Analyse der Potenziale betrieblicher Anwendungen des Web Content Mining

Diplomarbeit

zur Erlangung des Grades eines Diplom-Ökonoms an der Leibniz Universität Hannover

vorgelegt von

Naum Neuhaus

[redacted]

Erstprüfer: Prof. Dr. M. H. Breitner

Hannover, 11.11.2008

Inhaltsverzeichnis

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, im Internet veröffentlichter Inhalte reißt nicht ab, denn jeden Tag kommen etwa eine Milliarde weiterer Seiten hinzu. Die Vielfalt und der Umfang der im WWW zugänglichen Informationen ist enorm. Die Vermutung, dass sich in diesem Informationsdschungel wahre Schätze verbergen, ist daher mehr als berechtigt. Aus betrieblicher 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 zahlreiche Unternehmen bietet das Internet die Möglichkeit, Informationen in einem großen Umfang finden und sammeln zu lassen. Trotz technischen Fortschritts erfolgt die betriebliche Suche, Interpretation und Integration von Internetdokumenten noch überwiegend manuell mithilfe 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 wachsendem WWW immer aufwendiger und damit teuerer 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 mithilfe 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.

1.1 Forschungsfragen

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?

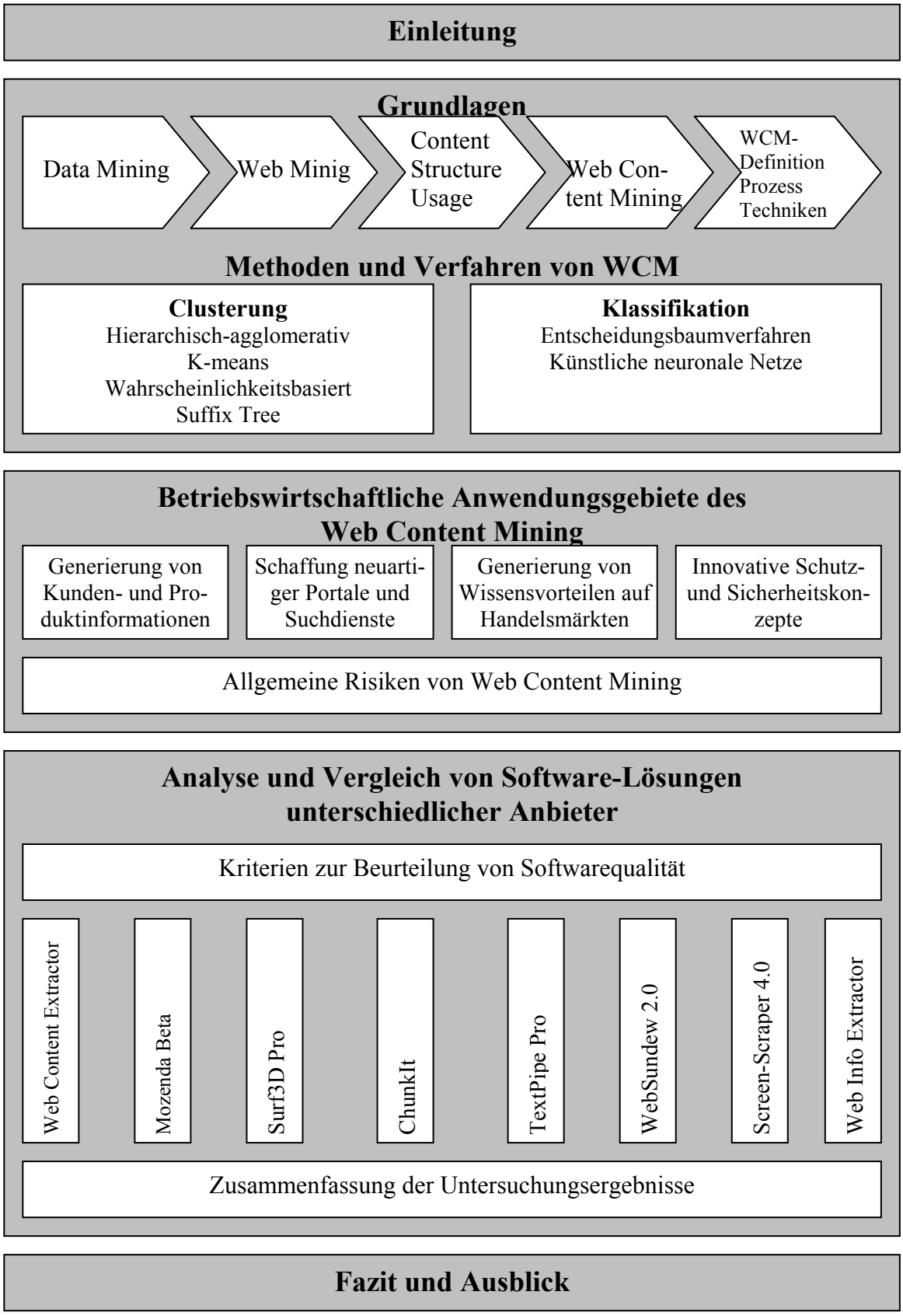
1.2 Ablauf der Untersuchung

Der Verlauf der Untersuchung ist in vier Kapitel unterteilt. Im Grundlagenteil erfolgt eine theoretische Auseinandersetzung mit WCM. Darin werden grundlegende Begriffe und Aspekte von WCM definitorisch erläutert und in einem Zusammenhang mit verwandten und übergeordneten Disziplinen dargestellt. Daraufhin werden einige ausgewählte Methoden und Verfahren von 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 erfolgt die Herleitung konkreter Geschäftskonzepte, die sich mithilfe der automatischen Inhaltsextraktion schon jetzt oder erst in Zukunft realisieren ließen. Dabei werden elf exemplarische Anwendungsgebiete auf ihr Potenzial für die Nutzung von WCM-Methoden untersucht. Es wird betrachtet, inwieweit WCM dort bereits jetzt verbreitet ist und welchen Mehrwert es künftig bringen könnte. Und zuletzt werden in dieses Kapitel Gefahren von WCM in Hinblick auf rechtliche, technische und werbewirksame Risiken thematisiert.

Im fünften 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 rausstellen, welche entwickelten Anwendungsideen bereits jetzt ausführbar sind und welches WCM-Programm für welche Anforderungen in Frage kommt.

Abbildung 1: Schematischer Aufbau der Diplomarbeit



Quelle: Eigene Erstellung

6 Fazit und Ausblick

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. Darin wurden alle verwandten Disziplinen definiert und in den Rahmen von WCM-Anwendungen eingeordnet. Unerlässlich war eine Abgrenzung von WCM zu den restlichen Web Mining-Disziplinen. Trotz der strikten Trennung zwischen WCM, Web Structure Mining und Web Usage Mining in der Literatur, haben sich dennoch einige Schnittflächen gezeigt. Das Kennen der Struktur einer oder verbundener Webseiten ist z. B. bei der Einstellung eines WCM-Agents unerlässlich. Darüber hinaus kann WCM dazu dienen, auf Basis von Web Usage Mining, generierte Nutzerprofile um personifizierte Inhalte aus sozialen Netzwerken, Blogs, etc. zu ergänzen.

Die ständige Präsenz zahlreicher Clusterung- und Klassifizierungsmethoden innerhalb der Fachliteratur erforderte eine rudimentäre Darstellung ausgewählter Algorithmen. Dabei konnte einige Algorithmen durchaus als Vorstufe zur ersten künstlichen Intelligenz identifiziert werden. Die bereits erfolgreich angewendeten künstlichen Neuronalen Netze bei der Bewertung von Finanzinstrumenten (vgl. Bartels 2008) sowie die Suffix Tree Clusterung im Rahmen von Text Mining verzeichnen derzeit den größten Erfolg.

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.

Seine Potenziale im Rahmen der Marktforschung konnten bestätigt und fundiert werden. Hierbei wurden u. a. verfügbare Daten innerhalb von Communities, den Nachfragern derartiger Informationen gegenübergestellt. Dadurch konnte die Extraktion zahlreicher Informationen über Mitglieder sozialer Netzwerke im Zusammenhang mit der Generierung von Kunden- und Produktinformationen hervorgehoben werden.

Mithilfe von WCM können neuartige Portale und Archive generiert sowie Online-Suchdienste optimiert werden. Bei der Betrachtung von Google & Co. wurde zwischen Sucheingabe und Ausgabe der Suchergebnisse unterschieden. Während im Rahmen der Sucheingabe, WCM-verwandte Methoden wie NLP von größter Bedeutung sind spielt bei der Präsentation der Ergebnisse eine inhaltsbasierte Sicht auf die Webdokumente eine wichtige Rolle. Hierfür müssen sämtliche indexierten Webdokumente bereits zuvor, mittels Klassifizierungs- und Clusterungsalgorithmen typisiert und geordnet werden. Vor dem Hintergrund des exponentiell wachsenden Internets, kann dem Nutzer in absehbarer Zeit nicht mehr zugemutet werden, sich auf der Suche nach Inhalten, durch den Webdschungel zu kämpfen. An Suchdiensten wie http://cuil.com oder http://quintura.com konnte gezeigt werden, dass nicht erst die Recherche innerhalb der vielen gefunden Seiten, sondern schon die Art in der die Resultate aufbereitet und präsentiert werden, Erkenntnisse über den Inhalt vermitteln kann. Ferner eignen sich WCM-Programme um Meta-Webseiten auf Basis zahlreicher bestehender Archive und Portale zu entwickeln. Meta-Nachrichtenarchive, -Produktvergleiche und –Netzwerke konnten hierbei hervorgehoben und genauer betrachtet werden.

Die Suche nach Inhalten innerhalb multimedialer Daten wie Bildern, Audio- und Video-Dateien stellt zwar derzeit einen stark erforschten und viel versprechenden, dennoch aber einen unausgereiften Mining-Bereich dar. Aufgrund des hohen Rechenaufwands und der mangelnden Reife von effizienten Methoden, sollten die Möglichkeiten von Image, Audio und Video Mining und ihre Anwendung im Kontext von WCM zum gegenwärtigen Zeitpunkt nicht überschätzt werden. Die Integration von Multimedia Mining in künftige WCM-Programme ist zwar derzeit noch nicht möglich, aber vor dem Hintergrund starker Verbreitung von Videos und Bildern im Internet auf lange Sicht unerlässlich.

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. Mithilfe 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 (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.

Die Identifikation und Einordnung von Webinhalten spielt auch in Verbindung mit Onlineschutz- und Sicherheitskonzepten eine große Rolle. Im Hinblick auf Kinder- und Jugendschutz versetzt eine Kombination aus Text Mining, Image Mining, NLP sowie moderne Klassifizierungsmethoden wie z. B. KNN eine Anwendung in die Lage effizient zwischen gefährdenden und ungefährdenden Seiten zu unterscheiden. Der Markt für zuverlässige Kinderschutzfilter ist noch lange nicht gesättigt und stellt in Zukunft ein großes Potenzial für den Einsatz von WCM-Anwendungen dar. Das, derzeit noch unausgereifte Multimedia Mining könnte in Verbindung mit WCM dazu beitragen, eine Vielzahl von Urheberrechtsverletzungen und Plagiaten im Internet ausfindig zu machen.

Als letztes Anwendungsgebiet konnte der Kampf gegen Spam betrachtet werden. Eine eigene Disziplin, die sich mit der Suche und Extraktion von Spam innerhalb großer unstrukturierter und stark fluktuierende Datenbestände, mithilfe von Methoden und Verfahren des Web Mining beschäftigt, könnte in Zukunft eventuell Spam Mining heißen. Als eine spezialisierte Art des WCM wäre es Aufgabe von Spam Mining zwischen echten und manipulierten Seiten, Links, Kommentaren, Profilen und sonstigen Einträgen zu unterscheiden.

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 Zu-

stimmung 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 betrieb liche 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. Mithilfe 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 Webinhal-

ten 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 Ana-

lysemöglichkeiten 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.