Verlust erzeugen wie auf dem internationalen Kapital- und Finanzmarkt. Zwar wird im Allgemeinen von einer Unprognostizierbarkeit von Finanzproduktpreisen ausgegangen, dennoch ist es streckenweise möglich, mit regelbasierten Ansätzen konstante Gewinne zu machen (vgl. Kühn 1998, S. 92). Zahlreiche Studien befassen sich derzeit mit der Frage, ob Methoden und Verfahren des WCM in der Lage sind bessere und effizientere Bewertung von Optionen und sonstigen Derivaten vorzunehmen, als es mit klassischen finanzwissenschaftlichen Ansätzen (z. B. Black/Scholes-Modell) möglich ist (vgl. Bartels 2008, S. 212). Mit Hilfe künstlicher neuronaler Netze ist es schon heute möglich, akkurate Bewertungen von Kapitalmarktprodukten anhand einer Vielzahl verfügbarer Internetquellen vorzunehmen und die gesammelten Information in Handlungsempfehlungen zu transformieren (vgl. Bartels 2008, S. 213).

Die automatische Aggregation entscheidungsrelevanter Kapitalmarktinformationen aus dem Internet birgt zahlreiche Vorteile gegenüber klassischen manuellen Ansätzen. Da Investitionsentscheidungen oftmals innerhalb kürzester Zeit gefällt werden müssten, bietet WCM dem Nutzer einen Zeitvorteil gegenüber der Konkurrenz. Insbesondere im Bereich des Daytrading kann WCM dem Händler zu höheren Arbitragegewinnen verhelfen. In Zukunft könnten intelligente Systeme nicht nur unterstützend sondern auch selbstständig tätig werden, um das Geld von Spekulanten streckenweise automatisch zu vermehren.

Langfristigen Anlegern, denen eine fundierte Analyse und konsequentes Risikomanagement wichtig ist, bietet das Web Content Mining eine zusätzliche Informationsbasis, die mindestens genauso gute Informationen liefert wie klassische Bewertungsansätze. Nicht zu vernachlässigen ist die enorme Zeitersparnis der automatischen Wissensextraktion. Zwar ist es i. d. R. ein langwieriger Prozess, bis das System eingerichtet und kalibriert ist. Sobald das System aber in einen betriebsbereiten Zustand versetzt wurde, ist kein Einsatz von Personal mehr erforderlich, um es am Laufen zu halten.

Eine Branche, die die Vorzüge des Internets sehr schnell erkannte und innerhalb weniger Jahre einen Großteil ihrer Beschaffungs- und Vertriebsaktivitäten dorthin verlagerte, ist der Gebrauchtwagenmarkt. Dieser wird im Folgenden exemplarisch für weitere Märkte mit einem ausgeprägten Bedarf an aktuellen Informationen bezüglich Preisen und Angeboten betrachtet. Die Aussagen lassen sich z. B. auch auf den Immobilien-, Elektronik- oder sonstige Märkte übertragen.

Vorbei sind die Zeiten, als Autohändler 500 DM und mehr bezahlten, um den „Heißen Draht“ am Vorabend seiner offiziellen Erscheinung zu bekommen. Heute werden die meisten Gebrauchtwagenkäufe über Internetplattformen wie z. B. Mobile.de, Internetscout24, Automatico.de oder Auto.de abgewickelt. Zahlreiche Automobilbörsen buhlen um die Gunst des Kunden, wobei die oben genannten etwa 90 Prozent des gesamten Marktes abdecken.

Ein Gebrauchtwagenhändler kann nur überleben, wenn er schneller und effektiver an Verkaufsanzeigen als seine Konkurrenz oder seine potenziellen Kunden kommt, denen er ihren Wunschwagen lieber mit Aufschlag verkaufen würde. Die manuelle Recherche innerhalb von Mobile.de und Co. ist sehr bequem und lässt in der Regel keine Wünsche offen. Dennoch ist es für ehrgeizige Autohändler extrem aufwendig, sämtliche relevanten Plattformen Tag und Nacht unter Aufsicht zu behalten. Zudem sind die Seiten für jedermann zugänglich, was den Wettbewerbsvorteil des Händlers gegenüber anderen Suchenden verringert.

Einige Softwareentwickler nahmen sich dem Problem an und entwarfen Anwendungen, mit derer Hilfe sich sämtliche relevanten Plattformen automatisch durchsuchen ließen. Zu den erfolgreichsten Anwendungen in diesem Bereich gehören:

- Carsnooper der Firma IT Advanced GmbH
- Autoreader der Firma VirtualOffice GmbH
- Autodetektor der Firma Profiler 3D GmbH
- AutoradarX der Firma Pachaev & Zheltov OHG
- Professionelle Auto-Suche (PAS) von PAS GbR

Die entscheidenden Anforderungen an ein Programm zur automatischen Suche von Gebrauchtwagenanzeigen im Internet sind im Folgenden aufgelistet:

Abdeckung der wichtigsten Gebrauchtwagenplattformen: zumindest die beiden Marktriesen Mobile.de[5] sowie Autoscout24 sollten vom Programm unterstützt werden.

Plattformübergreifende Suche: Ein detaillierter Suchfilter muss genügen, um alle unterstützten Webseiten simultan zu durchsuchen.

Sofortige Meldung relevanter Neuinserate: Wer schon einmal ein Auto bei Mobile.de inseriert hat, weiß, dass zahlreiche telefonische Händleranfragen schon wenige Sekunden nach dem Hochladen der Angaben eingehen. Das Ziel der Entwickler ist hier, die Anzeigen bereits vor ihrer Verfügbarkeit auf der Seite runterzuladen, weil in dieser Phase auch der kleinste Vorsprung entscheidend sein kann.

Benachrichtigung auch über Email oder Handy: Wegen des hohen Mehrwerts für ihre Kunden unterstützen die meisten Programme auch die Benachrichtigung per Email und SMS, sobald ein relevantes Inserat veröffentlicht wurde.

Extraktion wichtiger Anzeigendetails: Sämtliche relevanten Anzeigedetails (Kfz-, Kontaktdaten, Fotos) werden von einem qualitativ hochwertigen Programm erkannt und als Objektmerkmale gespeichert.

Vielfältige Eingrenzungs- und Sortiermöglichkeiten: Anhand der erkennbaren Merkmale kann die Suche vielfältig variiert werden.

Klassifizierung nach weiteren, auf den Plattformen nicht vorgesehenen Merkmalen: Einige der Programme gehen bei der Suchanfrage über die auf den Plattformen angebotenen Suchkriterien hinaus (z. B. Händler/Privat Unterscheidung oder Freitextsuche).

Bereitstellung regelmäßiger Updates sowie technischen Supports: Die Programme müssen laufend an die unterstützten Webseiten angepasst werden. Regelmäßige Updates sind hier besonders wichtig, um eine fehlerfreie Anwendung der Software zu gewährleisten.

Derartige Anwendungen sind relativ jung und können aufgrund ihrer strikten Spezialisierung und geringen Interpretationsfähigkeit zur Gruppe der Crawling-Programme gezählt werden. Die zuvor definierten Daten werden lediglich ausgewählt, extrahiert und in die Software integriert. Die für ein echtes Web Mining-Anwendungen erforderliche Interpretation der Resultate findet nur bedingt statt. Dabei ließen sich insbesondere Techniken des Text Mining und des NLP nutzen, um relevante Informationen aus Freitext-Inseraten zu extrahieren.

Deutlich textlastigere Inserate wie z. B. Anzeigen für Immobilien, Antiquitäten, Restposten, etc. könnten auf diesen Weise erkannt und durchsuchbar gemacht werden.

Web Content Mining könnte im Zusammenhang mit Gebrauchwagenbörsen auch verwendet werden, um relevante Preisänderungen oder Trends auf dem Gebrauchtwagenmarkt zu erforschen. Händler, Hersteller, Gutachter oder Automobilclubs sind i. d. R. sehr an aussagekräftigen Statistiken und Studien über ihre Kunden oder den Markt interessiert. Die vielen Kontaktdaten von privaten Autoverkäufern ließen sich von Autohändlern nutzen, um potenzielle Neukunden für ein neues Fahrzeug ausfindig zu machen. Aus den Informationen zum alten Fahrzeug könnten sogar Vorlieben und Wünsche der Person abgeleitet werden.

3.3 Allgemeine Risiken von Web Content Mining

Im folgenden Abschnitt werden zunächst rechtliche Risiken von WCM betrachtet. Darüber hinaus erfolgt die Untersuchung von Risiken aus der Konfrontation mit Seitenbetreibern sowie möglicher Einflüsse auf die Reputation von Unternehmen, die WCM betreiben.

Die Erhebung und Verarbeitung personenbezogener Daten ist in Deutschland durch das Recht auf informationelle Selbstbestimmung des Einzelnen eingeschränkt. Es ist das Grundrecht jedes Bürgers, selbst über die Preisgabe und Verwendung seiner persönlichen Informationen zu bestimmen. Das informationelle Selbstbestimmungsrecht ist im Grundgesetz nicht explizit erwähnt, wurde jedoch 1984 im so genannten Volkszählungsurteil vom Bundesverfassungsgericht als Ausprägung des allgemeinen Persönlichkeitsrechts (Art. 2 Abs. 1 i.V.m. Art. 1 Abs. 1 GG) anerkannt (vgl. Buchner 2006, S. 208-209). Die Erfassung und Verwendung von Personendaten ist für Unternehmen demnach ebenso lukrativ wie datenschutzrechtlich heikel. Es ist wichtig zu wissen, welche Daten verwendet werden dürfen und welche nicht, um etwaigen Imageschäden oder Klagen präventiv vorzubeugen.

[5]Mobile.de gehört mittlerweile zu Ebay Inc

Das Wissen wird relevant, wenn die Inanspruchnahme von Daten, die geeignet sind, eine Person zu bestimmen oder Rückschlüsse auf sie zu ziehen z. B. Name, Wohnort, Religion, Email, etc., über die reine Erfüllung des Vertragsverhältnisses hinausgeht. Die Erfassung dieser personenbezogenen Daten ist ausschließlich mit Zustimmung dieser Person zulässig. Bezogen auf die betrieblichen Anwendungen des WCM betrifft dies vor allem die Extraktion von Kundeninformationen aus sozialen Netzwerken, Blogs oder Handelsplattformen.

Ein Kompromiss zwischen den Marketinginteressen und der gewünschten Anonymität ist die Erstellung von Profilen mit Pseudonymen, wenn ein Rückschluss auf die Person nicht möglich ist (vgl. hierzu und zum Folgenden Koitz 2002, S. 285). In diesem Fall hat der Nutzer ein Widerspruchsrecht gegen die Verwendung seiner Daten, über die er zu unterrichten ist.

Die Datenklau-Affären der letzten Jahre veranlasste die Bundesparteien zur Ausweitung der individuellen informationellen Freiheit. Dies betrifft vor allem den Handel und die Weitergabe von personenbezogenen Daten. Brigitte Zypries (SPD) sagte im Interview mit dem Manager-Magazin „es muss geprüft werden, ob für die Übermittlung von Daten künftig eine Einwilligungserklärung vorliegen müsse. Zweitens müssten betroffene Kunden zum Beispiel von Banken und anderen Unternehmen umgehend informiert werden“ (vgl. Manager-Magazin 2008, S. 1). Derzeit regelt § 28 Abs. 4 BDSG, dass der Kunde die Weitergabe seiner Daten nur mittels Widerspruch verhindern kann.

Neben dem Gesetzgeber oder dem Nutzer ist auch der jeweilige Seitenbetreiber in der Lage sich gegen Crawling-Versuche zu wehren.

Das soziale Netz StudiVZ verbietet derartige Aktionen durch die AGBs (Abs. 5.4.2): „Jede Nutzung, die darauf abzielt, das StudiVZ-Netzwerk, über dieses zur Verfügung gestellte Anwendungen oder zugänglich gemachte Inhalte geschäftsmäßig, gewerblich oder sonstig kommerziell zu verwenden, ist untersagt“. Konkretisiert wird das Crawling-Verbot durch Abs. 5.4.3: „Untersagt ist insoweit auch der Einsatz von Computerprogrammen zum automatischen Auslesen von Daten, wie z.B. Crawlern, Spider, Robots“. Um zu verhindern, dass automatisierte Anwendungen, die Struktur oder die Inhalte einer Plattform auslesen können, wird oft das so genannte CAPTCHA verwendet. Die Idee hinter dem “Completely Automated Public Turing Test To Tell Computers and Humans Apart” ist, durch die geforderte Interpretation eines Bildes einen menschlichen Nutzer von einem Computer unterscheiden zu können. Die CAPTCHA-Abfragen erscheinen bei der Nutzung einer geschützten Seite regelmäßig auf dem Bildschirm und fordern den Nutzer bspw. auf, eine kaum lesbare Zeichenkette zu entziffern und einzugeben. Solange das CAPTCHA nicht richtig beantwortet wird, kann der Account nicht weiter genutzt werden. Mögliche Lösungsansätze im Kampf gegen CAPTCHA-Abfragen sind leistungsfähige OCR-Programme oder die Entlohnung von CAPTCHA-beantwortenden Menschen.

Abbildung 2: Captcha-Abfrage bei StudiVZ

Abbildung 3:Alternative Captcha-Abfrage

Als weiteres Risiko von Web Content Mining sollten manipulierte Blog- und Forenaktivitäten genannt werden. Mitbewerber und sonstige Missgönner sind theoretisch in der Lage, eine hinreichend große Anzahl von Seiten mit irreführenden Beiträgen und Kommentaren zu füllen. Durch derartige Manipulationsversuche können Suchresultate verfälscht und im schlimmsten Fall sogar unbrauchbar gemacht werden.

3.4 Zwischenfazit

In diesem Kapitel konnten einige interessante Anwendungsgebiete von WCM identifiziert und ausführlich betrachtet werden. WCM birgt ein hohes Potenzial in der Marktforschung. Gegenüber klassischen Marktforschungsmethoden weist die automatische Extraktion relevanter Daten aus sozialen Netzwerken, Foren oder Blogs viele Vorteile auf. Die Überwachung von Preisen, Angeboten und Aktivitäten der Konkurrenz können mit WCM automatisch erfolgen. Die gewonnenen Informationen können anschließend in konkrete Handlungsempfehlungen transformiert werden. Im Zusammenhang mit Internet-Suchdiensten eröffnet WCM eine neue Sicht auf die zugänglichen Webressourcen. Neben der Suche nach Webdokumenten kann WCM dazu beitragen, konkrete Inhalte aus den verfügbaren Daten zu extrahieren und bedarfsgerecht darzustellen. In diesem Zusammenhang konnten einige Schwächen derzeit gängiger Suchmaschinen festgestellt und den Stärken von neuartigen Suchdiensten gegenübergestellt werden. Eine ganz konkrete Nachfrage nach leistungsfähigen WCM-Anwendungen besteht bereits jetzt auf zahlreichen Märkten, die sich durch hohen Bedarf an aktuellen Marktinformationen auszeichnen. Um unterschiedliche Märkte abzudecken, wurde im Abschnitt 3.2 der Aktien- und der Gebrauchtwagenmarkt analysiert.

4 Analyse und Vergleich von Software-Lösungen unterschiedlicher Anbieter

Nachdem im vorigen Kapitel eine Vielzahl von teilweise futuristischen Anwendungsgebieten des Web Content Mining diskutiert wurden, soll im Folgenden untersucht werden, wozu aktuelle WCM-Software tatsächlich in der Lage ist. Der Fokus der Untersuchung liegt auf Standardsoftware, die im Internet erworben, zum Teil auch als Freeware erhalten werden kann. Unterschiedliche WCM-Anwendungen werden im Rahmen der Analyse einem standardisierten Alltagstest unterzogen. Das Ziel der Untersuchung ist, zu erfahren, welche Programme konkreten betrieblichen Anforderungen standhalten kann und welchen Mehrwert sie einem Unternehmen bringen.

4.1 Kriterien zur Beurteilung von Softwarequalität

Die Qualität einer Software lässt sich wissenschaftlich korrekt, gemäß der ehemaligen DIN-Norm 66272 anhand der folgenden sechs Qualitätsmerkmale bestimmen:

Tabelle 3: Qualitätskriterien von Software-Anwendungen gem. DIN 66272 im Bezug auf den Untersuchungsgegenstand

Qualitätskriterien	Untersuchungsgegenstand
Funktionalität	Es wird überprüft, welche Methoden und Verfahren des Web Content Mining unterstützt werden und inwieweit sie ausführbar sind
Zuverlässigkeit	In diesen Zusammenhang spielen vor allem die Verfügbarkeit des Programms und Korrektheit der Ergebnisse eine übergeordnete Rolle
Benutzbarkeit	Ein Augenmerk der Untersuchung liegt auf der Benutzerfreundlichkeit der Bedienung, Dokumentation und Ergonomie
Effizienz	Wie schnell liegen Ergebnisse vor und welche technischen und humanen Ressourcen werden benötig
Wartbarkeit und Übertragbarkeit	Die Fragen der Änderbarkeit und Übertragbarkeit spielen in vorliegender Untersuchung eine untergeordnete Rolle

Quelle: In Anlehnung an DIN 66272

Die Qualität von Internet-Suchprogrammen lässt sich am besten mit Hilfe konkreter Suchanfragen bestimmen. Ein gutes WCM-Programm sollte nach einer angemessenen Trainings- und Vorbereitungsphase in der Lage sein, Webseiten automatisch nach konkreten Informationen durchzusuchen und die ausgewerteten Ergebnisse bedarfsgerecht darzustellen. Entscheidend ist, dass diese Anwendungen, die sich Web Content Mining groß auf den Werbebanner geschrieben, neben der Datenextraktion und Mustererkennung auch eine bedarfsgerechte Interpretationsleistung vollbringen. Im Rahmen der Untersuchung sollen ganz konkrete Suchanfragen aus unterschiedlichen Bereichen und Quellen beantwortet werden.

Tabelle 4: Suchszenarien für die Untersuchung von WCM-Software

Suchszenario	Forschungsfrage	Quellen	Darstellung
Nachrichtensuche	Welche Wahlversprechen und Argumente stehen im USA Wahlkampf 2008 einander gegenüber?	Nachrichtenportale wie Spiegel.de, Reuters.de, Sueddeutsche.de, Handelsblatt.com, Ft.de, etc.	Eine integrierte Liste der Aussagen beider Kandidaten, die sich vielfältig sortieren und kategorisieren lässt.
Wettbewerbsanalyse	Wer konkurriert auf dem Onlinepoker-Markt und womit zeichnen sich die einzelnen Anbieter aus?	Google-Resultate sowie spezialisierte Foren, Blogs, etc.	Auflistung aller Anbieter mit ihren Besonderheiten. Beobachtung konkurrierender Angebote und Trends.
Produktbeschaffung	Wo finde ich einen maximal 5 Jahre alten VW T5 und was wird er ungefähr kosten?	Einschlägige Kfz-Börsen und Ebay.	Passende Angebote mit Angaben zum Fahrzeug und Verkäufer. Ggf. Kritiken und Meinungen zum Fahrzeug.
Meinungsforschung	Welche Filme und Bücher werden von den Nutzern favorisiert? Welche Hobbys und sonstige Vorlieben und Abneigungen sind den Profilen zu entnehmen?	Soziale Netzwerke: StudiVZ, Myspace, etc. Spezialisierte Seiten: Kinonews.de, cineastentreff.de, Zelluloid.de, Google-Resultate zum Thema.	Eine Zusammenfassende Meinung, ohne weitere Internetseiten besuchen zu müssen.

Quelle: Eigene Erstellung

Die Suchszenarien orientieren sich dabei an den im dritten Kapitel erörterten, betrieblichen Anwendungsgebieten des WCM. Die Anfragen erfordern zusammengefasste Ergebnisse aus unterschiedlichen unstrukturierten Datenbeständen des Internets. Diese Ergebnisse sollten vielfältig sortiert und kategorisiert werden können. Neben der reinen Datenextraktion muss hinter dem Programm eine, wenn auch geringe, Intelligenz erkennbar sein.

4.2 Untersuchung von Web Content Mining-Software

Acht unterschiedliche WCM-Anwendungen werden im Verlauf des folgenden Abschnitts auf Ihre Funktionen und Fähigkeiten getestet. Es ist das Ziel, alle vorliegenden Programme zu erlernen, um sie auf zuvor definierte Suchszenarien anwenden zu können. Anschließend erfolgt eine Gesamtbewertung und Gegenüberstellung aller getesteten Programme.

4.2.1 Web Content Extractor 3.1, Newprosoft

Web Content Extractor (WCE) 3.1 ist eine WCM-Software aus dem Hause Newprosoft. Die Lizenz von WCE kostet $129, wobei für die vorliegende Untersuchung eine Trial Version genutzt wurde. Mit Ausnahme einer Begrenzung auf maximal 150 Datensätze innerhalb einer Suche bietet die Demoversion den kompletten Funktionsumfang einer Vollversion. Newprosoft beschreibt seine Anwendung folgendermaßen: „Web Content Extractor harvests data from targeted sites automatically and delivers it with a touch of a button, just the way you wish! Web Content Extractor offers you a friendly, wizard-driven interface that will walk you through the process of building a data extraction pattern and creating crawling rules in a simple point-and-click manner. Not a single string of code is required! Web data extraction is completely automatic" (http://www.newprosoft.com).

Erste Eindrücke von Web Content Extracor

Die Installation verlief problemlos und sehr schnell. Es mussten keine Voreinstellungen vorgenommen werden, um das Programm zu starten. Es wird dringend empfohlen die drei angebotenen Tutorial-Videos anzuschauen. Denn be-

reits nach dieser halben Stunde ist die intuitive Bedienung von WCE verstanden und es kann sofort mit ersten Experimenten begonnen werden. WCE bietet einen Wizard, der die Vorbereitung zur Datenextraktion unterstützt und den Anwender durch die wichtigsten Schritte leitet.

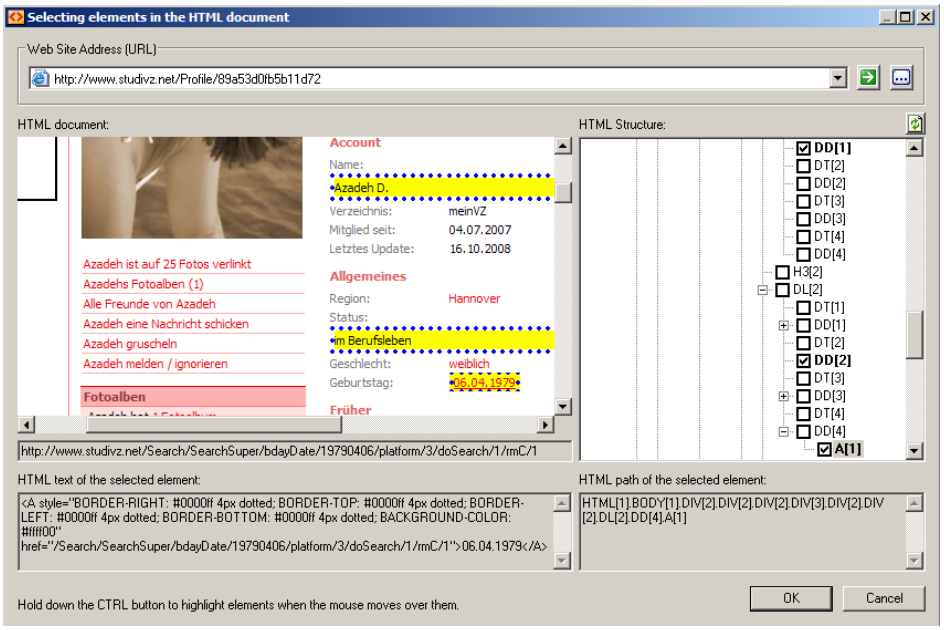


Abbildung 4: Auswahl relevanter Daten aus einem StudiVZ-Profil mit Hilfe von Web Content Extractor

Der erste entscheidende Schritt ist es, die erforderlichen Daten einer Webseite einmalig zu kennzeichnen. Hierfür muss ein exemplarischer Datensatz aufgerufen und die darin enthaltenden relevanten Daten markiert werden. Abbildung 5 zeigt, dass Name, Status und Geburtstag des StudiVZ-Mitglieds für die nachfolgende Extraktion ausgewählt wurden. Ein Klick auf das jeweilige Feld genügt dafür und es sind keine HTML-Kenntnisse erforderlich. Neben weiteren, weniger interessanten Arbeitsschritten wie Definition der Startseite und der Datenfelder erfolgt die entscheidende Konfiguration von Extraktions-Regeln. Im vorliegenden Fall soll das Programm die vielen Profile in der Gesamtansicht mit „>>“ durchblättern und dabei jedes Profil (URL: http://www.studivz.net/Profile/*) analysieren. Darüber hinaus konnten in diesem Schritt Ausschlusskriterien gebildet und die erforderliche Tiefe und Reihenfolge der Extraktion eingestellt werden. Bereits nach einem 15-minütegen Tutorial und weiteren zehn Minuten für die Definition der Suche konnten erste Ergebnisse abgerufen werden. Im vorliegenden Experiment hat die Extraktion auf Anhieb funktioniert. Im unteren Fenster des Screenshots, erscheinen die extrahierten Daten, wobei die Fotos unter Angabe der seiteninternen ID automatisch auf der Festplatte abgelegt werden.

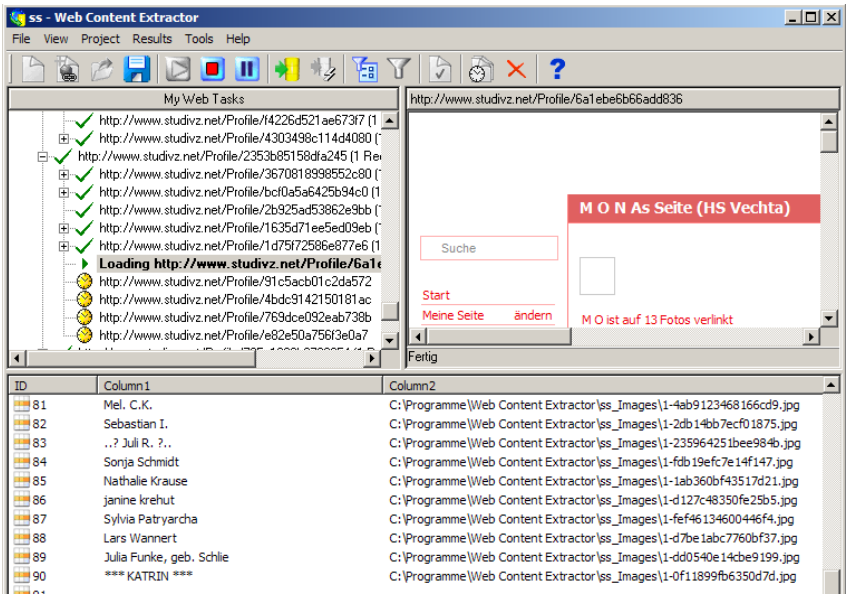


Abbildung 5: Web Content Extractor in Betrieb

Oben links sind die erfassten Dokumente zu sehen, die anhand ihrer Verlinkung in einer hierarchischen Struktur dargestellt sind. Oben rechts kann der Software beim Blättern der Dokumente in Echtzeit zugeguckt werden. Die Extraktion der 150 erlaubten Datensätze hat ca. sieben Minuten gedauert und verlief erfolgreich. Anschließend erlaubt das Programm die Daten in unterschiedlichen Formaten zu exportieren. Zur Wahl stehen Excel-, XML, HTML oder Textdateien sowie SQL und Access-Datenbanken.

Das Programm macht einen exzellenten ersten Eindruck. Bereits beim experimentieren konnte problemlos eine ganze Datenbank mit Informationen und Fotos von 150 StudiVZ-Mitgliedern runtergeladen werden. Nachdem WCE so einfach funktioniert, wird das Programm im Folgenden auf die Suchszenarien angewendet.

Suchszenarien mit Web Content Extractor

Ausgangspunkt für die Suche nach Nachrichten aus dem US-Wahlkampf ist das Auslandsarchiv von Spiegel-Online. Zunächst erfolgte die Definition der erforderlichen Daten innerhalb des Archivs. Abbildung 7 zeigt, wie die wesentlichen Informationen im Artikel markiert wurden.

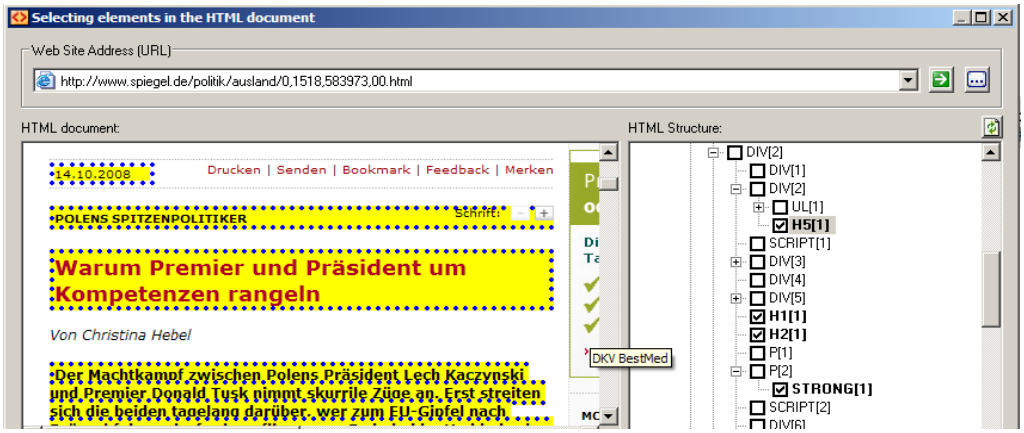


Abbildung 6: Definition der wesentlichen Inhalte eines Spiele-Online Artikels mit Web Content Extraktor

Im Folgenden sollen Erscheinungsdatum, Themengebiet, Überschrift und Einleitung des jeweiligen Artikels extrahiert werden. Der Suchagent wurde so konfiguriert, dass nur Artikel, deren Überschrift oder Einleitung Begriffe wie USA, Obama oder McCain enthalten gespeichert werden sollen. Darüber hinaus sollte das Programm dem Link „Weitere Artikel“ folgen, um das gesamte Archiv auslesen zu können. Externe und einige der internen URLs mussten im Verlauf der Untersuchung unterdrückt werden, um den Agenten von einer falschen Fährte abzubringen. Bereits nach ca. 30 Minuten war das Programm bereit, dem Spiegel-Online-Archiv die Nachrichten zum US-Wahlkampf automatisch entnehmen zu können. Nach etwa zehn Minuten war die Beschränkung von 150 Ergebnissen erreicht. Die 150 Ergebnisse enthielten zahlreiche leere oder unvollständige Datensätze, weil sich der Agent trotz der Feineinstellungen manchmal verwirrt hat. Dennoch konnten dem Spiegel-Archiv letztendlich über 50 absolut brauchbare Artikel zum gewünschten Thema entnommen werden. Abbildung 8 zeigt die Ergebnisse der Suche.

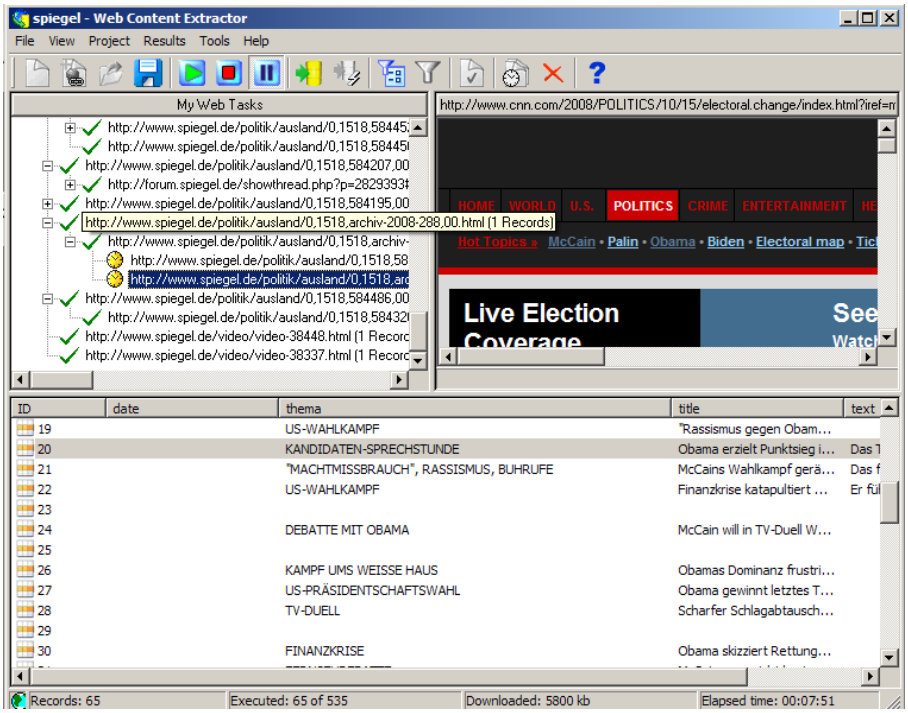


Abbildung 7: Extrahierte Nachrichten zum Thema US-Wahlkampf mit Hilfe von Web Content Extractor

Auch die Anwendung von WCE auf Online-Nachrichtenarchive von Handelsblatt und Financial Times hat problemlos funktioniert. Die voreingestellten Agenten können daraufhin gespeichert und regelmäßig mit der Suche nach relevanten Artikeln beauftragt werden.

Im Rahmen der Wettbewerbsanalyse sollten Onlinepoker-Anbieter gefunden, sowie deren speziellen Angebote und Besonderheiten einander gegenübergestellt werden. Der Identifikation der Onlinecasino-Betreiber dienen Google-Resultate zur Suchanfrage poker+online. WCE wurde dabei so eingestellt, dass die URL und Kurzbeschreibung des jeweiligen Anbieters gespeichert werden konnten. Abbildung 9 zeigt die markierten Felder innerhalb der Google-Resultate sowie einen Ausschnitt der erhaltenen Ergebnisse.

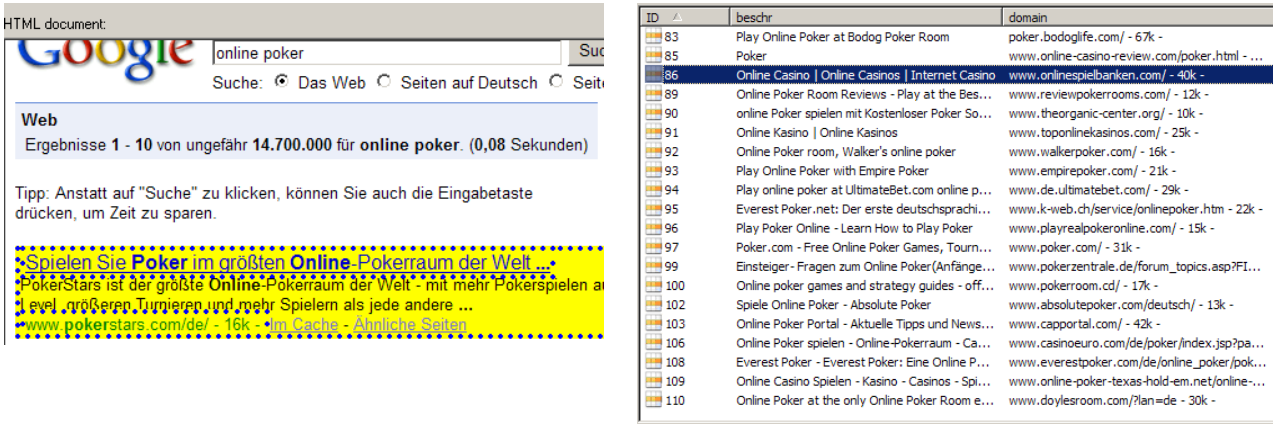


Abbildung 8: Konfiguration und Resultate im Rahmen der Identifikation von Onlinepoker-Anbietern mit Web Content Extracor 3.1

Nach Löschung von Duplikaten konnten mit dieser Methode über 100 unterschiedliche Anbieter identifiziert werden. Der Zeitaufwand hierfür betrug inklusive Voreinstellungen ca. 15 Minuten.

Darüber hinaus sollten weitere Informationen zu diesen Webseiten gefunden werden. Da die Strukturen der einzelnen Betreiberseiten sehr unterschiedlich sind, ist WCE nicht in der Lage ihnen direkt die nötigen Informationen zu entnehmen. Hierzu war es erforderlich solche Webseiten zu finden, die eine detaillierte und stetige Bewertung diverser Onlinepoker-Räume vornehmen. Die Seite http://www.pokerlistings.de/online-pokerraume beschäftigt sich genau mit dieser Frage und nimmt eine Echtzeitbewertung von insgesamt 29 Online-Pokerräumen auf Basis von Be-

fragungen und eigenen Analysen vor. Diese Rezensionen konnten mit Hilfe von WCE vollständig ausgelesen und in Tabellenform gebracht werden. Nachdem der Agent für die Extraktion von Rezensionen einmal konfiguriert ist, kann er regelmäßig auf die Webseite angewendet werden, damit die Daten stets aktuell bleiben.

ID	Column1	Column2	Column3	Column4
136	VW Transporter T5 Kasst...	9.900 E...		Kölner Landstr. 350 40589 DüsseldorfDeutschland ...
137	VW Transporter T5 ++Tr...	9.900 E...	VW T5 mit Tresorschließanlage Top Zus...	Frankfurterstraße 40 63571 Gelnhausen-RothDeut...
138	VW Transporter T5 7JL13...	9.999 E...	NETTO-Preis für Export Weißes Servic...	Details zum Anbieter
139	VW Transporter T5 ,1.9 ...	9.999 E...	Govorim : RUS;PL;CZ;SK;FR;NL; prix:1...	Details zum Anbieter
140	VW T5 Transporter Kaste...	9.999 E...	guter und sauberer Zustand!!! Angebo...	Eschersheimerstrasse 07 12099 BerlinDeutschl...
141	VW Transporter T5 TDI 6...	10.000 ...	Fahrzeug befindet sich in einem guten ...	Kirchbergring 12 74706 OsterburkenDeutschla...
142	VW Transporter T5 KLIM...	10.499 ...	NUR 2 SITZHE Vollständige Beschreibung	Goethering 66 63067 OffenbachDeutschland ...
143	VW T5 Transporter Kombi...	11.576 ...	Volkswagen T5 Transporter Kombi 9 Sit...	Banmolen 1 5768 ET Meijel - NiederlandeNieder...
144	VW Transporter T5 7JL13...	11.590 ...	Erste Hand, MwSt ausweisbar, Digialer...	Anton-Sommer-Str. 1 88046 FriedrichshafenDe...
145	VW Transporter T5 7JL13...	11.590 ...	VW T5 1.9 TDI Transporter, Servolenk...	Südstrasse 136-138 74072 HeilbronnDeutschland ...
146	VW Transporter T5 1,9T...	11.662 ...	ABS,AHK,ASR,RC,St.Heizung,1-Hand,...	Eichenstrasse 71 65933 Frankfurt/MainDeutsc...
147	VW Transporter T5 2.5 T...	11.675 ...	Zentralverriegelung mit Fernbedienung...	Nardter Weg 6 02977 HoyerswerdaDeutschlan...
148	VW Transporter T5 Lang ...	11.880 ...	T5 Lang --> 9 Sitzer, Kunstlederaussta...	Bahnhofsstr. 37 24143 KielDeutschland 1....
149	VW T5 Bus — EUR 11.900	11.900 ...	Volkswagen T 5 Bus, 8-Sitze, ABS, ASR...	An der Silbergrube 1 07551 GeraDeutschland ...
150	VW T5 Kasten — EUR 11....	11.900 ...	Anhängevorrichtung Flügeltüren V...	Veldener Strasse 12 84169 AltfraunhofenDeut...

Abbildung 9: Extraktion von relevanten Daten aus Kfz-Anzeigen von Mobile.de mit Hilfe von Web Content Extractor 3.1

Im Rahmen der Beschaffung eines VW T5 wurde WCE auf Mobile.de und Autoscout24.de angewendet. In beiden Fällen wurde das Programm beauftragt, Titel, Beschreibung, Preis, Kontaktdaten und Bild der jeweiligen Anzeige runterzuladen. Ausgangspunkt der Suche waren jeweils zuvor definierte Suchergebnisse mit relevanten T5-Anzeigen. Auf keiner der beiden Seiten traten Schwierigkeiten auf. Die 150 Ergebnisse waren nach jeweils 10 Minuten voll. Abbildung 10 zeigt die Ergebnisse der Datenextraktion von Mobile.de.

Bereits im Rahmen des ersten Eindrucks konnte gezeigt werden, dass WCE in der Lage ist, persönliche Informationen aus den Profilen von StudiVZ-Mitgliedern auszulesen. Um den Schwierigkeitsgrad zu steigern, sollten neben Name und Foto auch Vorlieben wie z. B. Lieblingsbuch, -film oder –musik gespeichert werden. Hierbei traten erhebliche Schwierigkeiten auf. Die Struktur der Profile variiert in Abhängigkeit davon, wie viele Informationen der Nutzer von sich offenbart hat. Damit befinden sich die markierten Felder an unterschiedlichen Stellen, was zu einer hohen Anzahl leerer oder unzutreffender Datenfelder führt. Dieses Problem ließ sich nach zahlreichen Versuchen leider nicht beheben. Schließlich waren nur ca. 25 Prozent der Resultate brauchbar. Zwar können auch diese 25 Prozent aussagekräftige Ergebnisse liefern, solange die Stichprobe hinreichend groß ist, doch auch hier macht StudiVZ dem Programm einen Strich durch die Rechnung. Nach ca. 100 Anfragen an den StudiVZ-Server schaltet sich eine Captcha-Abfrage ein. Damit ist der schnelle und automatische Suchvorgang vorerst beendet. Die Extraktion kann nur fortgesetzt werden, wenn das Captcha in einem parallelen Browserfenster manuell beantwortet wird. Dieser manuelle Vorgang ist sehr mühsam und macht die versprochene Schnelligkeit und den Automatismus von WCE zu Nichte.

Fazit zu Web Content Extractor 3.1

WCE ist eine sehr einfache und bedienerfreundliche WCM-Anwendung. Ohne große Vorbereitung konnte das Programm auf Anhieb für schwierige Aufgaben erfolgreich eingesetzt werden. Die Steuerung erfolgt ausschließlich über eine graphische Oberfläche und es sind keinerlei Programmierkenntnisse erforderlich. Die Resultate können in vielen unterschiedlichen Formaten exportiert oder im Programm zwischengespeichert werden. Der große Nachteil von WCE ist, dass er nur schwer mit heterogenen Dokumenten fertig wird. Um eine erfolgreiche Extraktion mit Hilfe von WCE zu gewährleisten, müssen sich die Daten stets an der gleichen Stelle jedes Dokuments befinden. Damit ist WCE für die Analyse unterschiedlicher Webseiten ungeeignet. Nur schwachstrukturierten Inhalte eines einheitlichen Portals oder einer Seite konnten mit Hilfe von WCE effizient durchsucht und extrahiert werden.

4.2.2 Mozenda Beta, Mozenda Inc.

Mozenda Beta aus dem Hause Mozenda Inc. ist die neueste aller getesteten WCM-Programme. Die Entwickler beschreiben ihre Software folgendermaßen: „Mozenda includes everything you need to create and manage a web data extraction project. Our simple point and click interface enables the most novice user the ability to get data quickly and easily from the web" (http://www.mozenda.com/mozenda-products.php). Mozenda ist an gewerbliche Nutzer wie z. B. Händler, IT- und Webentwickler gerichtet. Ihnen verspricht der Betreiber ein hohes Maß an Bedienerfreundlichkeit mit einer intuitiven grafischen Oberfläche, die sich ohne jegliche Programmierkenntnisse steuern

lässt. Der Internetauftritt von Mozenda Inc. (http://www.mozenda.com) beinhaltet eine Vielzahl zielgruppenorientierter Szenarien und Anwendungsmöglichkeiten. Diese sind multimedial aufbereitet und laden ein, die Software sofort auszuprobieren. Das Programm kann in einem Monats- oder Jahresabonnement für $39 bzw. $395 bezogen werden. Für die vorliegende Untersuchung wurde eine auf 30 Tage begrenzte Demoversion von Mozenda Beta genutzt.

Erste Eindrücke von Mozenda

Nach problemloser Registrierung und Installation der Software konnte bereits die erste Besonderheit von Mozenda erkannt werden. Um das Programm zu starten, muss sich der Anwender zunächst über seinen Webbrowser ins Back-End von Mozenda.com einloggen. Der Server des Betreibers und die Anwendung stehen in einer hohen Abhängigkeit zu einander. Die auf dem Rechner installierte Anwendung dient ausschließlich der Konfiguration und dem Training eines Agents. Alle Anfragen erfolgen vom Server des Betreibers und erfordern damit keine eigenen Internetressourcen. Auch die Speicherung mit dem Anschließenden Export der Daten erfolgt über die Administrationsoberfläche im Internet.

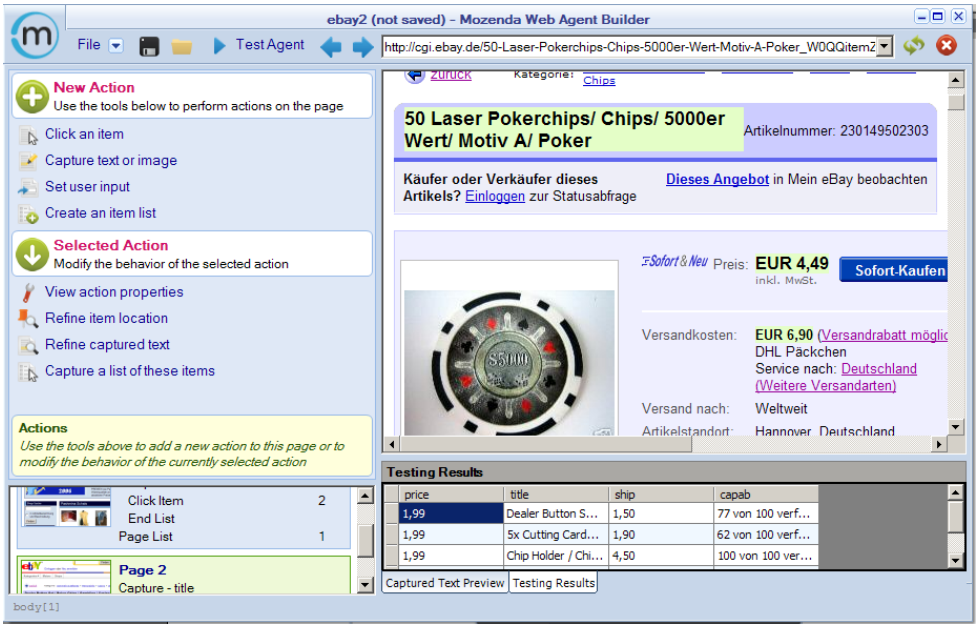


Abbildung 10: Bedienungsoberfläche von Mozenda Beta

Zur Einweisung in die Bedienung stellt der Betreiber eine Vielzahl von Videos sowie ein ausführliches Tutorial bereit. Die multimedialen Hilfsmittel sind sehr gut geeignet, um einem Neuanwender die wichtigsten Funktionen von Mozenda näher zu bringen. Abbildung 11 zeigt die Bedienungsoberfläche des installierten Programms, mit dessen Hilfe die Konfiguration der Agents erfolgt.

In der oberen linken Ecke sind die grundlegenden Funktionen zu sehen. Das Training des Agents erfolgt, indem diese Funktionen auf die im Browserfenster dargestellte Seite angewendet werden. Alle vom Nutzer ausgeführten Klicks und Markierungen werden vom Agent in ihrer Reihenfolge aufgezeichnet und schrittweise im unteren linken Feld Abb. 38 dargestellt. Ein einfaches Beispiel soll dieses intuitive Prinzip verdeutlichen.

Im vorliegenden Fall sollen konkrete Verkaufsaktivitäten bei Ebay ausgelesen werden. Mit dem Befehl „Set user input“ und dem entsprechenden Klick in das Texteingabefeld der Seite kann der Agent angewiesen werden, eine Identifikation oder eine Sucheingabe zu tätigen. Um den Inhalt der erschienenen Dokumente zu erfassen, müssen die relevanten Funde zunächst markiert werden. Die gewünschten Artikel können für den Agent kenntlich gemacht werden, indem eine „Item List“ kreiert wird. Abbildung 12 zeigt, wie die einzelnen Objekte einfach durch das Anklicken markiert werden.

Abbildung 11: Definition relevanter Artikel mit Mozenda Beta

Sollten alle Artikel untersucht werden, reicht es aus, den ersten und letzten anzuklicken. In einem weiteren Schritt kann der Agent auf Links wie z. B. „weiter“ oder „next“ hingewiesen werden, um alle Seiten der Suchergebnisse erfassen zu können. Daraufhin muss ein exemplarischer Artikel mit „Click an item“ geöffnet werden. Mit „Capture Text“ werden die relevanten Stellen markiert und benannt (siehe Abbildung 13)

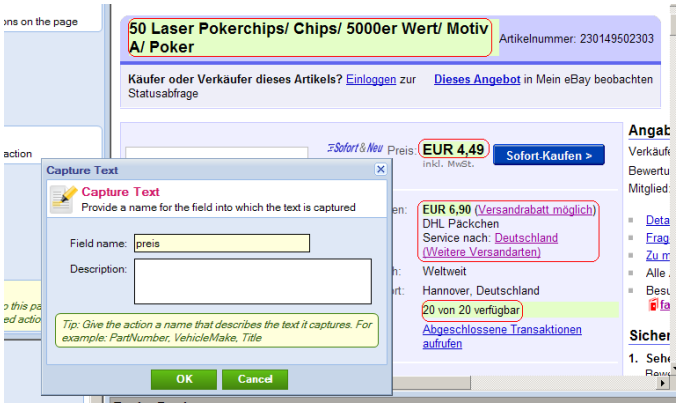


Abbildung 12: Markierung relevanter Datenfelder eines exemplarischen Artikel mit Mozenda Beta

Für die vorliegende Aufgabe wäre die Konfiguration des Agents somit abgeschlossen. Zuletzt kann ein Testlauf durchgeführt werden, bei dem die einzelnen Schritte in Echtzeit simuliert und die markierten Daten in eine vorläufige Tabelle extrahiert werden (siehe Abbildung 14, Feld unten rechts). Wenn der Test erfolgreich verlief, kann der Agent gespeichert werden. Das Programm wird daraufhin geschlossen, weil die restlichen Schritte auf dem Mozenda-Server ablaufen.

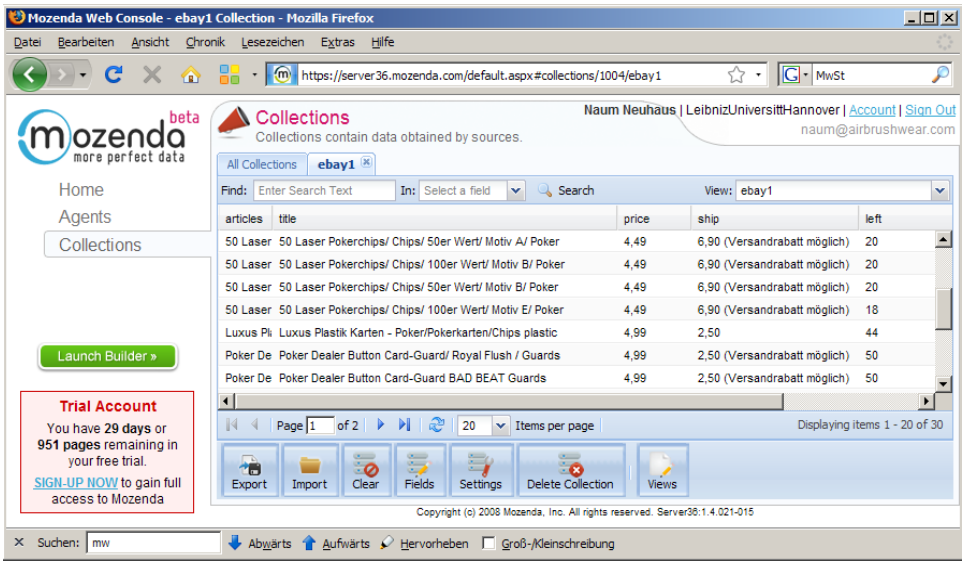


Abbildung 13: Extrahierte Resultate in der Online-Administrationsoberfläche von Mozenda Beta

In der Online-Administrationsoberfläche muss der Agent daraufhin ausgewählt und gestartet werden. Bereits nach kurzer Zeit gehen im Ordner „Collections“ (Abbildung 14) die Ergebnisse ein. Diese können nun betrachtet, verändert oder in CSV-, TSV- oder XML-Dateien exportiert werden. Auf den ersten Blick fällt der große Vorteil dieser

Serverbasierten WCM-Art auf. Damit wird der eigene Computer für die eigentliche Suche überflüssig. Der Agent kann dadurch programmiert werden, die Extraktion zu jeder Tages- oder Nachtzeit zu starten und die Suche beliebig oft wiederholen.

Die Screenshots der Beispielaufgabe stammen aus der Eingewöhnungsphase mit dem Programm. Tatsächlich konnten alle Artikelinformationen bereits beim ersten Durchlauf erfolgreich extrahiert und auf dem Server abgelegt werden. Dies geschah bereits nach etwa 15 Minuten. Dieser großartige erste Eindruck lädt ein, Mozenda Beta an den standardisierten Suchszenarien auszuprobieren.

Suchszenarien mit Mozenda Beta

Im Rahmen der Suche nach aktuellen Nachrichten aus dem US-Wahlkampf soll Mozenda darauf eingestellt werden, Spiegel-Online, Handelsblatt und Financial Times regelmäßig nach relevanten Meldungen zu durchforsten. Hierzu hat sich die Funktion „Set user input“ als besonders sinnvoll herausgestellt. Damit lassen sich Suchbegriffe eingeben oder ggf. nötige Identifikationen bzw. Autorisierungen durchführen. Als Ausgangspunkt der Suche nach neuen und relevanten Beiträgen dient in jedem der drei Nachrichtenportale die jeweilige Startseite. Mit einem Klick auf das Suchbegrifffeld und der Eingabe „obama mccain“ ist der erste Schritt des Agents eingestellt.

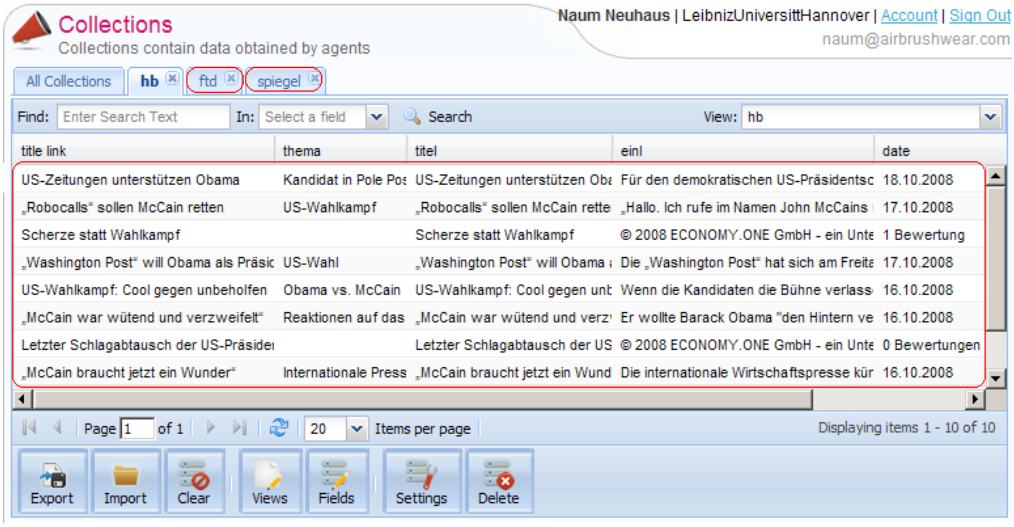


Abbildung 14: Resultate der Nachrichtenextraktion aus Spiegel-Online, Handelsblatt und Financial Times mit Mozenda Beta

Mit „Create an item list“ werden daraufhin die verlinkten Titel der Beiträge ausgewählt. Obwohl sich theoretisch alle erschienen Nachrichten extrahieren ließen, beschränkte sich die Extraktion nur auf die ersten bzw. aktuellsten Resultate.

Nachdem schließlich ein exemplarischer Beitrag geöffnet und die nötigen Daten (Titel, Datum, Einleitung) markiert wurden, war das Training des Agents abgeschlossen. Nach jeder fertigen Agent-„Ausbildung“ empfiehlt es sich einen Test durchzuführen, um etwaige Fehler schon im Vorfeld identifizieren und ausräumen zu können. Dieses Prozedere konnte innerhalb von 20 Minuten auf alle drei Nachrichtenportale angewendet werden. Schon fünf Minuten nach dem Start des Agents standen alle extrahierten Daten zur Verfügung. Alle Spiegel Artikel konnten fehlerfrei entnommen werden. Einige der Beiträge von Handelsblatt oder FTD enthielten Videostreams anstatt von Text. Diese Dokumente konnten erfolgreich unterdrückt, nicht aber gespeichert oder erkannt werden. Der Export erfolgt ausschließlich auf die Festplatte des eingeloggten Rechners. Die Dokumente können aus dem Programm nicht automatisch auf einen Webserver geladen oder per Email verschickt werden. Besonders praktisch wäre das in Verbindung mit der Schedule-Funktion, die benötigt wird, wenn der Agent regelmäßig oder zeitversetzt ausgeführt werden soll.

Für das vorliegende Experiment wurde eingestellt, dass alle drei Agents täglich gestartet werden sollen. Exakt zur voreingestellten Uhrzeit wurden auch am kommenden Tag neue Beiträge von allen drei Seiten erfolgreich auf den Server geladen. Es wäre alternativ möglich gewesen, alle drei Agents in einem einzigen zu verbinden, der ihm bekannte Seiten sukzessive durchsucht.

Um möglichst viele Anbieter von Online-Pokerräumen zu identifizieren, wurden zunächst Google-Resultate zu den Begriffen „Online Poker“ nach Domains mit Mozenda durchforstet. Die Anzahl von falschen Seiten und Duplikaten war, wie auch zuvor beim Web Content Extraktor, sehr hoch. Das Überspringen von doppelten und ähnlichen (po-

kerstars.de und pokerstars.com/de) Resultaten kann im Programm nicht eingestellt werden. Die nötige Präzision kann nur durch die Variation der Suchanfragen bei Google und in der anschließenden manuellen oder automatischen Datenauswertung außerhalb des Programms erfolgen. Sehr praktisch ist, dass auch die Variation der Anfragen hinter die erste Suche geschaltet werden kann. So können beispielsweise nach Abschluss der Analyse von Ergebnissen zu „Poker Online“ viele weitere Suchanfragen getätigt und anschließend analysiert werden. Mit diesem Verfahren wurde zwar eine Menge Fehldaten extrahiert, es kann aber zumindest von einer annähernden Vollständigkeit ausgegangen werden. Zudem stehen die Rohdaten bereits nach wenigen Minuten auf dem Server verfügbar.

Nach nur einem Durchlauf und der anschließenden groben manuellen Säuberung der Daten konnten über 300 relevante Domains identifiziert werden. Wie auch zuvor beim Web Content Extraktor, können diese Seiten leider auch mit Mozenda nicht weiter automatisch analysiert werden, weil deren Strukturen sehr unterschiedlich sind. Der Agent muss für jede dieser Seiten einzeln voreingestellt werden. Eine exemplarische Onlinepokerseite wurde gewählt, um daraus künftig die aktuellen Angebote und Events entnehmen zu können. Dies erwies sich als eine einfache Aufgabe und der Agent war bereits nach wenigen Klicks bereit. Sollte sich jemand die Mühe machen, Mozenda auf alle Konkurrenzseiten einzustellen, hätte er stets eine aktuelle und verlässliche Informationsbasis bezüglich ihrer Aktivitäten. Wie auch bereits mit WCE zuvor, konnte das Bewertungsportal http://www.pokerlistings.de problemlos nach Beiträgen durchsucht werden. Auch Blogs und andere spezialisierte Foren konnten ganz einfach ausgelesen werden. Mit Mozenda stehen dem Anwender somit stets aktuelle Kommentare, Bewertungen, Angebote und Events bekannter Online-Pokerräume in Tabellenform zur Verfügung.

Die Anwendung von Mozenda auf Kfz-Onlinebörsen und Ebay bei der Suche nach Informationen zum gewünschten Fahrzeug verlief problemlos. Alle nötigen Informationen sowie das Foto jedes Autos konnten erfolgreich runtergeladen werden. Leider unterstützte Mozenda keine detaillierten Anfragen auf Mobile.de und Autoscout24.de, weil das Programm mit Dropdown-Menüs nicht zurechtkommt. Daher mussten die potenziellen Fahrzeuge zunächst mit der seiteneigenen Suchfunktion aufgerufen und erst damit für den Agent zugänglich gemacht werden.

Zuletzt wurde Mozenda an sozialen Netzwerken ausprobiert. Die Profile von StudiVZ-Mitgliedern zu crawlen, ist auch mit Mozenda eine große Herausforderung. Besonders schwierig war es, konkrete persönliche Daten wie z B. Lieblingsbücher, Hobbys etc. auszulesen. Diese befinden sich stets an unterschiedlichen Stellen und konnten daher vom Programm nicht erkannt werden. Es war möglich die gesamten Profiltexte zu extrahieren, allerdings sind sie dadurch nicht strukturierter als zuvor. Kein Problem waren hingegen allgemeine Daten wie Name, Geburtstag, Mitgliedschaftsdauer sowie Status, Geschlecht und Hochschule. Diese stehen im Profil fast immer an der gleichen Stelle und waren für Mozenda einfach zu finden. Zudem lassen sich Anzahl der Freunde und Fotoalben sowie für besonders eifersüchtige Freunde die hinterlassenen Kommentare samt Absender extrahieren. Wie auch schon zuvor, stellt die begrenzte Anzahl von Anfragen an den StudiVZ-Server mit dem darauf folgenden Captcha auch für Mozenda ein unlösbares Problem dar.

Fazit zu Mozenda Beta

Mozenda ist besonders einfach zu bedienen und dennoch sehr fortschrittlich und leistungsfähig. Wie bereits oft erwähnt, die Konfiguration der Agents gelang immer auf Anhieb und ein anschließender Test zeigte sofort, ob die Einstellungen korrekt waren. Das Tutorial und die dazugehörigen Beispiel-Streams sind absolut ausrechend, um bereits nach wenigen Versuchen erste eigene Agents auszubilden und zu starten. Die Idee, sämtliche Anfrage- und Speicheraktivitäten ausschließlich vom Mozenda-Server zu tätigen, eröffnet im Vergleich zu anderen WCM-Anwendungen ganz neue Perspektiven. Nur weil das Mining zeitversetzt oder planmäßig erfolgen kann, lohnt es sich am Beispiel von Online-Pokerräumen tatsächlich, einen Agent pro Konkurrenzseite zu erstellen, der ohne weiteren Zeitaufwand regelmäßig die benötigten Informationen liefert. Leider verfügt das Programm über keine „Intelligenz“, was die anschließende Analyse und Interpretation von Resultaten anbetrifft. Mozenda eignet sich nicht, um bspw. Texten einen Sinn zu entnehmen oder diese zu kategorisieren. Es ist leider auch nicht möglich aus einer Reihe homogen strukturierter Dokumente nur diejenigen auszulesen, die über bestimmte Eigenschaften (Begriffe, URL, etc.) verfügen. Der Entwickler löste dieses Problem, indem er erlaubte die seiteneigenen Suchfunktionen mit Mozenda steuern und nutzen zu lassen. Der Gesamteindruck der Software ist sehr positiv. Dafür, dass es sich hier um eine ganz frische Beta Version handelt, wirkt Mozenda sehr ausgereift.

4.2.3 Surf3D Pro, Navagent

Surf3D Pro ist eine Beta-Version der Firma Navagent, die sich auf Visualisierung von Suchergebnissen, Webseiten und Inhalten spezialisiert. Der Anbieter beschreibt das Programm folgendermaßen: “We provide professional and consumer solutions for time-saving personalized web browsing, smart web crawling, visual web mining, site pre-

viewing, event detection, custom alerts, and content visualization and retrieval. Our products reduce search time by over 80% in comparison to what it normally takes you to click through and evaluate search engine results“. Das letzte Release von Surf3D Pro stammt aus dem Jahr 2002. Die genutzte Version ist eine frei erhältliche Freeware.

Erste Eindrücke von Surf3D Pro

Nach einer schnellen und problemlosen Installation öffnet sich das Hauptfenster von Surf3D Pro. Die Ansicht überrascht durch ihre Einfachheit und Übersichtlichkeit. Mit insgesamt 15 grundlegenden Befehlen und sieben unterschiedlichen Visualisierungsarten macht die Menüführung einen sehr geordneten und transparenten ersten Eindruck.

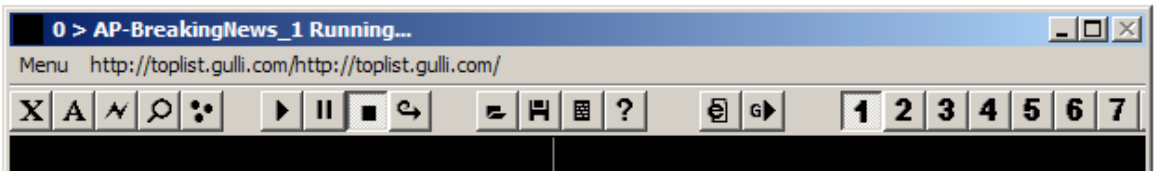


Abbildung 15: Menüführung von Surf3D Pro

Surf3D Pro funktioniert auf Basis von Agents. Jede Suche beginnt daher mit der Einstellung und dem Training eines solchen. Hierfür steht dem Anwender eine Suchmaske zur Verfügung, in der die Suche präzisiert und gesteuert werden kann.

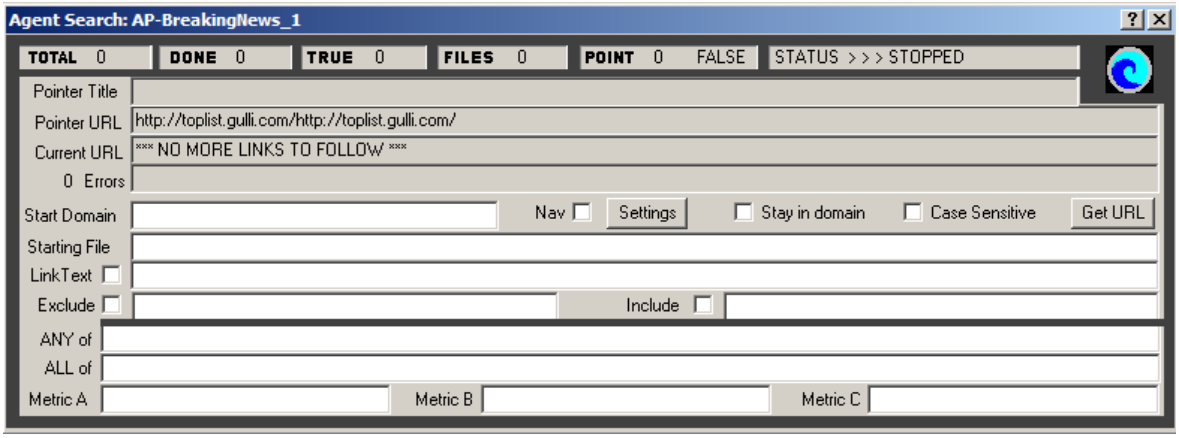
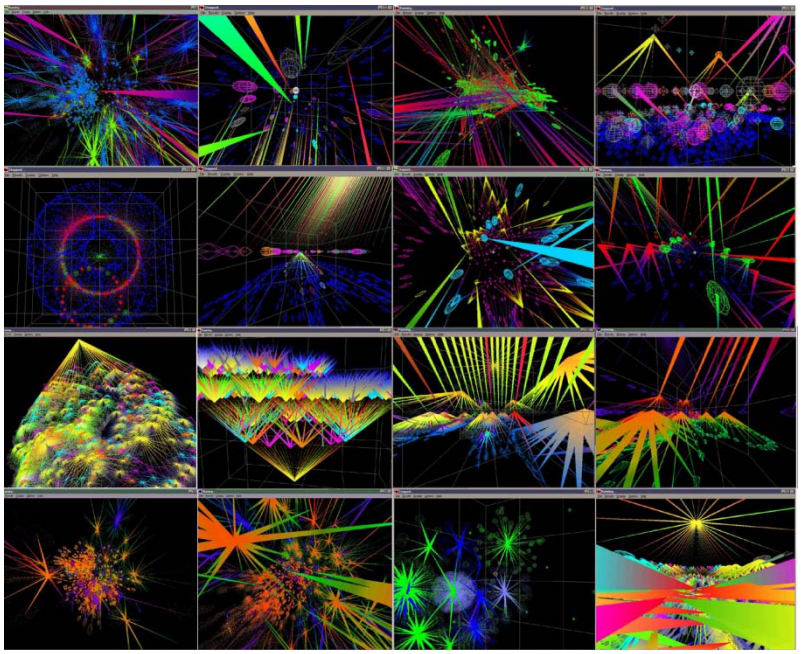


Abbildung 16: Konfiguration eines Suchagenten mit Surf3D Pro

Das Suchfeld bietet die Möglichkeit, eine geeignete Quelle sowie Ein- und Ausschlusskriterien anzugeben. Zudem werden hier mit Hilfe des booleschen Operators die Suchbegriffe vorgegeben. Der Software liegen praktischerweise 40 voreingestellte Agents bei. Zwar sind einige der Quell-URLs bereits veraltet, dennoch bieten Sie einem unerfahrenen Anwender die Möglichkeit, sich an den Voreinstellungen zu orientieren.

Im Mittelpunkt der Anwendung steht die Visualisierung der Resultate in einem euklidischen Raum. Hierfür müssen unter Metric A, B, C mögliche Kategorisierungsvorschläge durch Begriffe vorgegeben werden. Die Kategorien entsprechen im Folgenden den drei Achsen des Raumes, indem die Ergebnisse abgebildet werden. Es kann frei zwischen sieben unterschiedlichen Visualisierungsarten gewählt werden. Die Visualisierungseffekte sind farbenprächtig, erinnern aber an alte Computerspiele. Mit Ausnahme künstlerischer Eindrücke kann ihnen auf den ersten Blick keine sinnvolle Information entnommen werden.

Abbildung 17: Visualisierung der Suchergebnisse mit Hilfe von Surf3D Pro



Quelle: http://www.navagent.com/products/gallery/animation/

Abbildung 18 zeigt einige ausgewählte Visualisierungseffekte. Bereits bei den ersten Sucheingaben mit Surf3D Pro konnten beeindruckende Strukturen zum Vorschein gebracht werden. Nach kurzer Suche werden die meisten Darstellungen jedoch sehr unübersichtlich. Die veraltete Grafik und die fehlende Möglichkeit, innerhalb der Gebilde frei zu navigieren, um einzelne Strukturen genauer anzusehen erschwert die Interpretation erheblich.

Suchszenarien mit Surf3D Pro

Für die Suche nach Nachrichten stellt das Programm mehrere voreingestellte Agents bereit. Mit Hilfe dieser Agents wurde das Programm mit der Suche nach Argumenten im US-Wahlkampf beauftragt. Als Startdomain diente das Nachrichtenportal von Spiegel Online. Es wurden einschlägige Suchbegriffe zum diesem Thema verwendet (USA demokraten republikaner obama mccain wahlen, etc.). Zur Kategorisierung und Abbildung der Resultate im dreidimensionalen Raum konnten nach mehreren Anläufen die Bereiche Innenpolitik (innenpolitik steuern abgaben innenminister, etc.), Außenpolitik (aussenpolitik iraq afghanistan al qaida terrorismus bin laden osama krieg bagdad, etc) und die aktuelle Wirtschaftskrise (finanzkrise banken zinsen fonds zertifikate börse aktien etc.) als sinnvoll identifiziert werden. Bei einer sinnvollen Klasseneinteilung decken die Resultate alle drei Klassen ab, ohne eine davon leer zu lassen. Nach einigen Fehlanläufen konnte die Suche auf Spiegel Online schließlich gestartet werden. Dabei wurde das Programm besonders oft von der spiegeleigenen Suchfunktion und den vielen Texteinblendungen um die Artikel herum sehr irritiert. Die Suchfunktion und die mit ihr verbundenen URLs konnten mit der Zeit manuell unterdrückt werden. Die Tatsache aber, dass das Programm die eigentlichen Artikel nicht von den vielen Texten auf der Seite (Themenvorschläge, Auszüge aus anderen Artikeln, Werbung, etc.) unterscheiden konnte, hat sehr gestört. Das beste Suchergebnis lieferte nach etwa 15 Minuten 499 Treffer und acht verwirrende Animationen.

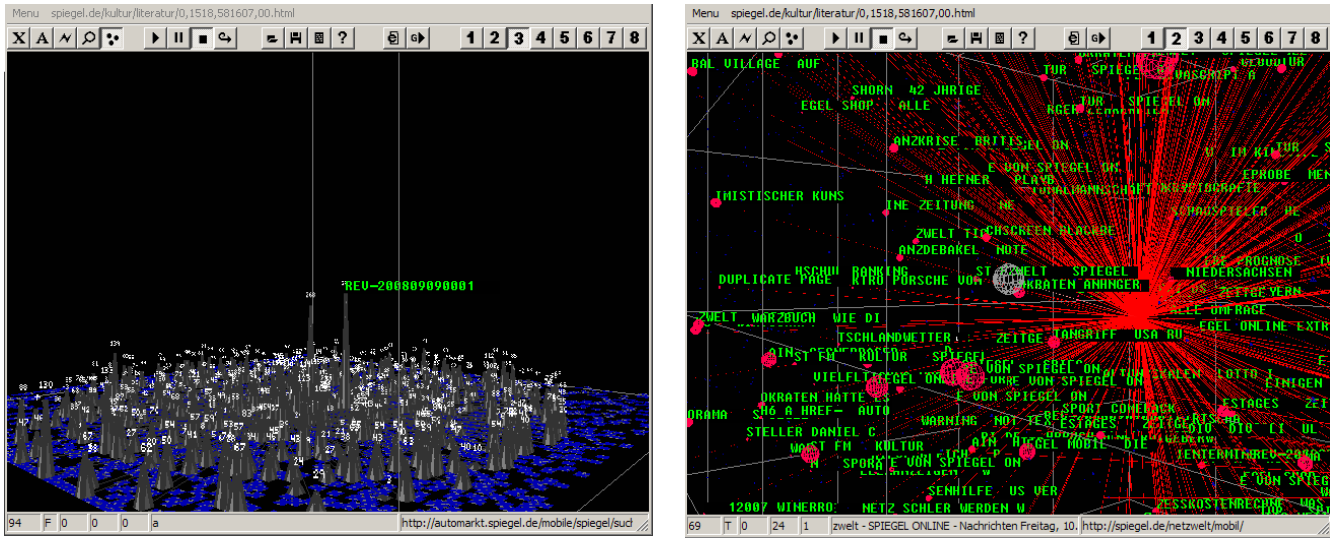


Abbildung 18: Visualisierung der Suchergebnisse auf Spiegel Online mit Hilfe von Surf3D Pro

Die versprochene einfache und intuitive visuelle Navigation innerhalb der Resultate kann nicht bestätigt werden. Zwar können die Objekte durch anklicken im Browserfenster angezeigt werden, eine Struktur oder ein Ranking der Objekte ist jedoch kaum zu finden.

Die Funktion „Result Report“ bietet eine Auflistung sämtlicher Ergebnisse, wobei unter „Results Format“ Einfluss auf das Ranking genommen werden kann. Hier kann die Gewichtung der Kategorien und die Darstellung der Ergebnisse eingestellt werden. Trotz vieler Durchläufe konnte zu keiner Zeit eine sinnvolle Anordnung der gefundenen Dokumente erreicht werden. Viele der gelisteten Objekte enthielten die Suchbegriffe nur am Rand oder als Link und waren damit irrelevant.

Fazit zu Surf3D Pro

Das Programm erlaubt zwar nur die Eingabe von einer Startdomain, möglich ist aber die parallele Suche in mehreren Fenstern. Eine Integration der Daten unterschiedlicher Suchresultate ist im Programm leider nicht vorgesehen. Die Anwendung des Programms auf weitere Nachrichtenportale (Handelsblatt, Financial Times, SZ) bestätigte die mangelhafte Sortierung und Kategorisierung der Ergebnisse. Lediglich die Suche innerhalb von Google-Ergebnissen zum Thema US-Wahlkampf 2008 führte bei relativ niedrigen Fehlerrate nach 45 Minuten zu einer Vielzahl zutreffender Dokumente (178 Treffer). Zwar konnten auch diese Resultate nur unzureichend kategorisiert werden, stellten aber im Vergleich zur Suche innerhalb von Spiegel Online und Co. einen Fortschritt dar. Da Surf3D Pro bereits beim einfachsten aller Suchszenarien hoffnungslos scheiterte, wurde von weiteren Experimenten abgesehen. Zusammenfassend lässt sich sagen, dass Surf3D Pro nur wenig mit WCM zu tun hat. Da das letzte Release bereits über 6 Jahre zurückliegt, ist das Programm in vielerlei Hinsicht veraltet. Die Idee, die Visualisierung der Suchergebnisse in den Mittelpunkt zu stellen ist sehr progressiv, ist aber in der Beta-Version lediglich als Bildschirmschoner zu gebrauchen.

4.2.4 ChunkIt 1.1.1.0021

Eine weitere Anwendung, die sich mit der Fähigkeit zum WCM schmückt ist ChunkIt der Firma TigerLogic. Es handelt sich hierbei um ein kleines (8,3 MB) einfaches und kostenloses Tool welches nach Angaben der Entwickler, die Inhaltsextraktion aus dem Internet unterstützen und vereinfachen soll: “ChunkIt searches and extracts the valuable chunks of information burried within the numerous hyperlinks. Search web pages and discover information conventional search tools may have never revealed. In addition to mining content on a webpage, ChunkIt! will mine all of the links on that page for information relevant to your search” (http://www.tigerlogic.com/ChunkIt/what.html).

Unter Chunks werden in der Psychologie Informationseinheiten verstanden, die wiederum aus zusammengefassten Einzelinformationen bestehen können (vgl. Riemenschneider 2006, S. 73). Die Chunk-Theorie enthält Aussagen über den Umfang des Kurzzeitgedächtnisses. Ursprünglich stammt der Begriff von George A. Miller, der herausfand, dass das menschliche Kurzzeitgedächtnis Informationen in Blöcken bzw. Chunks abspeichert, wobei das Limit 7±2 Chunks beträgt. Die Informationseinheit bezüglich eines Fahrzeugs kann z. B. aus Einzelinformationen Farbe,

Marke und Größe bestehen. Für einen Fachmann stellt das ganze Auto einen Chunk dar, während ein Laie die Details als einzelne Chunks in seinem Gedächtnis speichert und dementsprechend langsamer lernt.

Erste Eindrücke von ChunkIt

Die Installation verlief problemlos und sehr schnell. ChunkIt ist ein kostenloses Firefox-Add-on und kann ausschließlich in Verbindung mit dem Browser Mozilla Firefox genutzt werden. ChunkIt ist somit keine eigenständige Anwendung, sondern eine Art ständiger Unterstützer bei allen Suchvorgängen innerhalb vieler oder einer bestimmten Webseite. Nach der Installation erscheint in der Symbolleiste des Browsers eine zusätzliche ChunkIt-Befehlleiste.

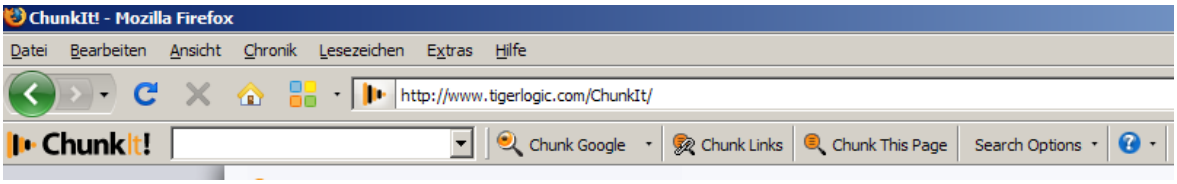


Abbildung 19: Befehlsleiste von ChunkIt in Mozilla Firefox

Chunk Google: Hier können Resultate von Google, Yahoo, Ask, Livesearch und AOL aufgerufen und zusätzlich durchsucht und geordnet werden. Dabei folgt ChunkIt den einzelnen Links und fertigt eigene Snippets an, die sofort neben den Suchmaschinenergebnissen angezeigt werden (siehe. Abbildung 20).

Chunk Links: Hiermit erhält ChunkIt den Befehl, alle von der aufgerufenen Seite verlinkten Dokumente nach Suchbegriffen zu durchforsten. Dafür eignen sich insbesondere Portale, Webshops und andere Webseiten mit vielen Verknüpfungen.

Chunk This Page: Mit diesem Befehl durchsucht das Programm lediglich den Text der aufgerufenen Seite ohne den Links zu folgen. Dabei fertigt ChunkIt eine Liste eigener Snippets an, die eine einfachere Navigation innerhalb langer Texte ermöglichen sollen.

Search Options: Hier können die Suchbegriffe miteinander verknüpft werden. Zur Wahl stehen AND- und OR-Operatoren sowie die Suche nach ganzen Sätzen. Hier kann auch das Stemming, die Suche nach allen Formen des Suchbegriffs, ein- oder ausgestellt werden.

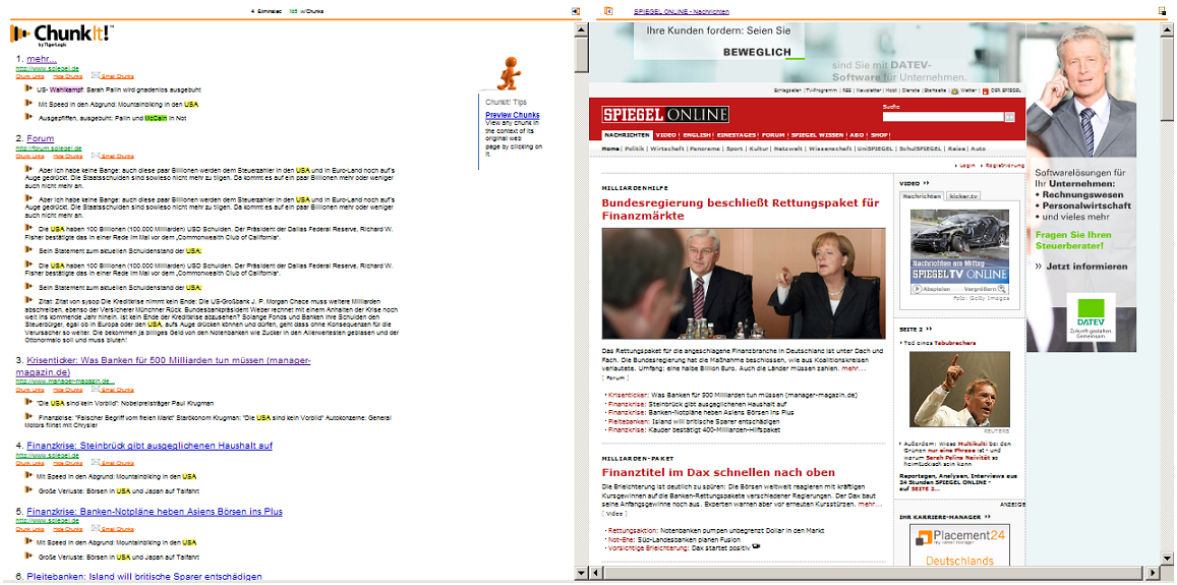


Abbildung 20: Anwendung von ChunkIt auf Spiegel-Online

Nach Eingabe von Suchbegriffen teilt sich das Browserfenster in zwei Teile. Rechts bleibt die aufgerufene Webseite und links kommen die Suchergebnisse von ChunkIt hinzu. Die Suche erfolgt in Echtzeit und kann je nach Anzahl der Links einige Minuten dauern. Schließlich liefert das Programm eine Liste von Dokumenten unter Angabe von Überschrift und URL. Jedes Ergebnis wird durch relevante Absätze, in denen die Suchbegriffe auftauchen präzisiert. Nachdem die Resultate schließlich nach einem dem Autor nicht bekannten Ranking sortiert wurden, erfolgt die Navigation entlang der extrahiert Snippets, vom Entwickler Chunks genannt. Ein Klick auf einen Absatz führt sofort

zum betreffenden Dokument, indem dieser Absatz bereits markiert ist. Das Prinzip und die Bedienung sind sehr einfach und bereits nach kurzer Eingewöhnung können ziemlich schnell reelle Suchszenarien durchgespielt werden.

Suchszenarien mit ChunkIt

Als erstes erfolgte die Nachrichtensuche auf Haldeslblatt.com. Nach mehreren Anläufen konnten zum Thema US-Wahlkampf, 20 relevante Snippets von der Startseite aus extrahiert werden. Davon waren zwölf Stück zutreffend und acht nicht. Ausschlussbegriffe (z. B. –Banken), nach denen die unzutreffenden Resultate unterdrückt würden, können bei ChunkIt leider nicht eingegeben werden. Zudem tauchen mehrere identische Snippets aus unterschiedlichen Dokumenten auf.

Eine Identische Suche auf Spiegel Online kam zu deutlich schlechteren Ergebnissen. Die zahlreichen Texteinblendungen um die Artikel herum (Meldungen aus anderen Ressorts, Zum Thema auf Spiegel Online, Tags Abb. 22, 23) kann ChunkIt nicht vom eigentlichen Text unterscheiden. ChunkIt wirft demnach stets eine zu hohe Anzahl von Treffen aus, die nichts mit dem gewünschten Ergebnis zu tun haben. Nach einigen Anläufen konnte das Problem nicht behoben werden, da die nötigen Einstellungen hierzu fehlen. Das gleiche Problem betrifft auch Financial Times, Handelsblatt und Reuters.

Abbildung 21: Der Link: US-Schlammschlacht..., in jedem Artikel von Spiegel.de der zu unzutreffenden Suchergebnissen führt

Abbildung 22: Das gleiche Problem auf Ftd.de

Im Rahmen der Wettbewerbsanalyse auf dem Onlinepokermarkt stellte sich folgendes vorgehen als sinnvoll heraus. Zunächst erfolgte eine Suche nach „online+poker“ mit Google. Daraufhin wurden die Resultate mit ChunkIt noch mal durchsucht. Mit dem Begriff Bonus konnten alle relevanten Absätze über die angebotenen Boni und deren Struktur aus den Google-Ergebnissen extrahiert werden. Mit der Suche nach „spezial events“ konnten Informationen zu besonders werbewirksamen Turnieren und Veranstaltungen erfasst werden. Teilweise sind die erfassten Absätze zu lang, geben jedoch meist genug Auskunft ohne die betreffende Seite besuchen zu müssen. Leider analysiert ChunkIt jeweils nur eine Google-Ergebnisseite auf einmal. Auch ohne die einzelnen Links besuchen zu müssen, würde es demnach sehr lange dauern alle 14.500.000 Google-Ergebnisse zu durchsuchen. Ebenso fehlt eine Funktion zum Speichern und Exportieren der Snippets und Links. Dennoch funktioniert das ChunkIt-Prinzip im Rahmen der Suche nach konkreten Informationen zu Online-Poker-Anbietern überraschenderweise gut. Beinahe alle extrahierten Absätze liefern relevante Inhalte und ein Überblick über die speziellen Angebote ist mit diesem Tool schnell gewonnen.

Anschließend wurde ChunkIt im Rahmen der Produktsuche eingesetzt. Die Suche nach einem VW T5, ab Baujahr 2003 führte zu einschlägigen Kfz-Börsen wie Mobile.de und Autoscout24. Zunächst musste die seiteneigene Suche nach passenden Fahrzeugen vollzogen werden. Daraufhin galt es, die gewünschten Inhalte aus den gefundenen Fahrzeugangeboten zu extrahieren. Auf Anhieb gelang es mit den Suchbegriffen „endpreis+tel“, Preise, Kontaktdaten sowie jeweils eine kurze Beschreibung aus den Suchergebnissen zu extrahieren, ohne die einzelnen Angebote besuchen zu müssen. Die Suche auf Mobile.de misslang, da ChunkIt derzeit noch keine Flash-Inhalte identifizieren kann.

ChunkIt kann in sozialen Netzwerken nur bedingt eingesetzt werden. Da das Programm lediglich direkte Verlinkungen der aufgerufenen Seite durchsucht, müssen auch hier die zu analysierenden Seiten zunächst manuell ausgewählt werden. Das erfolgt bei StudiVZ oder Myspace mittels der Suchfunktion nach Gruppen und Personen. So könnten theoretisch alle Studenten-Profile der wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover nach konkreten Informationen durchsucht werden. Wodurch z. B folgende Fragen beantwortet werden. Welche Filme und Bücher werden von den Nutzern favorisiert? Welche Hobbys und sonstige Vorlieben und Abneigungen sind den

Profilen zu entnehmen? Leider ließ sich das Experiment weder auf StudiVZ- noch auf Myspace-Profile anwenden. Obwohl alle Voraussetzungen für eine erfolgreiche ChunkIt-Suche gegeben waren, wurden die nötigen Informationen nicht extrahiert. Es wird angenommen, dass sich die Betreiber mit Hilfe technischer Restriktionen vor derartigem Crawling geschützt haben. ChunkIt ist derzeit noch nicht in der Lage diese Schutzmechanismen zu umgehen.

Fazit zu ChunkIt

Der größte Vorteil von ChunkIt ist seine Einfachheit. Bereits wenige Minuten nach der Installation konnten erste Suchergebnisse produziert werden. Zudem ist die Software kostenlos und kann ohne Zeit- oder Funktionseinschränkungen genutzt werden. ChunkIt eignet sich insbesondere für die Analyse von Suchmaschinenergebnissen, Portalen oder langen Texten. Solange die Dokumente für ChunkIt lesbar sind, kann sehr viel Zeit gespart werden, weil die einzelnen Seiten oder Textstellen nicht mehr manuell besucht werden müssen. Bei der Bildung von Snippets orientiert sich ChunkIt an ganzen Absätzen. Die relevanten Absätze werden durch das Programm extrahiert und dem Anwender präsentiert. Obwohl die Absätze teilweise etwas lang sind, bieten sie dem Anwender dennoch einen besseren Überblick als die zu kurz geratenen Snippets von Google.

Die Einfachheit der Nutzung kann aber auch zu einem Problem werden, sobald einige entscheidende Funktionen erforderlich sind. Folgende Aspekte haben die Suche mit ChunkIt erschwert:

- Es kann leider kein Einfluss auf das Ranking der Ergebnisse genommen werden.
- Störende Begriffe, Quellen und doppelte Resultate können bei der Suche nicht unterdrückt werden.
- Das Programm kann nicht zwischen Haupt- und Nebeninhalten eines Dokuments unterscheiden.
- ChunkIt untersucht lediglich die Dokumente, die von einer Seite verlinkt sind ohne automatisch weiterzumachen
- Die ChunkIt-Ergebnisse können nicht gespeichert oder exportiert werden

4.2.5 TextPipe Pro i. V. m. WebPipe, Datamystic Inc.

TextPipe Pro ist eine Text Mining-Software der Firma Datamystics Inc. Nach Meinung der Entwickler ist die Anwendung für die Konvertierung, Transformation und Extraktion von Daten besonders großer Textdokumente entworfen worden. Sehr große Textbestände, Tabellen und Webdokumente sollen mit TextPipe Pro spielend leicht analysiert, aktualisiert oder bearbeitet werden können. TextPipe Pro wird ausschließlich in Verbindung mit WebPipe, einer Software des gleichen Herstellers, auf Webseiten angewendet. Der Gesamtpreis beider Programme beträgt $494,00.

Bereits nach dem ersten Eindruck konnte festgestellt werden, dass TextPipe Pro nur bedingt in die Reihe der zu testenden WCM-Anwendungen hineinpasst. Das Problem: um eine Webseite für TextPipe analysierbar zu machen, muss diese zunächst mit Hilfe von WebPipe runtergeladen werden. WebPipe ist in der Lage ganze Seiten, oder nur ihre Teile extern zu speichern und darzustellen. Über eine Schnittstelle können die Webdokumente erst dann an TextPipe übergeben werden. Die vorliegenden Suchszenarien scheiterten immer wieder daran, dass im Vorfeld zu viele Seiten gespeichert werden mussten. Was bei einzelnen kleinen Webshops, Blogs oder CMS relativ erträglich ist, stößt bei großen Portalen, Archiven oder Netzwerken sofort an Grenzen. Der aktuellen Version von TextPipe Pro fehlt eine eigene Schnittstelle zum Internet, wo das Programm die Seiten analysieren kann ohne sie zuvor runterladen zu müssen.

Es mag sein, dass TextPipe Pro eine leistungsfähige Text Mining-Software ist, mit der Fähigkeit zum WCM zu werben, sollte von Datamystic zunächst noch unterlassen werden. Nach vielen fruchtlosen Versuchen einige aussagekräftige Ergebnisse zu produzieren, wird von einer ausgiebigen Anwendung von diesem Programm abgesehen und vor der Nutzung für Zwecke des WCM abgeraten.

4.2.6 WebSundew 2.0, Sundewsoft

WebSundew 2.0 ist eine WCM-Software der Firma Sundewsoft. Die Entwickler beschreiben ihr Programm folgendermaßen: “WebSundew is the solution that allows you to handle web content without using scripts. It was developed for those who wish to use scripts' functionality for web data extraction and not bother for code writing” (http://www.sundewsoft.com/products).

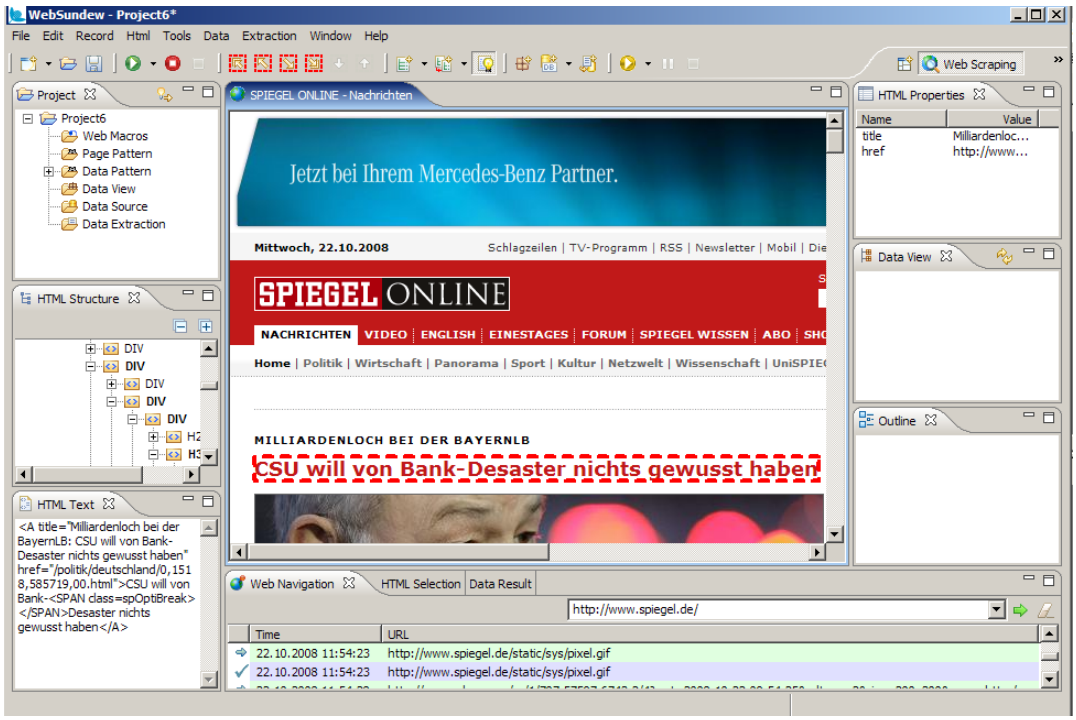


Abbildung 23: Grafisches Interface von WebSundew 2.0

Das Programm ist mit einem Preis von $249,00 vor allem an gewerbliche Kunden gerichtet und bietet ein grafisches Interface, dessen Bedienung ohne jegliche Programmierkenntnisse möglich ist. Für die vorliegende Untersuchung konnte eine Trial-Verson von WebSundew genutzt werden, diese ist auf 14 Tage begrenzt und erlaubt die Extraktion von maximal drei Seiten einer Website. Vor allem diese Funktionsbeschränkung macht es schwierig das Programm vollwertig auszuprobieren und zu testen. Glücklicherweise konnte bereits im Rahmen der Tutorials und des ersten Eindrucks festgestellt werden, dass WebSundew sehr an Web Content Extractor erinnert, welches bereits im Abschnitt 5.2.1 ausführlich untersucht und getestet wurde.

Die einzige wesentliche Verbesserung gegenüber Web Content Extractor ist, dass mit Hilfe von WebSundew vorprogrammierte Texteingaben in Such- oder Identifikationsfelder möglich sind. Der große Nutzen dieser Funktion konnten bereits im Zusammenhang mit Mozenda Beta verdeutlicht werden. Der restliche Funktionsumfang ist nahezu identisch. Von der Präsentation der Suchresultate wird daher abgesehen, weil sie ebenso dem Abschnitt zum Web Content Extractor entnommen werden können.

Nach Angaben des Herstellers soll noch in diesem Jahr ein neues Release auf den Markt kommen. Eine Beta-Version von WebSundew 3 Beta wird auf der Betreiberseite zwar nicht angeboten, konnte aber im Internet gefunden werden. Ein Grund dafür, warum Sundewsoft die neue Beta-Version nicht offiziell anbietet, ist wahrscheinlich die mangelnde Marktreife. Auf keinem dem Autor verfügbaren Rechner konnte die neue Beta Version zum installiert werden. Eine wesentliche Neuerung gegenüber WebSundew 2.0 und Web Extractor konnte zumindest bei der Installation identifiziert werden. Das neue Release verfügt ebenso wie Mozenda über eine Schedule-Funktion und kann auch vom Server aus gestartet werden. Ob WebSundew 3.0 über weitere revolutionäre Funktionen verfügt konnte nicht in Erfahrung gebracht werden. Es bleibt abzuwarten bis eine funktionsfähige Demo- oder Vollversion davon auf den Markt gebracht wird.

4.2.7 Screen-Scraper 4.0 Basic Edition, Ekiwi LLC

Screen-Scrapper 4.0 Basic Edition ist die schwierigste und anspruchvollste WCM-Anwendung im Test. Sie ist an gewerbliche Kunden gerichtet, die über genügend Programmierwissen in HTML sowie in einer der folgenden Programmiersprachen verfügen: VBScript, JScript, Perl, Javascript, Python oder Java. Die Entwickler beschreiben Ihr Programm folgendermaßen: „Screen-scapper consists of a proxy server that allows the contents of HTTP and HTTPS requests to be viewed, and an engine that can be configured to extract information from Web sites using special patterns and regular expressions" (http://www.screen-scraper.com/products/all.php). Für die Untersuchung stand dem Autor eine kostenlose Basic-Version zur Verfügung. Darüber hinaus kann bei Ekiwi LLC auch eine Professional oder Enterprise-Edition für $399,00 bzw. $2499,00 käuflich erworben werden. Diese erweiterten Versionen verfügen über eine Vielzahl von Schnittstellen zu anderen Anwendungen, wie z. B. Java, .NET, COM, PHP,

ASP, Ruby, Python oder Cold Fusion. Die erweiterten Versionen können auch mit Hilfe der Schedule-Funktion zeitlich versetzt oder vom Server aus gestartet werden. Nach Wunsch kann auch ein ständig verfügbarer persönlicher Support gebucht werden.

Das größte Problem von Screen-Scrapper ist, dass es nur sehr schwer erlernt werden kann. Dem Anwender stehen äußerst lange und textlastige Tutorials zur Verfügung, die allerdings nur bedingt in Lage sind, das Programm verständlich zu machen. Im Gegensatz zu anderen getesteten Anwendungen, beginnt die Vorbereitung auf die Extraktion mit einer manuellen Markierung der gewünschten Daten aus dem HTML-Code einer Seite. Wenn z. B. die Warenbeschreibungen eines Webshops extrahiert werden müssen, reicht es nicht, die gewünschten Datenfelder im Browserfenster zu markieren. Alle relevanten HTML-Verweise müssen manuell aus dem Code in die Software übertragen werden. Noch schwieriger wird es, wenn sich die Artikel auf mehreren Seiten befinden. Denn auch diese Funktion muss im HTML-Code gefunden und manuell in Screen-Scrapper eingefügt werden.

Als ob das noch nicht aufwendig genug wäre, so muss auch der eigentliche Extraktionsprozess mit Hilfe von VBScript, JScript, Perl, Javascript, Python oder Java in einem Script manuell hinterlegt werden. Zwar konnten die einfachen Aufgaben der Tutorials erfolgreich absolviert werden, weil die Scripte vorgegeben waren. Im Rahmen eigener Suchszenarien scheiterten die Mining-Versuche an diesem Schritt, weil dem Autor das nötige Programmierwissen fehlt.

Die Leistungsfähigkeit und Qualität von Screen-Scrapper kann demnach nur anhand der Produktinformationen und Tutorials beurteilt werden. Eigene erfolgreiche Extraktionsversuche fanden leider nicht statt, weil kein einziger Agent fehlerfrei voreingestellt werden konnte. Ob die Entwickler künftig dem Trend der hundertprozentig grafischen Bedienung solcher Programme folgen, bleibt fraglich. Tatsache bleibt, dass Konkurrenzprodukte einen vergleichbaren Funktionsumfang bei erheblich einfacherer Bedienung anbieten und damit über einen großen Vorteil gegenüber der aktuellen Screen-Scrapper-Version verfügen.

4.2.8 Web Info Extractor 1.7.0, WebIESoft Corp. Ldt.

Web Info Extractor (WIE) 1.7.0 ist der neuste Release (2007) einer WCM-Anwendung der Firma WEBIESoft. Die Entwickler beschreiben ihr Programm folgendermaßen: „Web Information Extractor is a powerful tool for web data mining and content extraction, content analysis. It can extract structured or unstructured data from web page, reform into local file or save to database, post to web server" (http://www.webinfoextractor.com/). Die untersuchte Version von WIE ist eine Shareware, mit der sich maximal drei Resultate pro Anfrage speichern lassen. Die Funktionalität der käuflich erwerbbaren Versionen ist identisch. Die Basic, Standard oder Full Lizenzen unterscheiden sich lediglich durch Einschränkungen für die Anzahl der zu speichernden oder abzufragenden Daten. Eine Basic Version kostet $99,95, eine Standard-Lizenz kostet $199,00 und die unbeschränkte Nutzung einer Full Licence kostet $499,95 (vgl. http://www.webinfoextractor.com/purchase.htm).

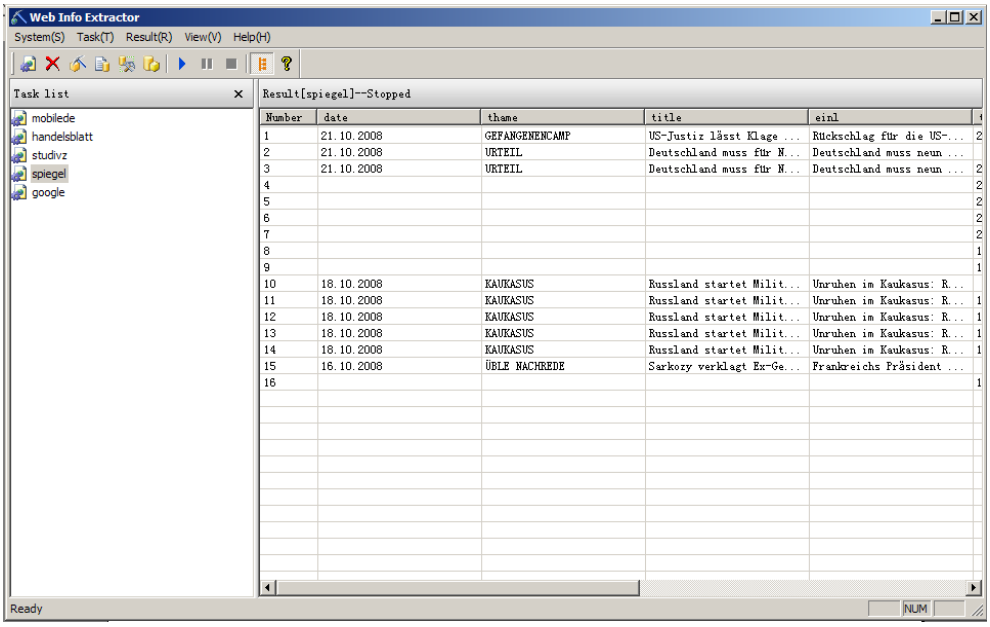


Abbildung 24: Bedienungsinterface von Web Info Extractor 1.7.0

WIE macht auf den ersten Blick einen sehr einfachen und geordneten, aber auch etwas veralteten Eindruck. Abbildung 25 zeigt das übersichtliche Bedienungsinterface des Programms. In der „Task List“ auf der linken Seite werden mit Hilfe eines Wizards die Projekte angelegt und auf der rechten Seite erscheinen die Resultate der jeweiligen Extraktion. Die Erstellung eines Projekts mit Hilfe von WIE erfordert keinerlei Programmierkenntnisse. Das Prinzip ist ähnlich den zuvor untersuchten Programmen WCE, Mozenda und WebSundew. Die Einstellungen erfolgen über ein integriertes Browserfenster. Darin werden die relevanten Webdaten ausgewählt und der Extraktionsprozess definiert. Aufgrund der im Laufe der Untersuchungen gesammelten Erfahrung mit ähnlichen Programmen konnte bereits nach wenigen Versuchen ein fertiger Agent für den Spiegel Online-Archiv erstellt werden.

Das erste erkannte Manko von WIE ist die mangelhafte Identifikation von verwandten Beiträgen. Anders als bei ähnlichen Programmen, konnte WIE schon aufgrund geringer struktureller Unterschiede von Dokumenten keine vollständigen Ergebnisse liefern. Dieses substanzielle Problem konnte im Laufe der Untersuchung nicht behoben werden. WIE kann demnach nur auf Seiten angewendet werden, deren strukturelle Ähnlichkeit besonders hoch ist. Erfreulich war, dass die extrahierten Daten vor der Speicherung sortiert und gefiltert werden konnten. Mit Hilfe der Filterfunktion können mit WIE bestimmte Resultate anhand ihrer Eigenschaften unterdrückt bzw. übernommen werden. Die Schedule-Funktion erlaubt dem Anwender, ein Projekt zeitlich zu koordinieren. Der Export erfolgt in eine Excel- oder Textdatei. Die Resultate können auch auf einen Webserver oder in SQL- und Access-Datenbanken überführt werden.

Abbildung 26 zeigt einen Screenshot des WIE-Wizards. Hier erfolgen sämtliche Einstellungen eines Projekts. Oben ist das integrierte Browserfenster indem relevante Daten selektiert werden. Unten links erscheinen Informationen zum ausgewählten Objekt und unten rechts werden die relevanten Datenfelder und die Extraktionsschritte definiert. Im Reiter „Option“ in der oberen Leiste kann die Sortierung und Filterung der Resultate eingestellt werden. Rechts liegen die Schedule-Funktionen und die restlichen Reiter zum Export der Daten.

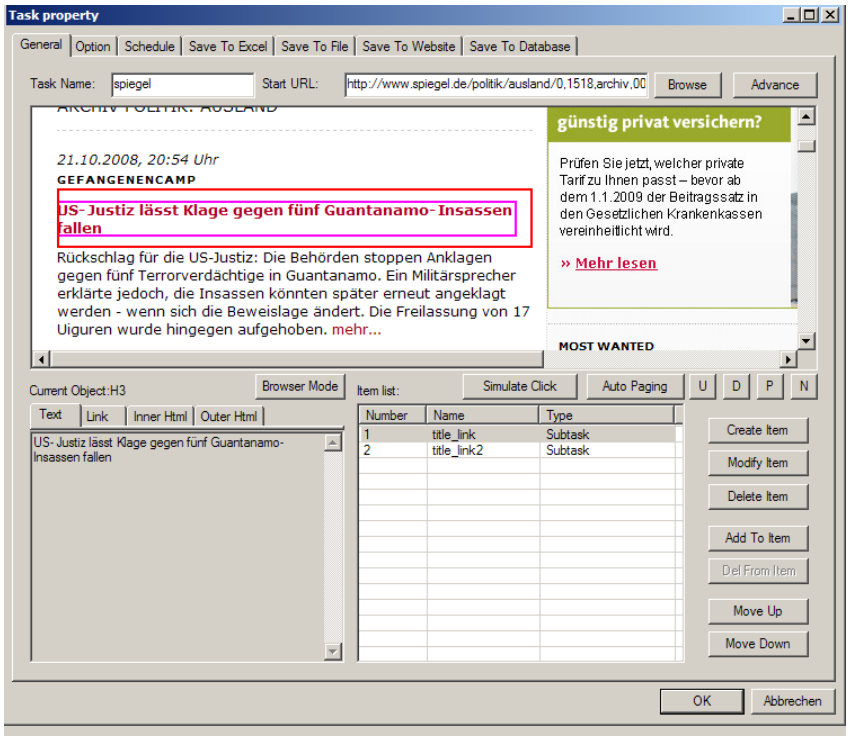


Abbildung 25: Screenshot vom Wizard von Web Info Extractor 1.7.0

WIE verfügt über einen vergleichsweise großen Umfang von nützlichen Funktionen. In der praktischen Ausführung zeigte das Programm aber Schwächen in der Erkennung schwachstrukturierter Inhalte. Dieses substanzielle Problem kann durch die einfache Bedienung und die Filter- und Sortierungsmöglichkeiten nicht ausgeglichen werden. Andere getestete Programme sind zuvor mit deutlich weniger geordneten Archiven und Webseiten fertig geworden. Obwohl der letzte Release 2007 relativ neu ist, empfiehlt sich auf die nächste Version von WIE zu warten und zu hoffen, dass sich die Entwickler von Konkurrenzprodukten produktiv inspirieren lassen.

4.3 Zusammenfassung der Untersuchungsergebnisse

Die Ergebnisse zu den untersuchten WCM-Programmen sollen im Folgenden zusammengefasst und einander gegenübergestellt werden. In Tabelle 5 sind alle getesteten Anwendungen gegen die zuvor im Abschnitt 5.1 definierten Kriterien der Softwarequalität abgetragen. Die relevanten Ausprägungen dieser Kriterien sind mit Schulnoten bewertet und nach ihrer Gesamtbewertung sortiert.

Tabelle 5: Gesamtbewertung der untersuchten Web Content Mining-Programme

	Mozenda	Web Content Extractor	Web-Sundew	Web Info Extractor	Screen-Scraper	ChunkIt	WebPipe + TextPipe Pro	Surf3D Pro
Umfang der praktischen Anwendung	ausgiebig	ausgiebig	ausgiebig	ausgiebig	nur bedingt möglich	ausgiebig	nur bedingt möglich	nur bedingt möglich
Funktionalität								
Datenauswahl	2	2	3	5	2	5	3	6
Inhaltsextraktion	2	2	3	2	2	4	4	6
Interpretation	5	5	5	3	3	3	6	6
Zuverlässigkeit								
Korrektheit der Ergebnisse	2	2	2	4	3	4	3	5
Verfügbarkeit des Programms	2	2	2	2	3	2	4	3
Benutzbarkeit								
Schulungsaufwand	1	2	2	2	5	2	3	3
Bedienung	1	2	2	2	5	2	4	3
Support	1	2	2	2	3	5	3	6
Effizienz								
Geschwindigkeit	3	3	2	3	2	1	3	6
Ressourcenbelastung	1	2	2	2	3	2	5	3
Preis-Leistungsverhältnis	2	1	3	3	2	3	3	4
Übertragbarkeit								
Export	1	1	1	1	2	5	2	6
Schnittstellen	3	3	2	3	3	5	4	6
Preis	**$39 monatl.**	**$129**	**$249**	**$99,95-$499,95**	**$0,00-$2499**	**$0,00**	**$494**	**$0,00**
Gesamtbewertung	**2**	**2,2**	**2,6**	**2,8**	**3,2**	**3,3**	**3,6**	**4,8**

Quelle: Eigene Erstellung

Das fortschrittlichste Programm im Test war Mozenda Beta. Mozenda überzeugte mit einem hohen Funktionsumfang bei gleichzeitig kinderleichter Bedienung. Bestnoten erreichte Mozenda vor allem in Benutzbarkeit und Ressourcenbelastung. Auch die Tatsache, dass für die Datenextraktion die Serverressourcen des Anbieters bereitgestellt werden begründet den hohen Preis von $39,95 im Monat.

Die Plätze zwei, drei und vier gehen an untereinander sehr ähnliche Anwendungen. Die Funktionsweisen von WCE, WIE und WebSundew sind nahezu identisch. In der Ausführung konnte sich WCE jedoch gegen seine Konkurrenz durchsetzen. Sowohl in der Korrektheit der Datenauswahl als auch in der Inhaltsextraktion lieferte WCE deutlich bessere Ergebnisse bei ähnlichem Schulungsaufwand und vergleichbaren Preisen. Trotz verheerender Schwächen in der Extraktion semistrukturierter Inhalte, überzeugte WIE mit einem umfangreichen Funktionspaket welches unter anderem Sortierung und Filterung von Ergebnissen sowie eine Schedule-Funktion enthält. WebSundew ist hingegen das einzige dieser drei Programme das vom Web Server gestartet werden kann.

Screen-scaper konnte aufgrund des hohen Schulungsaufwands nur bedingt angewendet werden. Die Entwickler haben den Trend zu einfachen und bedienerfreundlichen Software-Anwendungen in ihrem Bereich verschlafen. Die

Konfiguration eines Projekts dauert mit Screen-scraper deutlich zu lange und die Programmierkenntnisse des Anwenders sind dabei im hohen Maße gefordert.

ChunkIt kann nur schwer mit anderen WCM-Programmen verglichen werden weil es sich eines ganz anderen Prinzips bedient. Es handelt sich hierbei um kein vollwertiges Programm sondern um ein Mozilla Firefox-Addon welches die Suche nach bestimmten Inhalten im WWW erheblich erleichtert. ChunkIt extrahiert gewünschte Inhalte aus einer Liste von Webseiten, Dokumenten oder Beiträgen, ohne dass diese Quellen jeweils besucht werden müssen. Die Extraktion erfolgt in Form von Absätzen, die aus textbasierten Seiten gewonnen werden können.

TextPipe Pro ist eine Text Mining-Anwendung, die nach Meinung der Entwickler in Verbindung mit WebPipe, WCM beherrschen soll. Dabei ist es erforderlich, die zu analysierende Webseite zunächst mit WebPipe runterzuladen und lokal zu speichern. Erst dann kann TextPipe darauf angewandt werden. Gerade bei größeren Archiven oder umfangreichen Suchmaschinenergebnissen stellte sich diese Restriktion als eine als eine hohe Hürde heraus. Dem Programm fehlt eine eigene Schnittstelle zum Internet um relevante Daten online analysieren und identifizieren zu können. Damit ist die aktuelle Version von TextPipe Pro für WCM-Anwendungen ungeeignet.

Surf3D Pro ist das älteste Programm im Test. Das letzte Release aus dem Jahr 2002 beeindruckt mit der revolutionären Herangehensweise, Webinhalte und Webdokumente in einem dreidimensionalen Raum sichtbar und suchbar zu machen. Ähnlich den anderen getesteten Anwendungen, bietet auch Surf3D die Möglichkeiten einen Agent zu konfigurieren, der innerhalb eines bestimmten Suchraum relevante Inhalte finden soll. Die Resultate werden dann jedoch nicht wie üblich in Tabellenform, sondern in Form futuristischer Grafiken in einem euklidischen Raum abgebildet. Obwohl die, im Rahmen der Untersuchung erhaltenen Grafiken über kaum oder keine Aussagekraft verfügten, stellt die Visualisierung von Inhalten ein leider zu wenig beachtetes Forschungsfeld innerhalb von WCM dar. Surf3D kann nur schwer mit den anderen WCM-Programmen verglichen werden weil es einfach zu alt ist und daher mit den Anforderungen des heutigen Internets nicht fertig wird. Surf3D verdient jedoch einen Ehrenpreis für die, zwar nicht ausgereifte, aber dennoch sehr progressive Art Inhalte zu präsentieren.

Im Rahmen der Untersuchung konnten schließlich sechs brauchbare WCM-Programme gefunden und ausgiebig genutzt werden. In Tabelle 7 befinden die sich daher auf den ersten sechs Plätzen. Mit Ausnahme von ChunkIt sind diese Anwendungen gezielt für betriebliche Zwecke entwickelt worden. Mit Preisen von $99 bis $2499 oder $39 monatlich im Fall von Mozenda, sprechen sie untereinander verschiedene Zielgruppen an. Der Autor empfiehlt für gelegentliche Nutzung, die Programme WCE oder WebSundew. Die beiden Anwendungen enthalten fast alle nötigen Funktionen um mühelos an Webinhalte aus zahlreichen Archiven, Portalen, etc. ranzukommen. Mozenda produziert vergleichbare Resultate, überzeugt aber mit dem höchsten Grad an Bedienungsfreundlichkeit und Automatisierung. Der Preis von $39 im Monat rechtfertigt sich jedoch nur im Fall regelmäßiger und ausgiebiger Nutzung.

5 Fazit

Erst seit wenigen Jahren, stehen der Forschung einige fundierte und spezialisierte Fachbücher zum Thema WCM zur Verfügung. Darin setzen sich die Autoren jedoch überwiegend mit theoretischen Aspekten WCM-verwandter Methoden und Verfahren auseinander. Eine Vielzahl komplizierter Formeln und Algorithmen füllen diese Bücher. Damit ist diese Literatur in erster Reihe an Entwickler und nicht an potenzielle Anwender von WCM gerichtet. U. a. bieten Markov/Larose 2007 und Valasquez/Palade 2008 dem Leser einen fundierten Einblick hinter die Funktionsweise verwendeter Methoden. Nur wenige Autoren versuchen die Vorzüge und Potenziale von WCM-Software aus einer betriebswirtschaftlichen Sicht zu erforschen und dem Leser näher zu bringen. Motivation und Zielsetzung dieser Untersuchung war es daher die Erforschung gewinnbringender Potenziale mit der praktischen Anwendung bereits verfügbarer WCM-Programme zu verbinden.

Die vorliegende Untersuchung kann in drei gleichwertige Abschnitte unterteilt werden. Der erste Teil hatte die wissenschaftliche Grundlegung des Themas WCM zum Gegenstand. Der zweite entscheidende Abschnitt der Diplomarbeit widmete sich den betriebswirtschaftlichen Anwendungsgebieten von WCM. Das kommerzielle Potential von WCM ist sehr groß und im Laufe der Untersuchung konnten einige interessante Anwendungsgebiete identifiziert und untersucht werden.

Großes Potenzial für die Anwendung von WCM konnte im Zusammenhang mit Handelsmärkten identifiziert werden. Handelsakteure mit einem ausgeprägten Bedarf an umfangreichen und aktuellen Informationen bezüglich Preisen, Angeboten, etc. können in hohem Ausmaß von WCM-Programmen profitieren. Mit Hilfe künstlicher neuronaler Netze, ist es schon heute möglich akkurate Bewertungen von Kapitalmarktprodukten anhand einer Vielzahl ver-

fügbarer Internetquellen vorzunehmen und die gesammelten Information in Handlungsempfehlungen zu transformieren (Bartels 2008, S. 212). In Zukunft könnten intelligente WCM-Systeme nicht nur unterstützend sondern auch selbstständig tätig werden um das Geld von Spekulanten streckenweise automatisch zu vermehren. Am Beispiel des Gebrauchtwagenmarktes konnte aufgezeigt werden, wie Anwendungen zur Extraktion und Integration von Webinhalten zu beträchtlichen Wettbewerbsvorteilen führen können.

Das größte rechtliche Risiko für die Anwender von WCM liegt in der Extraktion bzw. Verwendung persönlicher Nutzerdaten. Um Imageschäden oder Klagen zu vermeiden ist es wichtig, die datenschutzrechtlichen Voraussetzungen für die Anwendung von WCM auf Communities, Foren, etc. zu kennen. Die Erfassung von Daten, die geeignet sind eine Person zu identifizieren (Name, Anschrift, Religion, Email, etc.) ist ausschließlich mit Zustimmung dieser Person zulässig. Pseudonymisierte Daten, die keine Rückschlüsse auf die Person zulassen, sind nur bis zum Widerspruch dieser Person zulässig. Ferner konnte momentan ein Trend zu Ausweitung des Schutzes personenbezogener Daten im Internet festgestellt werden.

Da es den Webseitenbetreibern wichtig ist, die Kontrolle über ihre und die Daten ihrer Nutzer zu behalten, wehren sie sich oft gegen die automatische Inhaltextraktion. Für sie ist CAPTCHA, die derzeit wirksamste Waffe gegen WCM. Mit der Fähigkeit, menschliche Nutzer von Crawlern bzw. Agents zu unterscheiden stellen CAPTCHA-Abfragen ein, nur mit großem Aufwand überwindbares Hindernis für WCM-Programme dar. Abhängig vom Schwierigkeitsgrad eines CAPTCHAs können dagegen derzeit entweder leistungsfähige OCR-Programme genutzt oder zu entlohnende Menschen beauftragt werden.

Nach Grundlegung und Anwendungsanalyse wurden schließlich derzeit erhältliche WCM-Programme vom Autor erlernt, angewendet und verglichen. Dabei konnten einige interessante Softwarelösungen identifiziert werden, die sich durch ihre Bedienungsfreundlichkeit und einen hohen Grad an Automatismus auszeichneten. In diesem Zusammenhang seien die Programme Mozenda, Web Content Extractor, WebSundew und Web Info Extractor zu erwähnen. Diese Anwendungen eignen sich insbesondere zur stetigen und regelmäßigen Extraktion gewünschter Inhalte aus einzelnen Webseiten oder schwachstrukturierte Archive oder Portalen. Ohne Programmierkenntnisse, können mit ihrer Hilfe in kürzester Zeit Agents entworfen werden, die zuvor definierte Webinhalte in strukturierte Datenbanken überführen. Schließlich stehen dem Anwender komprimierte Inhalte in Form von Tabellen zur Verfügung. Diese Tabellen können anschließend in gängige Formte exportiert werden.

Keines dieser Programme bot dem Anwender die Möglichkeit, die exportierten Daten einer weiteren Analyse zuzuführen. Die erhofften Anzeichen einer künstlichen Intelligenz, die durch Clusterung oder Klassifizierungsverfahren aus Daten Erkenntnisse generieren können, wurden nicht bestätigt. Um aus den gewonnenen strukturierte Daten schließlich relevantes Wissen erhalten zu können, müssen Text Mining- oder Data Mining Programme darauf angewendet werden. Automatisiert wird demnach derzeit nur der Extraktionsprozess, der die relevanten Daten in eine strukturierte Form bringt. In vielerlei Hinsicht ist das bereits eine sehr große Entlastung und stellt die wichtigste Voraussetzung für eine anschließende Analyse dar.

Der nächste Schritt zu besseren WCM-Programmen ist die Integration von Analyse- und Interpretationsinstrumenten. Es ist erforderlich, die theoretischen Erkenntnisse im Hinblick auf anwendbare Klassifizierung- und Clusterungsverfahren in der praktischen Anwendung umzusetzen. Vor dem Hintergrund des hohen Forschungsaufwands auf diesem Gebiet, kann mit den ersten „intelligenten“ WCM-Programmen noch diesem Jahrzehnt gerechnet werden. Um die Fähigkeiten künftiger WCM-Software abzurunden und den Anforderungen des modernen Internets gerecht zu werden, wird die Integration von Multimedia Mining erforderlich sein.

Zuletzt werden die, in der Einleitung gestellten Fragen aufgegriffen und einzeln beantwortet.

Inwieweit ist WCM in der Lage das betriebliche Informationsmanagement zu entlasten oder zu optimieren?

WCM-Programme wie z. B. Mozenda, WCE oder WIE sind bereits jetzt in der Lage Inhalte, Daten oder Bilder automatisch und regelmäßig einzelnen Webseiten, Archiven, Portalen, Suchresultaten, etc. zu entnehmen. Mit Hilfe von WCM-Programmen können Schnittstellen zwischen schwachstrukturierte Inhalten des Internets und eigenen Anwendungen geschaffen werden. Ganz ohne personellen Aufwand wird damit dem Informationsmanagement eine neue und breite Datenbasis aus vielfältigen Internetressourcen bereitgestellt.

Welche neuen Geschäftskonzepte lassen sich mit WCM künftig verwirklichen?

Das kommerzielle Potential von WCM sehr groß. Es konnten einige interessante Anwendungsgebiete identifiziert und untersucht werden. Besonders viel versprechend sind Anwendungen aus den Bereichen. Marktforschung,

Trendforschung, Wettbewerbsanalyse, Online Suchdienste, personalisierte Werbung sowie im Rahmen von Schutz- und Sicherheitskonzepten.

Welche Anspruchsgruppen profitieren von WCM-Anwendungen?

Unternehmen, die einen hohen Bedarf an aktuellen Informationen über Konkurrenten, Preise, Kunden etc. aufweisen. Betreiber oder Gründer von Webseiten, die ihren Besuchern aggregierte Inhalte bieten wollen. Online-Suchdienste die, die Suche nach Webinhalten revolutionieren wollen. Dazu kommen Webentwickler, Nachrichtendienste, Werbeagenturen, und Meinungsforschungsagenturen.

Welche konkreten Anwendungen bietet der Markt für WCM-Software?

Der Markt für WCM-Software wächst. Abschnitt 5.3 bietet eine Gegenüberstellung und Bewertung aller getesteten Programme.

Wie weit sind moderne WCM-Programme fortentwickelt?

Moderne WCM-Programme beschränken sich nur auf die Extraktion schwachstrukturierter Inhalte in strukturierte Tabellen oder Datenbanken. Die Interpretation und Analyse der gewonnen Daten erfolgt derzeit noch außerhalb dieser Programme. Die, im Grundlagenteil beschriebenen Typisierungsmethoden fanden in den getesteten Programmen keine Anwendung.

Welche Funktionsweise steckt hinter den jeweiligen WCM-Programmen?

Jede Inhaltsextraktion beginnt stets mit der Konfiguration eines Agents. Hierfür muss i. d. R. ein exemplarischer Datensatz aufgerufen und die darin enthaltenden relevanten Daten markiert werden. Die meisten Programme zeigen, dass diese Voreinstellungen ohne Programmierkenntnisse in geringer Zeit vorgenommen werden können. Daraufhin kann die Extraktion beliebig oft wiederholt, mit Mozenda und WebSundew sogar automatisch vom Server aus gestartet werden. Schließlich erhält der Anwender eine strukturierte Tabelle mit allen gekennzeichneten Inhalten. Mit dem Export dieser Daten ist der, von erhältlichen Programmen unterstützte WCM-Prozess abgeschlossen.

Wohin geht der Trend von WCM-Anwendungen?

Neuste WCM-Programme zeichnen sich durch ein hohes Maß an Bedienungsfreundlichkeit aus. Der Trend geht zu übersichtlichen grafischen Oberflächen, die sich ohne große Erfahrung und Programmierkenntnisse steuern lassen. Ein weiterer Trend ist, die Entlastung der Ressourcen des Anwenders. Mozenda bietet seinen Kunden für die Datenextraktion und –Speicherung eigene Serverkapazitäten. Damit läuft der Extraktionsvorgang automatisch und regelmäßig, ohne dabei die Ressourcen des Nutzers zu belasten. Mit Blick auf die Fachliteratur, kann gesagt werden, dass die Integration von Interpretations- und Analysemöglichkeiten in WCM-Programme künftig an Bedeutung gewinnen wird. Ebenso zukunftsweisend sind Methoden und Verfahren des Multimedia Mining, deren Integration für in WCM-Programme unerlässlich sein wird.

Literaturverzeichnis

Ackerman, R.: Intelligence Center Mines Open Soures. http://www.afcea.org/signal /articles/templates/SIGNAL_Article_Template.asp?articleid=1102&zoneid=31. Erstelldatum: 01.03.2006, Druckdatum: 01.10.2008

Adam, D.: Planung und Entscheidung: Modelle- Ziele- Methoden; mit Fallstudien und Lösungen. Gabler, Wiesbaden 1996

Adriaans, P., Zantinge, D.: Data Mining. Addison-Wesley, Harlow u. a. 1996

Agrawal, D.: Proceedings of the Fourth International Workshop on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems. IEEE Computer Soc. Press, Los Alamitos u. a. 1996

Ahlert, D. u.a. (Hrsg): Customer Relationship Management im Handel. Springer, Berlin 2002

Akerkar, R., Lingras, P.: Building an Intelligent Web, Theory and Practice. Jones an d Bartlett Publishers, Boston u. a. 2008

Alpar, P., Niedereichholz, J.: Einführung zu Data Mining. Vieweg & Sohn Verlag, Braunschweig/Wiesbaden 2000

Alpert, J., Hajaj, N.: We knew the web was big… .(http://googleblog.blogspot.com/2008/ 07/we-knew-web-was-big.html. Erstelldatum: 25.07.2008, Druckdatum: 25.09.2008

Back, A., Maertens, P.: Lexikon Der Wirtschaftsinformatik. Springer, Berlin u. a. 2001

Backhaus, K.: Multivariate Analysemethoden: Eine anwendungsorientierte Einführung. Springer, Berlin u. a. 2000

Bartels, P.: Echtzeit-Bewertung von Optionen mit Marktpreisen durch Web-Mining und Neurosimulation. Josef Eul, Lohmer/Köln 2008

Behrendt, B. u. a.: Web Mining: From Web to Semantic Web: First European Web Mining Forum. Springer Berlin u. a. 2003

Bentele, G. u. a.: Markenwert und Markenwertermittlung: Eine systematische Modelluntersuchung und ~bewertung. DUV, Wiesbaden 2005

Beyer, J.: Pfadabhängigkeit: Über institutionelle Kontinuität, anfällige Stabilität und fundamentalen Wandel. BoD, Berlin 2006

Buchner, B.: Informationelle Selbstbestimmung im Privatrecht. Mohr Siebeck, Tübingen 2006

Catwright, H., Sztandera, L.: Soft Computing Approaches in Chemistry. Springer, Berlin u. a. 2003

Computer Bild 3/2007

Dehmer, M.: Strukturelle Analyse Web-basierter Dokumente. DUV, Wiesbaden 2006

Deinhard, M., Oswald, J.: Das Geheimnis des Web Mining: Die Suche nach verborgenen Schätzen. http://www.it-daily.net/content/view/433/32/. Erstelldatum: 26.02.2008, Druckdatum 01.10.2008

Dolata, U.: Neue Technologien verschlafen. http://www.focus.de/kultur/musik/ musikbranche-neue-technologien-verschlafen_aid_318852.html. Erstelldatum: 20.07.2008, Druckdatum: 01.10.2008

Dürr, H.: Anwendungen des Data Mining in der Praxis. Universität Ulm 2003

Eckstein, P.: Angewandte Statistik mit SPSS: Praktische Einführung für Wirtschafts-wissenschaftler. Gabler, Wiesbaden 2008

Elder, J., Pregibon, D.: A statistical perspective on knowledge discovery in databases. AAAI Press, Menlo Park 1996

Elleithy, K., Sobh T. (Hrsg): Advances in Systems, Computer Sciences and Software Engineering. Springer, Dordrecht 2006

Ensthaler, J.: Gewerblicher Rechtsschutz und Urheberrecht. Springer, Berlin u. a. 2003

Etzioni, O.: The World Wide Web: Quagmire or Goldmine. In: Communications-ACM, 39 (1996) S. 65-98

Feibel, T.: Was macht der Computer mit dem Kind?: Kinder im Medienzeitalter begleiten, fördern und schützen. Family Media, Frankfurt 2002

Freyer, W.: Tourismus-Marketing: Marktorientiertes Management im Mikro- und Makrobereich der Tourismuswirtschaft. Oldenbourg, München u. A. 2006

Fritsch, H.: StudiVZ. Inoffizielle Statistiken Dezember 2006. http://studivz.irgendwo.org. Erstelldatum: 03.01.2007 Druckdatum: 01.10.2008

Furht, B., Marques, O.: Handbook of video databases : design and application. CRC Press, Boca Raton 2003

Gentsch, P.: Web-personalisierung und Web-mining für eCRM. Oxygon-Verlag, Würzburg 2002

Glos, M.: Sicher Surfen, Mailen, Daten tauschen. Espresso. Franzis, Poing 2004

Gluchowski, P., Gabriel, R., Dittmar, C.: Management Support Systeme und Business Intelligence: Computergestützte Informationssysteme für Fach- und Führungskräfte. 2. Auflage, Springer, Berlin u. a. 2008

Gulla, J., Borch, H., Ingvaldsen, J.: Contextualized Clustering in Exploratory Web Search. In: Zhou, Z. u. a.: Advances in Knowledge Discovery and Data Mining. Springer, Berlin u. a. 2007. S. 184-207

Gutheim, P.: Der Webdesign Praxisguide, Professionelle Konzeption von der Planung bis zur Promotion. Springer, Berlin/Heidelberg 2008

Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. MIT-Press, Cambridge 2001

Herrmann, H.: Data Mining. Books on Demand, Nordestedt 2008

Hiller, A.: Einführung in den Einsatz von Data Mining. GRIN, München/Ravensburg 2007

Hippner, H., Merzenich, M., Wilde, K.: Handbuch Web Mining im Marketing: Konzepte, Systeme, Fallstudien. Vieweg+Teubner Verlag, Wiesbaden 2002

Hoffmann, A.: Advances in knowledge acquisition and management. Springer, Guili u. a. 2006

Hornig, F., Müller, M., Weingarten, S.: Die Datensucht. In: Spiegel, 33 (2008) S. 82-90

Jung, H.: Allgemeine Betriebswirtschaftslehre. Oldenbourg, Wiesbaden 2006

Kähler, W.: Statistische Datenanalyse: Verfahren verstehen und mit SPSS gekonnt einsetzen. Vieweg+Teubner, Wiesbaden 2008

Kernahan, M., Capretz, L. F.: Different Strategies for Web Mining. In: Advances in Systems, Computer Sciences and Software Engineering S. 83-88 2006, S. 84

Kleinz, T.: Suchmaschine Cuil: Größer als Google. http://www.focus.de/digital /internet/tid-11272/suchmaschine-cuil-groesser-als-google_aid_320781.html. Erstelldatum: 28.07.2008 Druckdatum: 01.10.2008

Kneip, M.: Data Mining. GRIN, München/Ravensburg 2008

Koitz, R.: Informatikrecht : schnell erfasst. Springer, Berlin u. a. 2002

Kollmann, T.: E-entrepreneurship: Grundlagen der Unternehmensgründung in der net Economy. Gabler, Wiesbaden 2006

Kollmann, T.: Online-Marketing : Grundlagen der Absatzpolitik in der Net Economy. Kohlhammer, Stuttgart 2007

Kretschmar, O., Dreyer, R.: Medien-datenbank- und Medien-logistik-systeme: Anforderungen und praktischer Einsatz. Oldenbourg, München u. a. 2004

Kühn, M.: Moderne Software-Werkzeuge. http://www.wallst.de/handel9.pdf. Erstelldatum: 01.05.1998, Druckdatum: 01.10.2008

Langner, S.: Viral Marketing: Wie sie Mundpropaganda gezielt auslösen und Gewinn bringend nutzen. Gabler, Wiesbaden 2007

Linder, A., Wehrli, P.: Web Mining- die Fallstudie Swarovski: Theoretische Grundlagen und praktische Anwendungen. DUV, Wiesbaden 2005

Linoff, G., Berry, M.: Data mining techniques: for marketing, sales, and customer relationship management. 2. Auflage, Indianapolis Ind., Wiley 2004

Madlberger, M.: Electronic Retailing: Marketinginstrumente und Marktforschung im Internet. DUV, Wiesbaden 2004

Markoff, J.: Entrepreneurs See a Web Guided by Common Sense. http://www.nytimes.com /2006/11/12/business/12web.html?_r=1&oref=slogin. Erstelldatum: 12.11.2006, Druckdatum: 01.10.2008

Markov, Z., Larose, D.: Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage. Wiley, New Jersey 2007

Marks, P.: Pentagon sets its sights on social networking websites. http://www.newscientist.com/article/mg19025556.200. Erstelldatum: 09.06.2006, Druckdatum: 01.10.2008

Meusers, R.: Personalisierte Werbung für freizügige Nutzer. http://www.spiegel.de/netzwelt /web/0,1518,druck-506646,00.html. Erstelldatum: 19.09.2007, Druckdatum: 01.10.2008

Mühlenbeck, F., Skibicki, K.: Community Marketing Management: Wie man online-communities im Internet-Zeitalter des Web 2.0 zum Erfolg führt. Books on Demand, Nordestedt 2008

Nauck, D. u. a.: Neuro-Fuzzy-Systeme: Von den Grundlagen künstlicher neuronaler Netze zur Kopplung mit Fuzzy-Systemen. Vieweg+Teubner, Braunschweig/Wiesbaden 2003

Needleman, R.: New search engine Cuil takes aim at Google. http://news.cnet.com/new-search-engine-cuil-takes-aim-at-google. Erstelldatum: 27.08.2008 Druckdatum: 01.10.2008

o. A.: Anti-Mafia-Methoden gefordert. http://www.manager-magazin.de/it/artikel/ 0,2828,573108,00.html. Erstelldatum: 19.08.2008, Druckdatum: 01.10.2008

o. A.: Google Inc. http://de.wikipedia.org/wiki/Google_Inc. Erstelldatum: 03.04.2008, Druckdatum: 01.10.2008

o. A.: Stören Sie zunehmen Werbe-E-mails die man unaufgefordert bekommt? http://de.statista.org/statistik/diagramm/studie/22881/umfrage/zunehmende-stoerung-durch-werbe-e-mails-(spam). Erstelldatum: 01.02.2008. Druckdatum: 01.10.2008

o. A.: Одноклассники.ру вскрыли структуру базирования ВС РФ. http://www.rndcnews.ru/army/news/top/index_science.shtml?2008/02/22/289393. Erstelldatum: 22.02.2008, Druckdatum: 01.10.2008

o.A.: Entschädigung für Myspace. http://www.manager-magazin.de/it/artikel/ 0,2828,553154,00.html. Erstelldatum: 14.05.2008, Druckdatum: 01.10.2008

Patalong, F., Stöcker, C.: Web-Gemeinde zwingt Google zu Chromer-Korrekturen. http://www.spiegel.de/netzwelt/tech/0,1518,576186,00.html. Erstelldatum: 04.08.2008, Druckdatum: 01.09.2008

Perner, P.: Machine Learning and Data Mining in Pattern Recognition. Springer, Berlin u. a. 2007

Peterson, H.: Data Mining: Verfahren, Prozesse, Anwendungsarchitektur. Oldenbourg, München u.a. 2005

Pietsch, T., Memmler, T.: Balanced Scorecard erstellen: Kennzahlenermittlung mit Data Mining. Erich Schmidt Verlag, Berlin 2003

Prado, H., Ferneda, E.: Emerging Technologies of Text Mining: Techniques and Applications. Idea Group Inc., Hershey 2007

Preißner, A.: Promotionsratgeber. 4. Auflage, Oldenbourg, München u.a. 2001

Remmert, J.: Fälscher machen auch vor Bremsen und Viagra nicht halt. http://www.faz.net/s/RubBEFA4EA6A59441D98AC2EC17C392932A/Doc~EA6D9DF2230004609AEDD31E7493B3A8F~ATpl~Ecommon~Scontent.html. Erstelldatum: 06.01.2006, Druckdatum: 01.10.2008

Roche, J.: Handbuch Mediendidaktik Fremdsprachen. Hueber, Ismaning 2008

Salmen, S.: Electronic Relationship Marketing im Bankgeschäft: Individualisierte Kundenbeziehungen- Schlüssel zum private Internet-banking. Gabler, Wiesbaden 2003

Salmen, S.: Handbuch Electronic Customer Care: Der Weg zur digitalen Kundennähe. Springer, Heidelberg 2004

Säuberlich, F.: Web Mining: Effektives Marketing im Internet. In: Wiedmann, P (Hrsg.).: Neuronale Netze im Marketing-Management: Praxisorientierte Einführung in modernes Data-mining. Gabler, Wiesbaden 2003. S. 129-146

Scheffer, S., Bickel, T.: Multi-View Clusterung. Humboldt Universität zu Berlin, Berlin 2004

Schildhauer, T (Hrsg.).: Lexikon Electronic Business. Oldenbourg, München 2003

Schmitz, R.: Kompendium Medieninformatik: Medienpraxis. Springer, Berlin u. a. 2007

Schumacher, J., Meyer, M., Amberg, M.: Customer Relationship Management strukturiert dargestellt: Prozesse, Systeme, Technologien. Springer, Berlin 2003

Scime, A.: Web Mining: Applications and Techniques. Idea Group Inc., Hershey 2005

Segaran, T.: Kollektive Intelligenz analysieren, programmieren und nutzen. Lassen sie User-Daten für sich arbeiten. O'Reilly, Köln 2008

Smith, T.: Power to the people: Social Media Tracker WAVE3. http://www.universalmccann.com/Assets/wave_3_20080403093750.pdf. Erstelldatum:14.04.2008 Druckdatum:01.09.2008

Srivastava, J. u. a.: Web usage Mining: Discovery and applications of Web usage patterns from Web data. In: ACM exporations, 2 (2000) S. 12-23

Stegbauer, C.: Grenzen virtueller Gemeinschaft. VS, Wiesbaden 2001

Stock, W.: Information Retrieval: Informationen suchen und finden. Oldenbourg, München 2007

Thuraisingham, S., Koutroumbas, K.: Pattern recognition. Academic Press, Amsterdam u. a. 2001

Thurau, V.: Algorithmische Graphentheorie. Oldenbourg, München 2004

Uno, Y., Ota, Y., Uemichi, A.: Web Structure Mining by Isolated Stars. Springer, Berlin u. a. 2008

Vakali, A., Pallis, G.: Web Data Management Practices: Emerging Techniques and Technologies. Idea Group Inc., Hershey 2007

Velasquez, J., Palade, V.: Adaptive Web Sites: A Knowledge Extractions from Web Data Approach. IOS Press, Amsterdam u. a. 2008

Warnow, T., Zhou, B.: Computing and Combinatorics: 9th Annual International Conference. Springer, Berlin u. a. 2003

Wieschowski, S.: Studenten demonstrieren gegen SchnüffelVZ. http://www.spiegel.de/ netzwelt/web/0,1518,523906,00.html. Erstelldatum: 18.12.2007, Druckdatum: 01.09.2008

Wilkens, S.: Optionsbewertung und Risikomanagement unter gemischten Verteilungen: Theoretische Analyse und empirische Evaluation am europäischen Terminmarkt. DUV, Wiesbaden 2003

Zaiane, O. (Hrsg.): MiningWeb Data for Discovering Usage Patterns and Profiles. Springer, Berlin u. a. 2002

Zhou, Z., Li, H., Yang, Q.: Advances in Knowledge Discovery and Data Mining. Springer, Berlin u. a. 2007

Ziser, S.: Was ist Web 2.0? GRIN, München/Ravensburg 2007

Riemenschneider, M.: Der Wert von Produktvielfalt Wirkung großer Sortimente auf das Verhalten von Konsumenten: Wirkung großer Sortimente auf das Verhalten von Konsumenten. DUV, Wiesbaden 2006

IWI Discussion Paper Series / Diskussionsbeiträge

ISSN 1612-3646

Michael H. Breitner, *Rufus Philip Isaacs and the Early Years of Differential Games*, 36 p., #1, January 22, 2003.

Gabriela Hoppe and Michael H. Breitner, *Classification and Sustainability Analysis of e-Learning Applications*, 26 p., #2, February 13, 2003.

Tobias Brüggemann und Michael H. Breitner, *Preisvergleichsdienste: Alternative Konzepte und Geschäftsmodelle*, 22 S., #3, 14. Februar, 2003.

Patrick Bartels and Michael H. Breitner, *Automatic Extraction of Derivative Prices from Webpages using a Software Agent,* 32 p., #4, May 20, 2003.

Michael H. Breitner and Oliver Kubertin, *WARRANT-PRO-2: A GUI-Software for Easy Evaluation, Design and Visualization of European Double-Barrier Options,* 35 p., #5, September 12, 2003.

Dorothée Bott, Gabriela Hoppe und Michael H. Breitner, *Nutzenanalyse im Rahmen der Evaluation von E-Learning Szenarien*, 14 S., #6, 21. Oktober, 2003.

Gabriela Hoppe and Michael H. Breitner, *Sustainable Business Models for E-Learning*, 20 p., #7, January 5, 2004.

Heiko Genath, Tobias Brüggemann und Michael H. Breitner, *Preisvergleichsdienste im internationalen Vergleich*, 40 S., #8, 21. Juni, 2004.

Dennis Bode und Michael H. Breitner, *Neues digitales BOS-Netz für Deutschland: Analyse der Probleme und mögliche Betriebskonzepte,* 21 S., #9, 5. Juli, 2004.

Caroline Neufert und Michael H. Breitner, *Mit Zertifizierungen in eine sicherere Informationsgesellschaft*, 19 S., #10, 5. Juli, 2004.

Marcel Heese, Günter Wohlers and Michael H. Breitner, *Privacy Protection against RFID Spying: Challenges and Countermeasures*, 22 p., #11, July 5, 2004.

Liina Stotz, Gabriela Hoppe und Michael H. Breitner, *Interaktives Mobile(M)-Learning auf kleinen End-geräten wie PDAs und Smartphones*, 31 S., #12, 18. August, 2004.

Frank Köller und Michael H. Breitner, *Optimierung von Warteschlangensystemen in Call Centern auf Basis von Kennzahlenapproximationen,* 24 S., #13, 10. Januar, 2005.

Phillip Maske, Patrick Bartels and Michael H. Breitner, *Interactive M(obile)-Learning with UbiLearn 0.2*, 21 p., #14, April 20, 2005.

Robert Pomes and Michael H. Breitner, *Strategic Management of Information Security in State-run Organizations,* 18 p., #15, May 5, 2005.

Simon König, Frank Köller and Michael H. Breitner, *FAUN 1.1 User Manual,* 134 p., #16, August 4, 2005.

Christian von Spreckelsen, Patrick Bartels und Michael H. Breitner, *Geschäftsprozessorientierte Analyse und Bewertung der Potentiale des Nomadic Computing,* 38 S., #17, 14. Dezember, 2006.

Stefan Hoyer, Robert Pomes, Günter Wohlers und Michael H. Breitner, *Kritische Erfolgsfaktoren für ein Computer Emergency Response Team (CERT) am Beispiel CERT-Niedersachsen,* 56 S., #18, 14. Dezember, 2006.

Christian Zietz, Karsten Sohns und Michael H. Breitner, *Konvergenz von Lern-, Wissens- und Personalmanagementssystemen: Anforderungen an Instrumente für integrierte Systeme,* 15 S., #19, 14. Dezember, 2006.

Christian Zietz und Michael H. Breitner, *Expertenbefragung „Portalbasiertes Wissensmanagement“: Ausgewählte Ergebnisse*, 30 S., #20, 5. Februar, 2008.

IWI Discussion Paper Series/Diskussionsbeiträge

ISSN 1612-3646

Harald Schömburg und Michael H. Breitner, *Elektronische Rechnungsstellung: Prozesse, Einsparpotentiale und kritische Erfolgsfaktoren*, 36 S., #21, 5. Februar, 2008.

Halyna Zakhariya, Frank Köller und Michael H. Breitner, *Personaleinsatzplanung im Echtzeitbetrieb in Call Centern mit Künstlichen Neuronalen Netzen*, 35 S., #22, 5. Februar, 2008.

Jörg Uffen, Robert Pomes, Claudia M. König und Michael H. Breitner, *Entwicklung von Security Awareness Konzepten unter Berücksichtigung ausgewählter Menschenbilder*, 14 S., #23, 5. Mai, 2008.

Johanna Mählmann, Michael H. Breitner und Klaus-Werner Hartmann, *Konzept eines Centers der Informationslogistik im Kontext der Industrialisierung von Finanzdienstleistungen*, 19 S., #24, 5. Mai, 2008.

Jon Sprenger, Christian Zietz und Michael H. Breitner, *Kritische Erfolgsfaktoren für die Einführung und Nutzung von Portalen zum Wissensmanagement*, 44 S., #25, 20. August, 2008.

Finn Breuer und Michael H. Breitner, *„Aufzeichnung und Podcasting akademischer Veranstaltungen in der Region D-A-CH": Ausgewählte Ergebnisse und Benchmark einer Expertenbefragung*, 30 S., #26, 21. August, 2008.

Harald Schömburg, Gerrit Hoppen und Michael H. Breitner, *Expertenbefragung zur Rechnungseingangsbearbeitung: Status quo und Akzeptanz der elektronischen Rechnung*, 40 S., #27, 15. Oktober, 2008.

Hans-Jörg von Mettenheim, Matthias Paul und Michael H. Breitner, *Akzeptanz von Sicherheitsmaßnahmen: Modellierung, Numerische Simulation und Optimierung,* 30 S., #28, 16. Oktober, 2008.

Markus Neumann, Bernd Hohler und Michael H. Breitner, *Bestimmung der IT-Effektivität und IT-Effizienz serviceorientierten IT-Managements,* 20 S., #29, 30. November, 2008.

Matthias Kehlenbeck und Michael H. Breitner, *Strukturierte Literaturrecherche und -klassifizierung zu den Forschungsgebieten Business Intelligence und Data Warehousing,* 10 S., #30, 19. Dezember, 2009.

Michael H. Breitner, Matthias Kehlenbeck, Marc Klages, Harald Schömburg, Jon Sprenger, Jos Töller und Halyna Zakhariya, *Aspekte der Wirtschaftsinformatikforschung 2008,* 128 S., #31, 12. Februar, 2009.

Sebastian Schmidt, Hans-Jörg v. Mettenheim und Michael H. Breitner, *Entwicklung des Hannoveraner Referenzmodels für Sicherheit und Evaluation an Fallbeispielen,* 30 S., #32, 18. Februar, 2009.

Sissi Eklu-Natey, Karsten Sohns und Michael H. Breitner, *Buildung-up Human Capital in Senegal - E-Learning for School drop-outs, Possibilities of Lifelong Learning Vision,* 39 p., #33, July 1, 2009.

Horst-Oliver Hofmann, Hans-Jörg von Mettenheim und Michael H. Breitner, *Prognose und Handel von Derivaten auf Strom mit Künstlichen Neuronalen Netzen,* 34 S., #34, 11. September, 2009.

Christoph Polus, Hans-Jörg von Mettenheim und Michael H. Breitner, *Prognose und Handel von Öl-Future-Spreads durch Multi-Layer-Perceptrons und High-Order-Neuronalnetze mit Faun 1.1 ,* 55 S., #35, 18. September, 2009.

Jörg Uffen und Michael H. Breitner, *Stärkung des IT-Sicherheitsbewusstseins unter Berücksichtigung psychologischer und pädagogischer Merkmale*, 37 S., #36, 24. Oktober, 2009.

Christian Fischer und Michael H. Breitner, *MaschinenMenschen – reine Science Fiction oder bald Realität?,* 36 S., #37, 13. Dezember, 2009.

Tim Rickenberg, Hans-Jörg von Mettenheim und Michael H. Breitner, *Plattformunabhängiges Softwareengineering eines Transportmodells zur ganzheitlichen Disposition von Strecken- und Flächenverkehren,* 38 S., #38, 11. Januar, 2010.

IWI Discussion Paper Series / Diskussionsbeiträge

ISSN 1612-3646

Björn Semmelhaack, Jon Sprenger und Michael H. Breitner, *Ein ganzheitliches Konzept für Informationssicherheit unter besonderer Berücksichtigung des Schwachpunktes Mensch*, 56 S., #39, 03. Februar, 2009.

Markus Neumann, Achim Plückebaum, Jörg Uffen und Michael H. Breitner, *Aspekte der Wirtschaftsinformatikforschung 2009*, 70 S., #40, 12. Februar, 2010.

Markus Neumann, Bernd Hohler und Michael H. Breitner, *Wertbeitrag interner IT – Theoretische Einordnung und empirische Ergebnisse*, 38 S., #41, 31. Mai, 2010.

Daniel Wenzel, Karsten Sohns und Michael H. Breitner, *Open Innovation 2.5: Trendforschung mit Social Network Analysis*, 46 S., #42, 1. Juni, 2010.