(19) **Bundesrepublik Deutschland**
Deutsches Patent- und Markenamt

(10) **DE 603 15 506 T2** 2008.04.17

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 1 573 660 B1**
(21) Deutsches Aktenzeichen: **603 15 506.5**
(86) PCT-Aktenzeichen: **PCT/US03/39356**
(96) Europäisches Aktenzeichen: **03 790 448.9**
(87) PCT-Veröffentlichungs-Nr.: **WO 2004/053771**
(86) PCT-Anmeldetag: **11.12.2003**
(87) Veröffentlichungstag der PCT-Anmeldung: **24.06.2004**
(97) Erstveröffentlichung durch das EPA: **14.09.2005**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **08.08.2007**
(47) Veröffentlichungstag im Patentblatt: **17.04.2008**

(51) Int Cl.[8]: ***G06K 9/62*** (2006.01)

(30) Unionspriorität:
317438 **11.12.2002** **US**

(84) Benannte Vertragsstaaten:
AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LI, LU, MC, NL, PT, RO, SE, SI, SK, TR

(73) Patentinhaber:
Attenex Corp., Seattle, Wash., US

(74) Vertreter:
HOFFMANN & EITLE, 81925 München

(72) Erfinder:
KNIGHT, William, Bainbridge Island, WA 98110, US

(54) Bezeichnung: **IDENTIFIZIERUNG VON KRITISCHEN MERKMALEN IN EINEM GEORDNETEN SKALA-RAUM**

Beschreibung

TECHNISCHES GEBIET

[0001] Die vorliegende Erfindung bezieht sich im Allgemeinen auf eine Merkmalserkennung und Kategorisierung, und insbesondere auf ein System und Verfahren zum Identifizieren kritischer Merkmale in einem geordneten Skala-Raum innerhalb eines multidimensionalen Merkmal-Raums.

STAND DER TECHNIK

[0002] Beginnend mit Gutenberg in der Mitte des 15. Jahrhunderts hat sich das Volumen von bedruckten Materialien stetig mit einer explosionsartigen Geschwindigkeit erhöht. Heutzutage enthält sogar die Kongressbibliothek alleine über 18 Millionen Bücher und 54 Millionen Manuskripte. Ein wesentlicher Teil von bedrucktem Material ist ebenfalls in elektronischer Form verfügbar, größtenteils aufgrund der weit verbreiteten Einführung des Internets und des Personal-Computing.

[0003] Nichtsdestotrotz verbleibt eine wirksame Erkennung und Kategorisierung von bemerkenswerten Merkmalen innerhalb eines vorgegebenen Hauptteils von bedruckten Dokumenten eine gewaltige und komplexe Aufgabe, und zwar sogar dann, wenn durch eine Automatisierung unterstützt wird. Es existierten wirksame Suchstrategien lange Zeit für Datenbanken, Tabellenkalkulationen und ähnliche Formen von geordneten Daten. Der Hauptteil von bedruckten Dokumenten ist jedoch eine unstrukturierte Sammlung von einzelnen Wörtern, welche, auf einem semantischen Pegel, Ausdrücke und Konzepte bilden, es ihnen jedoch im Allgemeinen an einer regulären Ordnung oder Struktur mangelt. Ein Extrahieren oder „Fördern" von einer Bedeutung von unstrukturierten Dokumentensätzen erfordert konsequenterweise ein Ausnutzen der inhärenten oder „latenten" semantischen Struktur, welche Sätzen und Wörtern unterliegt.

[0004] Ein Erkennen und Kategorisieren eines Textes innerhalb von unstrukturierten Dokumentensätzen stellt Probleme dar, welche analog zu weiteren Formen einer Datenorganisation sind, welche eine latente Bedeutung haben, welche in der natürlichen Ordnung von einzelnen Merkmalen eingebettet ist. Beispielsweise bilden Genom- und Proteinsequenzen Muster aus, welche Datenförderungs-Verfahren zugänglich sind, und welche leicht bestimmt und analysiert werden können, um einzelne genetische Charakteristiken zu identifizieren. Jede Genom- und Proteinsequenz enthält eine Folge von Großbuchstaben und Ziffern, welche eindeutig einen genetischen Code für DNA Nukleotide und Aminosäuren identifizieren. Generische Marker, das heißt Gene oder weitere identifizierbare Abschnitte einer DNA, deren Vererbung gefolgt werden kann, treten natürlicherweise innerhalb eines vorgegebenen Genoms oder einer Proteinsequenz auf, und können zur Unterstützung einer Identifikation und Kategorisierung helfen.

[0005] Eine wirksame Verarbeitung eines Merkmal-Raums, welcher Ausdrücke und Konzepte enthält, welche aus einem unstrukturierten Text oder generischen Markern extrahiert sind, welche aus Genom- und Proteinsequenzen extrahiert sind, leiden beide unter dem Fluch der Dimension: Die Dimension des Problembereiches wächst proportional zur Größe des Körpers von einzelnen Merkmalen. Beispielsweise können Ausdrücke und Konzepte von einem unstrukturierten Dokumentensatz gefördert werden, und die Auftritts-Häufigkeit von einzelnen Ausdrücken und Konzepten kann leicht bestimmt werden. Jedoch steigt die Auftrittsläufigkeit linear mit jedem aufeinanderfolgenden Ausdruck und Konzept an. Das exponentiale Wachstum des Problembereiches erstellt eine Analyse schnell hartnäckig, obwohl ein Großteil des Problembereiches auf einem semantischen Pegel konzeptionell unwesentlich ist.

[0006] Die hohe Dimension des Problembereiches resultiert aus dem reichen Merkmal-Raum. Die Auftritts-Häufigkeit von jedem Merkmal über den gesamten Satz von Daten (Körper für Textdokumente) kann über statistische und ähnliche Mittel analysiert werden, um ein Muster einer semantischen Regelmäßigkeit zu bestimmen. Jedoch kann die bloße Anzahl von Merkmalen ein Identifizieren der relevantesten Merkmale durch redundante Werte und konzeptionell unwesentliche Merkmale übermäßig komplizieren.

[0007] Darüber hinaus versagen populärste Klassifikationstechniken im Allgemeinen beim Betreiben in einem Merkmal-Raum mit einer hohen Dimension. Beispielsweise arbeiten neurale Netzwerke, Bayesian-Klassifizierer und ähnliche Annäherungen am besten, wenn sie auf einer relativ kleinen Anzahl von Eingangswerten arbeiten. Diese Annäherungen versagen, wenn Hunderte oder Tausende von Eingangsmerkmalen verarbeitet werden. Neurale Netzwerke enthalten beispielsweise eine Eingangsschicht, eine oder mehrere Zwischenschichten und eine Ausgangsschicht. Durch geführtes Lernen werden die Gewichtungen, welche diese Schichten zwischenverbinden, modifiziert, indem sukzessive Eingangssätze und eine Fehlerverbreitung durch das Netzwerk angewendet werden. Ein Umtrainieren mit einem neuen Satz von Eingaben erfordert ferner ein Trainieren auf diese Art. Ein Merkmal-Raum mit hoher Dimension bewirkt, dass ein solches Umtrainieren viel Zeit verbraucht und undurchführbar ist.

[0008] Ein Abbilden eines Merkmal-Raumes mit hoher Dimension auf niedrigere Dimensionen ist ebenfalls schwierig. Eine Annäherung der Abbildung ist in unserer U.S-Patentanmeldung Serial No.

09/943,918, eingereicht am 31. August 2001, beschrieben. Diese Annäherung verwendet statistische Verfahren, um es einem Benutzer zu ermöglichen, relevante Merkmale, welche in Gruppen zur Anzeige in einem zweidimensionalen Konzeptraum ausgebildet sind, zu modellieren und auszuwählen. Jedoch werden logisch bezogene Konzepte nicht geordnet, und konzeptionell unwesentliche und redundante Merkmale innerhalb eines Konzeptraumes werden in einer Projektion mit einer niedrigeren Dimension beibehalten.

[0009] Eine anverwandte Annäherung zum Analysieren eines unstrukturierten Textes ist beschrieben in N.E. Miller et al., „Topic Islands: A Wavelet-Based Text Visualization System" IEEE Visualization Proc., 1998. Das Textvisualisierungssystem analysiert automatisch einen Text, um Unterbrechungen in einem erzählerischen Fluss zu lokalisieren. Es werden Wavelets dazu verwendet, um es zu ermöglichen, dass der erzählerische Fluss in eindeutige Kanäle konzeptionalisiert wird. Jedoch beschreiben die Kanäle nicht einzelne Merkmale und verdauen nicht einen gesamten Hauptteil von mehreren Dokumenten.

[0010] Ähnlich sind eine Vielzahl von Dokumenten-Einlagerungs- und Textförderungs-Techniken in D. Sullivan, „Document Warehousing and Text Mining-Techniques for Improving Business Operations, Marketing, and Sales" Parts 2 and 3, John Wiley & Sons (Februar 2001) beschrieben. Jedoch sind die Annäherungen ohne eine Fokussierung der Identifizierung eines Merkmal-Raums innerhalb eines größeren Körpers oder der Neuordnung von Merkmalsvektoren einer hohen Dimension, um eine latente semantische Bedeutung zu extrahieren, beschrieben.

[0011] Das U.S.-Patent No. 6,070,133, Brewster et al., offenbart ein Informationserlangungssystem unter Verwendung von einer Wavelet-Transformation. Das System führt eine spektrale Analyse auf eine Wellenform oder ein digitales Signal durch, um eine Dokumenten-Charakterisierung bereitzustellen, wobei das digitale Signal eine numerische Darstellung der Wörter ist, welche innerhalb des Dokuments enthalten sind. Es wird eine spektrale Analyse durchgeführt, um das digitale Signal zu verstärken und ein Rauschen zu reduzieren, um es einem Benutzer zu erlauben, eine visuelle Darstellung von dem semantischen Aufbau zu erzeugen, welcher die Reihenfolge ist, in welcher die Themen in dem Dokument erzählerisch diskutiert werden. Die spektrale Analyse wird bereitgestellt, indem eine Wavelet-Transformation auf das digitale Signal durchgeführt wird, und die Ausgabe von der Wavelet-Transformation kann dazu verwendet werden, um eine visuelle Darstellung des semantischen Aufbaus zu erzeugen, welche eine textbasierte, grafische oder zusammengefasste Darstellung sein kann. Die Ausgabe von der Wavelet-Transformation kann ebenfalls dazu verwendet werden, um das Dokument gemäß einem semantischen Inhalt auf einem einzelnen Pegel oder auf mehreren Pegeln zu partitionieren, um einen Umriss oder einen „Fuzzy"-Umriss von dem Dokument zu erzeugen. Die numerische Darstellung der Wörter innerhalb des Dokuments kann von Worthäufigkeits-Zählungen, Funktionen von Worthäufigkeits-Zählungen und statistischen Korrelationen von Wörtern innerhalb des gesamten Dokuments oder Gruppen von Wörtern oder Teilmengen von Wörtern in einem Dokument hergeleitet werden. Das digitale Signal behält die Wort-Reihenfolge bei, welche in der Erzählung gefunden wird.

[0012] M. Slaney et al., „Multimedia Edges: Finding Hierarchy in all Dimensions" Proc. 9th ACM International Conference an Multimedia, Seiten 29-40 (30. September 2001), offenbart Techniken zum Analysieren der temporalen Eigenschaften von Audio- und Bilddaten in einem Video, um eine hierarchische Segmentierung des Videos oder eine Inhaltstabelle von dem Audio, eine Semantik und Bilddaten zu erzeugen. Änderungen in dem Video oder des semantischen Inhaltes von dem Video werden als eine Funktion der Zeit erfasst. Bilddaten werden über eine Schwenkerfassung analysiert und mit einer Information von dem Audiosignal zusammengefasst, um Änderungen in dem Inhalt oder dem Ton zu finden, welche Strukturen mit höherem Pegel innerhalb des Videos anzeigen. Techniken, wie beispielsweise eine latente Semantik-Indexierung, werden dazu verwendet, um anverwandte Dokumente zu gruppieren oder um ein Dokument zu finden, welches einer Warteschlange am nächsten gleicht, welches dazu verwendet werden kann, um den semantischen Pfad von einer Kopie eines Videos als ein Signal in einem Skala-Raum zu beschreiben. Das Signal kann analysiert werden, um semantische Unterbrechungen in dem Audio zu finden und um eine semantische Inhaltstabelle für das Video zu erzeugen. Jeglicher Zeitpunkt in dem Video kann durch eine Position in einem Akustik-Farb-Semantik-Vektor-Raum beschrieben werden.

[0013] Die U.S.-Patentanmeldung Veröffentlichung No. US 2002/0016798, von Sakai et al., offenbart eine Textinformations-Analyseeinrichtung und ein Verfahren, welches eine Mehrzahl von Text gemäß einem Inhalt anordnet. Eine Kategorie-Positionseinheit ist dazu konfiguriert, um einen Text in eine von einer Mehrzahl von vorbestimmten Kategorien zu klassifizieren. Eine Gruppenerzeugungseinheit ist dazu konfiguriert, um Texte zu gruppieren, welche ähnliche Inhalte haben. Eine Steuereinheit ist dazu konfiguriert, um die Kategorieentscheidungs- und Gruppenerzeugungseinheiten dazu zu steuern, um gleichzeitig eine Kategorieentscheidung und Gruppierung auszuführen. Eine morphologische Analyse wird für jeden Text ausgeführt, und jegliche Wörter, welche enthalten sind, werden vor der Entscheidung identifi-

ziert, ob der Text gemäß einer Kategorieentscheidungsregel klassifiziert ist. Jedoch überspannen die Kategorien einen Skala-Raum und es mangelt ihnen daran, mehrere Detailpegel bereitzustellen, M. Kurimo, „Fast Latent Semantic Indexing of Spoken Documents by using Self-Organizing Maps", IEEE International Conference an Acustics, Speech and Signal Proc., Vol. 5, Seiten 2425-2428 (5. Juni 2000), offenbart ein latentes Semantik-Indexierungsverfahren für gesprochene Audiodokumente. Dokumente werden als Vektoren von Wortzählungen dargestellt, deren Dimension schnell durch eine willkürliche Abbildung reduziert wird, und welche in einen latenten semantischen Teilraum projiziert werden. Die Vektoren werden durch eine selbstorganisierende Abbildung geglättet, welches einen einfachen Weg bereitstellt, um Index- und Warteschlangen-Ergebnisse zu visualisieren und die Datenbank zu untersuchen. Jedoch stellen die selbstorganisierenden Abbildungen nicht-lineare Daten einer hohen Dimension in einer Anzeige mit einer niedrigen Dimension dar, und es mangelt ihnen daran, mehrere Pegel von Detail und Merkmalen bereitzustellen.

[0014] Es gibt daher eine Notwendigkeit nach einer Annäherung zum Bereitstellen eines geordneten Satzes von extrahierten Merkmalen, welche von einem multidimensionalen Problembereich bestimmt sind, welcher Textdokumente und Genom- und Proteinsequenzen enthält. Vorzugsweise wird eine solche Annäherung kritische Merkmalsräume isolieren, während Null-Wert, konzeptionell unwesentliche und redundante Merkmale innerhalb des Konzeptraumes herausgefiltert werden.

[0015] Es gibt eine weitere Notwendigkeit nach einer Annäherung, welche den Merkmal-Raum in einen geordneten Skala-Raum transformiert. Vorzugsweise wird eine solche Annäherung einen skalierbaren Merkmal-Raum bereitstellen, welcher dazu in der Lage ist, in variierenden Detailpegeln über eine Mehrfachauflösungs-Analyse zu abstrahieren.

BESCHREIBUNG DER ERFINDUNG

[0016] Die vorliegende Erfindung stellt ein System und ein Verfahren gemäß der folgenden Ansprüche zum Transformieren eines multidimensionalen Merkmal-Raumes in eine geordnete und priorisierte Skala-Raum-Darstellung bereit. Der Skala-Raum wird im Allgemeinen im Hilbert-Funktion-Raum bestimmt. Eine Mehrheit von einzelnen Merkmalen wird von einer Vielzahl von diskreten Datensammlungen extrahiert. Jedes einzelne Merkmal stellt einen latenten Inhalt dar, welcher der semantischen Struktur von der Datensammlung inhärent ist. Die Merkmale werden in einem Satz von Mustern auf einer Datensammlungsbasis organisiert. Jedes Muster wird auf Ähnlichkeiten analysiert und nahe anverwandte Merkmale werden zu einzelnen Gruppen gruppiert. In der beschriebenen Ausführungsform werden die Ähnlichkeits-Messungen anhand einer Distanz-Metrik erzeugt. Die Gruppen werden dann in einem geordneten Skala-Raum projiziert, in welchem die einzelnen Merkmalsvektoren nacheinander folgend als Wavelet- und Skalierungs-Koeffizienten unter Verwendung einer Mehrfachauflösungs-Analyse encodiert werden. Die geordneten Vektoren enthalten ein „Semantik"-Signal, welches Signalverarbeitungstechniken, wie beispielsweise eine Komprimierung, zugänglich ist.

[0017] Eine Ausführungsform stellt ein System und ein Verfahren zum Identifizieren kritischer Merkmale in einem geordneten Skala-Raum innerhalb eines multidimensionalen Merkmal-Raumes bereit. Merkmale werden von einer Mehrzahl von Datensammlungen extrahiert. Jede Datensammlung wird durch eine Sammlung von Merkmalen, welche semantisch durch eine Grammatik bezogen sind, charakterisiert. Jedes Merkmal wird dann normalisiert und es werden Auftritts-Häufigkeiten und Nebenauftritte für die Merkmale für jede der Datensammlungen bestimmt. Die Auftritts-Häufigkeiten und die Nebenauftritts-Häufigkeiten für jedes der extrahierten Merkmale werden in einem Satz von Auftritts-Häufigkeitsmustern und einem Satz von Nebenauftritts-Häufigkeitsmustern abgebildet. Das Muster für jede Datensammlung wird ausgewählt, und es werden Ähnlichkeits-Messungen zwischen jeder Auftritts-Häufigkeit in dem ausgewählten Muster berechnet. Die Auftritts-Häufigkeiten werden auf ein eindimensionales Dokumentsignal projiziert, um eine Ähnlichkeit unter Verwendung der Ähnlichkeits-Messungen relativ zu verringern. Fälle von Merkmalsvektoren einer hohen Dimension können dann als ein eindimensionaler Signalvektor behandelt werden. Wavelet- und Skalierungs-Koeffizienten werden aus dem eindimensionalen Dokumentsignal hergeleitet.

[0018] Eine weitere Ausführungsform stellt ein System und Verfahren zum Abstrahieren von semantisch latenten Konzepten, welche von einer Mehrzahl von Dokumenten extrahiert sind, bereit. Ausdrücke und Phrasen werden von einer Mehrzahl von Dokumenten extrahiert. Jedes Dokument enthält eine Sammlung von Ausdrücken, Phrasen und nicht beweiskräftigen Wörtern. Die Ausdrücke und Phrasen werden zu Konzepten bestimmt und auf eine Einzelstamm-Wortform reduziert. Eine Auftritts-Häufigkeit wird für jedes Konzept angesammelt. Die Auftritts-Häufigkeiten für jedes der Konzepte werden in einem Satz von Mustern von Auftritts-Häufigkeiten abgebildet, und zwar eines solchen Musters pro Dokument, angeordnet in einer zweidimensionalen Dokument-Merkmals-Matrix. Jedes Muster wird iterativ von der Dokumentenmerkmals-Matrix für jedes Dokument ausgewählt. Ähnlichkeits-Messungen zwischen jedem Muster werden berechnet. Die Auftrittswahrscheinlichkeiten, beginnend von einem im We-

sentlichen maximalen Ähnlichkeitswert, werden in ein eindimensionales Signal in einer skalierbaren Vektorform transformiert, und zwar geordnet in einer Sequenz von einer relativ abnehmenden Ähnlichkeit. Wavelet- und Skalierungs-Koeffizienten werden von dem eindimensionalen Skala-Signal hergeleitet.

[0019] Eine weitere Ausführungsform stellt ein System und ein Verfahren zum Abstrahieren semantisch latenter genetischer Teilsequenzen bereit, welche von einer Mehrzahl von genetischen Sequenzen extrahiert werden. Generische Teilsequenzen werden von einer Mehrzahl von genetischen Sequenzen extrahiert. Jede genetische Sequenz enthält eine Sammlung von zumindest einem von genetischen Codes für DNA Nukleotide und Aminosäuren. Eine Auftritts-Häufigkeit für jede genetische Teilsequenz wird für jede der genetischen Sequenzen, von welchen die genetischen Teilsequenzen hergeleitet sind, angesammelt. Die Auftritts-Häufigkeiten für jede der genetischen Teilsequenzen werden in einem Satz von Mustern von Auftritts-Häufigkeiten abgebildet, und zwar eines von einem solchen Muster pro genetischer Sequenz, angeordnet in einer zweidimensionalen genetischen Teilsequenz-Matrix. Jedes Muster wird iterativ aus der genetischen Teilsequenz-Matrix für jede genetische Sequenz ausgewählt. Ähnlichkeits-Messungen zwischen jeder Auftritts-Häufigkeit in jedem ausgewählten Muster werden berechnet. Die Auftritts-Häufigkeiten, beginnend von einer im Wesentlichen maximalen Ähnlichkeits-Messung, werden auf ein eindimensionales Signal in skalierbarer Vektorform, geordnet in einer Sequenz von einer relativ abnehmenden Ähnlichkeit, projiziert. Wavelet- und Skalierungskoeffizienten werden aus dem eindimensionalen Skala-Signal hergeleitet.

[0020] Noch weitere Ausführungsformen von der vorliegenden Erfindung werden dem Fachmann anhand der folgenden detaillierten Beschreibung leicht deutlich, wobei in ihr Ausführungsformen von der Erfindung mittels Darstellung des besten Modus, welcher zur Ausführung der Erfindung in Erwägung gezogen ist, beschrieben sind. Wie erkannt werden wird, ist die Erfindung zu weiteren und unterschiedlichen Ausführungsformen fähig, und ihre mehreren Details sind zur Modifikation auf verschiedene offensichtliche Hinsichten fähig, wobei sie alle nicht vom Umfang der vorliegenden Erfindung abweichen. Demgemäß sind die Zeichnungen und ist die detaillierte Beschreibung als natürlich darstellhaft und nicht als beschränkend anzusehen.

KURZE BESCHREIBUNG DER ZEICHNUNGEN

[0021] Fig. 1 ist ein Blockdiagramm, welches ein System zum Identifizieren von kritischen Merkmalen in einem geordneten Skala-Raum innerhalb eines multidimensionalen Merkmal-Raumes gemäß der vorliegenden Erfindung zeigt.

[0022] Fig. 2 ist ein Blockdiagramm, welches mittels Beispiel ein Satz von Dokumenten zeigt.

[0023] Fig. 3 ist ein Venn-Diagramm, welches mittels Beispiel die Merkmale, welche von dem Dokumentensatz von Fig. 2 extrahiert sind, zeigt.

[0024] Fig. 4 ist ein Datenstruktur-Diagramm, welches mittels Beispiel Projektionen von Merkmalen, welche von dem Dokumentensatz von Fig. 2 extrahiert sind, zeigt.

[0025] Fig. 5 ist ein Blockdiagramm, welches die Softwaremodule zeigt, welche den Datensammlungs-Analysator von Fig. 1 implementieren.

[0026] Fig. 6 ist ein Prozess-Ablaufdiagramm, welches die Stufen der Merkmalsanalyse zeigt, welche durch den Datensammlungs-Analysator von Fig. 1 durchgeführt werden.

[0027] Fig. 7 ist ein Ablaufdiagramm, welches ein Verfahren zum Identifizieren von kritischen Merkmalen in einem geordneten Skala-Raum innerhalb eines multidimensionalen Merkmal-Raums gemäß der vorliegenden Erfindung zeigt.

[0028] Fig. 8 ist ein Ablaufdiagramm, welches die Routine zur Durchführung einer Merkmalsanalyse zur Verwendung in dem Verfahren von Fig. 7 zeigt.

[0029] Fig. 9 ist ein Ablaufdiagramm, welches die Routine zur Bestimmung einer Konzepthäufigkeit zur Verwendung in der Routine von Fig. 8 zeigt.

[0030] Fig. 10 ist ein Datenstrukturdiagramm, welches eine Datenbankaufzeichnung für ein Merkmal, welches in der Datenbank von Fig. 1 gespeichert ist, zeigt.

[0031] Fig. 11 ist ein Datenstrukturdiagramm, welches mittels Beispiel eine Datenbanktabelle zeigt, welche ein Lexikon von extrahierten Merkmalen enthält, welche in der Datenbank von Fig. 1 gespeichert sind.

[0032] Fig. 12 ist ein Kurvenverlauf, welcher mittels Beispiel ein Histogramm der Häufigkeiten von Merkmalsauftritten zeigt, welches durch die Routine von Fig. 9 erzeugt ist.

[0033] Fig. 13 ist ein Kurvenverlauf, welcher mittels Beispiel eine Zunahme in einer Anzahl von Merkmalen, bezogen auf eine Anzahl von Datensammlungen, zeigt.

[0034] Fig. 14 ist eine Tabelle, welche mittels Beispiel eine Matrix-Abbildung von Merkmals-Häufigkeiten, welche durch die Routine von Fig. 9 erzeugt ist, zeigt.

[0035] Fig. 15 ist ein Kurvenverlauf, welcher mittels Beispiel einen Korpus-Kurvenverlauf der Häufigkeit von Merkmalsauftritten, welche durch die Routine von Fig. 9 erzeugt ist, zeigt.

[0036] Fig. 16 ist ein Ablaufdiagramm, welches eine Routine zum Transformieren eines Problem-Raumes in einem Skala-Raum zur Verwendung in der Routine von Fig. 8 zeigt.

[0037] Fig. 17 ist ein Ablaufdiagramm, welches die Routine zum Erzeugen von Ähnlichkeits-Messungen und Ausbilden von Gruppen zur Verwendung in der Routine von Fig. 16 zeigt.

[0038] Fig. 18 ist eine Tabelle, welche mittels Beispiel die Merkmalsgruppen zeigt, welche durch die Routine von Fig. 17 erzeugt sind.

[0039] Fig. 19 ist ein Ablaufdiagramm, welches eine Routine zum Identifizieren von kritischen Merkmalen zur Verwendung in dem Verfahren von Fig. 7 zeigt.

MODUS bzw. MODI ZUR DURCHFÜHRUNG DER ERFINDUNG

Wortliste

Dokument:	Eine Basiskollektion von Daten, welche zur Analyse als ein Datensatz verwendet werden.
Vorgang:	Eine Basiskollektion von Daten, welche zur Analyse als ein Datensatz verwendet werden. In der beschriebenen Ausführungsform ist ein Vorgang im Allgemeinen äquivalent zu einem Dokument.
Dokumentvektor:	Ein Satz von Merkmalswerten, welche ein Dokument beschreiben.
Dokumentsignal:	Äquivalent zu einem Dokumentvektor.
Skala-Raum:	Im Allgemeinen als ein Hilbert-Funktion-Raum H bezeichnet.
Schlüsselwort:	Ein wortgetreuer Suchausdruck, welcher in einem Dokument oder einer Datensammlung entweder vorliegt oder nicht vorliegt. Schlüsselwörter werden nicht in der Auswertung von Dokumenten und Datensammlungen, wie hier beschrieben, verwendet.
Ausdruck:	Ein Wurzelstamm von einem einzelnen Wort, welches in dem Hauptteil von zumindest einem Dokument oder einer Datensammlung erscheint. Analog ein generischer Marker in einer Genom- oder Proteinsequenz.
Phrase:	Zwei oder mehrere Wörter, welche in dem Hauptteil von einem Dokument oder einer Datensammlung nebeneinander auftreten. Eine Phrase kann Stopp-Wörter enthalten.
Merkmal:	Eine Sammlung von Ausdrücken oder Phrasen mit gemeinsamen semantischen Bedeutungen, ebenfalls bezeichnet als Konzept.
Thema:	Zwei oder mehrere Merkmale mit einer gemeinsamen semantischen Bedeutung.
Gruppe:	Alle Dokumente oder Datensammlungen, welche bei einer vorbestimmten Messung von einer Ähnlichkeit fehlschlagen.
Korpus:	Alle Textdokumente, welche den gesamten Rohdatensatz bestimmen.

[0040] Die vorhergehenden Ausdrücke werden über dieses Dokument hinweg verwendet, und werden, es sei denn anders angezeigt, den oben dargelegten Bedeutungen angehängt. Ferner, obwohl mit Bezug auf eine Dokumentenanalyse beschrieben, werden die Ausdrücke analog an weitere Formen von strukturierten Daten angewendet, welche Genom- und Proteinsequenzen enthalten, und ähnlichen Datensammlungen, welche ein Vokabular, eine Grammatik und atomare Dateneinheiten haben, wie durch den Fachmann anerkannt werden wird.

[0041] Fig. 1 ist ein Blockdiagramm, welches ein System **11** zum Identifizieren von kritischen Merkmalen in einem geordneten Skala-Raum innerhalb eines multidimensionalen Merkmal-Raums gemäß der vorliegenden Erfindung zeigt. Der Skala-Raum ist ebenfalls als ein Hilbert-Funktion-Raum bekannt. Mittels einer Darstellung arbeitet das System **11** in einer verteilten Berechnungsumgebung **10**, welche eine Mehrzahl von heterogenen Systemen und Datensammelquellen enthält. Das System **11** implementiert einen Datensammel-Analysator **12**, wie im Folgenden, beginnend mit Bezug auf Fig. 4, ferner beschrieben, zur Auswertung von latenten semantischen Merkmalen in unstrukturierten Datensammlungen. Das System **11** ist an einer Speichervorrichtung **13** gekoppelt, welche eine Datensammlungs-Quelle **14** zum Archivieren der Datensammlungen und eine Datenbank **30** zum Beibehalten von einer Datensammlungs-Merkmalsinformation speichert.

[0042] Der Dokumenten-Analysator **12** analysiert Datensammlungen, welche von einer Mehrzahl von lokalen Quellen erlangt werden. Die lokalen Quellen enthalten Datensammlungen **17**, welche in einer Speichervorrichtung **16** beibehalten werden, welche an einen lokalen Server **15** gekoppelt ist, und Datensammlungen **20**, welche in einer Speichervorrichtung **19** beibehalten werden, welche an einem lokalen Client **18** gekoppelt ist. Der lokale Server **15** und der lokale Client **18** sind über ein Intranetzwerk **21** zum System **11** zwischenverbunden. Zusätzlich kann der Datensammlungs-Analysator **12** Datensammlungen von entfernten Quellen über ein Internetzwerk **22**, welches das Internet enthält, über ein Gateway **23**, welches mit dem Intranetzwerk **21** eine Schnittstelle bildet, identifizieren und erlangen. Die entfernten Quellen enthalten Datensammlungen **26**, welche in einer Speichervorrichtung **25** beibehalten werden, welche an einen entfernten Server **24** gekoppelt ist, und Datensammlung **29**, welche in einer Speichervorrichtung **28** beibehalten werden, welche an einen entfernten Client **27** gekoppelt ist.

[0043] Die einzelnen Datensammlungen **17**, **20**, **26**, **29** bilden jeweils eine semantisch bezogene Sammlung von gespeicherten Daten, welche alle Formen und Typen von unstrukturierten und halb strukturierten (textbasierten) Daten enthalten, welche elektronische Meldungsvorräte, wie beispielsweise elektronische Mail (E-Mail)-Ordner, Textverarbeitungsdokumente oder Hypertext-Dokumente, enthalten, und ebenfalls Grafik- oder Multimediadaten enthalten können. Die unstrukturierten Daten enthalten ebenfalls Genom- und Proteinsequenzen und ähnliche Datensammlungen. Die Datensammlungen enthalten eine bestimmte Art eines Vokabulars, mit welchem atomare Dateneinheiten bestimmt werden, und Merkmale semantisch durch eine Grammatik bezogen werden, wie der Fachmann anerkennen wird. Eine atomare Dateneinheit ist analog zu einem Merkmal und enthält eine oder mehrere suchbare Charakteristiken, welche, wenn einzeln oder in Kombination genommen, eine Gruppierung darstellen, welche eine allgemeine semantische Bedeutung hat. Die Grammatik erlaubt es, dass die Merkmale syntaktisch und semantisch zusammengefasst werden, und ermöglicht die Entdeckung von latent-semantischen Bedeutungen. Die Dokumente können ebenfalls in der Form von strukturierten Daten sein, wie beispielsweise in einer Tabellenkalkulation oder einer Datenbank gespeichert. Ein Inhalt, welcher von diesen Typen von Dokumenten gefördert wird, wird keine Vorverarbeitung erfordern, wie im Folgenden beschrieben.

[0044] In der beschriebenen Ausführungsform enthalten die einzelnen Datensammlungen **17**, **20**, **26**, **29** elektronische Meldungsordner, wie beispielsweise durch die Outlook- und Outlook Express-Produkte beibehalten, lizenziert durch Microsoft Corporation, Redmond, Washington. Die Datenbank ist eine SQL-basierte relationale Datenbank, wie beispielsweise das Oracle Datenbank Verwaltungssystem, Ausgabe **8**, lizenziert durch die Oracle Corporation, Redwood Shores, Kalifornien.

[0045] Die einzelnen Computersysteme, welche das System **11**, den Server **15**, den Client **18**, den entfernten Server **24** und den entfernten Client **27** enthalten, sind Vielzweckprogrammierte digitale Berechnungsvorrichtungen, welche eine zentrale Verarbeitungseinheit (CPU), einen Arbeitsspeicher (RAM), einen nicht-flüchtigen sekundären Speicher, wie beispielsweise eine Festplatte oder ein CD-ROM-Laufwerk, Netzwerk- und Drahtlos-Schnittstellen und Peripherie-Vorrichtungen enthalten, welche ein Benutzer-Schnittstellenmittel enthalten, wie beispielsweise eine Tastatur und eine Anzeige. Ein Programmcode, welcher Softwareprogramme enthält, und Daten werden in den RAM zur Ausführung und Verarbeitung durch die CPU geladen und Ergebnisse werden zur Anzeige, Ausgabe, Übertragung oder Speicherung erzeugt.

[0046] Der vollständige Satz von Merkmalen, welche von einem vorgegebenen Dokument oder einer Datensammlung extrahierbar sind, kann in einem logischen Merkmal-Raum, ebenfalls als ein Hilbert-Funktion-Raum H bezeichnet, modelliert werden. Die einzelnen Merkmale bilden einen Merkmalssatz, aus welchem Themen extrahiert werden können. Aus Gründen der Darstellung ist Fig. 2 ein Blockdiagramm, welches mittels Beispiel einen Satz **40** von Dokumenten **41**-**46** zeigt. Jedes einzelne Dokument **41**-**46** enthält eine Datensammlung, welche einzelne Ausdrücke enthält. Beispielsweise enthalten Dokumente **42**, **44**, **45** und **46** jeweils „Mäuse", „Mäuse", „Maus" und „Mäuse", wobei der Wurzelstamm davon gleich „Maus" ist. Ähnlich enthalten Dokumente **42** und **43** beide „Katze"; Dokumente **43** und **46**

enthalten jeweils „Mann" und „Männer", der Wurzelstamm davon ist „Mann"; und Dokument **43** enthält „Hund". Jeder Satz von Ausdrücken enthält ein Merkmal. Dokumente **42**, **44**, **45** und **46** enthalten den Ausdruck „Maus" als ein Merkmal. Ähnlich enthalten Dokumente **42** und **43** den Ausdruck „Katze", Dokumente **43** und **46** enthalten den Ausdruck „Mann" und Dokument **43** enthält den Ausdruck „Hund" als ein Merkmal. Somit bilden die Merkmale „Maus", „Katze", „Mann" und „Hund" den Korpus von dem Dokumentensatz **40**.

[0047] Fig. 3 ist ein Venn-Diagramm **50**, welches mittels Beispiel die Merkmale **51**-**54** zeigt, welche von dem Dokumentsatz **40** von Fig. 2 extrahiert sind. Das Merkmal „Maus" tritt vier Mal in den Dokumentensatz **40** auf. Ähnlich treten die Merkmale „Katze", „Mann" und „Hund" jeweils zwei Mal, zwei Mal und ein Mal auf. Ferner treten die Merkmale „Maus" und „Katze" konsistent zusammen in dem Dokumentensatz **40** nebeneinander auf und bilden ein Thema „Maus und Katze". „Maus" und „Mann" treten ebenfalls nebeneinander auf, um ein zweites Thema „Maus und Mann" auszubilden. „Mann" und „Hund" treten nebeneinander auf, um ein drittes Thema „Mann und Hund" auszubilden. Das Venn-Diagramm stellt diagrammartig die Zwischenbeziehungen von den thematischen Nebenauftritten in zwei Dimensionen dar und spiegelt wider, dass „Maus und Katze" das stärkste Thema in dem Dokumentensatz **40** ist.

[0048] Venn-Diagramme sind zweidimensionale Darstellungen, welche lediglich eine thematische Überlappung entlang einer einzelnen Dimension abbilden können. Wie im Folgenden, beginnend mit Bezug auf Fig. 19, ferner beschrieben, können die einzelnen Merkmale genauer als Gruppen in einem multidimensionalen Merkmal-Raum modelliert werden. Wiederum können die Gruppen auf geordnete und priorisierte eindimensionale Merkmalsvektoren oder Projektionen projiziert werden, modelliert im Hilbert-Funktion-Raum H, welcher die relativen Stärken von den Zwischenbeziehungen zwischen den jeweiligen Merkmalen und Themen widerspiegelt. Die geordneten Merkmalsvektoren bilden ein „semantisches" Signal, welches für Signalverarbeitungstechniken, wie beispielsweise eine Quantisierung und Encodierung, zugänglich ist.

[0049] Fig. 4 ist ein Datenstrukturdiagramm, welches mittels Beispiel Projektionen **60** von den Merkmalen zeigt, welche von dem Dokumentensatz **40** von Fig. 2 extrahiert sind. Die Projektionen **60** sind in vier Detailpegel **61**-**64** im Skala-Raum gezeigt. In dem höchsten oder detailliertesten Pegel **61** sind alle anverwandten Merkmale in Reihenfolge einer abnehmenden Zwischenbeziehung beschrieben. Beispielsweise ist das Merkmal „Maus" am meisten dem Merkmal „Katze" als den Merkmalen „Mann" und „Hund" anverwandt. Ähnlich ist das Merkmal „Maus" ebenfalls am meisten dem Merkmal „Mann" als dem Merkmal „Hund" anverwandt. Das Merkmal „Hund" ist das am wenigsten anverwandte Merkmal.

[0050] Auf dem zweithöchsten Detailpegel **62** ist das Merkmal „Hund" ausgelassen. Ähnlich sind in dem dritten und vierten Detailpegel **63**, **64** die Merkmale „Mann" und „Katze" jeweils ausgelassen. Der vierte Detailpegel **64** spiegelt das relevanteste Merkmal wider, welches in dem Dokumentensatz **40** vorliegt, nämlich „Maus", welches vier Mal auftritt und daher den Korpus auf einem minimalen Pegel abstrahiert.

[0051] Fig. 5 ist ein Blockdiagramm, welches die Softwaremodule **70** zeigt, welche den Datensammlungs-Analysator **12** von Fig. 1 implementieren. Der Datensammlungs-Analysator **12** enthält sechs Module: einen Speicher- und Erlangungsverwalter **71**, einen Merkmals-Analysator **72**, einen unüberwachten Klassifizierer **73**, eine Skala-Raum-Transformation **74**, einen Kritisches-Merkmal-Identifizierer **75** und eine Anzeige und Visualisierung **82**. Der Speicher- und Erlangungsverwalter **71** identifiziert und erlangt Datensammlungen **76** in der Datenquelle **14**. Die Datensammlungen **76** werden von verschiedenen Quellen erlangt, welche lokale und entfernte Clients- und Server-Lager enthalten. Der Merkmals-Analysator **72** führt den Hauptteil der Merkmalsförderverarbeitung durch. Der unüberwachte Klassifizierer **73** verarbeitet Muster von häufigen Auftritten, welche im Merkmal-Raum in neu geordneten Vektoren ausgedrückt werden, welche im Skala-Raum ausgedrückt werden. Die Skala-Raum-Transformation **74** abstrahiert die Skala-Raum-Vektoren in variierende Detailpegel mit beispielsweise Wavelet- und Skalierungs-Koeffizienten über eine Mehrfachauflösungs-Analyse. Die Anzeige und Visualisierung **82** vervollständigt die Betriebe, welche durch den Merkmals-Analysator **72**, unüberwachten Klassifizierer **73**, die Skala-Raum-Transformation **74** und den Kritisches-Merkmal-Identifizierer **75** durchgeführt sind, indem visuelle Darstellungen von der Information dargestellt werden, welche von den Datensammlungen **76** extrahiert ist. Die Anzeige und Visualisierung **82** kann ebenfalls eine grafische Darstellung von den vermischten und verarbeiteten Merkmalen erzeugen, welches unabhängig variable Beziehungen erhält, wie in der gemeinsam zugewiesenen U.S.-Patentanmeldung Serial No. 09/944,475, eingereicht am 31. August 2001, beschrieben.

[0052] Während der Textanalyse identifiziert der Merkmals-Analysator **72** Ausdrücke und Phrasen und extrahiert Merkmale in der Form von Nominalphrasen, Genom- oder Protein-Markern oder ähnlichen atomaren Dateneinheiten, welche dann in einem Lexikon **77** gespeichert werden, welches in der Datenbank **30** beibehalten ist. Nach dem Normalisieren der extrahierten Merkmale erzeugt der Merk-

mals-Analysator **72** eine Merkmals-Häufigkeits-Tabelle **78** von Merkmalsauftritten im Dokument und eine geordnete Merkmals-Häufigkeits-Abbildungsmatrix **79**, wie im Folgenden mit Bezug auf Fig. 14 ferner beschrieben. Die Merkmals-Häufigkeits-Tabelle **78** bildet die Auftritte von Merkmalen auf einer Pro-Dokument-Basis ab, und die geordnete Merkmals-Häufigkeits-Abbildungsmatrix **79** bildet die Auftritte von allen Merkmalen über den gesamten Korpus oder die Datensammlung ab.

[0053] Der unüberwachte Klassifizierer **73** erzeugt logische Gruppen **80** der extrahierten Merkmale in einem multidimensionalen Merkmal-Raum zur Modellierung einer semantischen Bedeutung. Jede Gruppe **80** gruppiert semantisch bezogene Themen, basierend auf relativen Ähnlichkeits-Messungen, beispielsweise in Ausdrücken von einer ausgewählten L^2-Distanzmetrik.

[0054] In der beschriebenen Ausführungsform sind die L^2-Distanzmetriken im L^2-Funktionsraum bestimmt, welcher der Raum von absoluten quadratischen integrierbaren Funktionen ist, wie beispielsweise beschrieben in B.B. Hubbard, „The World According to Wavelets, the Story of a Mathematical Technique in the Making", Seiten 227-229, A.K. Peters (2d ed. 1998). Die L^2-Distanzmetrik ist äquivalent der euklidischen Distanz zwischen zwei Vektoren. Weitere Distanz-Messungen enthalten eine Korrelation, Richtungs-Kosinusse, Minkowski-Metriken, Tanimoto-Ähnlichkeits-Messungen, Mahanobis-Distanzen, Hamming-Distanzen, Levenshtein-Distanzen, Maximalwahrscheinlichkeits-Distanzen und ähnliche Distanzmetriken, wie sie im Stand der Technik bekannt sind, wie beispielsweise beschrieben in T. Kohonen, „Self Organizing Maps", Kapitel 1.2, Springer-Verlag (3. Ausgabe 2001).

[0055] Die Skala-Raum-Transformation **74** bildet Projektionen **81** von den Gruppen **80** in einem eindimensional geordneten und priorisierten Skala-Raum aus. Die Projektionen **81** werden unter Verwendung von Wavelet- und Skalierungs-Koeffizienten (nicht gezeigt) ausgebildet. Der Kritisches-Merkmal-Identifizierer **75** erlangt Wavelet- und Skalierungs-Koeffizienten von dem eindimensionalen Dokumentsignal. Schließlich erzeugt die Anzeige und Visualisierung **82** ein Histogramm **83** von Merkmalsauftritten pro Dokument oder Datensammlung, wie im Folgenden mit Bezug auf Fig. 13 ferner beschrieben, und einen Korpus-Kurvenverlauf **84** von Merkmalsauftritten über alle Datensammlungen, wie im Folgenden mit Bezug auf Fig. 15 weiter beschrieben.

[0056] Jedes Modul ist ein Computerprogramm, eine Prozedur oder ein Modul, welches als Quellencode in einer herkömmlichen Programmiersprache, wie beispielsweise die C++ Programmiersprache, geschrieben ist, und wird zur Ausführung durch die CPU als Objekt oder Byte-Code, wie im Stand der Technik bekannt, dargelegt. Die verschiedenen Implementierungen des Quellencodes und der Objekt- und Byte-Codes können auf einem computerlesbaren Speichermedium gehalten werden oder auf einem Übertragungsmedium in einer Trägerwelle ausgeführt sein. Der Datensammlungs-Analysator **12** arbeitet gemäß einer Sequenz von Prozessschritten, wie im Folgenden mit Bezug auf Fig. 7 weiter beschrieben.

[0057] Fig. 6 ist ein Prozess-Ablaufdiagramm, welches die Stufen **90** der Merkmalsanalyse, welche durch den Datensammlungs-Analysator **12** von Fig. 1 durchgeführt wird, zeigt. Die einzelnen Datensammlungen **76** werden vorverarbeitet, und Nominalphrasen, Genom- und Protein-Marker oder ähnliche atomare Dateneinheiten werden als Merkmale (Übergang **91**) im Lexikon **77** extrahiert. Die Merkmale werden normalisiert und aufgereiht (Übergang **92**), um die Merkmals-Häufigkeits-Tabelle **78** zu erzeugen. Die Merkmals-Häufigkeits-Tabelle **78** identifiziert einzelne Merkmale und jeweilige Auftritts-Häufigkeiten innerhalb jeder Datensammlung **76**. Die Häufigkeiten der Merkmalsauftritte werden in der geordneten Merkmals-Häufigkeits-Abbildungsmatrix **79** abgebildet (Übergang **93**), welche die Auftritts-Häufigkeiten von jedem Merkmal auf einer Pro-Datensammlung-Basis über alle Datensammlungen in Zusammenhang stellt. Die Merkmale werden in Gruppen **80** von semantisch bezogenen Themen, basierend auf einer relativen Ähnlichkeit, beispielsweise in Ausdrücken von der Distanz-Messung gemessen, ausgebildet (Übergang **94**). Schließlich werden die Gruppen **80** in Projektionen **81** projiziert (Übergang **95**), welche in eindimensionale Dokumentsignalvektoren neu geordnet und priorisiert werden.

[0058] Fig. 7 ist ein Ablaufdiagramm, welches ein Verfahren **100** zum Identifizieren von kritischen Merkmalen in einem geordneten Skala-Raum innerhalb eines multidimensionalen Merkmal-Raums **40** (in Fig. 2 gezeigt) gemäß der vorliegenden Erfindung zeigt. Als ein vorläufiger Schritt wird der Problem-Raum definiert, indem die zu analysierende Datensammlung identifiziert wird (Block **101**). Der Problem-Raum kann jegliche Sammlung von strukturierten oder unstrukturierten Datensammlungen sein, welche Dokumente oder Genom- oder Proteinsequenzen enthalten, wie durch den Fachmann anerkannt werden wird. Die Datensammlungen **41** werden von der Datenquelle **14** (in Fig. 1 gezeigt) erlangt (Block **102**).

[0059] Sobald identifiziert und erlangt, werden die Datensammlungen **41** nach Merkmalen analysiert (Block **103**), wie im Folgenden mit Bezug auf Fig. 8 weiter beschrieben. Während der Merkmalsanalyse wird eine geordnete Matrix **79**, welche die Auftritts-Häufigkeit von extrahierten Merkmalen abbildet

(unten in Fig. 14 gezeigt), konstruiert, um den semantischen Inhalt, welcher den Datensammlungen **41** inhärent ist, zusammenzufassen. Schließlich kann der semantische Inhalt, welcher von den Datensammlungen **41** extrahiert ist, optional grafisch dargestellt und visualisiert werden (Block **104**), wie in unserer U.S.-Patentanmeldung Serial No. 09/944,475, eingereicht am 31. August 2001; U.S.-Patentanmeldung Serial No. 09/943,918, eingereicht am 31. August 2001; und U.S.-Patentanmeldung Serial No. 10/084,401, eingereicht am 25. Februar 2002, beschrieben. Das Verfahren schließt dann ab.

[0060] Fig. 8 ist ein Ablaufdiagramm, welches die Routine **110** zur Durchführung einer Merkmalsanalyse zur Verwendung in dem Verfahren **100** von Fig. 7 zeigt. Der Grund für diese Routine liegt in der Extrahierung und Indexierung von Merkmalen von den Datensammlungen **41**. In der beschriebenen Ausführungsform werden Ausdrücke und Phrasen typischerweise aus Dokumenten extrahiert. Dokument-Merkmale können ebenfalls eine Absatz-Zählung, Sätze, ein Datum, einen Titel, eine Datei, einen Autor, ein Thema, eine Zusammenfassung usw. enthalten. Für Genom- und Proteinsequenzen werden Marker extrahiert. Für weitere Formen von strukturierten oder unstrukturierten Daten werden atomistische Dateneinheiten, welche einem semantischen Inhalt charakteristisch sind, extrahiert, wie durch den Fachmann erkannt werden wird.

[0061] Zuvor wird jede Datensammlung **41** in dem Problem-Raum vorverarbeitet (Block **111**), um Stopp-Wörter oder ähnlich atomare, nicht-beweiskräftige Dateneinheiten zu entfernen. Bei Datensammlungen **41**, welche Dokumente enthalten, enthalten Stopp-Wörter allgemein auftretende Wörter, wie beispielsweise unbestimmte Artikel („ein" und „eine"), bestimmte Artikel („der", „die", „das"), Pronomen („ich", „er" und „sie"), Verbindungsworte („und" und „oder") und ähnliche nichtsubstantive Wörter. Für Genom- und Proteinsequenzen enthalten Stopp-Wörter Nicht-Marker-Nachfolgekombinationen. Weitere Formen von Stopp-Wörtern oder nicht-beweiskräftigen Dateneinheiten können eine Entfernung oder Filterung erfordern, wie durch den Fachmann anerkannt werden wird.

[0062] Der Vorverarbeitung folgend, wird die Auftritts-Häufigkeit von Merkmalen für jede Datensammlung **41** bestimmt (Block **112**), wie im Folgenden mit Bezug auf Fig. 9 weiter beschrieben. Optional wird ein Histogramm **83** von der Häufigkeit von Merkmalsauftritten pro Dokument oder Datensammlung (in Fig. 4 gezeigt) logisch erzeugt (Block **113**). Jedes Histogramm **83**, wie im Folgenden mit Bezug auf Fig. 13 weiter beschrieben, bildet die relative Auftritts-Häufigkeit von jedem extrahierten Merkmal auf einer Pro-Dokument-Basis ab. Als Nächstes wird die Auftritts-Häufigkeit von Merkmalen für alle Datensätze **41** über den gesamten Problem-Raum abgebildet (Block **114**), indem eine geordnete Merkmals-Häufigkeits-Abbildungsmatrix **79** erzeugt wird, wie im Folgenden mit Bezug auf Fig. 14 weiter beschrieben. Optional wird ein Merkmals-Häufigkeits-Auftritt-Kurvenverlauf **84** (in Fig. 4 gezeigt) logisch erzeugt (Block **115**). Der Korpus-Kurvenverlauf, wie im Folgenden mit Bezug auf Fig. 15 weiter beschrieben, wird für alle Datensätze **41** erzeugt und bildet grafisch die semantisch bezogenen Konzepte, basierend auf den kumulativen Auftritten von den extrahierten Merkmalen ab.

[0063] Eine Mehrfachauflösungs-Analyse wird auf der geordneten Häufigkeits-Abbildungsmatrix **79** (Block **116**) durchgeführt, wie im Folgenden mit Bezug auf Fig. 16 weiter beschrieben. Eine Gruppen-Neuordnung erzeugt einen Satz von geordneten Vektoren, wobei jeder ein „semantisches" Signal enthält, welches herkömmlichen Signalverarbeitungstechniken zugänglich ist. Somit können die geordneten Vektoren analysiert, wie beispielsweise über eine Mehrfachauflösungs-Analyse, quantisiert (Block **117**) und encodiert (Block **118**), wie im Stand der Technik bekannt, werden. Die Routine kehrt dann zurück.

[0064] Fig. 9 ist ein Ablaufdiagramm, welches die Routine **120** zum Bestimmen einer Häufigkeit von Konzepten zur Verwendung in der Routine von Fig. 8 zeigt. Der Zweck von dieser Routine liegt in der Extrahierung von einzelnen Merkmalen von jeder Datensammlung und in der Erzeugung einer normalisierten Darstellung von den Merkmalsauftritten und Nebenauftritten auf einer Pro-Datensammlung-Basis. In der beschriebenen Ausführungsform werden Merkmale für Dokumente auf der Basis von den extrahierten Nominalphrasen bestimmt, obwohl einzelne Nomen oder Tri-Grams (Wort-Dreiergruppen) anstelle von Nominalphrasen verwendet werden können. Ausdrücke und Phrasen werden typischerweise von den Dokumenten unter Verwendung des LinguistX-Produktes extrahiert, welches durch Inxight-Software, INC., Santa Clara, California, lizenziert. Weitere Dokument-Merkmale können ebenfalls extrahiert werden, welche eine Absatz-Zählung, Sätze, ein Datum, einen Titel, ein Verzeichnis, eine Datei, einen Autor, ein Thema, eine Zusammenfassung, Verbphrasen, usw. enthalten. Die Genom- und Proteinsequenzen werden ähnlich unter Verwendung von erkannten Protein- und Amino-Markern, extrahiert, wie im Stand der Technik bekannt.

[0065] Jede Datensammlung wird iterativ wie folgt verarbeitet (Blöcke **121**-**126**). Zuerst werden einzelne Merkmale, wie beispielsweise Nominalphrasen oder Genom- und Proteinsequenz-Marker, von jeder Datensammlung **41** extrahiert (Block **122**). Sobald extrahiert, werden die einzelnen Merkmale in Aufzeichnungen geladen, welche in der Datenbank **30** gespeichert sind (in Fig. 1 gezeigt) (Block **123**). Die in der

Datenbank **30** gespeicherten Merkmale werden normalisiert (Block **124**), so dass jedes Merkmal lediglich ein Mal als eine Aufzeichnung erscheint. In der beschriebenen Ausführungsform werden die Aufzeichnungen in die dritte normale Form normalisiert, obwohl weitere Normalisierungs-Schemata verwendet werden können. Eine Merkmals-Häufigkeits-Tabelle **78** (in Fig. 5 gezeigt) wird für die Datensammlung **41** erzeugt (Block **125**). Die Merkmals-Häufigkeits-Tabelle **78** bildet die Anzahl von Auftritten und Nebenauftritten von jedem extrahierten Merkmal für die Datensammlung ab. Eine iterative Verarbeitung setzt sich für jede verbleibende Datensammlung **41** fort (Block **126**), nach welcher die Routine zurückkehrt.

[0066] Fig. 10 ist ein Datenstruktur-Diagramm, welches eine Datenbank-Aufzeichnung **130** für ein Merkmal zeigt, welches in der Datenbank **30** von Fig. 1 gespeichert ist. Jede Datenbank-Aufzeichnung **130** enthält Felder zum Speichern einer Kennung **131**, eines Merkmals **132** und einer Häufigkeit **133**. Die Kennung **131** ist ein monoton ansteigender Ganzzahlwert, welcher das Merkmal **132**, welches in jeder Aufzeichnung **130** gespeichert ist, eindeutig identifiziert. Die Kennung **131** kann gleichwertig jegliche weitere Form von einer unverwechselbaren Kennzeichnung sein, wie durch den Fachmann anerkannt werden wird. Die Auftritts-Häufigkeit von jedem Merkmal wird in die Häufigkeit **133** sowohl auf einer Pro-Fall-Sammlung und von gesamten Problem-Raum-Basen aufgegangen.

[0067] Fig. 11 ist ein Datenstruktur-Diagramm, welches mittels Beispiel eine Datenbank-Tabelle **140** zeigt, welche ein Lexikon **141** von extrahierten Merkmalen enthält, welche in der Datenbank **30** von Fig. 1 gespeichert sind. Das Lexikon **141** bildet die einzelnen Auftritte von identifizierten Merkmalen **143** ab, welche für jegliche vorgegebene Datensammlung **142** extrahiert sind. Mittels Beispiel enthält die Datensammlung **142** drei Merkmale, welche mit 1, 3 und 5 nummeriert sind. Das Merkmal **1** tritt ein Mal in der Datensammlung **142** auf, das Merkmal **3** tritt zwei Mal auf und das Merkmal **5** tritt ebenfalls ein Mal auf. Das Lexikon wird die Auftritts-Häufigkeit von den Merkmalen **1**, **3** und **5** über alle Datensammlungen **44** in dem Problem-Raum hinweg zusammenzählen und darstellen.

[0068] Die extrahierten Merkmale in dem Lexikon **141** können grafisch visualisiert werden. Fig. 12 ist ein Kurvenverlauf, welcher mittels Beispiel ein Histogramm **150** von den Häufigkeiten von Merkmalsauftritten, welche durch die Routine von Fig. 9 erzeugt werden, zeigt. Die X-Achse bestimmt die einzelnen Merkmale **151** für jedes Dokument und die Y-Achse bestimmt die Auftritts-Häufigkeiten von jedem Merkmal **152**. Die Merkmale werden in Reihenfolge von abnehmender Häufigkeit **153** abgebildet, um einen Kurvenverlauf **154** zu erzeugen, welcher den semantischen Inhalt von dem Dokument **44** darstellt. Demgemäß haben Merkmale, welche auf dem zunehmenden Ende von dem Kurvenverlauf **154** erscheinen, eine hohe Auftritts-Häufigkeit, während Merkmale, welche am abnehmenden Ende von dem Kurvenverlauf **154** erscheinen, eine niedrige Auftritts-Häufigkeit haben.

[0069] Wieder Bezug nehmend auf Fig. 11, spiegelt das Lexikon **141** die Merkmale für einzelne Datensammlungen wider und kann eine wesentliche Anzahl von Merkmalsauftritten in Abhängigkeit von der Größe von der Datensammlung enthalten. Die einzelnen Lexika **141** können logisch zusammengefasst werden, um einen Merkmal-Raum über alle Datensammlungen auszubilden. Fig. 13 ist ein Kurvenverlauf **160**, welcher mittels Beispiel eine Zunahme in einer Anzahl von Merkmalen in Relation zu einer Anzahl von Datensammlungen zeigt. Die X-Achse bestimmt die Datensammlungen **161** für den Problem-Raum und die Y-Achse bestimmt die Anzahl von extrahierten Merkmalen **162**. Eine Abbildung des Merkmal-Raumes (Anzahl von Merkmalen **162**) über den Problem-Raum (Anzahl von Datensammlungen **161**) erzeugt einen Kurvenverlauf **163**, welcher die kumulative Anzahl von Merkmalen darstellt, welcher proportional zur Anzahl von Datensammlungen **161** zunimmt **163**. Jedes zusätzliche extrahierte Merkmal erzeugt eine neue Dimension innerhalb des Merkmal-Raumes, welcher, ohne eine Ordnung und Priorisierung, kaum einen semantischen Inhalt auf eine wirksame Weise abstrahiert.

[0070] Fig. 14 ist eine Tabelle, welche mittels Beispiel eine Matrix-Abbildung von Merkmals-Häufigkeiten **170** zeigt, welche durch die Routine von Fig. 9 erzeugt wird. Die Merkmals-Häufigkeits-Abbildungs-Matrix **170** bildet Merkmale **173** entlang einer horizontalen Dimension **171** und Datensammlungen **174** entlang einer vertikalen Dimension **172** ab, obwohl die Zuweisung von jeweiligen Dimensionen willkürlich ist und invers neu zugewiesen werden kann, wie durch den Fachmann anerkannt werden wird. Jede Zelle **175** innerhalb der Matrix **170** enthält die kumulative Anzahl von Auftritten für jedes Merkmal **173** innerhalb einer vorgegebenen Datensammlung **174**. Demgemäß bildet jede Merkmals-Spalte einen Merkmalssatz **176**, und jede Datensammlungs-Zeile bildet einen Fall oder ein Muster **177**. Jedes Muster **177** stellt ein eindimensionales Signal in skalierbarer Vektorform dar, und konzeptionell unwesentliche Merkmale innerhalb des Musters **177** stellen Rauschen dar.

[0071] Fig. 15 ist ein Kurvenverlauf, welcher mittels Beispiel einen Korpus-Kurvenverlauf **180** von der Häufigkeit von Merkmalsauftritten zeigt, welche durch die Routine von Fig. 9 erzeugt wird. Der Kurvenverlauf **180** visualisiert die extrahierten Merkmale, wie in der Merkmals-Häufigkeits-Abbildungs-Ma-

trix **170** (in Fig. 14 gezeigt) aufgezählt. Die X-Achse bestimmt die einzelnen Merkmale **181** für alle Datensammlungen und die Y-Achse bestimmt die Anzahl von Datensammlungen **41**, welche auf jedes Merkmal **182** Bezug nehmen. Die einzelnen Merkmale werden in Reihenfolge einer abnehmenden Auftritts-Häufigkeit **183** abgebildet, um einen Kurvenverlauf **184** zu erzeugen, welcher die latenten Semantiken von dem Satz von Datensammlungen **41** darstellt. Der Kurvenverlauf **184** wird dazu verwendet, um Gruppen zu erzeugen, wobei diese auf geordnete und priorisierte eindimensionale Projektionen im Hilbert-Funktion-Raum projiziert werden.

[0072] Während der Gruppenausbildung wird ein Mittelwert **185** ausgewählt, und es werden Kantenbedingungen **186a**-b aufgebaut, um zwischen Merkmalen, welche zu häufig auftreten, gegen Merkmale, welche zu unhäufig auftreten, zu unterscheiden. Jene Datensammlungen, welche innerhalb der Kantenbedingungen **186a**-b fallen, bilden einen Teilsatz von Datensammlungen aus, welche latente Merkmale enthalten. In der beschriebenen Ausführungsform ist der Mittelwert **185** Datensammlungstyp abhängig. Aus Gründen der Wirksamkeit wird die Obere-Kante-Bedingung **186b** auf 70% eingestellt, und ein Teilsatz von den Merkmalen, welche unmittelbar der Obere-Kante-Bedingung **186b** folgen, wird ausgewählt, obwohl weitere Formen von einer Schwellwert-Unterscheidung ebenfalls verwendet werden können.

[0073] Fig. 16 ist ein Ablaufdiagramm **190**, welches eine Routine zur Transformation eines Problem-Raums in einen Skala-Raum zur Verwendung in der Routine von Fig. 8 zeigt. Der Zweck von dieser Routine liegt in der Erzeugung von Gruppen **80** (in Fig. 4 gezeigt), welche dazu verwendet werden, um eindimensionale Projektionen **81** (in Fig. 4 gezeigt) in einen Skala-Raum auszubilden, von welchem aus kritische Merkmale identifiziert werden.

[0074] Kurz gesagt wird eine einzelne Gruppe anfänglich erzeugt, und zusätzliche Gruppen werden unter Verwendung von einer bestimmten Form von einer unüberwachten Gruppenbildung hinzugefügt, wie beispielsweise eine einfache Gruppenbildung, eine hierarchische Gruppenbildung, Aufteilungsverfahren und Zusammenführverfahren, wie beschrieben in T. Kohonen, Ibid. at CH. 1.3. Die Form der Gruppenbildung, welche verwendet wird, ist nicht kritisch und kann jegliche weitere Form von einem unüberwachten Training sein, wie im Stand der Technik bekannt. Jede Gruppe enthält jene Datensammlungen, welche anverwandte Merkmale gemeinsam benutzen, wie durch eine bestimmte Distanzmetrik gemessen, welche im multidimensionalen Merkmal-Raum abgebildet wird. Die Gruppen werden auf eindimensionale geordnete Vektoren projiziert, welche als Wavelet- und Skalierungs-Koeffizienten encodiert sind, und nach kritischen Merkmalen analysiert.

[0075] Anfangs wird eine Varianz, welche eine obere Grenze auf der Distanz-Messung in dem multidimensionalen Merkmal-Raum spezifiziert, bestimmt (Block **191**). In der beschriebenen Ausführungsform wird eine Varianz von 5% spezifiziert, obwohl weitere Varianzwerte, entweder größer oder kleiner als 5%, als geeignet verwendet werden können. Jene Gruppen, welche außerhalb der vorbestimmten Varianz fallen, werden in separate Gruppen gruppiert, so dass die Merkmale über einen aussagekräftigen Bereich von Gruppen verteilt werden, und jeder Fall in dem Problem-Raum in zumindest einer Gruppe erscheint.

[0076] Die Merkmals-Häufigkeits-Abbildungsmatrix **170** (in Fig. 14 gezeigt) wird dann abgerufen (Block **192**). Die geordnete Merkmals-Häufigkeits-Abbildungsmatrix **79** wird in einem multidimensionalen Merkmal-Raum ausgedrückt. Jedes Merkmal erzeugt eine neue Dimension, welche die Merkmal-Raum-Größe linear mit jedem aufeinanderfolgend extrahierten Merkmal erhöht. Demgemäß werden die Datensammlungen iterativ verarbeitet (Blöcke **193**-**197**), um den multidimensionalen Merkmal-Raum in einem einzeldimensionalen Dokumentenvektor (Signal), wie folgt, zu transformieren. Während jeder Iteration (Block **193**) wird ein Muster **177** für die derzeitige Datensammlung von der Merkmals-Haufigkeits-Abbildungsmatrix **170** extrahiert (Block **194**). Ähnlichkeits-Messungen werden von dem Muster **177** erzeugt und anverwandte Merkmale werden in Gruppen **80** (in Fig. 5 gezeigt) ausgebildet (Block **195**), indem eine bestimmte Form einer unüberwachten Gruppierung verwendet wird, wie oben beschrieben. Jene Merkmale, welche innerhalb der vorbestimmten Varianz fallen, wie durch die Distanzmetrik gemessen, werden in derselben Gruppe identifiziert und gruppiert, während jene Merkmale, welche außerhalb der vorbestimmten Varianz fallen, einer weiteren Gruppe zugewiesen werden.

[0077] Als Nächstes werden die Gruppen **80** im Merkmal-Raum jeweils auf ein eindimensionales Signal in skalierbarer Vektorform projiziert (Block **196**). Die geordneten Vektoren bilden ein „semantisches" Signal, welches Signalverarbeitungstechniken, wie beispielsweise eine Mehrfachauflösungs-Analyse, zugänglich ist. In der beschriebenen Ausführungsform werden die Gruppen **80** durch ein iteratives Ordnen der Merkmale, welche zu jeder Gruppe identifiziert sind, in den Vektor **61** projiziert. Alternativ kann eine Gruppenausbildung (Block **195**) und Projektion (Block **196**) in einem einzelnen Satz von Betrieben unter Verwendung einer selbstorganisierenden Abbildung, wie beispielsweise beschrieben in T. Kohonen, Ibid. at Ch. 3, durchgeführt werden. Weitere Verfahren zum Erzeugen von Ähnlichkeits-Messungen,

Ausbilden von Gruppen und Projizieren in einen Skala-Raum können gleichwertig angewendet werden und durch die vorhergehend beschriebenen Annäherungen ersetzt oder damit in Kombination durchgeführt werden, wie durch den Fachmann anerkannt werden wird. Eine iterative Verarbeitung fährt dann für jede verbleibende nächste Datensammlung fort (Block **197**), wonach die Routine zurückkehrt.

[0078] Fig. 17 ist ein Ablaufdiagramm **200**, welches die Routine zum Erzeugen von Ähnlichkeits-Messungen und Ausbilden von Gruppen zur Verwendung in der Routine von Fig. 16 zeigt. Der Zweck von dieser Routine liegt in der Identifizierung jener Merkmale, welche innerhalb des Merkmal-Raumes eine nächste Ähnlichkeit haben, und zum Gruppieren von zwei oder mehreren Sätzen von ähnlichen Merkmalen in einzelne Gruppen. Die Gruppen ermöglichen eine Visualisierung des multidimensionalen Merkmal-Raums.

[0079] Merkmale und Gruppen werden in einem Paar von verschachtelten Schleifen (Blöcke **201**-**212** und **204**-**209**) iterativ verarbeitet. Während jeder Iteration von der äußeren Verarbeitungsschleife (Blöcke **201**-**212**) wird jedes Merkmal i verarbeitet (Block **201**). Das Merkmal i wird zuerst ausgewählt (Block **202**) und die Varianz θ für das Merkmal i wird berechnet (Block **203**).

[0080] Während jeder Iteration von der inneren Verarbeitungsschleife (Block **204**-**209**) wird jede Gruppe j verarbeitet (Block **204**). Die Gruppe j wird ausgewählt (Block **205**) und der Winkel σ in Relation zum gemeinsamen Ursprung wird für die Gruppe j berechnet (Block **206**). Es ist zu erwähnen, dass der Winkel σ regelmäßig für jede Gruppe j neu berechnet werden muss, da Merkmale den Gruppen hinzugefügt werden oder aus diesen entfernt werden. Die Differenz zwischen dem Winkel θ für das Merkmal i und dem Winkel σ für die Gruppe j wird mit der vorbestimmten Varianz verglichen (Block **207**). Wenn die Differenz kleiner als die vorbestimmte Varianz ist (Block **207**), wird das Merkmal i in die Gruppe j gesetzt (Block **208**), und die iterative Verarbeitungsschleife (Block **204**-**209**) wird abgeschlossen. Wenn die Differenz größer als gleich der Varianz ist (Block **207**), wird die nächste Gruppe j verarbeitet (Block **209**), bis alle Gruppen verarbeitet wurden (Blöcke **204**-**209**).

[0081] Wenn die Differenz zwischen dem Winkel θ für das Merkmal i und dem Winkel σ für jede der Gruppen die Varianz übersteigt, wird eine neue Gruppe erzeugt (Block **210**) und der Zähler num_clusters wird erhöht (Block **211**). Die Verarbeitung fährt mit dem nächsten Merkmal i fort (Block **212**), bis alle Merkmale verarbeitet wurden (Blöcke **201**-**212**). Die Kategorisierung von Gruppen wird wiederholt (Block **213**), wenn notwendig. In der beschriebenen Ausführungsform wird die Gruppen-Kategorisierung (Blöcke **201**-**212**) zumindest ein Mal wiederholt, bis der Satz von Gruppen festgelegt ist. Schließlich können die Gruppen als ein optionaler Schritt finalisiert werden (Block **214**). Eine Finalisierung enthält eine Zusammenfügung von zwei oder mehreren Gruppen in eine einzelne Gruppe, eine Aufteilung einer einzelnen Gruppe in zwei oder mehrere Gruppen, eine Entfernung von Minimal- oder Ausreißer-Gruppen und ähnliche Betriebe, wie durch den Fachmann anerkannt werden wird. Die Routine kehrt dann zurück.

[0082] Fig. 18 ist eine Tabelle **210**, welche mittels Beispiel die Merkmalsgruppen zeigt, welche durch die Routine von Fig. 17 erzeugt werden. Idealerweise sollte jedes der Merkmale **211** in zumindest einer der Gruppen **212** erscheinen, wodurch sichergestellt wird, dass jede Datensammlung in einer bestimmten Gruppe erscheint. Die Distanzberechnungen **213a**-d zwischen den Datensammlungen für ein vorgegebenes Merkmal werden bestimmt. Jene Distanzwerte **213a**-d, welche innerhalb einer vorbestimmten Varianz fallen, werden jeder einzelnen Gruppe zugewiesen. Die Tabelle **210** kann dazu verwendet werden, um die Gruppen in einem multidimensionalen Merkmal-Raum zu visualisieren.

[0083] Fig. 19 ist ein Ablaufdiagramm, welches eine Routine zum Identifizieren von kritischen Merkmalen zur Verwendung in den Verfahren von Fig. 7 zeigt. Der Zweck von dieser Routine liegt in der Transformation des Skala-Raum-Vektors in variierende Detailpegel mit Wavelet- und Skalierungs-Koeffizienten über eine Mehrfachauflösungs-Analyse. Eine Wavelet-Dekomposition ist eine Form einer Signalfilterung, welche eine grobe Zusammenfassung von den ursprünglichen Daten und Details, welche während einer Dekomposition verloren gehen, bereitstellt, wodurch es erlaubt wird, dass der Datenstrom mehrere Detailpegel ausdrückt. Jeder Wavelet- und Skalierungs-Koeffizient wird durch eine Mehrfachauflösungs-Analyse ausgebildet, welche typischerweise den Datenstrom während jedes rekursiven Schrittes halbiert.

[0084] Somit wird die Größe des eindimensional geordneten Vektors **61** (in Fig. 4 gezeigt) durch die Gesamtanzahl von Merkmalen n im Merkmal-Raum bestimmt (Block **221**). Der Vektor **61** wird dann über jeden Mehrfachauflösungs-Pegel wie folgt iterativ verarbeitet (Blöcke **222**-**225**). Zunächst werden n/2 Wavelet-Koeffizienten und n/2 Skalierungsfunktionen φ von dem Vektor **61** erzeugt, um Wavelet-Koeffizienten und Skalierungs-Koeffizienten auszubilden. Inder beschriebenen Ausführungsform werden die Wavelet- und Skalierungs-Koeffizienten durch Falten der Wavelet Ψ und Skalierung φ Funktionen mit den geordneten Dokumentenvektoren in einen kontinuierlichen Satz von Werten in dem Vektor **61** erzeugt. Weitere Verfahren zum Falten von Wavelet Ψ und Skalie-

rung φ Funktionen können ebenfalls verwendet werden, wie durch den Fachmann erkannt werden wird.

[0085] Der ersten Iteration von der Wavelet- und Skalierungs-Koeffizientenerzeugung folgend, wird die Anzahl von Merkmalen n herunter abgetastet (Block **224**) und jeder verbleibende Mehrfachauflösungs-Pegel wird iterativ verarbeitet (Blöcke **222**-**225**), bis die gewünschte minimale Auflösung des Signals erzielt ist. Die Routine kehrt dann zurück.

[0086] Obwohl die Erfindung insbesondere bezogen auf die Ausführungsformen davon gezeigt und beschrieben wurde, wird der Fachmann verstehen, dass die vorgenannten und weiteren Änderungen in der Form und im Detail darin vorgenommen werden können, ohne vom Umfang der Erfindung abzuweichen.

Patentansprüche

1. System (**10**) zum Identifizieren semantisch bezogener Merkmale (**212**) in einem geordneten Skala-Raum innerhalb eines multidimensionalen Vektor-Raums, welche eine Mehrzahl von Dokumenten darstellen, welches enthält:
einen Merkmals-Analysator (**72**), welcher anfangs Merkmale (**173**) verarbeitet (**110**), welcher enthält:
einen Merkmals-Extraktor (**71**), welcher die Merkmale von einer Mehrzahl von Dokumenten (**21**) extrahiert (**122**) und jedes Merkmal normalisiert (**124**); und
einen Merkmals-Abbilder (**79**), welcher Häufigkeiten eines Auftritts (**183**) für jedes der Merkmale (**173**) in den Dokumenten (**21**) bestimmt und die Häufigkeiten eines Auftrittes (**183**) in Vektoren (**177**) mit einem Vektor (**177**) für jedes Dokument (**21**) abbildet (**114**);
einen unbeaufsichtigten Klassifizierer (**73**), welcher Ähnlichkeits-Messungen zwischen den Häufigkeiten eines Auftrittes (**183**) in jedem Vektor (**177**) erzeugt (**195**) und Gruppen (**80**) ausbildet, wobei jede ein oder mehrere der Merkmale (**173**) enthält, welche Ähnlichkeits-Messungen innerhalb einer vorbestimmten Varianz haben;
einen Skala-Raum Umformer (**74**), welcher die Merkmale (**173**) in jedem eindimensionalen Signal (**81**) ordnet, indem eine Ähnlichkeit verringert wird, und jede Gruppe (**80**) auf jedes eindimensionale Signal (**81**) unter Verwendung von Wavelets und Skalierungs-Koeffizienten projiziert wird; und
einen Merkmals-Identifizierer (**75**), welcher die Wavelets und Skalierungs-Koeffizienten von jedem eindimensionalen Signal (**81**) durch sukzessive Detailpegel rekursiv herleitet (**220**), indem die Anzahl von abgetasteten Merkmalen verringert wird und die Wavelets und Skalierungs-Koeffizienten für die Merkmale von den Dokumenten analysiert werden.

2. System nach Anspruch 1, welches ferner enthält:
einen Vorprozessor, welcher jedes der Dokumente (**21**) vor einer Merkmals-Extraktion zum Identifizieren und logischen Entfernen von Stopp-Wörtern vorverarbeitet (**111**).

3. System nach Anspruch 1, welches ferner enthält:
einen Datenbank-Rufzeichner (**130**), welcher ein einzelnes Auftreten von jedem Merkmal (**173**) in normalisierter Form speichert.

4. System nach Anspruch 1, welches ferner enthält:
einen Merkmals-Häufigkeits-Abbilder (**79**), welcher die Vektoren (**177**) in eine Dokument-Merkmals-Matrix (**170**) gemäß dem Dokument (**21**) anordnet, von welchem die Merkmale (**173**) in jedem Vektor (**177**) extrahiert sind.

5. System nach Anspruch 1, welches ferner enthält:
ein Ähnlichkeits-Modul, welches eine Distanz-Messung zwischen jeder Auftritts-Häufigkeit (**183**) als eine Ähnlichkeits-Messung berechnet.

6. System nach Anspruch 7, welches ferner enthält:
ein Eigenorganisations-Kennfeld des multidimensionalen Vektor-Raums, welches vor der Projektion ausgebildet ist.

7. System nach Anspruch 1, bei welchem der unbeaufsichtigte Klassifizierer (**73**) die Gruppen (**80**) kategorisiert, welcher enthält:
einen Varianz-Bestimmer, welcher eine Varianz für jedes der Merkmale (**173**) bestimmt (**203**);
einen Winkel-Bestimmer, welcher einen Winkel in Relation zu einem allgemeinen Ursprung für jede der Gruppen bestimmt (**206**); und
einen Gruppen-Zuweiser, welcher jene Merkmale in die Gruppe zuweist (**208**), bei welchen der Winkel und die Varianz innerhalb der vorbestimmten Varianz sind.

8. System nach Anspruch 7, bei welchem der unbeaufsichtigte Klassifizier (**73**) wiederholt die Gruppen (**80**) kategorisiert, und der Varianz-Bestimmer eine Varianz für jedes der Merkmale (**173**) neu bestimmt (**203**); wobei der Winkel-Bestimmer den Winkel in Relation zum allgemeinen Ursprung für jene Gruppen neu bestimmt (**206**), von welchen oder in welche ein oder mehrere der Merkmale jeweils entfernt oder zugewiesen sind; und der Gruppen-Zuweiser zumindest eines aus einem Entfernen jener Merkmale aus der Gruppe, bei welchen der Winkel und die Varianz außerhalb der vorbestimmten Varianz sind, und einem Zuweisen (**208**) jener Merkmale in die Gruppe, bei welchen der Winkel und die Varianz innerhalb der vorbestimmten Varianz sind, durchführt.

9. System nach Anspruch 7, bei welchem der un-

beaufsichtigte Klassifizierer (**73**) die Gruppen (**80**) fertigstellt (**214**), welcher zumindest eines aus einem Zusammenfassen einer Mehrzahl der Gruppen in eine einzelne Gruppe; Aufteilen einer Gruppe in eine Mehrzahl von Gruppen; und Entfernen von Ausreißer-Gruppen enthält.

10. Verfahren (**100**) zum Identifizieren semantisch bezogener Merkmale (**212**) in einem geordneten Skala-Raum innerhalb eines multidimensionalen Vektor-Raums, welche eine Mehrzahl von Dokumenten (**21**) darstellen, welches enthält:
Verarbeiten von Merkmalen durch Extrahieren (**122**) der Merkmale (**173**) von der Mehrzahl von Dokumenten (**21**), und Normalisieren (**124**) jedes Merkmals (**173**);
Bestimmen von Häufigkeiten (**183**) von einem Auftritt für jedes der Merkmale (**173**) in den Dokumenten (**21**), und Abbilden (**114**) der Häufigkeiten eines Auftritts (**183**) in Vektoren (**177**) mit einem Vektor (**177**) für jedes Dokument (**21**);
Erzeugen (**195**) von Ähnlichkeits-Messungen zwischen den Häufigkeiten eines Auftritts (**183**) in jedem Vektor (**177**), und Ausbilden von Gruppen (**80**), welche jeweils ein oder mehrere der Merkmale (**173**) enthalten, welche Ähnlichkeits-Messungen innerhalb einer vorbestimmten Varianz haben;
Ordnen der Merkmale (**173**) in jedem eindimensionalen Signal (**81**) durch Verringern einer Ähnlichkeit;
Projizieren (**196**) von jeder Gruppe (**80**) auf jedes eindimensionale Signal (**81**) unter Verwendung von Wavelets und Skalierungs-Koeffizienten; und rekursives Herleiten (**220**) der Wavelets und Skalierungs-Koeffizienten von jedem eindimensionalen Signal (**81**) durch sukzessive Detailpegel durch Verringern der Anzahl von abgetasteten Merkmalen, und Analysieren der Wavelets und Skalierungs-Koeffizienten für die Merkmale von den Dokumenten.

11. Verfahren nach Anspruch 10, welches ferner enthält:
Vorverarbeiten (**111**) jedes der Dokumente (**21**) vor einer Merkmals-Extraktion zum Identifizieren und logischen Entfernen von Stopp-Wörtern.

12. Verfahren nach Anspruch 10, welches ferner enthält:
Speichern eines einzelnen Auftretens von jedem Merkmal (**173**) in normalisierter Form.

13. Verfahren nach Anspruch 10, welches ferner enthält:
Anordnen der Vektoren (**177**) in eine Dokument-Merkmals-Matrix (**170**) gemäß dem Dokument (**21**), von welchem die Merkmale (**173**) in jedem Vektor (**177**) extrahiert wurden.

14. Verfahren nach Anspruch 10, welches ferner enthält:
Berechnen einer Distanz-Messung zwischen jeder Häufigkeit eines Auftritts (**183**) als eine Ähnlichkeits-Messung.

15. Verfahren nach Anspruch 10, welches ferner enthält:
Erzeugen eines Eigenorganisations-Kennfeldes des multidimensionalen Vektor-Raums vor einer Projektion.

16. Verfahren nach Anspruch 10, welches ferner enthält:
Kategorisieren der Gruppen (**80**), welches enthält:
Bestimmen (**203**) von einer Varianz für jedes der Merkmale;
Bestimmen (**206**) eines Winkels in Relation zu einem allgemeinen Ursprung für jede der Gruppen; und
Zuweisen (**208**) jener Merkmale in die Gruppe, bei welchen der Winkel und die Varianz innerhalb der vorbestimmten Varianz sind.

17. Verfahren nach Anspruch 16, welches ferner enthält:
wiederholtes Kategorisieren (**213**) der Gruppen (**80**), welches ferner enthält:
Neubestimmen (**203**) einer Varianz für jedes der Merkmale;
Neubestimmen (**206**) des Winkels in Relation zum allgemeinen Ursprung für jene Gruppen, von welchen oder in welche ein oder mehrere der Merkmale jeweils entfernt oder zugewiesen sind; und
Durchführen von zumindest einem aus:
Entfernen jener Merkmale aus der Gruppe, bei welchen der Winkel und die Varianz außerhalb der vorbestimmten Varianz sind; und
Zuweisen (**208**) jener Merkmale in die Gruppe, bei welchen der Winkel und die Varianz innerhalb der vorbestimmten Varianz sind.

18. Verfahren nach Anspruch 16, welches ferner enthält:
Fertigstellen (**214**) der Gruppen, welches zumindest eines enthält aus:
Zusammenfassen einer Mehrzahl der Gruppen in eine einzelne Gruppe;
Aufteilen einer Gruppe in eine Mehrzahl von Gruppen; und
Entfernen von Ausreißer-Gruppen.

19. Computerlesbares Speichermedium für eine Vorrichtung, welches einen Kode zum Durchführen des Verfahrens gemäß einem der Ansprüche 10 bis 18 enthält.

Es folgen 15 Blatt Zeichnungen

Anhängende Zeichnungen

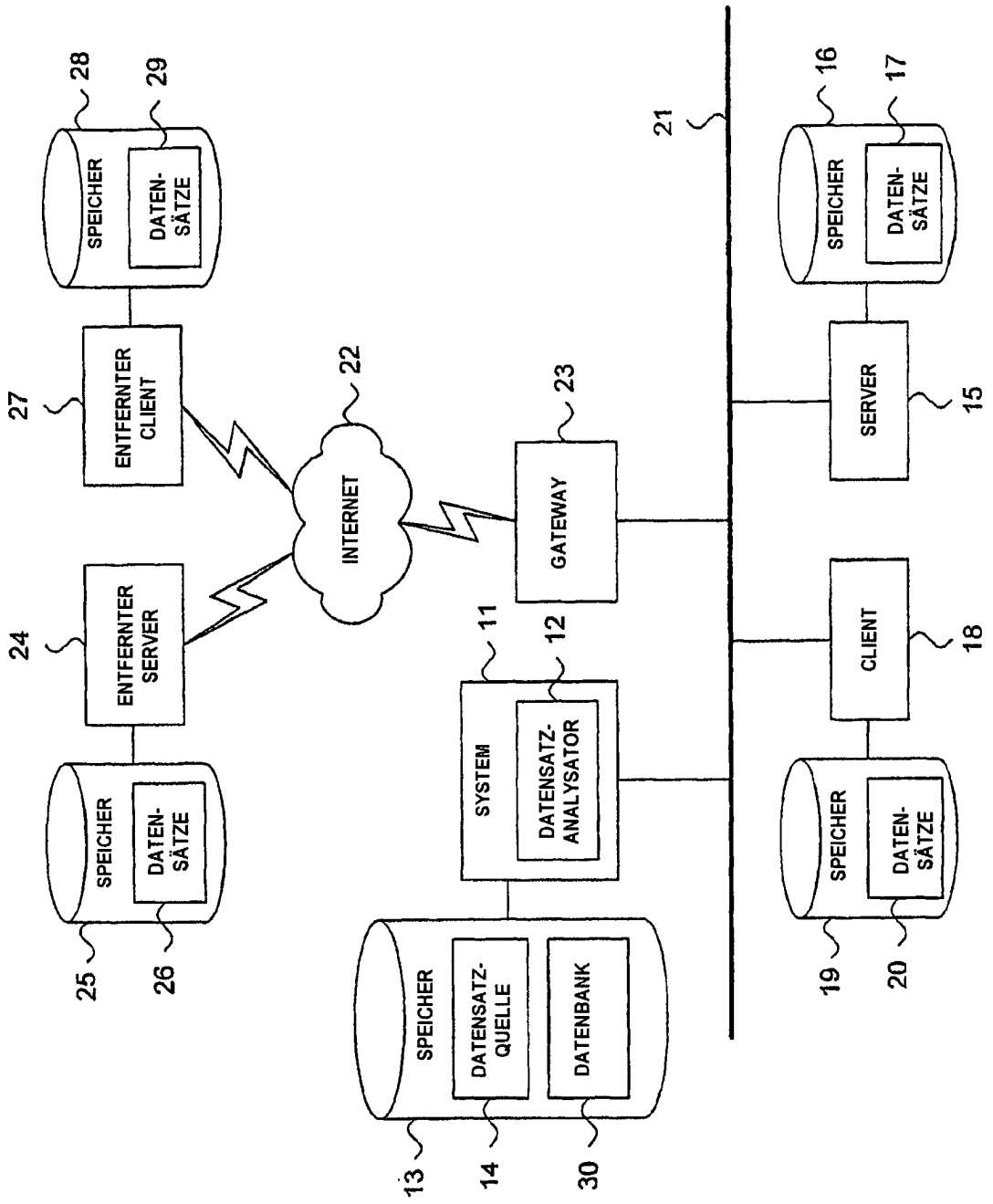


Fig. 1.

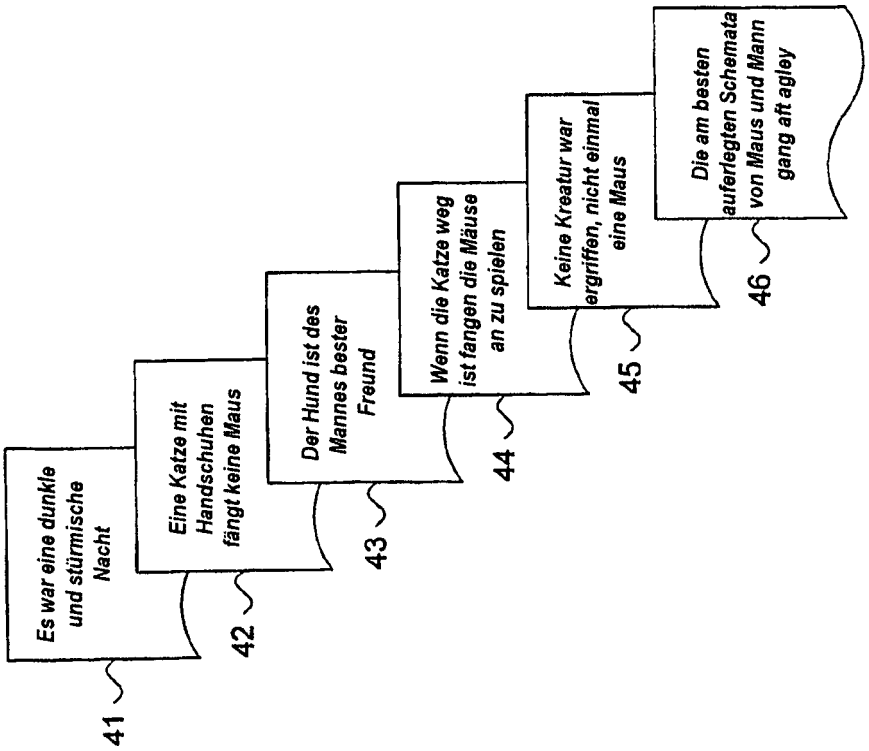

Es war eine dunkle und stürmische Nacht
41
Eine Katze mit Handschuhen fängt keine Maus
42
Der Hund ist des Mannes bester Freund
43
Wenn die Katze weg ist fangen die Mäuse an zu spielen
44
Keine Kreatur war ergriffen, nicht einmal eine Maus
45
Die am besten auferlegten Schemata von Maus und Mann gang aft agley
46

Fig. 2.

40

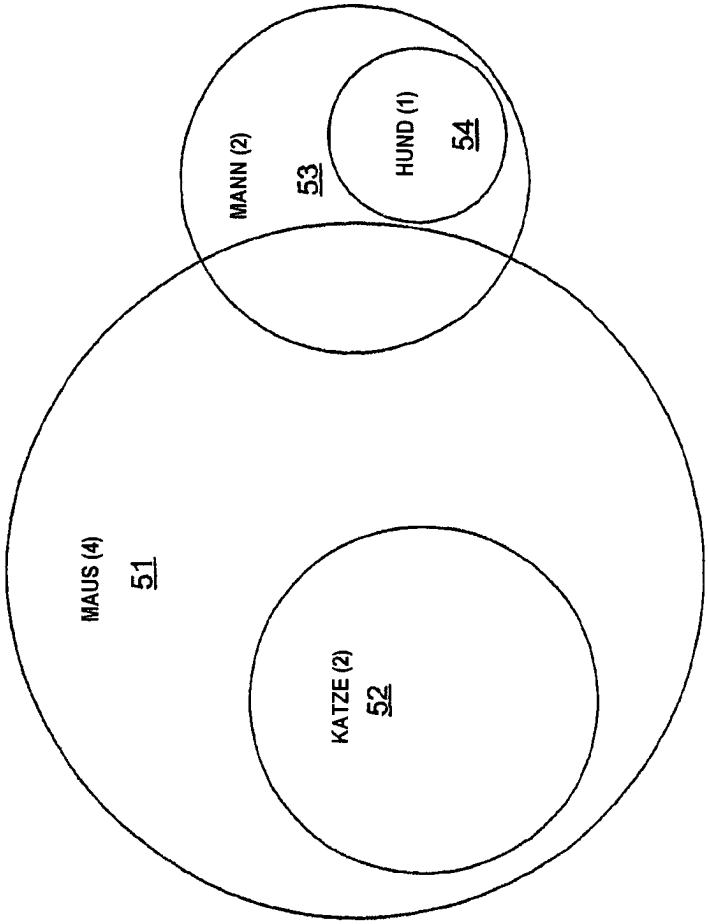


Fig. 3. 50

Fig. 4.

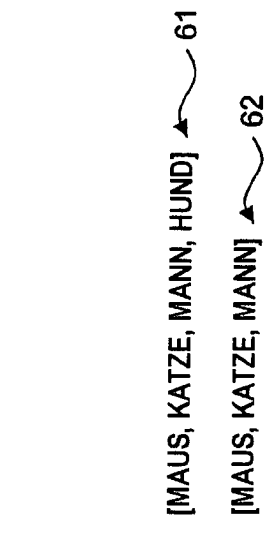

60
[MAUS, KATZE, MANN, HUND]
61
[MAUS, KATZE, MANN]
62
[MAUS, KATZE]
63
[MAUS]
64

Fig. 6.

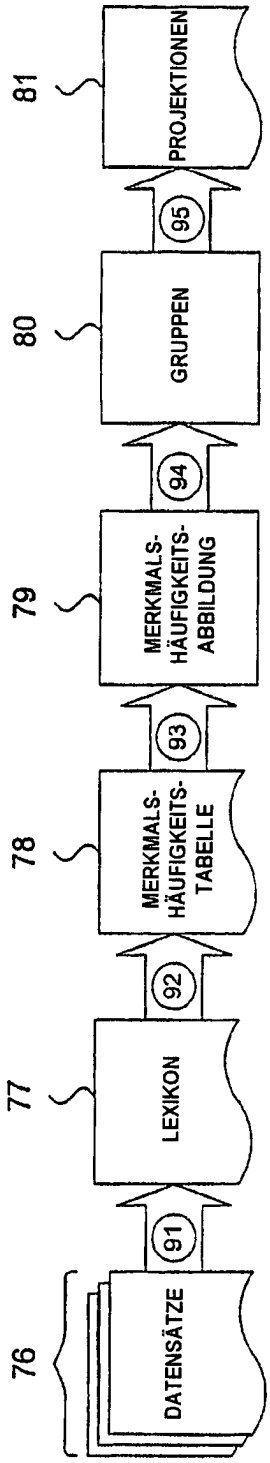

90
76
DATENSÄTZE
91
77
LEXIKON
92
78
MERKMALS-HÄUFIGKEITS-TABELLE
93
79
MERKMALS-HÄUFIGKEITS-ABBILDUNG
94
80
GRUPPEN
95
81
PROJEKTIONEN

Fig. 5.

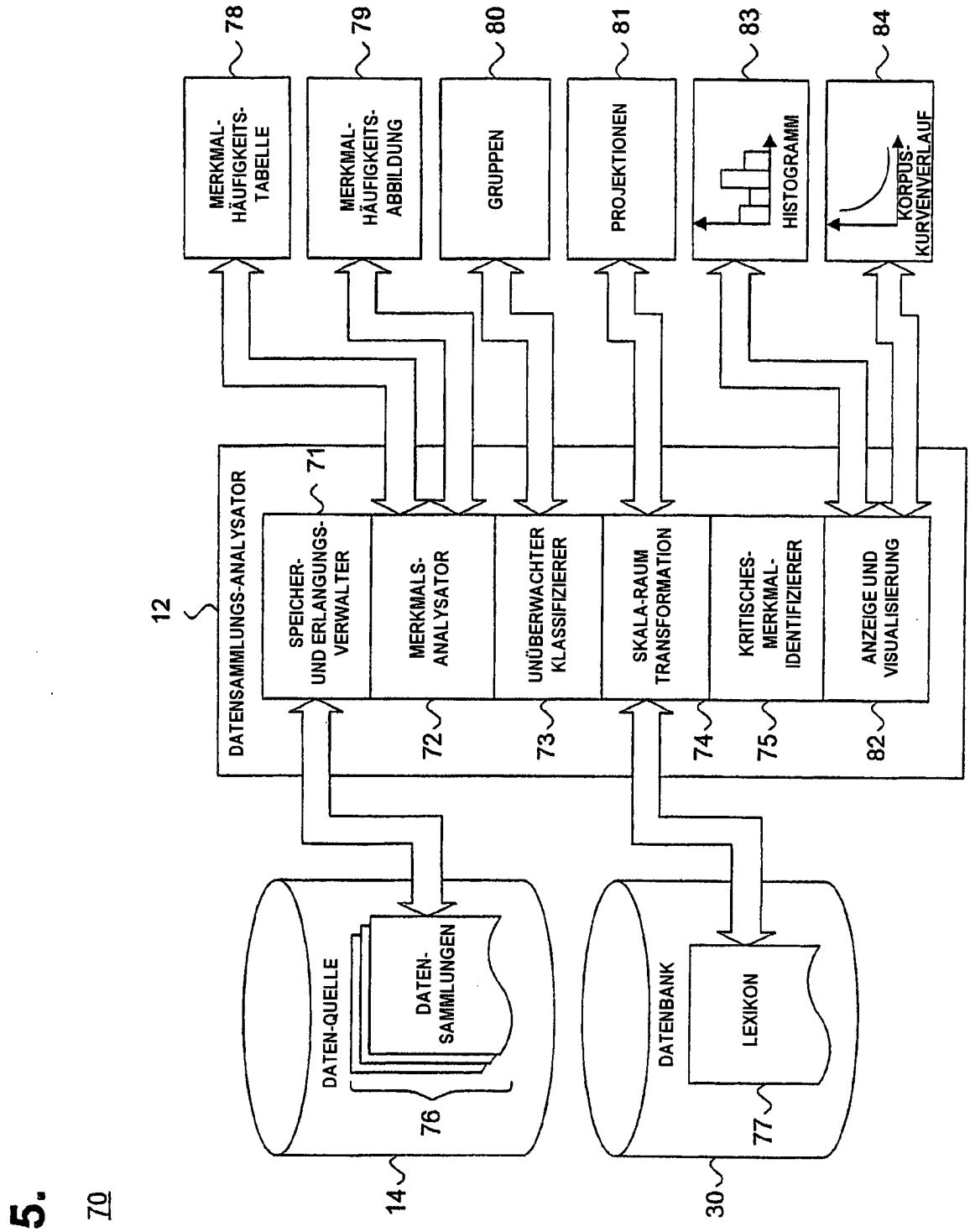

70
12
DATENSAMMLUNGS-ANALYSATOR
71
SPEICHER- UND ERLANGUNGS-VERWALTER
MERKMALS-ANALYSATOR
UNÜBERWACHTER KLASSIFIZIERER
SKALA-RAUM TRANSFORMATION
KRITISCHES-MERKMAL-IDENTIFIZIERER
ANZEIGE UND VISUALISIERUNG
72
73
74
75
82
MERKMAL-HÄUFIGKEITS-TABELLE
78
MERKMAL-HÄUFIGKEITS-ABBILDUNG
79
GRUPPEN
80
PROJEKTIONEN
81
HISTOGRAMM
83
KORPUS-KURVENVERLAUF
84
DATEN-QUELLE
DATEN-SAMMLUNGEN
76
14
DATENBANK
LEXIKON
77
30

Fig. 7.

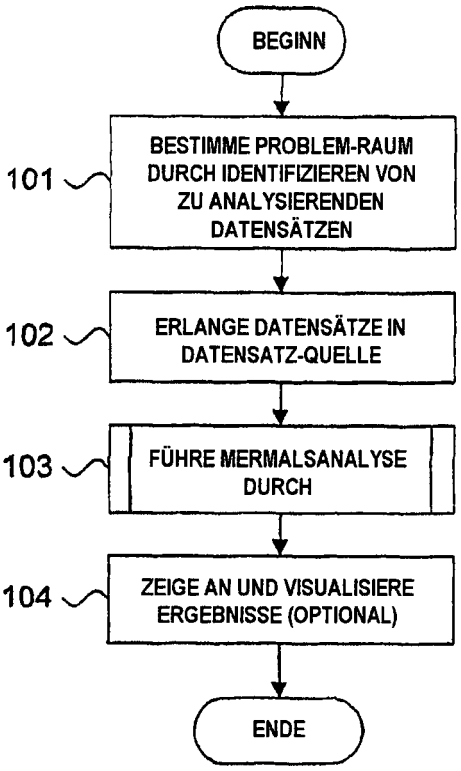

100
BEGINN
101
BESTIMME PROBLEM-RAUM DURCH IDENTIFIZIEREN VON ZU ANALYSIERENDEN DATENSÄTZEN
102
ERLANGE DATENSÄTZE IN DATENSATZ-QUELLE
103
FÜHRE MERMALSANALYSE DURCH
104
ZEIGE AN UND VISUALISIERE ERGEBNISSE (OPTIONAL)
ENDE

Fig. 8.

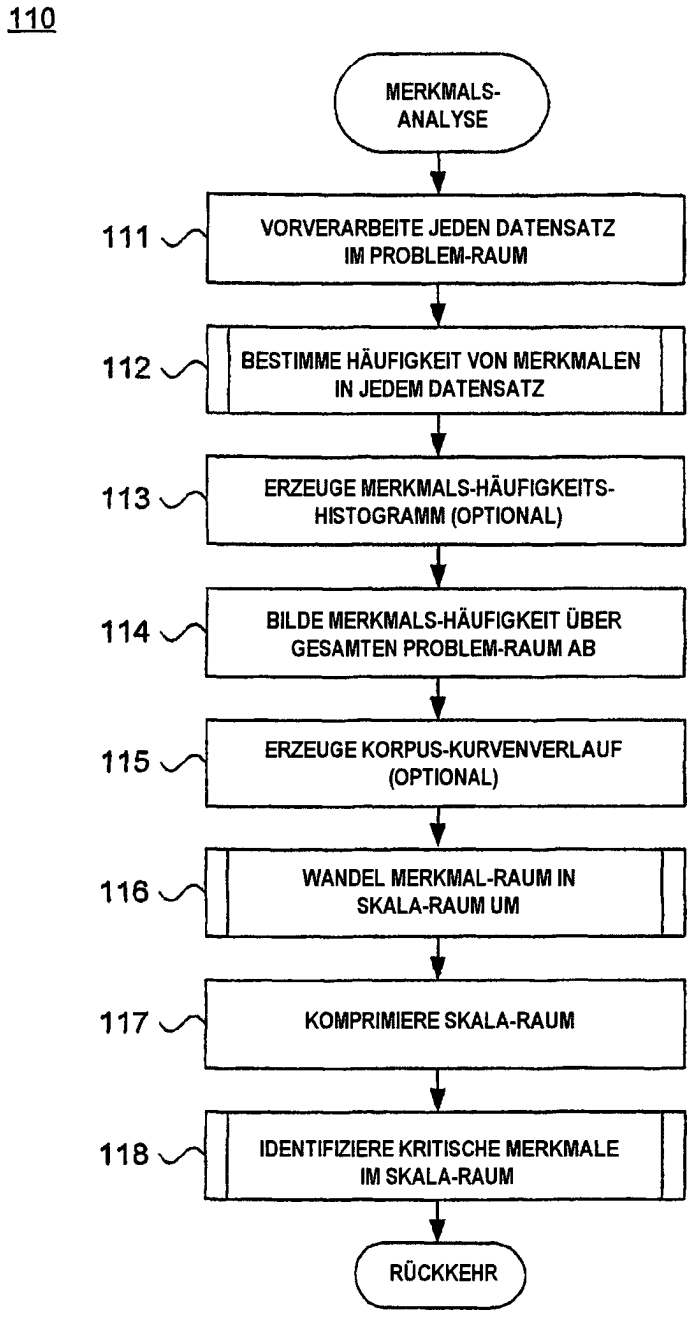

110
MERKMALS-ANALYSE
111
VORVERARBEITE JEDEN DATENSATZ IM PROBLEM-RAUM
112
BESTIMME HÄUFIGKEIT VON MERKMALEN IN JEDEM DATENSATZ
113
ERZEUGE MERKMALS-HÄUFIGKEITS-HISTOGRAMM (OPTIONAL)
114
BILDE MERKMALS-HÄUFIGKEIT ÜBER GESAMTEN PROBLEM-RAUM AB
115
ERZEUGE KORPUS-KURVENVERLAUF (OPTIONAL)
116
WANDEL MERKMAL-RAUM IN SKALA-RAUM UM
117
KOMPRIMIERE SKALA-RAUM
118
IDENTIFIZIERE KRITISCHE MERKMALE IM SKALA-RAUM
RÜCKKEHR

Fig. 9.

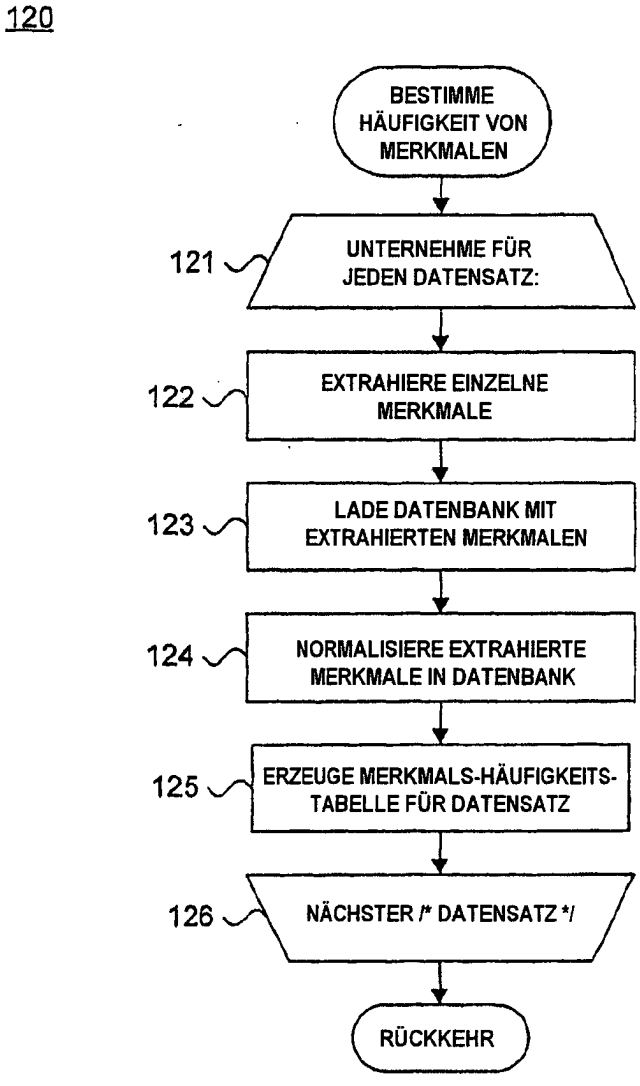

120
BESTIMME HÄUFIGKEIT VON MERKMALEN
121
UNTERNEHME FÜR JEDEN DATENSATZ:
122
EXTRAHIERE EINZELNE MERKMALE
123
LADE DATENBANK MIT EXTRAHIERTEN MERKMALEN
124
NORMALISIERE EXTRAHIERTE MERKMALE IN DATENBANK
125
ERZEUGE MERKMALS-HÄUFIGKEITS-TABELLE FÜR DATENSATZ
126
NÄCHSTER /* DATENSATZ */
RÜCKKEHR

Fig. 10.

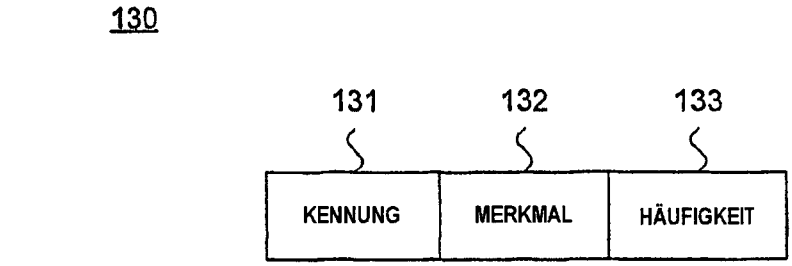

130
131
132
133
KENNUNG
MERKMAL
HÄUFIGKEIT

Fig. 11.

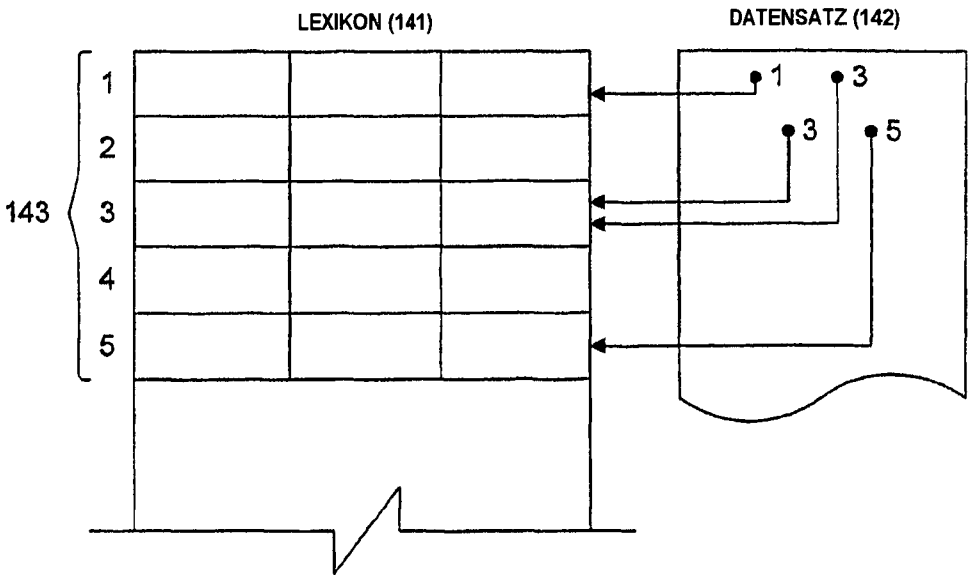

140
LEXIKON (141)
DATENSATZ (142)
143
1
2
3
4
5
1
3
3
5

Fig. 12.

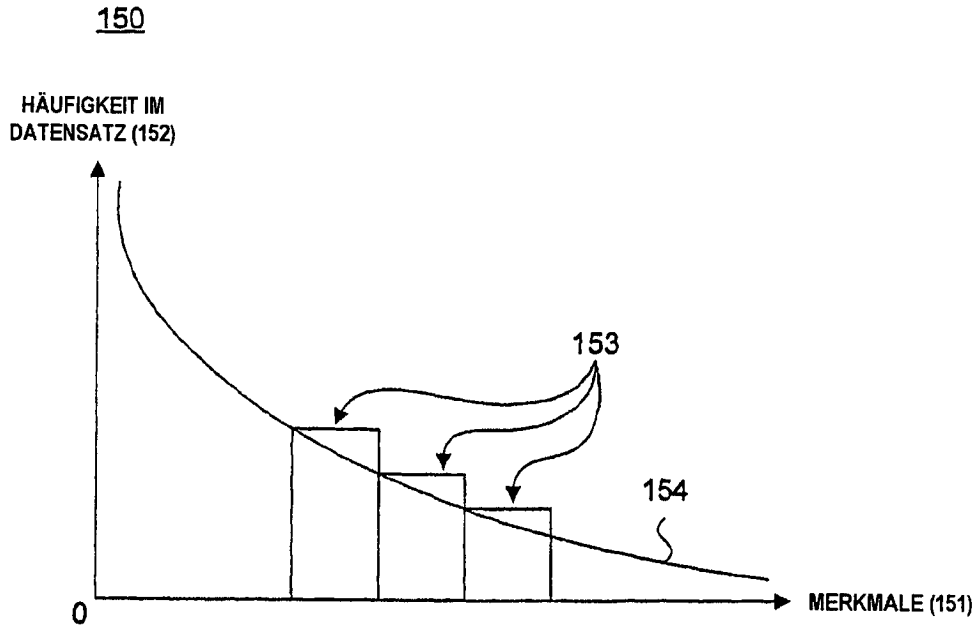

150
HÄUFIGKEIT IM DATENSATZ (152)
153
154
0
MERKMALE (151)

Fig. 13.

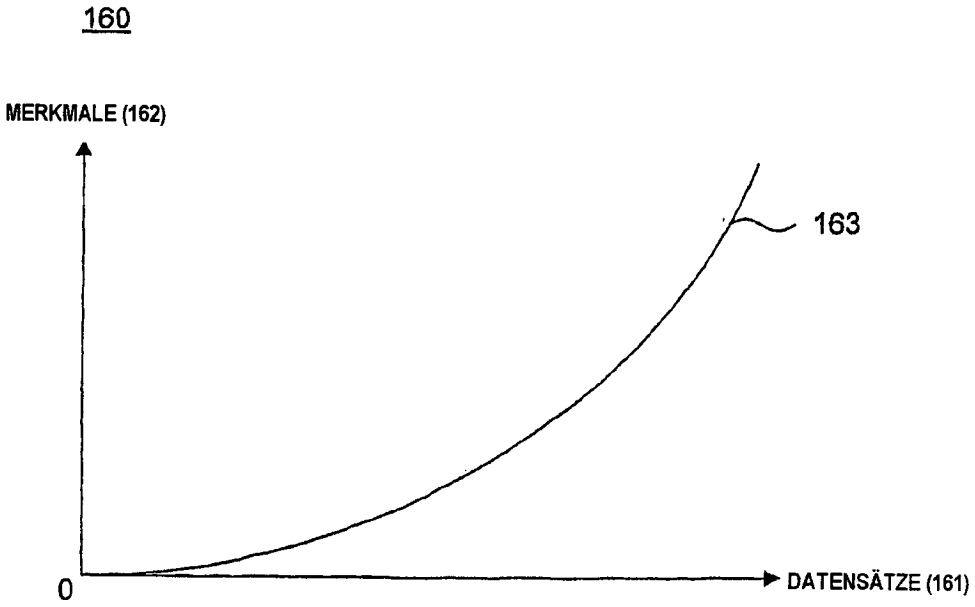

160
MERKMALE (162)
163
0
DATENSÄTZE (161)

Fig. 14.

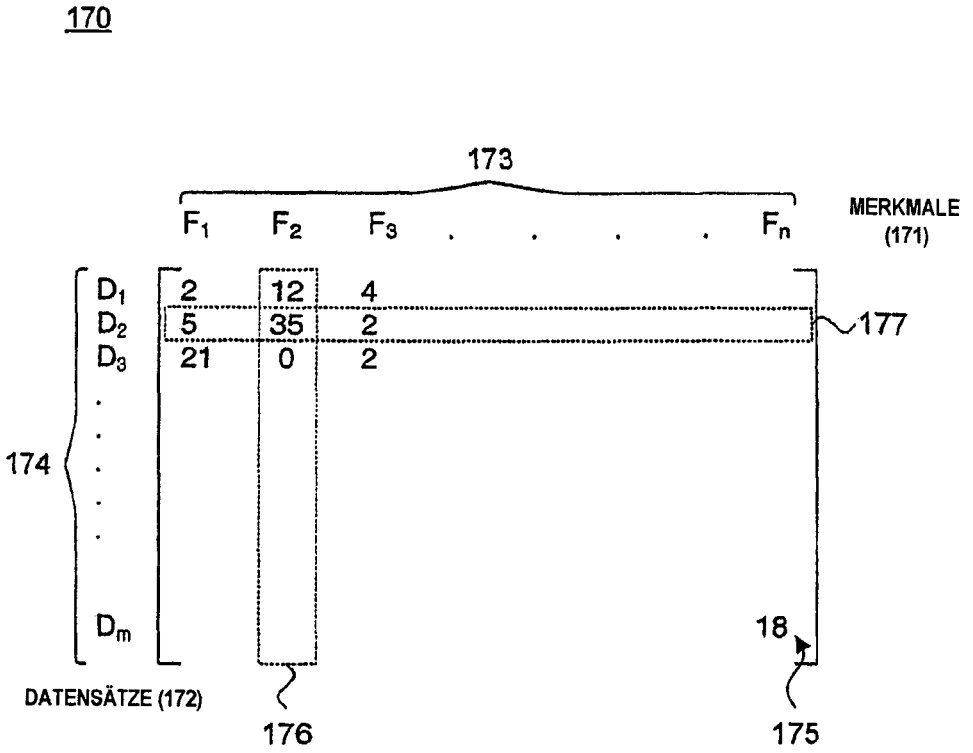

170
173
F1 F2 F3 . . . . Fn
MERKMALE (171)
D1 2 12 4
D2 5 35 2
D3 21 0 2
177
174
Dm
18
DATENSÄTZE (172)
176
175

Fig. 15.

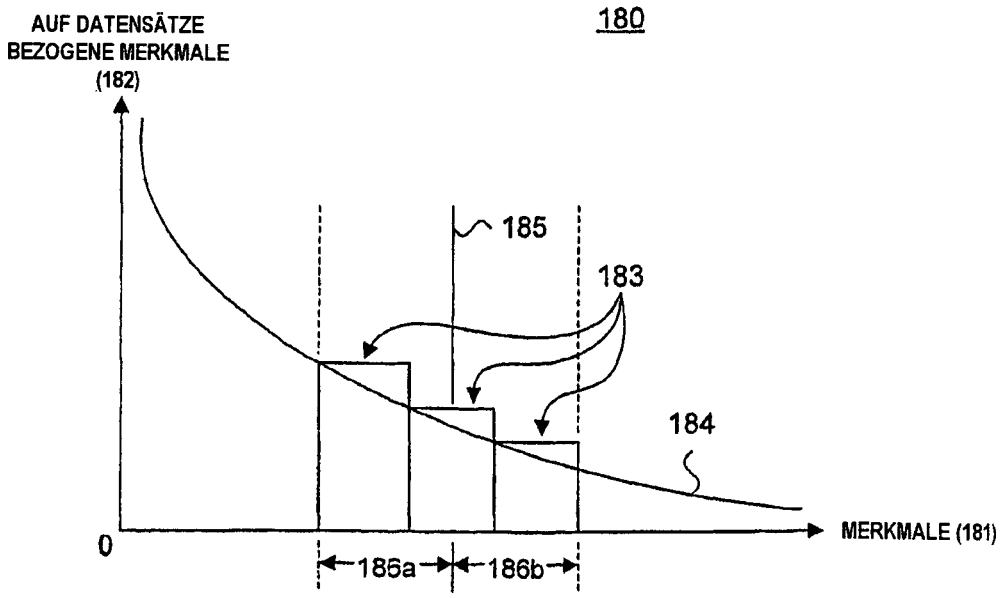

180
AUF DATENSÄTZE BEZOGENE MERKMALE (182)
185
183
184
0
MERKMALE (181)
186a
186b

Fig. 16.

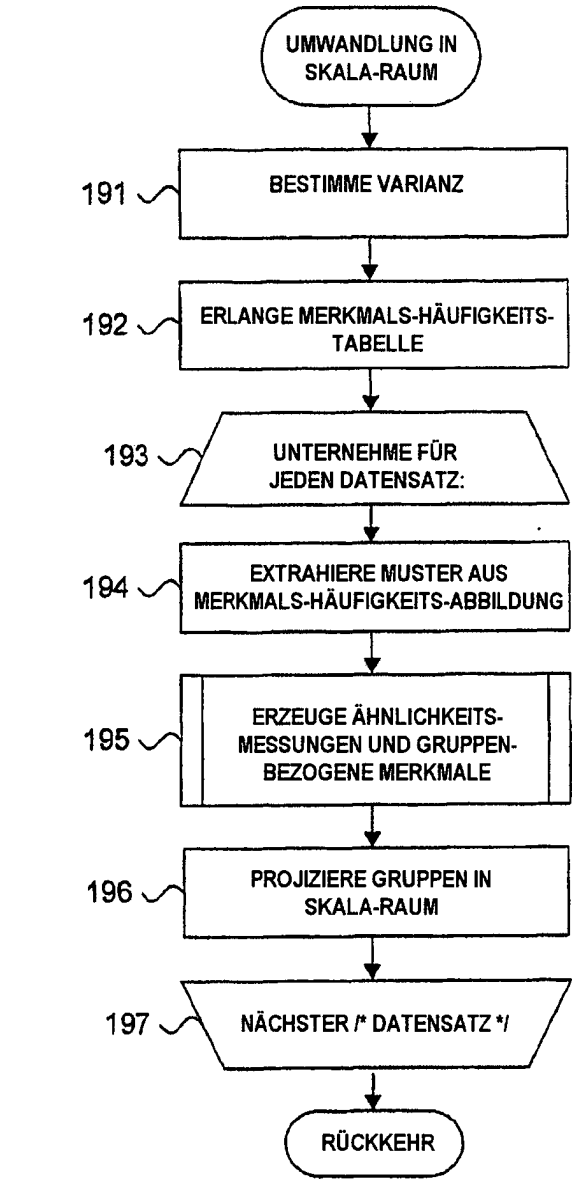

190
UMWANDLUNG IN SKALA-RAUM
191
BESTIMME VARIANZ
192
ERLANGE MERKMALS-HÄUFIGKEITS-TABELLE
193
UNTERNEHME FÜR JEDEN DATENSATZ:
194
EXTRAHIERE MUSTER AUS MERKMALS-HÄUFIGKEITS-ABBILDUNG
195
ERZEUGE ÄHNLICHKEITS-MESSUNGEN UND GRUPPEN-BEZOGENE MERKMALE
196
PROJIZIERE GRUPPEN IN SKALA-RAUM
197
NÄCHSTER /* DATENSATZ */
RÜCKKEHR

Fig. 17.

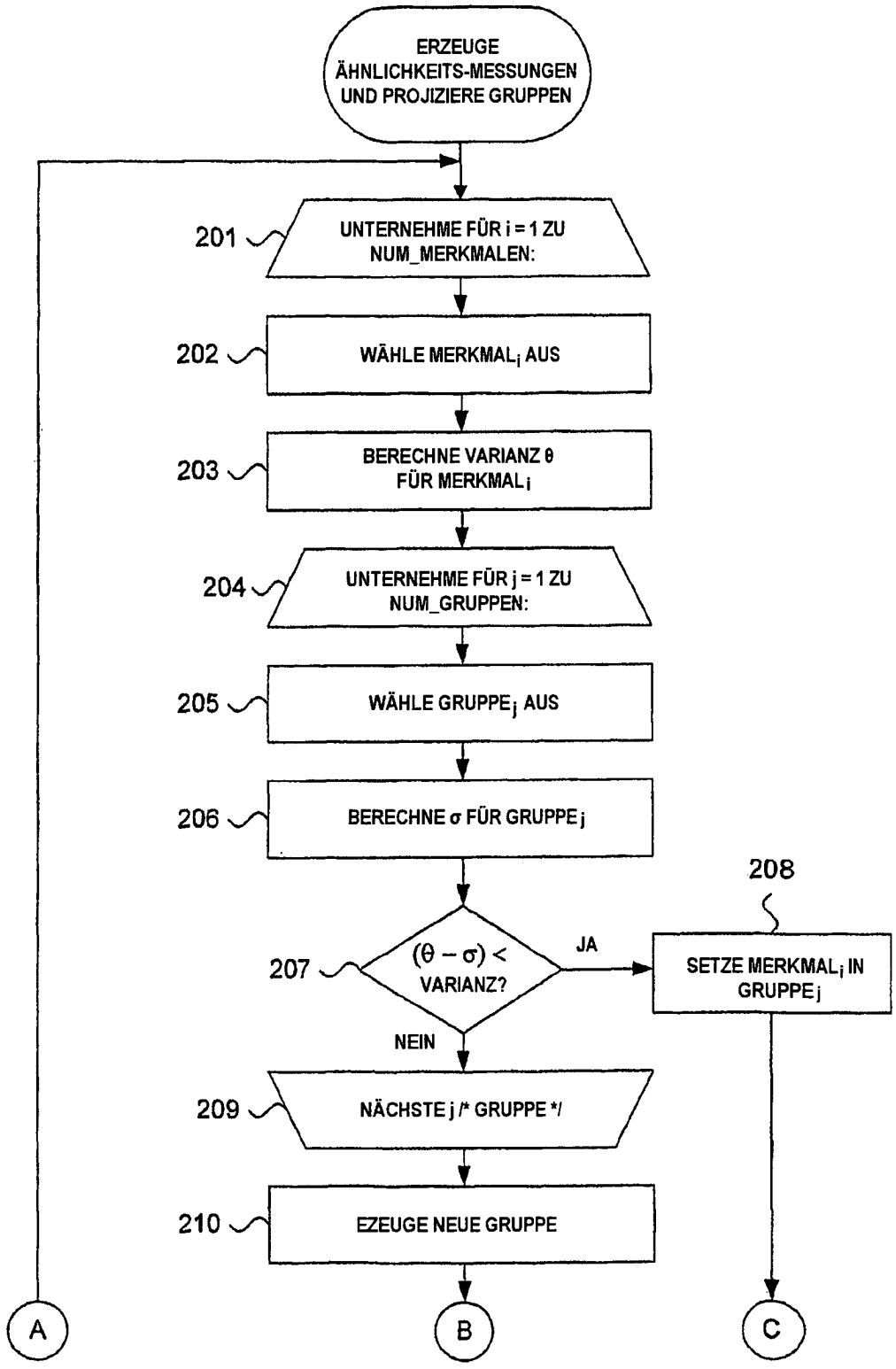

200
ERZEUGE ÄHNLICHKEITS-MESSUNGEN UND PROJIZIERE GRUPPEN
201
UNTERNEHME FÜR i = 1 ZU NUM_MERKMALEN:
202
WÄHLE MERKMAL$_i$ AUS
203
BERECHNE VARIANZ θ FÜR MERKMAL$_i$
204
UNTERNEHME FÜR j = 1 ZU NUM_GRUPPEN:
205
WÄHLE GRUPPE$_j$ AUS
206
BERECHNE σ FÜR GRUPPE$_j$
207
$(\theta - \sigma) <$ VARIANZ?
JA
208
SETZE MERKMAL$_i$ IN GRUPPE$_j$
NEIN
209
NÄCHSTE j /* GRUPPE */
210
EZEUGE NEUE GRUPPE
A
B
C

Fig. 10 (Fortsetzung)

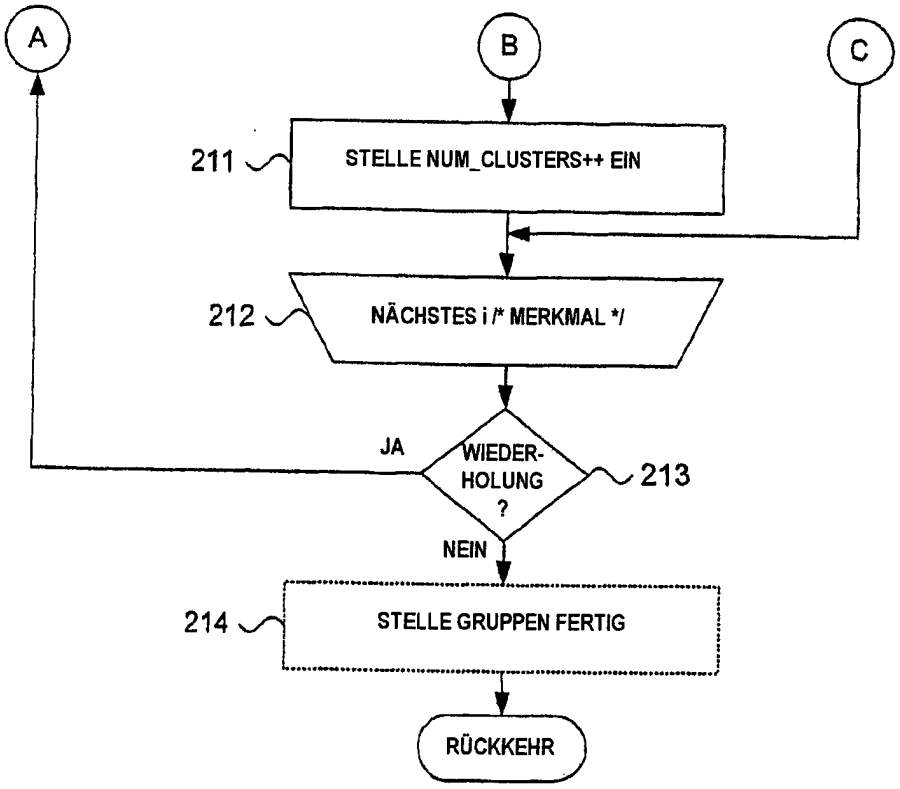


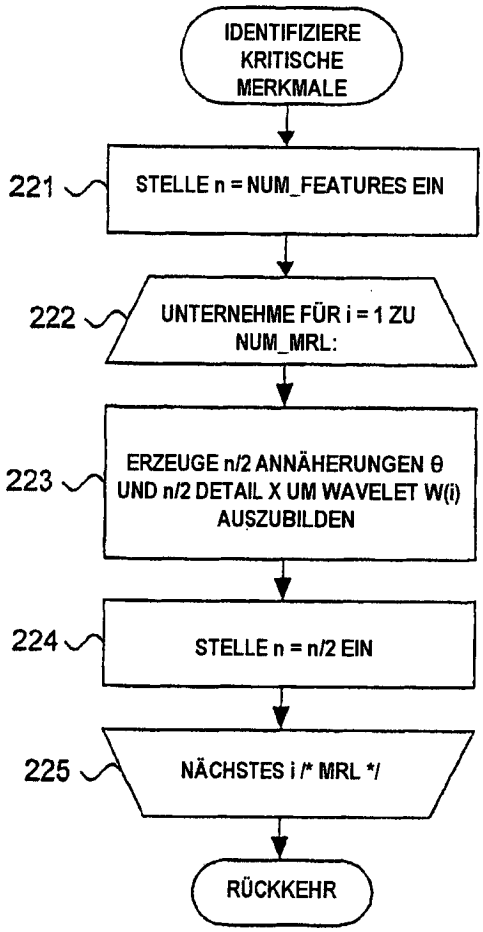
Fig. 18.

210

211

212	GRUPPE$_1$	GRUPPE$_2$	GRUPPE$_3$	GRUPPE$_4$	
MERKMAL$_1$	10	5	0	1	← 213a
MERKMAL$_2$	8	4	0	0	← 213b
MERKMAL$_3$	0	0	12	2	← 213c
.					
.					
.					
MERKMAL$_n$	0	0	17	3	← 213d

Fig. 19.


220
IDENTIFIZIERE KRITISCHE MERKMALE
221
STELLE n = NUM_FEATURES EIN
222
UNTERNEHME FÜR i = 1 ZU NUM_MRL:
223
ERZEUGE n/2 ANNÄHERUNGEN θ UND n/2 DETAIL X UM WAVELET W(i) AUSZUBILDEN
224
STELLE n = n/2 EIN
225
NÄCHSTES i /* MRL */
RÜCKKEHR