Informationsflüsse im Internet

Dimitri Lande, Sergei Braichevski, Kiew (Ukraine) und Dimitri Busch, Stuttgart

Im Artikel geht es um Entwicklungstrends und Probleme, die mit dem schnellen Informationszuwachs im Internet verbunden sind. Es wird gezeigt, dass die Hauptschwierigkeiten nicht mit der Leistung von Software und der Hardware, sondern mit Besonderheiten des Sachgebietes zusammen hängen. Im Rahmen des Konzeptes der Informationsflüsse werden aktuelle Probleme des Information Retrieval, der Strukturierung des Informationsraums und dessen Verhältnis zum semantischen Raum behandelt. Der Artikel zeigt einige Lösungen dieser Probleme, die auf dem Textmining und fraktalen Modellen basieren.

Information flows in the internet

The article discusses trends and problems arising from the rapid increase of information volumes. The main difficulties are caused not by the level of software or hardware but instead by specific features of the subject area. Current problems in information retrieval, information space structuring, and the connection of this with semantic space are analysed within the framework of information flows. The article describes some solutions of these problems that are based on the text mining and fractal models.

1 Einführung

Die Entwicklung der Informationstechnologien, insbesondere des Internet, verursacht in der letzten Zeit viele Probleme. Diese Probleme sind mit dem schnellen Anwachsen von Datenmengen verbunden, die zu speichern und zu verarbeiten sind.

In der Anfangsphase der Entwicklung des WWW publizierten wenige Websites Informationen von einzelnen Autoren für viele Besucher der Sites. Inzwischen änderte sich die Situation drastisch. Besucher der Websites nehmen selbst am Aufbau der Inhalte teil. Dies führt zu einem drastischen Zuwachs des Umfangs und der Dynamik des Informationsraums (Braichevski/Lande, 2005).

Andererseits veränderte sich auch das Verhältnis von Benutzern des Internet zur Arbeit mit dessen Ressourcen. Allmählich setzt sich die Auffassung durch, dass eine Vollständigkeit der Daten in jedem Falle unzugänglich bleibt, und die zugänglichen Datenauszüge viel Informationsrauschen enthalten. Statt nach „allem Nötigen" zu suchen, hofft man, etwas Nützliches zu finden.

Außer dem Umfang der Informationen spielt das Erneuerungstempo der Informationen im Netz eine bedeutende Rolle. Der Begriff eines *Informationsflusses* ist deswegen von besonderer Bedeutung (vgl. Del Corso et al., 2005). Obwohl dieser Begriff noch einer strikten Definition bedarf, wird er schon heute für ein breites Aufgabenspektrum verwendet, das mit der Dynamik von Informationen im Netz verbunden ist.

Heute verfügen wir über eine Informationsbasis, die für Experimente zugänglich ist und einen Umfang hat, der früher unvorstellbar war. Darüber hinaus übertrifft der Umfang dieser Basis alles, was vor zehn Jahren zugänglich war beträchtlich. Im August 2005 kündigte die Firma Yahoo an, dass sie ca. 20 Milliarden Dokumente indexierte. Im Jahre 2004 hatte Google weniger als zehn Milliarden Dokumente indexiert. Dies bedeutet, dass die Menge öffentlich zugänglicher Informationen sich im Laufe eines Jahres verdoppelte. Laut Web Server Survey[1], das vom Netcraft-Dienst erstellt wird, übertraf im Juni 2007 die Anzahl der Websites 120 Milliarden. Diese Angaben bestätigen das exponentielle Wachstum der Informationen im Web. Das Wachstum wird von einer Reihe von Problemen begleitet. Zu diesen Problemen gehören:

- unproportionales Wachstum von Informationsrauschen
- eine Fülle nicht abgerufener Informationen und von Spam;
- mehrfache Duplizität der Informationen;
- schwache Strukturierung.

Das konventionelle Web hat auch andere Nachteile, wie z.B. viel Informationsmüll, fehlende Unterstützung der semantischen Suche, eingeschränkter Zugang zum Invisible(versteckten) Web. Darüber hinaus ist es unmöglich, Dokumentenintegrität zu gewähren.

Zahlreiche Gruppen von Forschern und Spezialisten suchen nach Lösungen für die o.g. Probleme. Zu diesen Gruppen gehört u.a. W3C-Konsortium, welches ein Konzept des semantischen Web entwickelt (Berners-Lee et al., 2005). Zugleich wird ein allgemeinerer Einsatz, Web-2[2], entwickelt, welcher einen revolutionären Durchbruch verspricht. Im Rahmen des Web-2 wird eine Implementierung des semantischen Web beabsichtigt. Dazu gehören eine mehrschichtige Unterstützung von Metadaten, neue Ansätze zum Design und zu entsprechenden Werkzeugen, Textmining-Technologien und ein Konzept von Webdiensten.

Der mathematische Apparat und entsprechende Werkzeuge sind jedoch nicht immer in der Lage, die aktuelle Situation adäquat abzubilden. Es geht dabei weniger um die Analyse von endlichen Datensammlungen, sondern um die Navigation in dynamischen dokumentarischen Informationsflüssen. Im Folgenden werden die Probleme genauer betrachtet und einige Lösungen der Probleme vorgeschlagen.

2 Informationeller und semantischer Raum

Derzeit gibt es einen Grund für die Annahme, dass der Informationsbegriff, insbesondere dessen Beziehung zum Wissensbegriff, ein gewisses Überdenken erfordert. Der Begriff „Umwandlung von Informationen in Wissen", der früher häufig im Bereich künstlicher Intelligenz verwendet und später gründlich vergessen wurde, ist derzeit wieder von Interesse. Dieses Interesse ist durch Erfolge in der maschinellen Verarbeitung von Datenflüssen begründet, die nicht nur mehrsprachig sind, sondern auch zu verschiedenen soziokulturellen Kontexten gehören. Es ist klar, dass die Verarbeitung eines solchen Datenflusses, d.h. reiner Informationen, keine aktive Verwendung des Inhalts der Dokumente voraussetzt. Theoretische Überlegungen hierzu gehen davon aus, dass das eigentliche „Wissen" eine Schicht über den Informationsflüssen repräsentiert, die durch Beziehungen zwischen den Informationselementen bestimmt wird. Diese Beziehungen sind also in den Informationsflüssen selbst nicht vorhanden, sondern stellen einen bezüglich der Informationen externen Faktor dar.

Die Praxis zeigt, dass Informationen erfolgreich verarbeitet werden können, ohne dass dabei die Semantik der Infor-

1 http://news.netcraft.com/archives/2007/06/index.html [12.07.2007]
2 www.web2con.com [12.07.2007]

mation berücksichtigt wird. In diesem Zusammenhang entstand ein Interesse an Ansätzen, bei welchen die Information als ein Maß für die Ordnung innerhalb eines Systems verstanden wird. Einige Wissenschaftler und führende Teilnehmer des Informationsmarks, z.B. die Firma Autonomy, kehren zu Ursprüngen der Informationstheorie, zum Begriff der Entropie, Shannon-Theorie, Bolzmann-Gleichungen usw. zurück. In der Tat ähneln die Probleme der Bewegung von inhaltsreichen Daten über Netzwerkkanäle den Problemen der Signalübertragung über Kommunikationskanäle. Deswegen kann die Informationstheorie, die früher hauptsächlich im Bereich Informationsübertragungstechnik verwendet wurde, auch für die Analyse von inhaltsreichen Textflüssen nützlich sein.

Die moderne Informationstheorie geht wahrscheinlich zum ursprünglichen Ansatz zurück, allgemeine Merkmale aus Meldungen zu extrahieren, ohne die Semantik der Meldungen zu berücksichtigen. Diese Merkmalsextraktion ist auch unabhängig von unserer Fähigkeit, diesen Inhalt wahrzunehmen. Die Wissensextraktion aus Informationsflüssen im gewöhnlichen Sinne bildet ein eigenständiges Problem, das nach Methoden zu lösen ist, die eine separate Entwicklung erfordern. Die Erkennung dieser Tatsache wird zweifellos zur Weiterentwicklung solcher Methoden und entsprechender Werkzeuge beitragen.

Obwohl Informationen unabhängig von ihrem inhaltlichen Aspekt verarbeitet werden können, ist eine gegensätzliche Behauptung nicht wahr. Die Informationen können in jedem Falle als ein „Wissenssubstrat“ betrachtet werden. Es besteht höchstwahrscheinlich keine Möglichkeit, das „Wissensproblem“ nur mit technischen Mitteln zu lösen. Die Lösung des Problems erfordert viel Forschung inklusive theoretischer Arbeit auf einem hohen Niveau.

Eine der wichtigsten Fragen, die bis heute selten beachtet wurde, entsteht unserer Meinung nach für die Beziehung zwischen informationellen und semantischen Räumen. In der Literatur gelten diese Begriffe als identisch, ohne einen Grund für eine solche Annahme zu nennen. Die Tatsache, dass diese zwei Begriffe nicht identisch sind, folgt aus ihrer Natur: Während der Informationsraum aus Daten gebildet wird, die auf verschiedenen Datenträgern aufgezeichnet sind, wird der semantische Raum aus Konzepten erzeugt, die mit subjektiven menschlichen Einschätzungen verbunden sind. Der semantische Raum im Netz kann deswegen als eine Menge von semantischen Einheiten definiert werden, die in einem soziokulturellen Kontext aktuell sind und in einem Netzwerk dargestellt sind. Unter einer semantischen Einheit verstehen wir hier eine elementare Kategorie, die uns erlaubt, subjektiv bewertende Urteile über Dinge und Prozesse zu bilden, die zu unserer Welt gehören. In der Realität gibt es zwischen ihnen eine ganz bestimmte Beziehung, aber das Finden dieser Beziehung ist eine nicht triviale Aufgabe.

Das Verhältnis zwischen dem Informationsraum und dem semantischen Raum kann anhand des Referierens einer Menge von Textdokumenten verdeutlicht werden, die in verschiedenen Sprachen erstellt sind. Dabei entsteht gleich die Frage, ob es einen Algorithmus gibt, der es erlaubt, bedeutungsvolle Informationsfragmente aus einem beliebigen Dokument zu extrahieren, ohne dabei die Sprache des Dokuments zu „verstehen“ und sogar identifizieren zu können.

Es stellt sich heraus, dass ein solcher Algorithmus möglich ist, wenn die Eingangsdaten Zipf-Gesetzen entsprechen, d.h. von Menschen erstellt werden. Daraus ergeben sich andere „ketzerische“ Fragen: In welchem Maß ist der Begriff „Information“ mit dem Begriff „Semantik“ verbunden? Gibt es überhaupt eine Beziehung zwischen diesen Begriffen, wenigstens im allgemeingültigen Sinne? Inhaltsreiche und bedeutungsvolle Ergebnisse können beispielsweise unter der Verwendung ausschließlich statistischer Methoden erhalten werden, ohne dabei Methoden der künstlichen Intelligenz, umfangreiche semantische Formalisierungsmittel und die Arbeit von menschlichen Experten anzuwenden. Dies kann den Eindruck erwecken, dass die strukturlinguistische Ebene völlig ausreicht, um eine vollwertige Informationsarbeit durchzuführen.

Es ist ohne Zweifel so, dass der Informationsraum letztendlich vom semantischen Raum erzeugt wird. Die Entstehung von Informationsflüssen kann in der Tat als die Erzeugung und die Bewegung von Datenmengen verstanden werden, die mit einer bestimmten Meldung verbunden sind. Diese Meldung wird als ein semantischer Block verstanden. Einer Meldung kann dabei eine beliebige Anzahl separater Datensammlungen entsprechen. Zum Beispiel wird ein internationales Ereignis in vielen Medien vermeldet. Merkmale des Informationsraums werden also durch die Struktur des semantischen Raums bestimmt. Man spricht dabei von einer Struktur, weil die Meldungen Ereignisse der realen Welt abbilden, die einigermaßen geordnet ist.

3 Probleme des Information Retrieval

Folgende Wörter einer handelnden Person im Film „Wall Street“ können als ein Motto für das Thema dienen: „Tell me something I don't know“. Diese Wörter sind besonders aktuell, wenn sie sich auf die Suche im WWW beziehen, das als eine dynamische und abwechslungsreiche Datensammlung betrachtet wird. Bei der Suche in den Informationsflüssen entsteht ein separates Problem, das eine besondere Betrachtung erfordert.

Die Technologieentwicklungsversuche im Rahmen der modernen Theorie des Information Retrieval sind manchmal erfolglos und verschlechtern sogar die Situation. Zum Beispiel führt die Weiterentwicklung von technologischen Aspekten des Information Retrieval nur zu einer Steigerung der Anzahl von relevanten Daten, die für die Anwendung wenig geeignet sind. Moderne Technologien unterstützen sehr komplizierte Handlungen mit Daten, aber je effizienter diese Technologen verwendet werden, desto „ungenießbarer“ sind die Ergebnisse.

Hoffnungen, die früher auf die konsequente Verfeinerung der Suche gesetzt wurden, haben sicht aus folgenden zwei Gründen nicht erfüllt: Erstens kann das Dokument, das für Benutzer von Interesse ist, in primären Suchergebnissen fehlen, sodass die nachfolgende Iteration an Bedeutung verliert. Zweitens ist es häufig für einfache Benutzer zu schwierig und sogar die Kräfte übersteigend, eine präzisierende Anfrage zu formulieren, die sich qualitativ von der primären Anfrage unterscheidet.

Es ist zu erkennen, dass das ursprüngliche Paradigma von Information Retrievalsystemen, das vor Dekaden formuliert wurde, der realen Situation nicht entspricht. Deswegen sollten neue Methoden gefunden werden, um die umfangreichen dynamischen Datensammlungen zu verarbeiten. Wahrscheinlich wäre es sinnvoll, eine Navigation in einem Informationsfluss durchzuführen, statt die Suche in einer statischen Datensammlung. Eine solche Navigation besteht dabei in der Lokalisierung von separaten semantischen Segmenten im Informationsfluss. Dieser Vorgang hat eine bestimmte zeitliche Dauer und ist interaktiv. Vielversprechend wäre die Verwendung dynamischer Metadaten, die zur Einschränkung des Suchraums unter den gegebenen Bedingungen verwendet werden. Solche Metadaten können Benutzern auch helfen, die Lokalisierung von notwendigen Materialien zu beeinflussen. Adaptive Schnittstellen zu einer Verfeinerung der Suchanfragen, die eine Clusteranalyse unterstützen, finden in der letzten Zeit Verbreitung. In diesem Zusammenhang entstand der Begriff „Suchordner“ (Custom Search Folder), der keinen bestimmten Algorithmus voraussetzt und viele verschiedene Ansätze repräsentiert. Das Gemeinsame in diesen Ansätzen besteht in einem Versuch, die Daten zu gruppieren und die Cluster in einer benutzerfreundlichen Form darzustellen. Um Suchanfragen zu verfeinern, wurde von den Autoren im Rahmen der Info-Stream- Technologie ein Ansatz entwickelt, der als „Informationsbild“ bezeich-

net wird (Braichevski/Lande, 2005). Das Informationsbild (Abbildung 1) stellt eine Menge von Stichwörtern dar, die am genauesten Informationen repräsentiert, die in Suchergebnissen enthalten sind. Die Stichwörter werden aus gefundenen Dokumenten extrahiert, einer statistischen Verarbeitung unterzogen und dem Benutzer zugänglich gemacht. Der Benutzer kann dann diese Stichwörter zur Verfeinerung seiner Suchanfrage verwenden, ohne dabei neue Suchwörter zur Beschreibung der von ihnen gesuchten Begriffe zu suchen oder zu erfinden.

Abbildung 1: Ein Informationsbild im InfoStream-System

Das zentrale Problem von modernen Informationsflüssen besteht wahrscheinlich in dem qualitativen Unterschied zwischen den Begriffen „Relevanz“ und „Pertinenz“. Obwohl dieser Unterschied seit langem bekannt ist, wird er bei einem begrenzten Datenumfang nicht berücksichtigt. Bei einer kleineren Menge von Suchergebnissen kann ein Benutzer die relevanten Dokumente alleine durchsehen und aus diesen Dokumenten diejenigen auswählen, die ihm tatsächlich weiterhelfen. Eine solche Auswahl ist jedoch unmöglich, wenn die Suchergebnisse umfangreich sind. In diesem Falle tritt der Unterschied zwischen der Relevanz und der Pertinenz in den Vordergrund. Wenn z.B. ein Information Retrieval System 10000 Dokumente findet und alle diese Dokumente pertinent sind, wird der Benutzer zufriedengestellt, wenn er eine beliebige Anzahl dieser Dokumente durchliest. Die restlichen Dokumente können dann ignoriert werden, ohne damit irgendeinen Schaden zu verursachen. Diese Gesetzmäßigkeit wird in einigen Fällen auf eine effektive Weise verwendet. Zum Beispiel können Nachrichtensyndikationsdienste ihre Kunden zufrieden stellen, obwohl praktisch jeder solcher Dienst mit mehr als 50000 Informationsquellen arbeitet.

Der Nachteil aktueller Information Retrieval-Systeme besteht hauptsächlich darin, dass sie entwickelt werden, um die Relevanz von Suchergebnissen bezüglich formaler Anfragen zu gewährleisten.

Wir vermuten jedoch, dass moderne Informationstechnologien für den Zugang zu Daten im Netz abgeändert werden können. Diese Änderung kann man als einen Übergang von der Informationssuche zu einer Navigation im Netz definieren.

4 Strukturierungsprobleme

Es ist gut bekannt, dass der Informationsraum im Netz schwach strukturiert ist. Darüber hinaus kann die Evolution des gesamten Netzes und seiner Segmente als Beispiel eines stochastischen Vorganges betrachtet werden. Diese Tatsache ist die Hauptursache der niedrigen Effizienz des direkten Zuganges zu Informationseinheiten, über welche wir häufig nicht wissen, ob sie überhaupt zu einem gegebenen Zeitpunkt existieren.

Das oben Gesagte bedeutet nicht, dass der Informationsraum im Netz völlig chaotisch ist und nur in Termen des Informationsgeräusches vollständig beschrieben werden kann. In der Tat enthält dieser Raum Elemente einer Ordnungsmäßigkeit, die im Folgenden als Cluster bezeichnet werden. Die Anzahl der Cluster ist groß, und jedes von ihnen hat seine eigene Entwicklungsdynamik, die mit der Dynamiken anderer Cluster korreliert. Andererseits können diese Cluster intensiv aufeinander wirken. Die Cluster sind nicht immer stabil in der Zeit. Sie entstehen, ändern ihre Umrisse, verschwinden, migrieren usw. Darüber hinaus ist ihre Zusammenwirkung völlig stochastisch.

Der erste reale Schritt zur Lösung des Strukturierungsproblems im Informationsraum besteht offensichtlich in der Erzeugung eines sekundären Raums, der genügend geordnet und bei einer vernunftmäßigen Approximation dem Primärraum adäquat ist. Auf diese Weise entsteht die Aufgabe, eine ungeordnete Menge von Komponenten eines Informationsraums im Netz auf eine geordnete Menge entsprechender Muster abzubilden, die den Anforderungen gemäß, z.B. hierarchisch, organisiert ist.

Die Suche kann dann in einer strukturierten Menge der Muster durchgeführt werden, und die Präsentation der Suchergebnisse hat die Wiederherstellung der originellen Informationseinheiten einzuschließen. Dieser Ansatz kann auch in einigen Fällen helfen, ein immer noch offenes Problem des theoretischen Information Retrieval, das Problem von mehrfach vorhandenen Informationen (Dubletten), zu lösen. Das Problem kann beim Aufbau des Musterraums damit gelöst werden, dass Ketten von ähnlichen Informationseinheiten zuerst erzeugt und dann auf ein und dasselbe Muster abgebildet werden. Beim Aufbau des Musterraums können die Muster mit Metadaten versehen werden.

Eine der natürlichen Lösungen der o.g. Probleme wäre die Verlegung des Schwerpunktes von Daten, in denen die Suche durchgeführt wird, auf Metadaten, die mit diesen Daten verbunden sind und ein breites Spektrum von externen Merkmalen enthalten. Diese Merkmale können relativ einfach zum Erstellen eines „Wortbilds“ angeforderter Dokumente verwendet werden.

Der Kern der Suchanfrage muss aus formalen Parametern bestehen, die auf bestimmte Kategorien der Metadaten verweisen. Die konventionelle Anfrage, die Suchbegriffe enthält, kann dann als ein Hilfsmittel verwendet werden, um die ausgewählten Menge der pertinenten Dokumente zu verkleinern.

Die erwähnten Ausführungen beziehen sich natürlich nicht nur auf reine Informationssuche sonder auch auf andere Aufgaben, die mit der Suche verbunden sind, z.B. auf Profildienste.

5 Textmining

Die Effizienz des Information Retrieval kann erhöht werden, indem man Textmining, d.h. Technologien zur tiefgehenden automatischen Textanalyse, verwendet. Das Textmining kann auch Benutzern helfen, die Suchergebnisse schneller zu analysieren. Zu wichtigen Technologien des Textmining gehören u.a.

- Textklassifizierung
- Textclustern
- Informationsextraktion.

Die automatische Textklassifizierung bedeutet, dass zu einen Dokument Kategorien eines vordefinierten Ordnungssystems von einem Computer zugeordnet werden. Die Kategorien können dann verwendet werden, um die Dokumente wiederaufzufinden. Beim Textclustern werden aus Dokumentensammlungen Gruppen (Cluster) gebildet, die ähnliche Dokumente enthalten. Wenn ein Suchdienst mehrere Dokumente auf eine Anfrage findet, können aus diesen Suchergebnissen Cluster erzeugt und visualisier werden. Die Suchergebnisse werden auf diese Weise übersichtlicher und können von Benutzern schneller analysiert werden. Unter der Informationsextraktion

versteht man die Extraktion von strukturierten Daten, z.B. Attributwerten aus Texten. Dies kann u.a. ermöglichen, eine attributbasierte Suche in den Texten durchzuführen. Zu Aufgaben der Informationsextraktion gehört auch die Erkennung von Eigennamen, z.B. Personen-, und Firmennamen, in Texten. Wenn solche Namen erkannt und visualisiert werden, können die Texte von Benutzern schnell gelesen und analysiert werden. Einige weitere Aufgaben des Textmining sind z.B. die automatische Erzeugung von semantischen Netzwerken, die Prognostizierung eines Merkmalswertes in einem Objekt aufgrund der Werte anderer Merkmale und die Suche von Ausnahmen oder Anomalien, d.h. es werden Objekte gesucht, deren Merkmale sich erheblich von Charakteristiken der gesamten Menge der Objekte unterscheiden.

Da die meisten Dokumente im Web Texte enthalten, können auf diese Dokumente auch Textmining-Verfahren angewendet werden. Textmining für Web-Dokumente ist deswegen Aufgaben des Web Content Mining, d.h. der Wissensentdeckung in Webinhalten.

Bei der Verarbeitung und der Interpretation der Ergebnisse vom Textmining spielt die Visualisierung eine wichtige Rolle. Die Visualisierung auf der Basis des Textmining kann zur Inhaltsrepräsentation des gesamten Informationsflusses sowie für die Implementierung des Navigationsmechanismus verwendet werden, der bei Untersuchungen der Dokumente benutzt wird. Dies ist unserer Meinung nach eine der bedeutendsten Errungenschaften moderner Informationstechnologien mit dieser Zielrichtung. Wesentliche ist nämlich, dass die effektive Repräsentation von Datenflüssen in einer benutzerfreundlichen Form es erlaubt, die menschliche Intelligenz direkt einzusetzen, die letztendlich viel schneller als jeglicher Rechner zum Ziel führt.

Genaueres über das Text Mining kann man bei Heyer et al. (2006) und Weiss et al. (2005) erfahren. Das Problem der Wissensentdeckung in Texten befindet sich derzeit im Stadium einer gedanklichen Verarbeitung. In der näheren Zukunft wird es vermutlich eine stürmische Weiterentwicklung entsprechender Technologien geben.

6 Ranking von Informationsflüssen

Vor einiger Zeit wurde die Idee geäußert, dass das Information Retrieval im klassischen Sinne durch eine Sortierprozedur ersetzt werden kann, sofern sie effektiv genug ist. Diese Prozedur wird nach einem Satz von Parametern durchgeführt, die den Informationsbedarf eines Benutzers quantitativ repräsentieren. In diesem Falle gibt das Retrievalsystem alle Dokumente aus, die in der Datenbank enthalten sind. Die Dokumente, welche der Aufgabenstellung entsprechen, befinden sich dabei am Anfang der ausgegebenen Dokumentenliste.

Obwohl man diese Idee für zu radikal hält, enthält sie einen gewissen Anteil der Wahrheit. Eine konventionelle Retrievalprozedur kann in der Tat ausgeführt werden, wenn für jedes Dokument dessen Relevanz bezüglich der Suchanfrage berechnet wird.

Derzeit findet immer noch das Retrievalmodel Verbreitung, das invertierte Wörterbücher verwendet. Es ist bekannt, dass bei solcher Suche die Relevanz nur zwei Werte, 0 und 1, annehmen kann. Dies bedeutet, dass die klassische Suche aus der primären Datenmenge Dokumente auswählt, die in diesem Sinne gleichwertig und gleichbedeutend sind. Wenn wir die Suchergebnisse zur Abbildung der realen Informationssituation annähren wollen, ist die Suche durch eine Prozedur zu ergänzen, welche diese oder jene Verteilung nach Parametern unterstützt, die eine subjektive Bewertung und eine nachfolgende Sortierung der Suchergebnisse ermöglichen. Vielversprechend kann unserer Meinung nach die Verwendung von Mehrfachskalen sein, die auf Basis von einigen Metadaten aufgebaut werden.

Die Clusteranalyse hat eine neue Qualität zu erhalten. Die Daten werden von alleine ohne System erzeugt, d.h. sie werden aus verschiedenen Quellen ohne spezielle Aktionen und ohne Programme erzeugt. Die Clusteranalyse erlaubt, die Informationsflüsse auf permanente und sichere Weise zu systematisieren. Bei der Clusteranalyse entsteht jedoch das Problem, dass die meisten bekannten Methoden statische Objekte clustern, obwohl der Informationsraum ein dynamisches System ist (Del Corso et al., 2005). Das Problem verspricht paradoxerweise neue Möglichkeiten, die sich qualitativ vom statischen Clustern unterscheidet. Zu diesen Möglichkeiten gehört u.a. die Berücksichtigung von zeitlichen Zusammenhängen zwischen Hauptparametern der Informationsflüsse. Zum Beispiel ist es sehr wichtig, temporäre Stabilität von statischen Merkmalen der Flussdaten zu untersuchen.

Der o.g. Ansatz wird u.a. von den Autoren im Rahmen der InfoStream-Technologie verwendet, um thematische Ketten auf Basis von thematischen Retrievalergebnissen zu erzeugen (Braichevski/Lande, 2005).

Es ist zu bemerken, dass die Clusteranalyse von besonderer Bedeutung ist, weil die dynamischen Cluster ein konzeptionelles Netzwerk bilden, das für die Analyse des Informationsflusses verwendet wird. Dabei wird auch der menschliche Faktor verwendet, der in Expertenschätzungen repräsentiert wird, die als eine Rückkopplungsschleife von trainierten Systemen betrachtet werden. Die Erkennung eines Clusters setzt dessen Beschreibung voraus.

Im semantischen Ansatz zu dieser Beschreibung sind auch quantitative Schätzungen vorhanden, obwohl die Beschreibungsmerkmale komplex und vielfältig sind. Moderne Informationsflüsse enthalten auch das gesamte Wörterbuch der modernen Sprache sowie „fertige“ Spezialwörterbücher, wie z.B. Frequenzwörterbücher, invertierte Wörterbücher usw.

7 Automatische Textklassifizierung

Die automatische Klassifizierung gehört, wie erwähnt, zu wichtigen Technologien des Textmining. Bei der automatischen Klassifizierung werden zu einem Textdokument Kategorien (Notationen bzw. Deskriptoren) eines vordefinierten Ordnungssystems von einem Programm zugeordnet. Zu diesen Ordnungssystemen gehören sowohl Klassifikationssysteme als auch andere kontrollierte Wörterbücher, wie z.B. Schlagwortlisten und Thesauri. Die Kategorien, die einem Dokument zugeordnet wurden, können dann verwendet werden, um das Dokument wiederaufzufinden. Man kann nicht nur statische Datensammlungen, sondern auch dynamische Informationsflüsse automatisch klassifizieren. Zum Beispiel können Dokumente klassifiziert werden, die eine Suchmaschine auf Anfrage findet (Chen/Dumais, 2000; Kules et al., 2006). Solche Klassifizierung hilft Benutzern, in den Suchergebnissen zu navigieren und pertinente Dokumente zu finden.Abb Außer Retrievalzwecken kann die automatische Klassifizierung zu anderen Zielen, z.B. zum Herausfiltern von unerwünschten Nachrichten (Spam), verwendet werden. Im Folgenden werden Technologien zur automatischen Textklassifizierung genauer betrachtet.

In ersten Programmen zur automatischen Klassifizierung wurde ein Ansatz verwendet, derauf dem Einsatz von Expertensystemen basierte. Solche Klassifikatoren basieren auf Regeln, die intellektuell und manuell erstellt werden. Obwohl diese Klassifikatoren eine hohe Präzision erzielen, ist die intellektuelle und manuelle Erstellung der Klassifizierungsregeln zu teuer und zu zeitaufwendig. Heutzutage dominieren deswegen Ansätze, die auf maschinellen Lernverfahren basieren. Bei solchen Ansätzen werden Klassifikatoren automatisch aus Dokumenten abgeleitet, die bereits klassifiziert sind. Zu lernenden Klassifikationsverfahren gehören u.a. probabilistische und instanzbasierte Verfahren, Regelinduktionsverfahren, Rocchio, Online-Methoden und die Support-Vector-Machine. Es gibt auch kommerzielle Klassifizierungssoftware, z.B. Autonomy Classification und Insight SmartDiscovery. Eine genauere Beschreibung verschiede-

ner Verfahren und Programme zur automatischen Klassifizierung kann man bei Oberhauser(2005) und Weiss et al.(2005) finden.
Ein Prototyp, der die hierarchische Klassifizierung nach der Internationalen Patentklassifikation (IPC) unterstützt, wurde von den Autoren entwickelt. Abbildung 2 zeigt ein Bespiel der Klassifikation eines Dokuments nach der IPC. Das Programm erzeugt für das Dokument eine Liste von Kategorien (IPC- Hauptgruppen), die in absteigender Relevanz dem Dokument zugeordnet sind. Im o.g. Beispiel erfolgt die Klassifizierung nach einem probabilistischen Verfahren (Naive Bayes). Der Prototyp unterstützt auch andere Klassifizierungsverfahren, z.B. Rocchio und Support-Vector-Machinen.

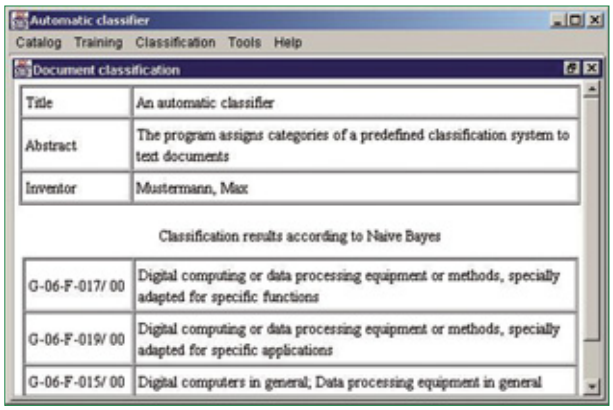

Abbildung 2: Automatische Klassifizierung eines Dokuments nach der IPC

In einigen Ordnungssystemen, z.B. Facettenklassifikationen, werden Attribute mit Kategorien verbunden, die es erlauben, Dokumente genauer zu beschreiben. Um eine vollständige Klassifikation eines Dokumentes durchzuführen, sind diesem Dokument nicht nur Kategorien zuzuordnen, sondern auch Werte der Attribute zu setzen, welche mit diesen Kategorien verbunden sind. Die Werte der Attribute können aus Dokumenten automatisch extrahiert werden. Die automatische Informationsextraktion kann ähnlich der automatischen Klassifizierung auf Expertenwissen basieren oder maschinelle Lernverfahren verwenden.
Einer der Autoren entwickelte ein prototypisches Programm, das Einträge von elektronischen Produktkatalogen nach ETIM, einem Klassifikationssystem mit Attributen, automatisch klassifiziert (Busch, 2005). Die automatische Klassifizierung und die Extraktion von Werten der Attribute erfolgt in diesem Prototyp nach regelbasierten Verfahren. Sowohl Klassifizierungs- als auch Extraktionsregeln können automatisch aus Katalogen abgeleitet werden, die bereits klassifiziert sind. Für die Ableitung der Extraktionsregeln müssen zusätzlich Attributwerte angegeben werden.
Für einfache Ordnungssysteme, die wenige Kategorien enthalten, können automatische Klassifikatoren hohe Präzision bis ca. 90 Prozent erzielen (vgl. Yang, 1999). Wenn ein Ordnungssystem komplex ist, z.B. tausende Kategorien und mehrere Hierarchieebenen enthält, ist dies jedoch problematisch (vgl. Oberhauser, 2005). Für solche komplexen Klassifikationssysteme wird deswegen die Zuordnung von Kategorien normalerweise nicht vollautomatisch sondern semi-automatisch durchgeführt, d.h. der Computer schlägt Kategorien vor, aber die tatsächliche Zuordnung der Kategorien erfolgt durch eine Fachkraft.

8 Semantisches Web

Zu vielversprechendsten Richtungen gehört ein aktueller Ansatz, der als synthetisch bezeichnet werden kann. Die grundlegende Idee des Ansatzes besteht in dem Versuch, komplizierte Aufgaben zu lösen, indem man aus einheitlichen Prinzipien der Erzeugung, des Transfers und der Verarbeitung von Daten ausgeht. Der Schwerpunkt dieses Ansatzes liegt dabei in der Abstimmung von Parametern von Objekten, die verarbeitet werden, und Werkzeugen für die Verarbeitung.
Als Beispiel kann das semantische Web (Semantic Web) genannt werden, das von dessen Entwicklern als absolut selbstgenügend betrachtet wird. Die Idee des semantischen Webs wurde zuerst von Berners-Lee et al. (2001) vorgeschlagen und besteht in einer Datenrepräsentation im Web, die erlaubt, diese Daten sowohl zu visualisieren als auch mit Programmen verschiedener Hersteller effizient zu verarbeiten. Anhand solcher radikalen Veränderungen im traditionellen Web-Konzept wird beabsichtigt, das Web in ein semantisches System umzuwandeln. Das semantische Web soll das automatische „Verstehen“ von Informationen, die Extraktion von Daten nach diesen oder jenen Kriterien und davon abgeleitet dann die Ausgabe der Informationen für Benutzer unterstützen.
Das semantische Web kann als eine Symbiose von zwei Bestandteilen verstanden werden. Der erste Teil umfasst Datenrepräsentationssprachen. Als wichtigste Datenrepräsentationssprachen gelten derzeit XML (Extensible Markup Language) und RDF(Resource Description Framework). Obwohl es auch andere Formatierungssprachen gibt, bieten XML und RDF mehr Möglichkeiten an, und werden deswegen von W3C-Konsortium empfohlen.
Der zweite, konzeptionelle Teil enthält theoretische Konzepte und Modelle von Sachgebieten, die in der Terminologie des semantischen Web als Ontologien bezeichnet werden. Um die Ontologien zu definieren, wurde vom W3C-Konsortium eine ontologische Sprache OWL (Web Ontology Language) entwickelt.
Die o.g. zwei Bestandteile des semantischen Webs verwenden also drei grundlegende Sprachen:

- XML- Spezifikation, die erlaubt, die Syntax und die Struktur der Dokumente zu erkennen
- RDF, der Mechanismus für die Beschreibung von Ressourcen, der ein Kodierungsmodell für Werte unterstützt, die in einer Ontologie definiert werden.
- OWL, die Ontologiesprache, die erlaubt, Begriffe und Beziehungen zwischen ihnen zu definieren.

Das semantische Web verwendet auch andere Sprachen, Technologien und Konzepte, z.B. universelle Kennzeichen für Ressourcen, digitale Signaturen und Systeme zur logischen Inferenz.
Fast jede Implementierung des semantischen Web hängt kritisch vom Vorhandensein von Web-Seiten ab, welche Metadaten enthalten, die nicht im Rahmen des Standardvorganges für die Web-Entwicklung erstellt wurden. Man kann von Web-Autoren wohl kaum erzwingen, ihre Webseiten mit terminologischen Wörterbüchern und mit Ontologien des semantischen Web zu indizieren. Web-Quellen, die bereits existieren, können offensichtlich nur automatisch in das semantische Web integriert werden. Diese Aufgabe ist sehr komplex und erfordert die Verwendung von Ansätzen, die Textmining-Technologien ähneln. Auf diesem Weg kann man wahrscheinlich unter den gegebenen Umständen die besten Ergebnisse erreichen.

9 Fraktale Eigenschaften des Informationsraums

In vielen Modellen des Informationsraums werden derzeit strukturelle Beziehungen zwischen separaten Objekten untersucht, die in diesem Raum enthalten sind. Bei der Modellierung des Informationsraums wird zunehmend der fraktale Ansatz verwendet, der auf der Selbstähnlichkeit des Informationsraums basiert. Die Selbstähnlichkeit bedeutet die Erhaltung der inneren Struktur von Mengen bei der Veränderung von Betrachtungsmaßstäben dieser Mengen.
Die Verwendung der Fraktaltheorie bei der Analyse des Informationsraums erlaubt, empirische Gesetze, die theoretische Grundlagen der Informationswissenschaft bilden, von einem gemeinsamen Standpunkt aus zu betrachten. Zum Bei-

spiel stellen thematische Informationssammlungen selbstentwickelnde und selbstähnliche Strukturen dar und können daher als stochastische Fraktale betrachtet werden (Van Raan, 1991).
Es ist bekannt, dass alle grundlegenden Gesetze der wissenschaftlichen Kommunikation, wie z.B. Pareto-, Lotka-, Bradford- und Zipf-Gesetze, im Rahmen der Theorie stochastischer Fraktale zusammengefasst werden können (Ivanov, 2002).
Die Selbstähnlichkeitseigenschaften von Fragmenten des Informationsraums können beispielsweise mit einer Benutzerschnittstelle verdeutlicht werden, die von der Website „News is Free“[3] unterstützt wird. In dieser Website wird der Zustand des Informationsraums in Form von Verweisen auf Nachrichtenquellen und separate Nachrichten dargestellt. Bei der Darstellung werden zwei Hauptparameter, der Popularitätsrang und die Aktualität der Informationen, berücksichtigt. Eine vergrößerte Präsentation einzelner Quellen und/oder Dokumente, die am populärsten und am aktuellsten sind, stellt die Selbstähnlichkeitseigenschaft auf anschauliche Weise dar.
Gegenwärtig wird die Fraktaltheorie weitgehend als ein Ansatz zur statistischen Forschung verwendet. Dieser Ansatz erlaubt es, wichtige Charakteristiken von Informationsflüssen zu erhalten, ohne die interne Struktur des jeweiligen Informationsflusses zu analysieren. Zum Beispiel ist die Anzahl von Internet-Meldungen, die eine Resonanz auf ein Ereignis in der realen Welt darstellen, proportional einer Potenz der Anzahl der Quellen(Websites). Wie bei der traditionellen wissenschaftlichen Kommunikation stellt die Anzahl von Meldungen zu einem ausgewählten Thema ein dynamisches Clustersystem dar.
Die fraktale Dimension in einem Clustersystem, das thematischen Informationsflüssen entspricht, ist ein Maß dafür, wie viele Meldungen den Informationsraum zu einem bestimmten Zeitpunkt ausfüllen:
$N_{pub}(\varepsilon t) = \varepsilon^{\rho} N^{k}(t)^{\rho}$,
wobei N_{publ}- Größe des Clustersystems (Gesamtzahl von elektronischen Publikationen in dem Informationsfluss); N_k- Anzahl der Cluster (z.B. Quellen); ρ- fraktale Dimension des Informationsarrays; ε- Maßstabsfaktor (vgl. Ivanov, 2002)
Im Zusammenhang mit der Entwicklung der Theorie stochastischer Fraktale wird heute häufig eine Zeitreihen-Charakteristik, Hurst- Exponent (Feder, 1988), verwendet. In seiner Zeit entdeckte Hurst auf experimentelle Weise, dass für viele Zeitreihen Folgendes galt:
$R/S = (N/2)^{H}$,
wobei R-„Spannweite“ der entsprechenden Zeitreihe, die auf eine bestimmte Weise berechnet wird; S-Standardabweichung.
Die Autoren beweisen, dass für thematische Informationsflüsse, die mächtig genug sind und auf iterative Weise gebildet werden, der Hurst-Exponent mit der traditionellen fraktalen Dimension (Θ) folgendermaßen zusammenhängt:
$\Theta = 2 - H$.
Es ist bekannt, dass der Hurst-Exponent ein Maß für die Persistenz, d.h. für die Neigung zu einer Vorzugsbewegung, darstellt (zum Unterschied von der gewöhnlichen „Brownschen Bewegung“). Im Falle von Informationsflüssen erlaubt der Hurst-Exponent (H), ihre Dynamik zu prognostizieren. Ein Wert von $H > 1/2$ bedeutet, dass die in eine bestimmte Richtung weisende Dynamik des Vorganges in der Vergangenheit die künftige Bewegung in dieselbe Richtung am wahrscheinlichsten zur Folge hat. Wenn $H < 1/2$, wird vorhergesagt, dass der Vorgang seine Richtung wechselt. $H = 1/2$ bedeutet eine Unbestimmtheit, d.h. „Brownsche Bewegung“.
Die Autoren untersuchten fraktale Eigenschaften von Informationsflüssen, indem sie einen Dokumentenkorpus vom InfoStream (Braichevski/Lande 2005), einem System für das Monitoring von Nachrichten im Internet, verwendeten. Thematiken von Informationsflüssen, die untersucht wurden, wurden durch typische Benutzeranfragen an InfoStream bestimmt. Man betrachtete Reihen, welche Anzahlen von Publikationen im Bezug auf Publikationsdaten für die Jahre 2004 bis 2006 zeigten. Werte des Hurst-Exponents stabilisierten sich für unterschiedliche Thematiken und betrugen von 0,68 bis 0,96. Dies deutet auf eine hohe Persistenz der untersuchten Zeitreihen hin.
Untersuchungen, die von den Autoren durchgeführt wurde, bestätigen also eine Vermutung über die Selbstähnlichkeit und die Iteration von Vorgängen im Informationsraum. Wiederholte Publikationen, Zitierungen, direkte Verweise usw. verursachen die Selbstähnlichkeit, die sich in stabilen statistischen Verteilungen und bekannten empirischen Gesetzen zeigt. Das Selbstähnlichkeitsprinzip wird auch mit der Mentalitätsähnlichkeit der Autoren erklärt, die ihre Meldungen im Internet publizieren. Zugleich führen verschiedene Marketing-, Werbe- und PR-Maßnahmen zu sprunghaften Veränderungen in stabilen statistischen Gesetzmäßigkeiten, heftigen Sprüngen und Verzerrungen im Vergleich mit statistischen Standardverteilungen.
Darüber hinaus bestätigen Topologien und Charakteristiken sowohl des bekannten Webraummodells nach Broder et al.(2000) als auch Modelle des Nachrichten-Webraums (Lande, 2006) die Beobachtung, dass die Struktureigenschaften des gesamten Webraums auch für dessen separate Untermengen gelten. Man kann also vermuten, dass Algorithmen, die die Struktur des Webraums und dessen dynamischen Teil beschreiben, auch für die einzelnen Untermengen des Webraums einsetzbar sind.

10 Das Invisible Web

Allgemeine Prinzipien der Netzwerkorganisation lassen die Existenz von geschlossenen Gebieten des Informationsraums zu, die für Standardwerkzeuge der Informationsverarbeitung nicht zugänglich sind. Dieser Faktor stellt beispielsweise aktuelle Bewertungen des Informationszuwachses im Netz in Frage. Es ist nicht ausgeschlossen, dass sichtbare Informationen nur einen kleineren Anteil an dem gesamten Informationsvolumen bilden. Ein anderes ernstes Problem besteht darin, das viele Informationen verloren gehen, wenn sie in ein unzugängliches Gebiet geraten. Bei der Verarbeitung von Informationsflüssen ist eine solche Situation besonders problematisch, weil es schwierig ist, Informationen zu kontrollieren, die ständig verändert werden.
Seit einiger Zeit wird der Begriff „Invisible Web“ verwendet. Das Invisible Web ist ein Teil des Web, das für konventionelle Information-Retrievalsysteme nicht zugänglich ist. Der Großteil der Inhalte von Websites bleibt häufig für Suchmaschinen unzugänglich, weil in vielen Webservern unterschiedliche Formate für Speicherung, Verarbeitung und äußerliche Gestaltung verwendet werden. Viele Webseiten werden dynamisch und nur nach Benutzeranfragen erzeugt. Traditionelle Suchmaschinen können solche Quellen nicht verarbeiten und ihre Inhalte nicht erkennen. Das Invisible Web umfasst hauptsächlich den Inhalt von Online-Datenbanken. Darüber hinaus sind Informationen verborgen, die schnell aktualisiert werden, z.B. Nachrichten, Konferenzen, Online-Zeitschriften. Ein Bericht der amerikanischen Firma BrightPlanet[4] behauptet, dass es im Web um hundertmal mehr Seiten gibt, als die Anzahl der Seiten, die von populären Suchmaschinen indexiert werden.
Man kann nicht sagen, dass es keine Schritte unternommen wurden, um Probleme zu lösen, die mit dem Invisible Web verbunden sind. Es gibt einige technologische Ansätze für bestimmte Quellen und Aufgabenklassen (vgl. Lewandowski, 2005). Dennoch gibt es keine umfassenden Lösungen der Probleme. Die Hauptschwierigkeit besteht darin, dass sich die Mechanismen mit denen Daten in geschlossene Bereiche geschleust werden, kaum umfassend und explizit berücksichtigen lassen. Es ist auch schwierig, Mechanismen des Entstehens und der Stabilisierung solcher Bereiche vo-

3 http://newsisfree.com [12.07.2007]
4 www.press.umich.edu/jep/07-01/bergman.html [12.07.2007]

DABIS.eu

Gesellschaft für Datenbank-InformationsSysteme mbH

Ihr Partner für Archiv-, Bibliotheks- und DokumentationsSysteme

BIS-C 2000

Archiv- und Bibliotheks-InformationsSystem

DABIS.eu · alle Aufgaben · ein Team

Synergien: Qualität und Kompetenz
Software: Innovation und Optimierung
Web · SSL · Warenkorb und Benutzeraccount
Lokalsystem zu Aleph-Verbünden

Software · State of the art · Open Source

Leistung	Sicherheit
Standards	Offenheit
Stabilität	Verläßlichkeit
Generierung	Adaptierung
Service	Erfahrenheit
Outsourcing	Support
Dienstleistungen	Zufriedenheit

GUI · Web · Wap · XML · Z 39.50

Archiv		Bibliothek
singleUser	System	multiUser
Lokalsystem		Verbund
multiDatenbank		multiServer
multiProcessing		multiThreading
skalierbar		stufenlos
Unicode		multiLingual
Normdaten		redundanzfrei
multiMedia		Integration

DABIS.com
Heiligenstädter Straße 213
1190 · Wien, Austria
Tel.: +43-1-318 9 777-10
Fax: +43-1-318 9 777-15
eMail: office@dabis.com
http://www.dabis.com

DABIS.de
Herrgasse 24
79294 · Sölden/Freiburg, Germany
Tel.: +49-761-40983-21
Fax: +49-761-40983-29
eMail: office@dabis.de
http://www.dabis.de

Zweigstellen: 61350 · Bad Homburg vdH, Germany / 1147 · Budapest, Hungary / 39042 · Brixen, Italy

rauszusehen. Deswegen ist es auch wenig wahrscheinlich, dass in der näheren Zukunft in dieser Richtung beträchtliche Erfolge erzielt werden.

11 Fazit

Es ist dringend erforderlich, eine multidisziplinäre Erforschung des Informationsraums durchzuführen. Eine der aktuellsten Aufgaben für Wissenschaftler und Forschern besteht darin, ein klares Modell des modernen Informationsraums zu entwickeln, das auf Erkenntnissen von Informationswissenschaft und Linguistik basiert. Bei der Erarbeitung eines solchen Modells werden auch strenge mathematische Werkzeuge und Methoden verwendet, die denen der theoretischen Physik ähnlich sind. Man hat u.a. maschinelle Lernverfahren zu entwickeln, die zum Unterschied von traditionellen Konzepten der künstlichen Intelligenz den Aufbau von Prozeduren ermöglichen, in welchen auch menschliche Intelligenz eingebunden wird. Diese Teilnahme kann sowohl implizit als auch explizit sein. Zum Beispiel können Benutzeranfragen berücksichtigt werden, die von Suchmaschinen bearbeitet werden. Darüber hinaus können automatisierte Prozeduren entwickelt werden, welche den Benutzern erlauben, Merkmale der gesuchten Objekte zu verfeinern.

Die Erforschung von Informationsflüssen kann andererseits für Linguisten, Mathematiker und Physiker von Interesse werden. Zu den Bereichen, die von multidisziplinärem Interesse sind, gehört z.B. die analoge Modellierung von statistischen Vorgängen einschließlich komplexer nichtlinearer Systemen mit Selbstorganisationselementen. Die Semantik des Informationsraums reizt auch zur Entwicklung von neuen Methoden für Kodierung und Komprimierung von Informationen inklusive der Technologien zur eindeutigen Entschlüsselung der Nachrichten. Für die neue Etappe der Entwicklung des Webraums werden voraussichtlich Technologien entscheidend, die sich für die Arbeit mit riesigen Informationsmengen im Internet eignen. Das Web der nächsten Generation wird durch einen Übergang von einem Dokumentennetz zu einem Netz der Daten, die nach Bedarf mit Hilfe von Webdiensten zu semantisch verbundenen Dokumenten zusammengefasst werden können. Es wird voraussichtlich ein gemeinsamer Informationsraum existieren, der aus einer Menge von Informationseinheiten besteht, die in zahlreichen Websites enthalten sind. Der Benutzer wird Dokumente erhalten, indem er die Informationseinheiten am Arbeitsplatz sammeln wird. Die Arbeit mit dem Informationsraum wird voraussichtlich von der Entwicklung einer effektiven Infrastruktur abhängig, die server- und clientseitige Programme unterstützen wird.

Da es eine umfangreiche und kostengünstige Basis für Experimente gibt, werden es sogar Teillösungen für die o.g. Aufgaben ermöglichen, nützliche und effiziente Werkzeuge zur Arbeit und zum Surfen in Informationsflüssen zu implementieren.

Literatur

Berners-Lee,T.; *Hendler*, J.; *Lassila*, O. (2001): The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In: Scientific American 284(2001)5, S. 34-43

Braichevski, S., *Lande*, D. (2005): Urgent aspects of current information flow. In: Scientific and Technical Information Processing 32(2005)6, S. 18-31

Broder, A.; *Kumar*, R.; *Maghoul*, F. et al. (2000): Graph structure in the Web. In: Computer Network 33(2000)1-6, S. 309-320

Busch, D. (2005): Automatische Klassifizierung von deutschsprachigen elektronischen Katalogen der Elektroindustrie nach dem Elektrotechnischen Informationsmodell (ETIM). Berlin: Mensch & Buch

Chen, H.; *Dumais*, S. (2000): Bringing order to the Web: automatically categorizing search results. In: Proceedings of the SIGCHI conference on Human factors in computing systems. The Hague, S. 145-152

Del Corso, G.; *Gull›*, A.; *Romani*, F.(2005): Ranking a stream of news. In: Processing of the 14th international World Wide Web conference, 2005. www 2005.org/cdrom/docs/p97.pdf [12.07.2007]

Feder, J. (1988): Fractals. New York: Plenum Press

Heyer, G.; *Quasthoff*, U.; *Wittig*, T. (2006): Text Mining: Wissensrohstoff Text. Bochum: W3L

Ivanov, S.(2002): Stochastic Fractals in Informatics. In: Automatic Documentation and Mathematical Linguistics 36(2002)4, S. 17-34

Kules, B.; *Kustanowitz*, J.; *Shneiderman*, B. (2006): Categorizing Web search results into meaningful and stable categories using fast-feature techniques. In: Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries, 2006. Chapel Hill. S. 210-219

Lande, D. (2006) Structure of the Web news space. In: Automatic Documentation and Mathematical Linguistics, 40(2006)4, S. 159-162

Lewandowski, D. (2005): Web Information Retrieval. In: Information: Wissenschaft und Praxis 56(2005)1, S. 5-12

Oberhauser, O. (2005): Automatisches Klassifizieren: Entwicklungsstand – Methodik – Anwendungsbereiche. Frankfurt: Peter Lang

Van Raan, A.(1991): Fractal geometry of information space as represented by cocitation clustering. In: Scientometrics 20(1991)3, S. 439-449.

Weiss, S.; *Indurkhya*, N.; *Zhang*, T.; *Damerau*, F. (2005): Text Mining: Predictive methods for analysing unstructured information. New York: Springer

Yang, Y. (1999): An evaluation of statistical approaches to text categorization. In: Journal of Information Retrieval 1(1999)1/2, S. 67-88

Klassifikation, Entwicklungstendenz, inhaltliche Erschließung, Forschung Textmining, Textanalyse, Web, Informationsnetz

Die Autoren

Dimitri Lande

Stellvertretender Direktor für wissenschaftliche Arbeit im Informationszentrum ElVisti, Kiew (Ukraine). Studium der Mathematik an der Staatsuniversität Kiew. Ph.D. in theoretischer Informatik am Institut für Kybernetik der Ukrainischen Akademie der Wissenschaften, Kiew. D.Sc.-Dissertation im Informationsmanagement in der Nationalbibliothek der Ukrainischen Akademie der Wissenschaften. 150 wissenschaftliche Publikationen über Information Retrieval, Volltext-Datenbanken und Informationsflusstheorie.

ElVisti Information Center
Vulitsja Maksima Krivonosa 2-A
03037 Kiev, Ukraine
dwl@visti.net
http://dwl.kiev.ua, www.visti.net

Sergei Braichevski

Leitender wissenschaftlicher Mitarbeiter im Informationszentrum ElVisti, Kiew (Ukraine). Studium der Kernphysik an der Staatsuniversität Kiew. Ph.D. in theoretischer Physik an der Staatsuniversität Minsk (Weißrussland). 20 wissenschaftliche Publikationen über Informationstechnologien. Einer der Entwickler des Information-Retrievalsystems InfoRes und des Content-Monitoring-Systems InfoStream.

ElVisti Information Center
Vulitsja Maksima Krivonosa 2-A
03037 Kiev, Ukraine
smb@visti.net, www.visti.net

Dimitri Busch

Freiberuflicher IT-Consultant. Studium der Wirtschaftsinformatik an der Hochschule für Volkswirtschaft Kiew (Ukraine) und Informationswissenschaft an der Universität Konstanz. Promotion zum Dr. phil. im Fach Informationswissenschaft an der Universität des Saarlandes. Fachliche Interessen: Textmining; Webmining; Information-Retrieval.

Marabustraße 35/30
70378 Stuttgart
Telefon/Fax: (07 11) 5 20 12 09
bips@gmx.de