Kilian Stoffel

Web + Data Mining = Web Mining

Dieser Beitrag gibt eine kurze Beschreibung, was unter dem Begriff Web Mining zu verstehen ist und wie Webdaten mit gängigen Data-Mining-Techniken kombiniert werden können, um das Web besser als Informationsquelle nutzbar zu machen. Des Weiteren wird ein Überblick über die gängigen Hauptachsen gegeben, entlang derer die meisten aktuellen Entwicklungen stattfinden. Diese Ausführungen werden mit einem Ausblick auf mögliche zukünftige Entwicklungen abgeschlossen, die vor allem durch die neusten Trends in der Benutzung des World Wide Web vorgezeichnet sind.

Inhaltsübersicht

1 Nutzbarmachung des Web

Das World Wide Web (oder kurz Web) ist heute zu einer der wichtigsten oder vielleicht sogar zur wichtigsten Informationsquelle avanciert. Neben dem Verteilen von Informationen aller Art hat sich das Web auch zu einer der wichtigsten Business-Plattformen (Infrastruktur zum Austausch von Geschäftsdaten, -informationen und -wissen) entwickelt.

Mit der Vielfalt an Information, die im Web gefunden werden kann, haben sich auch die Formate der verwendeten Dateien weiterentwickelt. Neben den traditionellen Textdateien können auch immer mehr Multimediadateien wie Bild-, Audio- und Videodateien gefunden werden. Der Dateityp beeinflusst entscheidend, wie die Benutzer mit dem Web interagieren.

Die häufigste Interaktion mit dem Web ist sicher das Auffinden und Konsumieren von Informationen. Informationen können entweder über Navigation gefunden werden oder mithilfe von Suchmaschinen wie z.B. Google und Yahoo. Mit der zunehmenden Größe des Web nimmt auch die Bedeutung der Suchmaschinen immer mehr zu. Durch einfaches Navigieren kann kaum mehr eine relevante Information gefunden werden. Auch die Suchmaschinen, die mit den traditionellen Hilfsmitteln wie Schlüsselwortsuche oder Ähnlichkeitsanfragen versuchen, die gesuchten Daten zu finden, gelangen immer mehr an ihre Grenzen. Hier setzt das Web Mining ein.

Web Mining versucht, das World Wide Web in eine nützlichere Umgebung umzuwandeln, in der ein Benutzer schneller und einfacher die gewünschten Informationen finden kann. Dies beinhaltet das Auffinden und Analysieren von Daten, Textdokumenten und Multimediadaten.

Web Mining kombiniert zwei Technologien: auf der einen Seite das Web und auf der anderen Seite das Data Mining. Diese beiden und alle dazugehörenden Begriffe werden im Folgenden etwas näher erörtert.

1.1 Web

Im Kontext vom Web Mining spielt das Web die Rolle der Datenquelle. Mit der rasanten Geschwindigkeit, mit der sich das Web in den letzten Jahren entwickelt hat, ist es wahrscheinlich zur größten öffentlich zugänglichen Datenquelle geworden. Das Web hat einige spezifische Eigenschaften, die die Suche (Mining) nach nützlicher Information sehr interessant, aber auch schwierig gestalten. Im Folgenden werden einige dieser Eigenschaften etwas näher betrachtet. Die hier gegebene Liste der Eigenschaften hält sich an den Vorschlag in [Bing 2007].

1. Die *Größe des Web*: Die Größe als solche ist schon eine große Herausforderung. Erschwerend kommt hinzu, dass das Web immer noch weiter wächst.
2. Die *Datentypen* sind *verschiedenster Art*: Die Datentypen erstrecken sich von strukturierten Tabellen über teils strukturierte HTML-Webseiten bis hin zu unstrukturierten Text- und Multimediadateien.
3. Die *Datenkonsistenz* ist *nicht garantiert*: Aufgrund der großen Anzahl von Autoren lässt die Konsistenz der Informationen häufig zu wünschen übrig. Es werden unterschiedliche Worte, aber auch verschiedene Formate verwendet.
4. Ein Großteil der *Daten* ist untereinander *vernetzt*: Hyperlinks sind eine der wesentlichen Charakteristiken des Web. Links ermöglichen das Vernetzen von Daten innerhalb einer Webseite, zwischen Webseiten auf dem gleichen Server oder zwischen verschiedenen Servern. Häufig werden diesen Links neben der Verknüpfung noch andere Bedeutungen zugeordnet, nämlich Autorität (authority) oder Vertrauen (trust).
5. Webdaten sind »*noisy*«. Auf vielen Seiten ist die wesentliche Information in einer Vielzahl von Zusatzinformationen wie z.B. Autorenrechten, Werbung und Links versteckt.
6. Es *fehlt* eine *Qualitätskontrolle*: Im Prinzip ist jeder frei, den Inhalt seiner Seiten selber festzulegen. Es gibt Webseiten, deren Inhalt kontrolliert ist. Dabei handelt es sich jedoch um eine kleine Minderheit.
7. Viele *Webseiten* stellen *Dienstleistungen* (services) zur Verfügung: Diese müssen von den Benutzern aktiviert werden, um die vorhandenen Informationen freizusetzen.
8. Das *Web* ist *dynamisch*: Die vorhandenen Informationen können sich ständig ändern oder ganz verschwinden. Diese Änderungen mitzuverfolgen ist eine große Herausforderung.
9. Das Web repräsentiert *virtuelle Gesellschaften*: Neben dem Bereitstellen von Informationen und Dienstleistungen wird das Web aber auch zur Kommunikation zwischen Personen verwendet. Diese Aktivität produziert ebenfalls eine große Menge an Information.

All diese Charakteristiken, die das Web als Datenquelle aufweist, repräsentieren auf der einen Seite Herausforderungen für die Suche nach Informationen, aber gleichzeitig auch eine Chance und nie da gewesene Möglichkeiten.

Um das Web Mining zu erörtern, ist es wichtig, nach der ersten Komponente, also dem Web, auch die zweite Komponente, nämlich das Data Mining, etwas genauer unter die Lupe zu nehmen.

1.2 Data Mining

Der Term Data Mining wird heute oft synonym mit dem Term Knowledge Discovery in Databases (KDD) verwendet. KDD beschreibt jedoch den gesamten Prozess, der zum Ziel hat, aus Datenquellen nützliche Regelmäßigkeiten (Patterns) herauszufinden. Data Mining ist ein Teil dieses Prozesses [Srivastava et al. 2004]. Bevor das Web Mining genauer beschrieben wird, werden zunächst einige Schlüsselbegriffe etwas näher betrachtet [Dunham 2003].

Regelmäßigkeiten/Patterns müssen einigen Kriterien genügen:

Gültigkeit (validity): Die Patterns müssen von den Daten unterstützt werden, d.h., sie müssen genügend oft in den zugrunde liegenden Daten vorkommen.

Nützlichkeit (usefulness): Die Nützlichkeit lässt sich nur über den Kontext definieren, in dem das Data-Mining-Projekt bzw. das Web-Mining-Projekt durchgeführt wird.

Verständlichkeit (understandability/meaningfulness): Dieses Kriterium geht in die gleiche Richtung wie das zweite. Es besteht immer die Möglichkeit, in großen Mengen Regelmäßigkeiten zu finden, vor allem wenn die Anforderungen an die Gültigkeit nicht zu hoch sind. Umso wichtiger ist es, dass die gefundenen Regeln im vorgegebenen Kontext auch Sinn machen und eventuell die Kenntnisse oder die Verhaltensmuster des Data Miner verändern können.

Es ist klar, dass eine solch anspruchsvolle Zielsetzung nicht nur mit einer einzelnen Methode erreicht werden kann. Aus diesem Grund ist Data Mining als Gebiet auch ausgesprochen interdisziplinär: Data Mining integriert Ideen aus den verschiedensten Gebieten wie Statistik, Machine Learning, Datenbanken, Künstliche Intelligenz und Visualisierung. Diese Liste ist bei Weitem nicht vollständig. Im Verlauf dieses Beitrags wird gezeigt werden, dass im Kontext des Web Mining z.B. das Information Retrieval eine wichtige Rolle spielt.

Genauso vielfältig wie die Gebiete, die ins Data Mining eingeflossen sind, sind auch die Aufgaben, die mittels Data Mining gelöst werden können. Dieser Beitrag konzentriert sich auf zwei Gruppen: *gezieltes/gerichtetes* (supervised) und *ungezieltes/ungerichtetes* (unsupervised) Data Mining.

Gezieltes Data Mining arbeitet mit Trainingsbeispielen. Von diesen Trainingsbeispielen werden dann die Regeln gelernt, die anschließend auf die übrigen Daten angewandt werden. Typische Beispiele für diese Art von Data Mining sind Klassifikationsprobleme.

Ungezieltes Data Mining versucht, Ordnung bzw. Struktur in eine Datenmenge zu bringen. Hier werden keine Trainingsdaten verwendet, sondern direkt die richtigen zu strukturierenden Daten. Ein typisches Beispiel aus dieser Familie von Aufgaben ist das Clustering. Diese Unterscheidung wird später verwendet, um eine Web-Mining-Taxonomie zu definieren.

Bevor das Web Mining genauer beschrieben werden kann, soll zuerst noch kurz der gesamte KDD-Prozess etwas detaillierter dargestellt werden.

Auch wenn dieser Prozess aus einer großen Anzahl kleiner Schritte zusammengesetzt ist, können drei wesentliche Etappen isoliert werden, nämlich *Vorbereitung* (pre-processing), *Data Mining* und *Nachbereitung* (post-processing).

Während der Vorbereitung geht es darum, die Daten an die Anforderungen des Data Mining anzupassen. Die meisten Datensets können nicht in ihrer Rohform direkt von Data-Mining-Algorithmen verwendet werden. Einige Probleme, die die direkte Verwendung von Data-Mining-Algorithmen sehr erschweren und in dieser ersten Phase des KDD-Prozesses behoben werden, sind nämlich Anpassung der Größe des Datensatzes, Auswahl der relevanten Attribute, Entfernen von Anomalien und das Vereinheitlichen der Codierungen.

Sobald die Daten gereinigt und vorbereitet sind, können sie den Data-Mining-Algorithmen zugeführt werden. Diese suchen nach Regelmäßigkeiten/Patterns. Diese Algorithmen, vor allem jene, die im Web Mining zum Einsatz kommen, werden in den nächsten Abschnitten etwas genauer erörtert.

Während der Nachbearbeitungsphase werden aus der Menge der Patterns diejenigen ausgesucht, die den oben genannten Anforderungen genügen. Hauptsächlich geht es darum, nützliche Regeln zu finden, wobei die Nützlichkeit über die Aufgabenstellung definiert ist. Um zu entscheiden, welche Regeln beibehalten

werden, können verschiedene Bewertungs- und Visualisierungshilfsmittel eingesetzt werden.

Dieser Prozess ist meist iterativ, d.h., häufig sind mehrere vollständige oder auch Teildurchgänge notwendig, um befriedigende Resultate zu erzielen, die dann auch in der realen Welt verwendet werden können.

Traditionelles Data Mining verwendet strukturierte Daten in Form von Tabellen, die meist von einer relationalen Datenbank stammen. Wie die Data-Mining-Techniken auf Webdaten mit den oben beschriebenen Eigenschaften angewandt werden können, wird nun im Folgenden dargestellt.

1.3 Web + Data Mining = Web Mining

Kombiniert man die beiden zuvor beschriebenen Technologien Web und Data Mining, so ergibt sich eine neue Umgebung, in der das Web die Daten liefert und das Data Mining nützliche Informationen und nützliches Wissen aus diesen Daten extrahiert.

Auch wenn die Idee dieser Kombination naheliegend scheint, ist sie nicht einfach zu realisieren. Dies hat im Wesentlichen mit der Art der Daten zu tun, die im Web gefunden werden können (siehe Charakteristiken in Abschnitt 1.1). Die Data-Mining-Algorithmen können nicht unverändert übernommen werden, sondern müssen an diese Gegebenheiten angepasst werden [Scime 2004]. Diese Anpassungen haben auch komplett neue Algorithmen hervorgebracht. Die neu entwickelten bzw. angepassten Algorithmen können namentlich in drei Gruppen entsprechend ihren Hauptfunktionen eingeteilt werden. Diese heißen Content Mining, Structure Mining und Usage Mining (z.B. [Hotho & Stumme 2007], [Kosala & Blockeel 2000]).

Content Mining extrahiert nützliche Informationen aus dem Inhalt von Webseiten. Als einfaches Beispiel kann das Gruppieren von Webseiten mit gleichen oder ähnlichen Inhalten in verschiedene Klassen angeführt werden.

Structure Mining entdeckt nützliche Informationen in den Hyperlinks, die die Struktur des Web festlegen. Links können z.B. dazu verwendet werden, die Wichtigkeit einer Seite zu bestimmen.

Usage Mining findet Patterns, die das Verhalten von Webnutzern beschreiben. Dieses Verhalten ist über die Klicks definiert, die ein Benutzer beim Navigieren ausführt und die in Log-Dateien gespeichert werden.

Kombiniert man nun das einfache Klassifikationssystem von Data-Mining-Algorithmen (gezielt bzw. ungezielt) mit den drei Hauptklassen für Web Mining (Content, Structure und Usage Mining) erhält man eine einfache Taxonomie, die nun im Weiteren detaillierter untersucht wird.

2 Web-Mining-Taxonomie

Um sich in der Vielzahl von Web-Mining-Algorithmen und -Anwendungen einfacher zurechtzufinden, wird nachfolgend eine Taxonomie präsentiert, die sich an den Hauptfunktionen im Bereich Web Mining orientiert und die jede Gruppe und die wesentlichsten Untergruppen beschreibt. Die Auflistung erhebt keinen Anspruch auf Vollständigkeit, sondern beschränkt sich auf die wichtigsten Methoden, Prozesse und Anwendungen.

2.1 Content Mining

Das Auffinden von relevanten Daten in Dokumenten aus dem Web ist wahrscheinlich die ursprünglichste Art des Web Mining. Im Prinzip kann jede Webseite als ein individuelles Dokument betrachtet werden. Die Menge aller Dokumente zusammen kann dann als Datenbank interpretiert werden, auf die die verschiedenen Data-Mining-Techniken angewandt werden können.

Ein sehr wichtiger Punkt ist das Unterscheiden, wie die Dokumente interpretiert werden. Dokumente können als eine *unstrukturierte* Menge von Worten aufgefasst werden oder als eine *halbstrukturierte* Menge, wobei die Struktur aus den HTML-Tags (Textstruktur, Links etc.) abgeleitet wird.

Unstrukturierte Dokumente werden meist als eine Menge von Worten interpretiert, d.h.,

die Reihenfolge der Worte spielt keine Rolle. Aus dieser Menge von Worten werden verschiedene Statistiken berechnet, die die individuellen Worte, aber auch die Dokumente beschreiben. Aus diesen Daten werden dann kompakte Repräsentationen berechnet, wie z.B. Vektoren, die dazu verwendet werden, Ähnlichkeitsgrade zwischen Worten, zwischen Dokumenten und Worten sowie zwischen Dokumenten zu berechnen. Für diese Vektoren werden häufig Indexstrukturen generiert, die dazu dienen, Beziehungen zwischen Worten und Dokumenten schnell abfragen zu können. Diese Techniken werden meist unter dem Begriff »Information Retrieval« zusammengefasst.

Bei *halbstrukturierten* Dokumenten (typischerweise HTML-Dokumenten) können weitere Informationen aus den Tags gewonnen werden. Die Tags geben zusätzliche Informationen über die Textelemente, aufgrund derer diesen Elementen verschiedene Gewichte zugewiesen werden können. Des Weiteren können Links verwendet werden. Diese definieren die eigentliche Struktur des Web, was das Thema des nächsten Abschnitts ist. Sie können aber auch verwendet werden, um die gesuchten Inhalte schneller und genauer zu finden, da die zugrunde liegende logische Struktur des Web mitverarbeitet werden kann.

Eine andere typische Content-Mining-Aufgabe ist die Informationsextraktion (information extraction), die zum Ziel hat, aus unstrukturierten oder halbstrukturierten Dokumenten strukturierte Informationen zu extrahieren. Es geht hier nicht darum, relevante Dokumente zu finden, sondern nur eine einzelne Information wie z.B. die Antwort auf eine Frage.

Beispiel: Content Mining

Als Beispiel für Content Mining wird hier eine Information-Retrieval-Anwendung beschrieben, die detailliert in [Manning et al. 2008] nachgelesen werden kann. Die Basisarchitektur dieses Systems ist in Abschnitt 3.1 unter »Information Retrieval« dargestellt.

Für dieses Beispiel wird angenommen, dass die Texte, deren Inhalte untersucht werden sollen, schon verarbeitet wurden und in Form von Tabellen (wie in Tab. 1) zur Verfügung stehen. Als Textdatenbank wurden die gesammelten Werke von Shakespeare verwendet.

Eine mögliche Aufgabe könnte darin bestehen, alle Werke zu finden, die von Brutus und Caesar, aber nicht von Calpurnia sprechen. Ein naiver Ansatz, dieses Problem zu lösen, wäre z.B., alle Dokumente zu suchen, die die Wörter Brutus und Caeser enthalten, und jene Dokumente von dieser Liste zu entfernen, die Calpurnia erwähnen. Dieses Vorgehen ist akzeptabel, solange die Textdatenbank relativ klein ist, kann aber nicht auf große Textdatenbanken, wie sie z.B. das Web darstellt, übertragen werden. Um die Abfragen schneller verarbeiten zu können, muss ein Index für die Textdatenbank erstellt werden, der die Form der Tabelle 1 annimmt.

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	…
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
Mercy	1	0	1	1	1	1	
Worser	1	0	1	1	1	0	
…							

Tab. 1: Dokument-Term-Index-Matrix

Mithilfe dieses Index kann die gleiche Anfrage wesentlich schneller beantwortet werden, da nur die Linien der Matrix, die zu den entsprechenden Wörtern gehören, verglichen werden müssen. In dem hier angegebenen Beispiel werden die Linien für Brutus und Caesar und das Komplement (d.h. 1 und 0 werden vertauscht) der Linie für Calpurnia miteinander verglichen.

110100 UND 110111 UND 101111 = 100100

Damit werden die Dokumente »Antony and Cleopatra« und »Hamlet« als Lösung zurückgegeben, die den Stellen entsprechen, an denen im Antwortvektor eine 1 steht.

Die hier präsentierte Version eines Index ist sehr einfach und kann natürlich weiter ausgebaut werden, indem z.B. anstelle von 0- und 1-Werten die Frequenzen, mit denen die Ausdrücke in einem Dokument auftreten, verwendet werden.

2.2 Structure Mining

Im Gegensatz zum Content Mining, bei dem es im Wesentlichen um den Inhalt und die Struktur innerhalb einer Webseite geht, richtet sich der Fokus beim Structure Mining nur auf die Struktur, die durch die Hyperlinks definiert ist. Es geht also um Beziehungen, die zwischen den Dokumenten existieren und somit für die Struktur des gesamten Web verantwortlich sind. Diese Struktur kann als ein gerichteter Graph oder ein Netzwerk interpretiert werden [Chakrabarti 2003].

Die Data-Mining-Algorithmen, die sich für diese Aufgaben einsetzen lassen, wurden von Algorithmen zur Analyse von Graphen und sozialen Netzwerken abgeleitet. Mittels dieser Techniken ist es möglich, spezifische Arten von Seiten zu identifizieren, basierend auf den ankommenden und abgehenden Links. Die wohl bekannteste Anwendung dieser Technologien findet sich in der Google-Suchmaschine wieder. Sie wird verwendet, um den PageRank zu berechnen.

All diese Algorithmen haben zum Ziel, die Topologie des Web zu berechnen und eventuell die Links mit inhaltlichen Informationen zu erweitern. Diese Modelle werden meist verwendet, um Qualitäts- oder Relevanzberechnungen durchzuführen.

In vielen Anwendungen werden Content- und Structure-Mining-Algorithmen miteinander kombiniert, um den Inhalt ganzheitlich betrachten zu können. In diesem Sinn bilden Structure- und Content-Mining-Algorithmen eine Gruppe, die sich klar von den Usage-Mining-Algorithmen unterscheiden, die in Abschnitt 2.3 vorgestellt werden.

Beispiel: Structure Mining

Abbildung 1 präsentiert das Resultat, das man erhält, wenn Googles Browsertool TouchGraph verwendet wird, um die Struktur des Web rund um die HMD-Website (*http://hmd.dpunkt.de*) zu untersuchen. Die HMD-Webseite befindet sich im Zentrum des Graphen, und die Seiten, die von Google als ähnlich definiert werden, sind rund um diese Seite angeordnet. Dabei werden nicht nur die ein- und ausgehenden Links berücksichtigt, sondern auch Faktoren wie z.B. andere Sites, die die zwei Webseiten als Anfangs- und Endpunkte eines Links erwähnen.

2.3 Usage Mining

Beim Usage Mining wird der Fokus vom Inhalt des Web auf seine Nutzung verlagert. Es geht also darum zu untersuchen, wie die Benutzer die Webseiten durchstöbern, d.h., in welcher Reihenfolge sie diese Seiten lesen und wie sie innerhalb des Web navigieren [Cooley 2000]. Während beim Content Mining untersucht wird, was die Autoren in die Webseiten hineingeschrieben haben, untersucht das Usage Mining, wie die Leser die Seiten konsumieren. Mittels dieser Untersuchungen können Beziehungen zwischen Seiten aufgedeckt werden, die von den Verfassern der Seiten nicht unbedingt vorgesehen waren.

Die Daten, die für diese Untersuchungen notwendig sind, finden sich in den Log-Dateien der Webserver. Diese Beziehungen, die durch das Navigieren definiert werden, können als

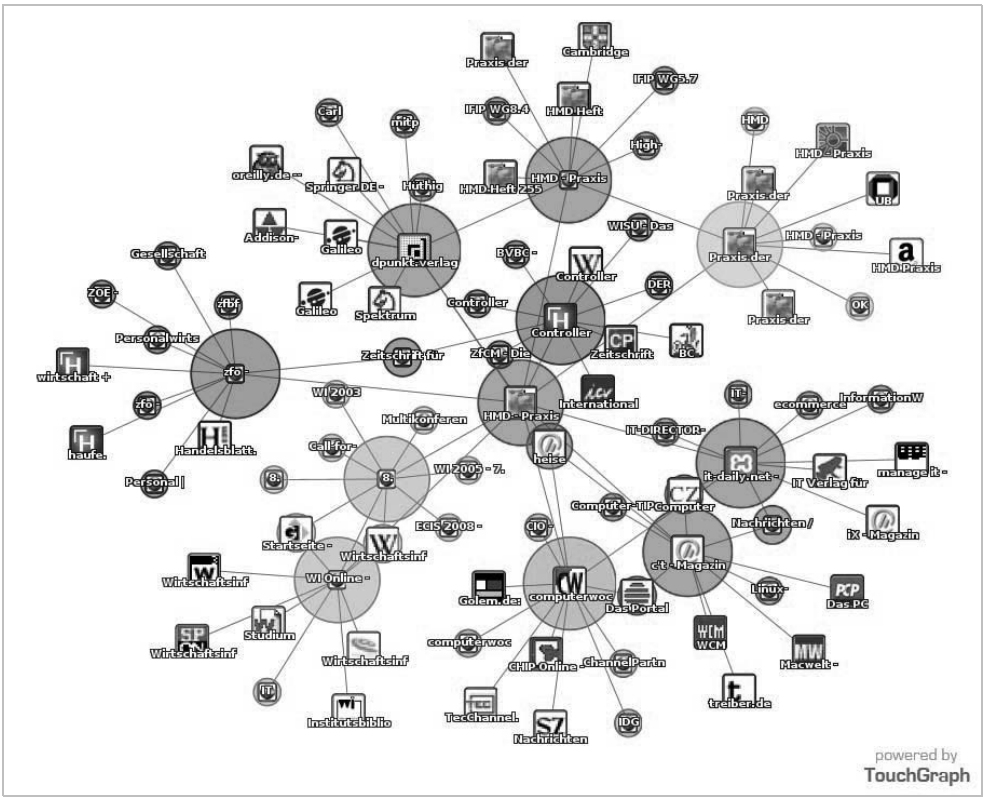

Abb. 1: Resultat der Analyse der Struktur von URL *http://hmd.dpunkt.de* mit Googles Browsertool TouchGraph

Graphen modelliert werden. Um die Usage-Mining-Algorithmen jedoch effizient zu verwenden, sind zusätzliche Informationen notwendig. Diese Zusatzinformationen umfassen Navigationstemplates, Inhaltsinformationen, Topologien der Webseiten, Konzepthierarchien usw.

Die Anwendungen des Web Usage Mining können in zwei große Kategorien eingeteilt werden. Die erste Gruppe besteht aus Anwendungen, bei denen es darum geht, das Profil von Benutzern zu lernen, meist mit dem Ziel, personalisierte Interfaces zu definieren, z.B. für eCommerce-Websites. In der zweiten Gruppe geht es um Navigationspatterns, die dazu verwendet werden können, die Websites effizienter zu gestalten.

Benutzer sind unter anderem daran interessiert, dass diese Techniken imstande sind, ihre Informationsbedürfnisse und -wünsche zu lernen, um die Suche nach Informationen für sie bestmöglich zu optimieren. Auf der anderen Seite haben die Designer von Websites Interesse daran, die gelernten Patterns dazu zu verwenden, ihre Websites effizienter zu gestalten.

Bespiel: Usage Mining

Usage-Mining-Systeme können viele verschiedene Strukturen annehmen. Eine gewisse Basisstruktur ist allen gemein. Sie beginnen mit dem Aufzeichnen der Navigationsvorgänge der Benutzer. Jeder Klick, der neue Information vom Webserver anfordert, wird in einer Log-Daten-

bank registriert. Diese Daten werden bereinigt, d.h., Einträge, die nicht direkt mit Benutzeroperationen zu tun haben, werden gelöscht, und die übrig bleibenden Daten werden in einem einheitlichen Format abgelegt. Die so erzeugte bereinigte Datenbank zeichnet alle Benutzertransaktionen auf. Diese ersten Schritte liegen jedem Structure-Mining-System zugrunde. Nun können die gewünschten Data-Mining-Algorithmen angewandt werden.

Welche Mining-Algorithmen zu verwenden sind, hängt ab von der Art der Benutzercharakteristiken, die als Resultat gefunden werden sollen. In Abbildung 2 werden drei Typen von Informationen gesucht:

- Mengen von Seiten, die die Benutzer häufig in einer Session anschauen,
- Sequenzen, die häufig in der gleichen Reihenfolge durchlaufen werden,
- Abhängigkeiten, die sich in Form eines Baums darstellen lassen.

3 Einige Techniken

Nachdem die wichtigsten Anforderungen im Bereich des Web Mining definiert wurden, soll nun anhand einiger Beispiele aufgezeigt werden, wie Data-Mining-Techniken verwendet werden können, um diese Probleme zu lösen.

3.1 Websuche

Als Illustration für die Websuche wollen wir hier zwei Techniken etwas näher betrachten, nämlich das Information Retrieval und das Webcrawling.

Information Retrieval

Das Durchsuchen des Web ist eine Tätigkeit, die nicht mehr näher beschrieben werden muss. Es ist wohl zur wichtigsten Informationsbeschaffungstechnik avanciert.

Die Websuche findet ihren Ursprung im »Information Retrieval«, kurz IR. Es ist dies eine Forschungsrichtung, die sich mit dem Auffinden relevanter Informationen in großen Sammlungen von Textdokumenten beschäftigt. Das Basiselement des IR ist ein Textdokument, d.h., alle Dokumente zusammen bilden die Textdatenbank. Die Analogie zum Web ist evident. Die Webseiten bilden die Basisdokumente, und das Web selber ist die Datenbank. Es ist sicher nicht übertrieben zu sagen, dass die Websuche heute die wichtigste Anwendung des IR ist.

Das Auffinden von Informationen im IR-Kontext bedeutet das Zusammenstellen einer Menge von Dokumenten, die die gesuchte Information enthalten. Die meistverwendete Anfrageform ist eine Liste von Schlüsselwörtern, die dann mit der indexierten Dokumentdaten-

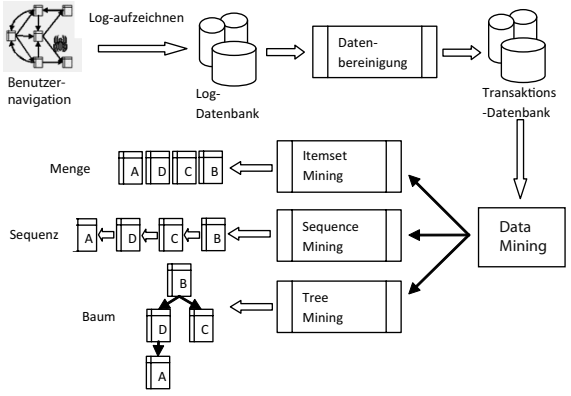

Abb. 2: Beispiel der Architektur eines Structure-Mining-Systems

bank abgeglichen wird, um die relevanten Dokumente zu finden. Webdokumente enthalten Informationen, die in einfachen Textdokumenten nicht vorkommen, wie z.B. Links und Ankertexte. Diese Zusatzinformationen werden von den angepassten IR-Algorithmen ebenfalls verwendet.

Ein großes Problem in der Websuche ist das sogenannte Spamming. Beim Spamming werden Wörter in die Seiten eingebaut, damit sie von den IR-Algorithmen gefunden werden. Beim Suchen mithilfe von IR-Techniken wird normalerweise eine große Anzahl von Dokumenten gefunden. Aus diesem Grund ist es wichtig, die Dokumente zu sortieren und ihnen einen Rang zuzuordnen, sodass die relevantesten Dokumente den höchsten Rang bekommen. Dokumente mit einem niederen Rang werden vom Benutzer meist ignoriert. Um diesen Rang einer Seite oder eines Dokuments zu verbessern, werden von den Autoren »unkorrekte« Schlüsselwörter in die Seite eingebaut, die nichts mit dem wirklichen Inhalt der Seite zu tun haben (spamming). Es ist somit eine der großen Herausforderungen für IR-Techniken, »echte« Schlüsselwörter von »unechten« zu unterscheiden.

Abbildung 3 zeigt die Basisarchitektur eines IR-Systems. Startend von einer Benutzerabfrage verwendet das IR-System den Dokumentenindex, um herauszufinden, welche Dokumente relevant sind, um diese anschließend in der Dokumentdatenbank zu finden und sie dann an den Benutzer zurückzuliefern.

Es gibt eine Vielzahl von Möglichkeiten, wie Anfragen formuliert werden können: Es sind dies unter anderem Schlüsselwörter, Sätze, Dokumente, natürliche Sprache oder boolesche Anfragen. Gemäß den Anfragemethoden müssen die entsprechenden IR-Modelle entwickelt werden. Es gibt vier Modelle, die hauptsächlich zum Einsatz kommen: boolesche Modelle, Vectorspace-Modelle, Sprachmodelle und Wahrscheinlichkeitsmodelle. Eine detaillierte Beschreibung dieser Modelle kann in [Manning et al. 2008] und [Chang et al. 2001] gefunden werden. All diese Modelle interpretieren jedoch Dokumente als eine Menge von Worten, und Anfragen werden über eine Wortliste definiert.

Schließlich muss mittels dieser Modelle den gefundenen Dokumenten ein Rang zugeordnet werden. Das Erstellen einer Ordnung dieser Art ist eines der Schlüsselelemente des gerichteten Data Mining. IR verwendet ebenfalls dieselben Techniken (siehe [Dunham 2003]).

Was in der bisherigen Beschreibung zu einem vollständigen System fehlt, ist ein Mechanismus, um die Textdatenbank zu kreieren. Hierzu werden Webcrawler verwendet. Diese unabhängige Technik wird kurz in den nächsten Abschnitten beschrieben.

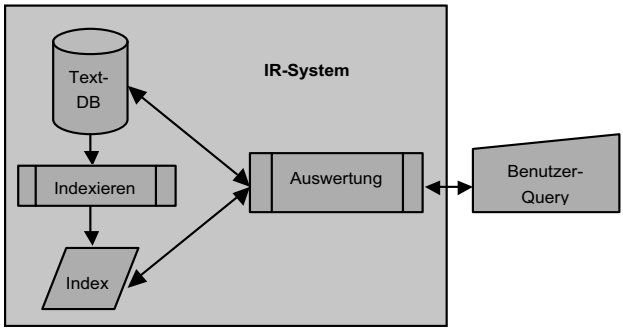

Abb. 3: Architektur eines IR-Systems

Webcrawler

Webcrawler, die oft auch Spiders oder Softbots genannt werden, sind Programme, die Webseiten oder zumindest den Inhalt von Webseiten automatisch in eine lokale Datenbank laden. Nachdem die Seiten heruntergeladen sind, untersuchen die Crawler die geladenen Webseiten nach neuen Links, die den Crawler zur nächsten Seite führen. Diese Seite wird wiederum heruntergeladen und nach weiteren Links untersucht. Wäre das Web statisch, könnte mithilfe von Crawlern ein Abbild des Web generiert werden. Dies ist jedoch bei Weitem nicht möglich, da das Web sich wesentlich schneller verändert, als es von Crawlern durchsucht werden kann.

Für Crawler präsentiert sich das Web als Netzwerk oder als Graph, der zu durchsuchen ist. Mit jeder heruntergeladenen Seite werden neue Links gefunden. Die Liste dieser Links wird »frontier« genannt und definiert die nächsten vom Crawler zu durchsuchenden Seiten.

Je nachdem, in welcher Reihenfolge die Liste abgearbeitet wird, kann man zwei Crawler-Arten unterscheiden: solche, die zuerst in die Breite suchen, und andere, die zuerst in die Tiefe suchen.

Wenn die Links in der Reihenfolge abgearbeitet werden, in der sie gefunden werden, so ergibt sich daraus eine Suche in die Breite. Durch das Abarbeiten der Seiten mittels dieser Strategie werden gewisse Seiten bevorzugt. Generell kann gesagt werden, dass diese Suchstrategie gut vernetzte Seiten bevorzugt und damit ein zufälliges Durchsuchen des Web ausschließt. Darüber hinaus spielt die Startseite eine sehr wichtige Rolle, da alle Seiten, die topologisch nahe bei dieser Seite liegen, bevorzugt werden. Diese Eigenschaften können positiv oder negativ sein, je nach der Crawling-Strategie, die verfolgt wird. Ist diese Strategie des Durchsuchens unerwünscht, kann zuerst in die Tiefe gesucht werden, indem die Präferenzen in der »frontier«-Liste dementsprechend umdefiniert werden.

Die Präferenzstrategie legt ein Kriterium fest, anhand dessen das Web durchsucht wird. Anstatt aus der »frontier«-Liste, d.h. aus der Liste der zu durchsuchenden Seiten, jene auszuwählen, die als Erste gefunden wurden, werden nun die Links ausgewählt, die dem neuen Präferenzkriterium genügen. Diese Kriterien können topologischer Natur sein, können aber auch inhaltliche Eigenschaften berücksichtigen. Am häufigsten werden inhaltsorientierte Topic-Crawler verwendet. Der Basisalgorithmus eines Crawlers ist in Abbildung 4 dargestellt.

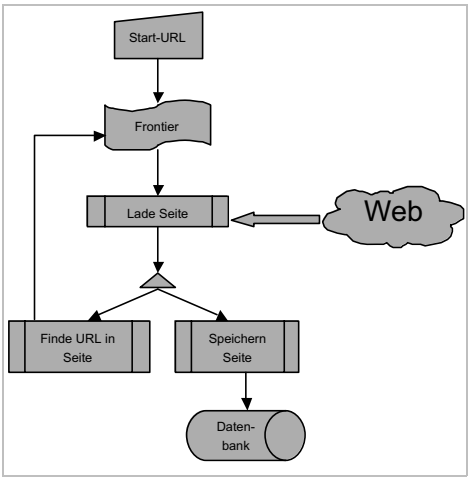

Abb. 4: Basisalgorithmus eines Crawlers

Da Webcrawler Informationen verwenden, die aus Links herausgelesen werden, sind sie auch ein ideales Hilfsmittel, um Datenbanken zu generieren, die von Link-Analyse-Algorithmen verwendet werden. Einen Überblick über diese Techniken geben wir im nächsten Abschnitt.

3.2 Link-Analyse

Links definieren die Topologie des Web und ermöglichen das Navigieren zwischen den Seiten. Aber mit dem enormen Anstieg der Anzahl der Webseiten in der Mitte der Neunzigerjahre wurde die Idee geboren, die Link-Topologie auch zum Suchen von relevanten Informationen zu verwenden, da die einfachen IR-Algorithmen an ihre Grenzen stießen. Vor allem das Problem, Seiten nach ihrer Rele-

vanz zu ordnen, wurde immer schwieriger, und auch das Problem des »Spammings«, d.h. das Vortäuschen des Vorhandenseins eines Inhalts, hatte zur Folge, dass das Suchen mittels IR-Techniken immer unbefriedigender wurde.

Ein Lösungsansatz, der gefunden wurde, war das Miteinbeziehen der Links in die Suchalgorithmen. Im Gegensatz zu reinen Textdokumenten, die keine wirklichen Beziehungen untereinander aufweisen, bieten die Links die Möglichkeit, Beziehungen zwischen Webdokumenten herzustellen. Dabei sind vor allem zwei Arten von Hyperlinks zu unterscheiden: auf der einen Seite jene Links, die dazu verwendet werden, den Inhalt innerhalb einer Website zu organisieren, und auf der anderen Seite die Links, die auf andere Websites verweisen. Es hat sich gezeigt, dass vor allem die zweite Art von Links für das Suchen nützlich ist. Die ausgehenden Links werden als Verweise auf Seiten interpretiert, deren Relevanz bzw. Qualität anerkannt ist.

Das Resultat dieser Forschung bestand hauptsächlich aus zwei Algorithmen, nämlich PageRank und Hits. Der PageRank-Algorithmus wird erfolgreich im Google-Suchmotor verwendet. Beide Algorithmen wurden aus Tools zur Analyse von sozialen Netzwerken abgeleitet. Sie gewichten Webseiten unter Verwendung von Hyperlinks, die als Indikatoren für das Prestige oder die Autorität einer Seite interpretiert werden.

Natürlich kann die Link-Analyse auch dazu verwendet werden, Web-Communities zu untersuchen. Hier werden für die soziale Netzwerkanalyse Algorithmen in ihrer ursprünglichen Form verwendet. Auf diese Verwendung wollen wir hier nicht weiter eingehen.

3.3 Usage Mining

Im Gegensatz zu den zwei zuvor beschriebenen Techniken geht es beim Usage Mining nicht primär um den Inhalt des Web (d.h. um den Inhalt der Webseiten bzw. die Links), sondern um die Art und Weise, wie die Benutzer durch die Seiten navigieren [Srivastava et al. 2000]. Diese Art von Untersuchungen haben vor allem mit dem Aufkommen von webbasierten Informationssystemen, eCommerce-Anwendungen und Webservices immer mehr an Bedeutung gewonnen.

Das Volumen der Daten, die sogenannten Clickstreams, die durch das Navigieren in Websites generiert werden, ist enorm. Die Analyse dieser Daten ist jedoch von großer Bedeutung, da sie den verschiedenen Institutionen die Möglichkeit geben, die Verhaltensmuster ihrer Kunden zu untersuchen. Einerseits können die Websites optimiert werden, und andererseits können sie an die individuellen Bedürfnisse der Benutzer angepasst werden (personalisieren).

Usage Mining bezieht sich also auf die Untersuchung von Clickstreams und der dazugehörigen Daten, die durch die Interaktion von Benutzern mit Webressourcen generiert werden. Das Ziel ist es, Modelle zu finden, die Verhaltensmuster bzw. Benutzerprofile beschreiben.

Der Usage-Mining-Prozess entspricht in großen Zügen genau dem Data-Mining-Prozess, wie er in Abschnitt 1.2 beschrieben wurde, d.h. im Bereitstellen der Daten, im Generieren von Modellen und schließlich in deren Analyse.

In der Vorbereitungsphase werden die Clickstream-Daten gereinigt und umorganisiert. Diese Daten stammen in den meisten Fällen aus den Log-Dateien der Webserver. Es handelt sich dabei um Textdaten, die sich in dieser Form nur schlecht für eine systematische Analyse eignen. Die Daten werden größtenteils in ein relationales Datenformat umgewandelt und nach Benutzern segmentiert. Diese Grunddaten werden in vielen Anwendungen mit zusätzlichen Informationen erweitert, wie z.B. Produktkataloge und Metadaten über Kunden. Diese Zusatzinformationen können in Form von Ontologien in den Prozess eingebaut werden.

Zum Generieren der Modelle können Data-Mining-Techniken, statistische Hilfsmittel und

auch Datenbank- und Machine-Learning-Operationen verwendet werden. Diese Modelle haben zum Ziel, sowohl typische Verhaltensmuster von Benutzern aufzuzeigen – eine der Hauptzielsetzungen von Data Mining schlechthin – wie auch Informationen über die Benutzung von Webressourcen und -sessions zu liefern.

Nachdem die Modelle generiert wurden, müssen sie in den meisten Fällen noch zusammengefasst werden. Dies hat damit zu tun, dass die Anzahl der gefundenen Modelle so groß ist, dass die Verwendung der individuellen Modelle keinen Sinn macht. Es können Aggregationen berechnet werden, die ähnliche Benutzer- bzw. ähnliche Verhaltensmuster zusammenfassen. Meist wird dies über Filter erreicht, die die irrelevanten Modelle entfernen.

Auch wenn der globale Prozess relativ einfach zu beschreiben ist, ist das Usage Mining sehr aufwendig, da, wie zuvor angedeutet, die Daten nicht in ihrer ursprünglichen Form verwendet werden können. Die Schlüsselelemente, die nach dem Kombinieren und Reinigen aller verwendeten Daten identifiziert werden können, sind die folgenden: die gelesenen Seiten, die Benutzeridentifikation und die Sessions. Meist werden diese Daten mittels statistischer Tools oder auch OLAP-Tools untersucht. Clustering und Segmentation von Benutzerdaten, Assoziations- und Korrelationsanalysen sowie Sequenzanalysen sind weitere wichtige Hilfsmittel.

4 Anwendungen

Die Anwendungen, die auf den oben beschriebenen Methoden und Techniken aufbauen, können in drei große Gruppen zusammengefasst werden.

Suchen

Wegen der Größe des Web gehören heute Suchmotoren zu den essenziellen Bestandteilen der Webnavigation. Alle Suchmotoren sind auf Information-Retrieval-Techniken angewiesen. Aber auch viele der übrigen zuvor angeführten Techniken werden verwendet. So benutzt der Marktführer Google unter anderen z.B. den PageRank-Algorithmus und auch Topic-Search-Crawler.

Personalisieren

Personalisierte Webseiten werden immer wichtiger, vor allem im Bereich des eCommerce. Die Seiten werden den spezifischen Bedürfnissen des Benutzers gemäß dynamisch generiert. Dazu werden zuerst Modelle für die spezifischen Verhaltensweisen mithilfe von Usage-Mining-Tools hergeleitet, die dann die Erstellung von Webseiten steuern. Diese Webseiten können den Gewohnheiten des Benutzers angepasst oder dazu verwendet werden, dem Benutzer Empfehlungen zu präsentieren, die wahrscheinlich für ihn interessant sind.

Gruppenidentifikation

Mit dem Aufkommen von Online-Communities (siehe Abschnitt 5) gewinnt die Analyse von sozialen Netzwerken immer mehr an Bedeutung. Das Hauptinteresse ist im Moment eher wissenschaftlicher Natur. Data Mining und soziale Netzwerkanalyse sind die bevorzugten Hilfsmittel zum Auffinden und zum Untersuchen der Entwicklung dieser Gruppen.

5 Zukünftige Entwicklungen

Die zukünftigen Entwicklungen des Web Mining gehen natürlich Hand in Hand mit den Weiterentwicklungen des Web selber. Vor allem die Entwicklung von Technologien und die veränderte Nutzung des Web, die unter dem Begriff »Web 2.0« zusammengefasst wird, haben einen ganz entscheidenden Einfluss auf die Art, wie sich das Web Mining weiterentwickelt. Außerdem hat auch die Entwicklung des semantischen Web einen maßgeblichen Einfluss auf die zukünftige Webnutzung.

5.1 Web 2.0 Mining

Durch die Liberalisierung der Erstellung von Webinhalten, die durch das Web 2.0 stark vorangetrieben wurde, wird die Qualität der Dokumente in die Hand des Benutzers gelegt. Darüber hinaus ist die Struktur der so generierten Dokumente meist sehr einfach, da deren Erstellung auch durch einen Nichtspezialisten möglich sein soll. Zudem haben Foto- und Videoportale stark an Bedeutung gewonnen. Die gemeinschaftlich generierten Daten werden nicht mehr zentral erstellt und verwaltet, sondern mithilfe von sozialer Software verteilt und von untereinander vernetzten Benutzern bearbeitet. Dies steht in einem krassen Gegensatz zu den Inhalten, die zentral von großen Medienunternehmen generiert und verwaltet werden. Des Weiteren gewinnen auch die Social-Bookmarking-Portale, virtuelle Welten wie auch die schon länger bekannten Tauschbörsen immer mehr an Bedeutung.

Diese Entwicklungen werden das Web Mining nicht unbeeinflusst lassen. Da die Struktur der Webinhalte meist relativ einfach ist, muss der Fokus bei Web 2.0 Mining auf das Netzwerk gelegt werden.

Als erste entscheidende Technologie sollen hier die sozialen Netzwerke genannt werden, die in den verschiedensten Formen existieren. Manche dieser Netzwerke haben ein vorgegebenes Ziel, wie z.B. XING/openBC, das zum Ziel hat, Professionals zu vernetzen. Andere sind weiter gefächert, wie z.B. Facebook und MySpace. Sie haben aber alle einen wichtigen Service gemein, nämlich die entstehenden Netzwerke zu analysieren und zu jedem Benutzer die »nächsten« Nachbarn zu finden und damit die Vernetzung zu verbessern. Diese Probleme bieten ein weites Anwendungsfeld für Link-Analyse-Tools.

Eine andere Technologie, die eine sehr wichtige Rolle spielt, sind die Wikis (z.B. Wikipedia), die als unstrukturierte Informationscontainer angeschaut werden. Um die relevante Information zu finden, sind Wikis, sobald sie eine gewisse Größe erreicht haben, sehr stark von der Suchmaschine abhängig. Neben dem Auffinden relevanter Textstellen können zukünftige Web-Mining-Anwendungen dazu dienen, Struktur in die unstrukturierten Wikis zu bringen.

Als letzte Technologie sollen hier noch die Bookmarking-Tools angeführt werden. Viele dieser Systeme basieren auf sogenannten Folksonomies, d.h. auf gemeinschaftlich erarbeiteten einfachen Ontologien (siehe Abschnitt 5.2). Im Gegensatz zu den zuvor betrachteten Systemen, bei denen Web Mining dazu verwendet wurde, zusätzlich Semantik für die existierenden Daten zu liefern, könnten qualitativ hochstehende Folksonomies dazu verwendet werden, semantische Informationen zu liefern, die in die Web-Mining-Anwendungen integriert werden könnten.

5.2 Semantic Web Mining

Das semantische Web ist eine Weiterentwicklung des Web mit der Zielsetzung, die Bedeutung der Informationen, die in Webseiten gefunden werden können, für Computer verwendbar zu machen. Die Informationen sollen direkt von Maschinen interpretiert und weiter verarbeitet werden können [Han & Chang 2002]. Diese zusätzlichen Informationen stellen interessante Ressourcen dar, die von Web-Mining-Tools untersucht werden können. Zusätzlich können sie aber auch als Metadaten für die Interpretation der von Web-Mining-Anwendungen gefundenen Modelle verwendet werden.

Die Hauptkomponente des semantischen Web sind Ontologien. Diese sind zwar vergleichbar mit den Folksonomies, die weiter oben beschrieben wurden. Sie sind aber meist wesentlich komplizierter und von Spezialisten in langwieriger Handarbeit zusammengestellt worden. Eine der großen neuen Herausforderungen für Web-Mining-Tools besteht darin, diese Ontologien aus dem Web zu extrahieren.

Andererseits kann natürlich auch die vom semantischen Web gelieferte Semantik von Web-Mining-Algorithmen verwendet werden. Die Ontologien können Hintergrundwissen liefern, das den Data-Mining-Prozess steuert und ihn damit effizienter gestalten kann. Zum Schluss können diese Zusatzinformationen während der Nachbereitung, d.h. während der dritten Phase des KDD-Prozesses, gebraucht werden, um aus der Vielzahl von Modellen die nützlichsten herauszufiltern.

6 Kombination von Bottom-up- mit Top-down-Zugängen

Dieser Beitrag beschreibt einige der Schlüsselbegriffe rund um das Thema Web Mining etwas genauer und zeigt auf, wie Web Mining als eine Kombination von Webdaten und Data-Mining-Techniken aufgefasst werden kann.

Wir haben in einem zweiten Teil auch aufgezeigt, wie die neuesten Entwicklungen im Bereich der Webtechnologien wie das Web 2.0 und das semantische Web neue Herausforderungen für Web-Mining-Tools darstellen. Diese zwei Technologien repräsentieren zwei gegensätzliche Tendenzen. Auf der einen Seite das Web 2.0, das eine Art Bottom-up-Zugang definiert, bei dem die Webbenutzer individuell oder in Zusammenarbeit mit anderen Benutzern frei neue Webinhalte generieren. Das semantische Web stellt eher einen Top-down-Zugang dar, bei dem normative Ontologien festgelegt werden, die Webinhalte effizienter nutzbar machen sollen. Die nächsten Entwicklungen zeichnen sich bereits ab, der Term »Web 3.0« wurde schon geprägt und steht für das Kombinieren vom Web 2.0 mit dem semantischen Web.

Die Herausforderung für das Web Mining besteht darin, klassische Data-Mining-Techniken an diese neuen Gegebenheiten anzupassen, sie zu erweitern oder eventuell auch neue Methoden zu entwickeln. Im Speziellen geht es darum, Top-down- und Bottom-up-Zugänge, die zum Generieren von Webinhalten verwendet werden, so ins Web Mining zu integrieren, dass die gesuchten Informationen effizienter und mit höherer Präzision gefunden werden können.

7 Literatur

[Bing 2007] *Bing, L.:* Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer-Verlag, Berlin, 2007.

[Chakrabarti 2000] *Chakrabarti, S.:* Data mining for hypertext: A tutorial survey. In: SIGKDD Explorations 1(1), 2000, S. 1-11.

[Chakrabarti 2003] *Chakrabarti, S.:* Mining the Web. Morgan Kaufmann, 2003.

[Chang et al. 2001] *Chang, G.; Healey, M. J.; McHugh, J.; Wang, J.:* Mining the World Wide Web. Kluwer Academic Publisher, 2001.

[Cooly 2000] *Cooley, R. W.:* Web Usage Mining: Discovery and Applications of Interesting Patterns from Web Data. 2000.

[Dunham 2003] *Dunham, M. H.:* Data Mining: Introductory and Advanced Topics. Prentice Hall, 2003.

[Han & Chang 2002] *Han, J.; Chang, K.:* Data Mining for Web Intelligence. In: Computer 35(11), 2002, S. 64-70.

[Hotho & Stumme 2007] *Hotho, A.; Stumme, G.:* Mining the World Wide Web – Methods, Applications, and Perspectives. In: Künstliche Intelligenz (3), 2007, S. 5-8.

[Kosala & Blockeel 2000] *Kosala, R.; Blockeel, H.:* Web Mining Research: A Survey. In: SIGKDD Explorations 2(1), 2000, S. 1-15.

[Manning et al. 2008] *Manning, C. D.; Raghavan, P.; Schütze, H.:* Introduction to Information Retrieval. Cambridge University Press, 2008.

[Scime 2004] *Scime, A.:* Web Mining: Applications and Techniques. IDEA Group Publishing, 2004.

[Srivastava et al. 2000] *Srivastava, J.; Cooley, R.; Deshpande, M.; Tan, P.:* Web usage mining: discovery and applications of usage patterns from Web data. In: SIGKDD Explorations 1(2), ACM, New York, NY, USA, 2000, S. 12-23.

[Srivastava et al. 2004] *Srivastava, J.; Desikan, P.; Kumar, V.:* Data Mining: Next Generation Challenges and Future Directions. MIT/AAAI Press, 2004, S. 399-417.

Prof. Dr. Kilian Stoffel
Université de Neuchâtel
Institut du management
de l`information (IMI)
Pierre-à-Mazel 7
CH-2000 Neuchâtel
kilian.stoffel@unine.ch
www.unine.ch/imi