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

(10) **DE 698 38 763 T2** 2008.10.30

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 1 498 827 B1**
(21) Deutsches Aktenzeichen: **698 38 763.5**
(96) Europäisches Aktenzeichen: **04 024 427.9**
(96) Europäischer Anmeldetag: **04.12.1998**
(97) Erstveröffentlichung durch das EPA: **19.01.2005**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **21.11.2007**
(47) Veröffentlichungstag im Patentblatt: **30.10.2008**

(51) Int Cl.[8]: ***G06F 17/27*** (2006.01)

(30) Unionspriorität:
987565 **11.12.1997** **US**

(84) Benannte Vertragsstaaten:
AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LI, LU, MC, NL, PT, SE

(73) Patentinhaber:
Microsoft Corp., Redmond, Wash., US

(72) Erfinder:
Powell, Robert David, Issaquah, WA 98027, US

(74) Vertreter:
BOEHMERT & BOEHMERT, 28209 Bremen

(54) Bezeichnung: **IDENTIFIZIERUNG DER SPRACHE UND DES ZEICHENSATZES AUS TEXT-REPRÄSENTIERENDEN DATEN**

Beschreibung

[0001] Die Erfindung liegt auf dem Gebiet der Textanalyse, insbesondere der statistischen Textanalyse.

Hintergrund der Erfindung

[0002] Es ist für moderne Computersysteme üblich, Einrichtungen zum Speichern und zum Verarbeiten von Daten bereitzustellen, die einen Text darstellen. Datenkörper, die durch ein Computersystem gespeichert sind und ein Textdokument darstellen, werden als "digitale Dokumentdarstellungen" (digital document representations) bezeichnet. Digitale Dokumentdarstellungen werden in einem Computersystem wie andere Daten gespeichert, nämlich als Reihenfolge von Werten, die "Bytes" genannt werden. Text wird in diese Bytewerte mittels einer "Zeichenfolge" (character set) konvertiert, eine Abbildung zwischen den Werten der verschiedenen Zeichen, die in dieser Schrift folgend als Zeichenglypt bezeichnet werden, und den verschiedenen Bytewerten. Zeichenfolgen, die auch als "Codeseiten" (code pages) bezeichnet werden, werden im allgemeinen von Standardisierungsorganisationen, wie beispielsweise das American National Standards Institute ("ANSI") oder der International Standards Organisation ("ISO"), definiert. Einige Zeichenfolgen, die "Zeichenfolgen mit mehreren Bytes" (multiple-byte character sets) genannt werden, bilden jeden Zeichenglypt auf einen Wert ab, der aus zwei oder mehr Bytes besteht. Es ist im allgemeinen möglich, das Dokument korrekt anzuzeigen, das durch eine digitale Dokumentendarstellung dargestellt ist, wenn nur die Zeichenfolge bekannt ist, die verwendet wurde, um die digitale Dokumentendarstellung zu erzeugen. Es ist im allgemeinen ebenso möglich, eine digitale Dokumentendarstellung von ihrer momentanen Zeichenfolge in eine andere Zielzeichenfolge zu konvertieren, wenn nur die momentane Zeichenfolge der digitalen Dokumentendarstellung bekannt ist.

[0003] Ein Text umfaßt im allgemeinen eine Folge von Worten, die jeweils aus einer oder mehreren Sprachen stammen. Natursprachenverarbeitungswerkzeuge, wie beispielsweise eine Rechtschreibüberprüfung, eine Grammatiküberprüfung und Zusammenfasser, können auf solche Dokumente angewendet werden. Um ein Dokument jedoch richtig verarbeiten zu können, müssen diese Werkzeuge jedoch von der Sprache oder den Sprachen unterrichtet werden, aus denen die Worte in dem Dokument stammen. Wenn beispielsweise eine Rechtschreibprüfung das Wort "Bitte" in einem Dokument entdeckt, von dem bekannt ist, daß es in Deutsch verfaßt ist, betrachtet diese das Wort nicht als falsch geschrieben. Wenn jedoch die Rechtschreibprüfung dasselbe Wort in einem Dokument entdeckt, von dem bekannt ist, daß es in Englisch verfaßt ist, so betrachtet diese das Wort als eine fehlerhafte Schreibweise des Wortes "bitter". Einige Werkzeuge, die Informationen wiedergewinnen, wie beispielsweise Worttrenner (die die Grenzen zwischen Worten erkennen) und Wortstammbilder (die Zusätze entfernen, um verschiedene Wort aufeinander abzustimmen, die denselben Stamm haben), müssen ebenso von der Sprache oder den Sprachen unterrichtet werden, die in den digitalen Dokumentdarstellungen auftreten, an denen diese Werkzeuge angewendet werden. Zusätzlich zu den Bedürfnissen von automatisierten Werkzeugen, ist die Kenntnis der Sprache, in der das Dokument verfaßt ist, für menschliche Leser nützlich, die nur eine oder eine geringe Anzahl der großen Vielzahl von Sprachen lesen können, in denen Dokumente verfaßt werden, um zu bestimmen, ob sie das Dokument lesen können oder nicht.

[0004] Somit ist es im allgemeinen für jede digitale Dokumentendarstellung, die gespeichert wird, wünschenswert, von einer expliziten Anzeige der Zeichenfolge, die verwendet wurde, um diese zu erzeugen, und der Sprache oder den Sprachen, von der die Worte für diese stammen, begleitet zu werden. Während eine solche Information für viele digitale Dokumentendarstellungen, insbesondere den Darstellungen, die in der jüngeren Vergangenheit erzeugt wurden, gespeichert wird, ist diese für viele andere digitale Dokumentendarstellungen nicht verfügbar. Zum Beispiel können viele der HTML-Dokumente, die über das World Wide Web verfügbar sind, nicht ihre Zeichenfolgen und Sprachen anzeigen.

[0005] Bei einigen digitalen Dokumentendarstellungen wurde eine Information, die die Zeichenfolge und die Sprache der digitalen Dokumentendarstellung identifiziert, nie mit der digitalen Dokumentendarstellung verbunden. Dies ist häufig der Fall, wenn diese Information ursprünglich in dem Computer abgelegt wurde, auf dem diese gespeichert wurde. Zum Beispiel ist diese Information implizit in digitalen Dokumentendarstellungen enthalten, die ursprünglich in einer einzelnen Sprache und in einer einzelnen Zeichenfolgeumgebung erzeugt wurden. Wenn derartige digitale Dokumentendarstellungen in ein Computersystem überführt werden, das verschiedene Sprachen und Zeichenfolgen verwendet, oder solchen Computersystemen über ein Netzwerk, wie beispielsweise das Internet, zur Verfügung gestellt wird, so ist die Zeichenfolge und die Sprache von solchen digitalen Dokumentendarstellungen nicht verfügbar.

[0006] Bei anderen digitalen Dokumentendarstellungen, wurde eine Information, die in der Zeichenfolge und der Sprache der digitalen Dokumentendarstellung identifiziert wurde, zu einem gewissen Zeitpunkt mit der di-

gitalen Dokumentendarstellung verbunden, ist jedoch momentan nicht verfügbar. Zum Beispiel kann eine solche Information in einer getrennten Datei gespeichert worden sein, die zu einem gewissen Zeitpunkt gelöscht wurde. Auf der anderen Seite kann diese Information noch vorhanden sein, jedoch ebenso nicht zur Verfügung stehen. Zum Beispiel kann die Datei, die die Information enthält, für den Benutzer nicht zugänglich sein, oder ein Programm, das versucht, die Zeichenfolge und Sprache der digitalen Dokumentendarstellung zu bestimmen. Eine derartige Information kann ferner zugänglich sein, jedoch in einem Format vorliegen, das für den Benutzer unverständlich ist, oder ein Programm, das versucht, die Zeichenfolge und Sprache der digitalen Dokumentendarstellung zu bestimmen. Daher kann aus einer Vielzahl von Gründen die Zeichenfolge und die Sprache einer digitalen Dokumentendarstellung nicht verfügbar sein.

[0007] Da die erforderliche Sprache und Zeichenfolge, um digitale Dokumentendarstellungen anzuzeigen und zu verarbeiten, häufig nicht verfügbar ist, wäre ein automatisierter Zugang zum Bestimmen der Zeichenfolge und Sprache oder Sprachen einer digitalen Dokumentendarstellung, insbesondere einer, die akzeptable Speicheranforderungen mit sich bringt und einfach auf neue Zeichenfolgen und Sprachen erweiterbar ist, von großem Nutzen.

[0008] G. Kikui, und andere, beschreibt in "Cross-lingual Information Retrieval an the WWW", ECA196, 12th European Conference an Artificial Intelligence, MULSAIC96 Workshop, 1996, Seiten 1–6, ein sprachenübergreifendes Suchsystem, das auf zwei AI-basierenden Modulen beruht: ein Sprachidentifizierungsmodul und ein Maschinenübersetzungsmodul. Das Sprachidentifizierungsmodul setzt statistische und regelbasierte Musterabgleiche zur Identifizierung der Sprache eines Dokumentes ein, während das Maschinenübersetzungsmodul sowohl Sucheinträge des Anwenders als auch die Titel von gefundenen Dokumenten übersetzt.

[0009] G. Kikui beschreibt in "Identifying the coding system and language of on-line documents an the Internet", 16th International Conference of Computational Linguistics (COLING), August 1996, Seiten 652–657, einen Algorithmus, der gleichzeitig das Codierungssystem und die Sprache eines Codestringes identifiziert, der aus dem Internet geholt wurde. Der Algorithmus verwendet statistische Sprachmodelle, um den korrekt decodierten String auszuwählen und um die Sprache zu bestimmen.

[0010] US-Patent Nr. 5,418,951 beschreibt ein Verfahren zum Identifizieren, Wiedergewinnen oder Sortieren von Dokumenten nach Sprache oder Thema. Das Verfahren umfaßt das Erzeugen eines n-Gramm-Arrays für jedes Dokument in einer Datenbank, Parsen eines nicht identifizierten Dokumentes oder eines Sucheintrages in n-Gramme, Zuweisen eines Gewichtes an jedes n-Gramm, Entfernen der Gemeinsamkeiten der n-Gramme, Vergleichen des nicht identifizierten Dokuments oder Sucheintrages mit jedem Dokument der Datenbank, Auswerten der Ähnlichkeit jedes Vergleiches und basierend auf den Ähnlichkeitswerten, Identifizieren, Wiedergewinnen oder Sortieren des Dokumentes oder des Sucheintrags.

Zusammenfassung der Erfindung

[0011] In einem ersten Aspekt stellt die vorliegende Erfindung das Verfahren nach Anspruch 1 zum Identifizieren einer unbekannten Sprache und eines unbekannten Zeichensatzes einer Textkette bereit.

[0012] In einem zweiten Aspekt stellt die vorliegende Erfindung das Verfahren nach Anspruch 5 zum Identifizieren einer jeden einer Vielzahl von Sprachen, die in einer Darstellung eines digitalen Dokuments vorkommen, bereit.

[0013] Die Erfindung ermöglicht das Vorsehen einer Softwareeinrichtung ("die Einrichtung"), die, wenn ihr Rohdaten gegeben werden, die ein Textdokument darstellen, das in irgendeiner Sprache mittels irgendeiner Zeichenfolge verfaßt ist, automatisch die Sprache und die Zeichenfolge identifiziert. Die Einrichtung erzeugt zuerst ein statistisches Modell des Dokumentes in einer jeden einer Anzahl von bekannten Sprachen und Zeichenfolgen in einer "Trainingsphase", und wendet diese Modelle sodann an, um die Sprache und Zeichenfolge eines eingegebenen Dokumentes in einer "Erkennungsphase" zu identifizieren.

[0014] Die von der Einrichtung verwendeten statistischen Modelle der Sprachen und Zeichenfolgen sind angepaßt, um Zeichenwerte hervorzuheben, die dazu neigen, zwischen verschiedenen Sprachen und Zeichenfolgen unterschiedlich zu sein, während gleichzeitig die Datenmenge in den Modellen, die den Zeichenwerten zugeordnet ist, die nicht dazu neigen, zwischen Sprachen und Zeichenfolgen unterschiedlich zu sein, mittels einer spezialisierten, reduktiven Abbildung verringert wird.

[0015] In der Trainingsphase erzeugt die Einrichtung die statistischen Modelle der Sprachen und Zeichenfol-

gen, die in der Erkennungsphase verwendet werden, um die Sprache und Zeichenfolge der eingegebenen Dokumente zu identifizieren. Für jede zu untersuchende Kombination aus Sprache und Zeichenfolge liest die Einrichtung Beispieldokumente, von denen die Sprache und Zeichenfolge bekannt ist, die jeweils eine Folge von Bytewerten umfassen. Die Einrichtung führt sodann die Schritte aus: (A) Abbilden der Bytewerte aus der Folge von 256 verschiedenen, möglichen Werten, die in einem Byte dargestellt werden können auf eine kleinere Anzahl von möglichen Werten, (B) Aufzeichnen der Häufigkeiten, mit der jede verschiedene, gleichlange Folge von abgebildeten Bytewerten, oder n-Grammen, in der abgebildeten Version des Dokumentes auftritt, (C) Zusammenfassen dieser Häufigkeitsverteilungen für jede Sprache und jede Zeichenfolge, und (D) Normalisieren der Häufigkeitsverteilungen über die Sprachen und die Zeichenfolgen. Bevorzugt werden n-Gramme verschiedener Länge wie auch verschiedene Abbildungen verwendet, abhängig von den Charakteristika einer jeden Sprache und einer jeden Zeichenfolge. Die Einrichtung stellt ferner die Häufigkeitsverteilungen ein, um n-Gramme mit hoher Häufigkeit zu betonen, deren Auftreten dazu neigt, ein Dokument in einer Sprache und einer Zeichenfolge von Dokumenten in anderen Sprachen und anderen Zeichenfolgen zu unterscheiden, und häufig auftretende n-Gramme abzuwerten, deren Auftreten nicht dazu führt, ein Dokument einer Sprache und einer Zeichenfolge von Dokumenten anderer Sprachen und Zeichenfolgen zu unterscheiden.

[0016] In der Erkennungssphase verwendet die Einrichtung die normalisierten und eingestellten Haufigkeitsverteilungen, die in der Trainingsphase erzeugt wurden, um die Sprache und Zeichenfolge für ein eingegebenes Dokument zu identifizieren, dessen Sprache und/oder Zeichenfolge unbekannt ist. Die Einrichtung bildet zuerst die Bytewerte eines eingegebenen Dokumentes auf eine kleine Anzahl von möglichen Werten ab, wobei dieselbe(n) Abbildung(en) wie in der Trainingsphase verwendet wird (werden). Die Einrichtung bestimmt sodann die Häufigkeit, mit der jedes n-Gramm in dem abgebildeten, eingegebenen Dokument auftritt. Für jede Sprache wird die Häufigkeit, mit der jedes n-Gramm in dem abgebildeten, eingegebenen Dokument auftritt, mit der Häufigkeit desselben n-Gramm, die in den Trainingshäufigkeitsverteilungen für die momentane Sprache ermittelt wurden, multipliziert und diese Produkte werden addiert. Die Summe für jede Sprache bildet die relative Wahrscheinlichkeit, daß die Sprache diejenige ist, in der das eingegebene Dokument geschrieben ist. Nachdem die Sprache mit der größten Summe bestimmt wurde, wird dieses Erkennungsverfahren wiederholt, wobei die Zeichenfolgeverteilungen für diese Sprache verwendet werden, um festzustellen, in welcher der bekannten Zeichenfolgen für die identifizierte Sprache das eingegebene Dokument geschrieben ist.

[0017] Die Einrichtung analysiert bevorzugt aufeinanderfolgende Einheiten von Trainingsdokumenten und eingegebenen Dokumenten, deren Länge der Länge eines typischen Absatzes entspricht. Wenn ein Dokument Text in mehr als einer Sprache oder einer Zeichenfolge enthält, kann die Einrichtung auf diese Weise jeder dieser Sprachen und Zeichenfolgen identifizieren.

[0018] Die Einrichtung verwendet ferner bevorzugt ähnliche statistische Analysetechniken, um die Sprache oder Sprachen der digitalen Dokumentendarstellungen zu identifizieren, die in der großen Unicodezeichenfolge erstellt sind. Beim Identifizieren einiger Gruppen von ähnlichen Sprachen in digitalen Unicodedokumentendarstellungen verwendet die Einrichtung bevorzugt anwenderspezifische reduktive Abbildungen, die automatisch erzeugt werden, um zwischen Sprachen in der Gruppe zu unterscheiden.

Beschreibung von bevorzugten Ausführungsbeispielen

[0019] Die Erfindung wird im folgenden anhand von Ausführungsbeispielen unter Bezugnahme auf Figuren einer Zeichnung näher erläutert. Hierbei zeigen:

[0020] Fig. 1 ist ein detailliertes Blockdiagramm, daß das Computersystem zeigt, auf dem die Einrichtung bevorzugt ausgeführt wird.

[0021] Fig. 2 ist ein Übersichtsflußdiagramm, das die Phasen darstellt, in denen die Einrichtung arbeitet.

[0022] Fig. 3 ist ein Flußdiagramm, das die Schritte zeigt, die von der Einrichtung in der Trainingsphase bevorzugt ausgeführt werden.

[0023] Fig. 4 ist ein Flußdiagramm, das die Schritte zeigt, die von der Einrichtung bevorzugt ausgeführt werden, um die Modelle für die Zeichenfolge und die primäre Sprache zu erweitern, die für eine bestimmte digitale Beispieldokumentendarstellung entsprechend Schritt **304** identifiziert wurden.

[0024] Fig. 5 ist ein Flußdiagramm, das die Schritte zeigt, die von Einrichtung bevorzugt ausgeführt werden, um eine Charakterisierung einer digitalen Beispieldokumentendarstellung zu erzeugen.

[0025] Fig. 6 ist ein Flußdiagramm, das die Schritte zeigt, die von der Einrichtung bevorzugt ausgeführt werden, um die Zeichenfolge und Sprache einer eingegebenen digitalen Dokumentendarstellung, für die diese Information nicht bekannt ist, zu identifizieren.

[0026] Fig. 7 ist ein Flußdiagramm, das die Schritte zeigt, die bevorzugt von der Einrichtung ausgeführt werden, um die Zeichenfolge und/oder Sprache der eingegebenen digitalen Dokumentendarstellung zu erkennen.

[0027] Fig. 8 ist ein Flußdiagramm, das die Schritte zeigt, die von der Einrichtung bevorzugt ausgeführt werden, um mehrere Sprachen wie auch Zeichenfolgen einer eingegebenen digitalen Dokumentendarstellung zu erkennen.

[0028] Fig. 9 ist ein Flußdiagramm, das die Schritte zeigt, die von der Einrichtung bevorzugt ausgeführt werden, um eine Zeichenfolge für die eingegebene, digitale Dokumentendarstellung von den Zeichenfolgen auszuwählen, die für Segmente der eingegebenen, digitalen Dokumentendarstellung bestimmt wurden.

[0029] Fig. 10 ist ein Flußdiagramm, das die Schritte zeigt, die bevorzugt von der Einrichtung in der Trainingsphase ausgeführt werden, um eine anwenderspezifische reduktive Abbildung für eine Unicodesprachengruppe zu erstellen, die verwendet wird, um zwischen Sprachen der Sprachengruppe zu unterscheiden.

[0030] Fig. 11 ist ein Flußdiagramm, das die Schritte zeigt, die von der Einrichtung bevorzugt aus geführt werden, um die Sprachen zu erkennen, die in einer eingegebenen digitalen Dokumentendarstellung verwendet werden, die in einer Unicodezeichenfolge erstellt ist.

[0031] Fig. 12 ist ein Flußdiagramm, das die Schritte zeigt, die bevorzugt von der Einrichtung ausgeführt werden, um die Sprache eines Segmentes einer eingegebenen digitalen Dokumentendarstellung zu erkennen, die in der Unicodezeichenfolge erstellt ist.

[0032] Die vorliegende Erfindung stellt eine Softwareeinrichtung ("die Einrichtung") bereit, die, wenn ihre Rohdaren zugeführt werden, die ein Textdokument darstellen, das in irgendeiner Sprache mit irgendeiner Zeichenfolge erstellt wurde, automatisch diese Sprache und diese Zeichenfolge identifiziert. Die Einrichtung erzeugt zuerst ein statistisches Modell der Dokumente in jeder aus einer Vielzahl von bekannten Sprachen und Zeichenfolgen in einer "Trainingsphase" und wendet diese Modelle dann an, um die Sprache und Zeichenfolge eines eingegebenen Dokumentes in einer "Erkennungsphase" zu identifizieren.

[0033] Die statistischen Modelle der Sprachen und Zeichenfolgen, die von der Einrichtung verwendet werden, sind angepaßt, um Zeichenwerte zu betonen, die dazu führen, daß zwischen verschiedenen Sprachen und Zeichenfolgen unterschieden werden kann, während gleichzeitig der Speicherplatz in den Modellen, der Zeichenwerten zugewiesen ist, die nicht dazu führen, daß zwischen Sprachen oder Zeichenfolgen unterschieden werden kann, durch die Verwendung von spezialisierten reduktiven Abbildungen minimiert wird.

[0034] In der Trainingsphase erzeugt die Einrichtung die statistischen Modelle der Sprachen und Zeichenfolgen, die in der Erkennungsphase verwendet werden, um die Sprache und Zeichenfolge der eingegebenen Dokumente zu identifizieren. Für jede Kombination aus Sprache und Zeichenfolge, die untersucht werden soll, liest die Einrichtung Beispieldokumente, von denen die Sprache und Zeichenfolge bekannt ist, die jeweils eine Reihe von Bytewerten umfassen. Die Einrichtung führt sodann die folgenden Schritte aus: (A) Abbilden der Bytewerte aus der Folge von 256 verschiedenen, möglichen Werten, die in einem Byte dargestellt werden können, auf eine kleinere Anzahl von möglichen Werten, (B) Aufzeichnen der Häufigkeiten, mit der jede verschiedene Folge von abgebildeten Bytewerten fester Länge, oder "n-Gramme", in der abgebildeten Version des Dokumentes auftritt, (C) Zusammenführen dieser "Häufigkeitsverteilungen" für jede Sprache und Zeichenfolge und (D) Normalisieren der Häufigkeitsverteilungen über die Sprachen und Zeichenfolgen. Bevorzugt werden n-Gramme verschiedener Längen wie auch verschiedene Abbildungen verwendet, abhängig von den Charakteristika jeder Sprachen und Zeichenfolge. Die Einrichtung stellt ferner die Häufigkeitsverteilungen ein, um n-Gramme mit hohen Häufigkeitswerten zu betonen, deren Auftreten dazu führt, daß ein Dokument einer Sprache und Zeichenfolge von Dokumenten anderer Sprachen und Zeichenfolgen unterscheidbar ist, und n-Gramme mit hohen Häufigkeiten abzuwerten, deren Auftreten nicht dazu führt, daß ein Dokument einer Sprache und Zeichenfolge von Dokumenten anderer Sprachen und Zeichenfolgen unterscheidbar ist.

[0035] In der Erkennungsphase verwendet die Einrichtung die normalisierten und eingestellten Häufigkeitsverteilungen, die in der Trainingsphase erzeugt wurden, um die Sprache und Zeichenfolge für ein eingegebenes Dokument zu identifizieren, dessen Sprache und/oder Zeichenfolge unbekannt ist. Die Einrichtung bildet

zunächst die Bytewerte des eingegebenen Dokumentes auf eine kleinere Anzahl von möglichen Werten mittels der gleichen Abbildung(en) ab, die in der Trainingsphase verwendet wurden. Die Einrichtung bestimmt sodann die Häufigkeiten, mit der jedes n-Gramm in dem abgebildeten, eingegebenen Dokument auftritt. Für jede Sprache wird die Häufigkeit, mit der jedes n-Gramm in dem abgebildeten, eingegebenen Dokument vorkommt, multipliziert mit der Häufigkeit, mit der dasselbe n-Gramm in den Trainingshäufigkeitsverteilungen für dieselbe Sprache vorkam, und diese Produkte werden addiert. Die Summe für jede Sprache stellt die relative Wahrscheinlichkeit dar, daß die Sprache diejenige Sprache ist, in der das eingegebene Dokument geschrieben ist. Nachdem die Sprache mit der höchsten Summe identifiziert wurde, wird dieser Erkennungsprozeß wiederholt, wobei die Zeichenfolgenverteilungen für diese Sprache verwendet werden, um festzustellen, in welcher der bekannten Zeichenfolgen für die identifizierte Sprache das eingegebene Dokument geschrieben ist.

[0036] Die Einrichtung analysiert bevorzugt aufeinanderfolgende Einheiten von Trainingsdokumenten und eingegebenen Dokumenten, deren Länge der Länge eines typischen Absatzes entspricht. Wenn ein Dokument Text in mehr als einer Sprache oder einer Zeichenfolge enthält, kann die Einrichtung auf diese Weise jede dieser Sprachen und Zeichenfolgen identifizieren.

[0037] Die Einrichtung verwendet ferner bevorzugt ähnliche statistische Analysetechniken, um die Sprache oder Sprachen der digitalen Dokumentendarstellungen zu identifizieren, die in der großen Unicodezeichenfolge erstellt sind. Beim Identifizieren einiger Gruppen ähnlicher Sprachen in den digitalen Unicodedokumentendarstellungen verwendet die Einrichtung bevorzugt anwenderspezifische reduktive Abbildungen, die automatisch erzeugt werden, um zwischen Sprachen in der Gruppe zu unterscheiden.

[0038] Fig. 1 ist ein detailliertes Blockdiagramm, daß das Computersystem zeigt, auf dem die Einrichtung bevorzugt ausgeführt wird. Wie in Fig. 1 gezeigt, umfaßt das Computersystem **100** eine zentrale Prozessoreinheit (CPU) **110**, Eingabe/Ausgabe-Einrichtungen **120** und einen Computerspeicher (Speicher) **130**. Ein Element der Eingabe/Ausgabe-Einrichtungen ist eine Speichereinrichtung **121**, wie beispielsweise ein Festplattenlaufwerk, ein weiteres Element ist ein computerlesbares Medienlaufwerk **122**, das verwendet werden kann, um Softwareprodukte zu installieren, umfassend die Einrichtung, die auf einem computerlesbaren Medium, wie beispielsweise einer CD-ROM, bereitgestellt sind, ein weiteres Element ist eine Netzwerkverbindung **123**, um das Computersystem **100** mit anderen Computersystemen (nicht dargestellt) zu verbinden. Der Speicher **130** umfaßt bevorzugt eine Einrichtung **131**, digitale Beispieldokumentendarstellungen **132**, die verwendet werden, um Modelle der Zeichenfolgen und Sprachen **133** zu erstellen, und eine eingegebene digitale Dokumentendarstellung, deren Zeichenfolge und Sprache(n) durch die Einrichtung zu bestimmen ist. Während die Einrichtung bevorzugt auf einem Computersystem für allgemeine Zwecke implementiert ist, das wie zuvor beschrieben konfiguriert ist, werden Fachleute erkennen, daß diese ebenso auf Computersystemen mit anderen Konfigurationen implementiert sein kann.

[0039] Fig. 2 ist ein Übersichtsflußdiagramm, das die Phasen zeigt, in denen die Einrichtung arbeitet. In Schritt **201**, in der Trainingsphase, erstellt die Einrichtung Modelle der Zeichenfolgen und Sprachen, die untersucht werden sollen, wobei digitale Beispieldokumentdarstellungen verwendet werden, von denen bekannt ist, daß diese in diesen Zeichenfolgen erzeugt wurden, und von denen bekannt ist, daß diese hauptsächlich diese Sprachen enthalten. Die digitalen Beispieldokumentdarstellungen, die in Schritt **201** verwendet werden, sind bevorzugt Zeitungsausschnitte oder rechtliche Meinungen, die Themen mit im wesentlich einem Gegenstand diskutieren, für den dessen primäre Sprache kennzeichnend ist. Das heißt es werden digitale Beispieldokumentendarstellungen gesucht, die einen begrenzten Inhalt haben, wie beispielsweise Zeitungsartikel und juristische Meinungen, da das Vokabular, das in solchen digitalen Dokumentendarstellungen verwendet wird, wahrscheinlich dazu geeignet ist, die primäre Sprache der digitalen Beispieldokumentendarstellung von anderen Sprachen zu unterscheiden. Der Schritt **201** wird detaillierter weiter unten in Verbindung mit Fig. 3 beschrieben.

[0040] Die Einrichtung führt den Schritt **201** einmal aus. Nachdem der Schritt **201** ausgeführt wurde, wird die Trainingsphase beendet. Die in der Trainingsphase im Schritt **201** erzeugten Modelle werden in der Erkennungsphase verwendet, um die Zeichenfolge und Sprache(n) in jeder aus irgendeiner Anzahl von eingegebenen digitalen Dokumentendarstellungen zu erkennen. In Schritt **202** empfängt die Einrichtung eine neue eingegebene digitale Dokumentendarstellung. In Schritt **203** untersucht die Einrichtung die Zeichenfolge und Sprache(n) der eingegebenen digitalen Dokumentendarstellung, die in Schritt **202** empfangen wurde, wobei die Modelle verwendet werden, die in Schritt **201** erzeugt wurden. Die Details des Schritt **203** werden weiter unten genauer in Verbindung mit Fig. 3 beschrieben. Nachdem die Zeichenfolge und Sprache(n) der eingegebenen digitalen Dokumentendarstellung in Schritt **203** untersucht wurde, fährt die Einrichtung in Schritt **202** fort, die nächste eingegebene digitale Dokumentendarstellung zu empfangen, wodurch diese in der Erken-

nungsphase verbleibt. Während Fig. 2 die Einrichtung so darstellt, als ob diese niemals erneut in die Trainingsphase eintreten würde, werden Fachleute erkennen, daß es wünschenswert sein kann, die Trainingsphase zu wiederholen, um zusätzliche digitale Beispieldokumentendarstellungen den Modellen, die in der Trainingsphase erzeugt wurden, hinzuzufügen, entweder um die Einrichtung zu erweitern, so daß diese zusätzliche Sprachen und Zeichenfolgen erkennen kann, oder um die Modelle für existierende Sprachen und Zeichenfolgen zu verbessern, wobei der Text von zusätzlichen digitalen Beispieldokumentendarstellungen verwendet wird.

[0041] Fig. 3 ist ein Flußdiagramm, das die Schritte zeigt, die von der Einrichtung in der Trainingsphase bevorzugt ausgeführt werden. In Schritt **301** wählt die Einrichtung digitale Beispieldokumentendarstellungen aus, deren Zeichenfolgen und primäre Sprachen bekannt sind. Da die Anzahl der digitalen Beispieldokumentendarstellungen direkten Einfluß auf die statistische Genauigkeit der Modelle, die durch die Einrichtung erzeugt werden, hat, und somit auf das Performanceniveau der Einrichtung, wird bevorzugt eine große Anzahl von digitalen Beispieldokumentendarstellungen im Schritt **301** ausgewählt. In den Schritten **302** bis **304** durchläuft die Einrichtung für jede digitale Beispieldokumentendarstellung, die in Schritt **301** ausgewählt wurde, eine Schleife. In Schritt **303** erweitert die Einrichtung das Modell für die Zeichenfolge und die primäre Sprache, die für die digitale Beispieldokumentendarstellung identifiziert wurden. Schritt **303** wird weiter unten genauer in Verbindung mit Fig. 4 beschrieben. Wenn in Schritt **304** eine oder mehrere ausgewählte digitale Dokumentendarstellungen zur Verarbeitung übrigbleiben, führt die Einrichtung mit Schritt **302** fort, die nächste ausgewählte digitale Beispieldokumentendarstellung zu verarbeiten. Sind alle ausgewählten digitalen Beispieldarstellungen verarbeitet, fährt die Einrichtung mit Schritt **305** fort. In Schritt **305** normalisiert die Einrichtung die Modelle, die als Ergebnis der Erweiterung in Schritt **303** erzeugt wurden. Die im Schritt **305** ausgeführte Normalisierung beinhaltet bevorzugt ein proportionales Erhöhen oder Erniedrigen der Zähler eines jeden Modells, so daß der gemittelte Zählwert in jedem Modell gleich ist. Im Schritt **306** stellt die Einrichtung die normalisierten Modelle ein, um deren Zeichenfolgen und Sprachen effektiver unterscheiden zu können, wobei für jedes Modell die folgenden Schritte ausgeführt werden: (a) Erhöhen der Häufigkeiten für n-Gramme, deren Häufigkeiten im Vergleich zu den Häufigkeiten für das gleiche n-Gramm in anderen Modellen am höchsten ist, und (b) Erniedrigen der Häufigkeiten für n-Gramme, deren Häufigkeiten im Vergleich mit den Häufigkeiten für dieses Modell hoch sind, jedoch nicht hoch sind im Vergleich zu den Häufigkeiten für dasselbe n-Gramm in den anderen Modellen. In einer bevorzugten Ausführungsform ist dieses Einstellen begleitet von einem Verdoppeln einer jeden Häufigkeit eines jeden Modells und folgendem Subtrahieren des Mittelwertes der entsprechenden Häufigkeiten über alle Modelle. Nach Schritt **306** sind diese Schritte und die Trainingsphase beendet.

[0042] Fig. 4 ist ein Flußdiagramm, das die Schritte zeigt, die von der Einrichtung bevorzugt ausgeführt werden, um die Modelle für die Zeichenfolge und primäre Sprache zu erweitern, die für eine bestimmte digitale Beispieldokumentendarstellung in Übereinstimmung mit Schritt **304** identifiziert wurden. Die Einrichtung verwendet bevorzugt verschiedene reduktive Abbildungen und behält Häufigkeitsdaten für verschieden große n-Gramme, abhängig von der Art der Zeichenfolge, die verwendet wurde, um die betrachtete digitale Beispieldokumentendarstellung zu erzeugen. Wenn die betrachtete digitale Beispieldokumentendarstellung in Schritt **401** eine romanische Einzelbytezeichenfolge ist, fährt die Einrichtung in Schritt **402** fort, die Schritte **402** bis **405** auszuführen, andernfalls ist die Zeichenfolge der betrachteten digitalen Beispieldokumentendarstellung entweder eine Zeichenfolge mit mehreren Bytes (bei der jedes Zeichen durch zwei oder mehr Bytewerte dargestellt ist) oder eine Einzelbytezeichenfolge für eine nicht romanische Sprache und die Einrichtung führt in Schritt **406** fort, die Schritte **406** und **407** auszuführen. Beispiele von romanischen Einzelbytezeichenfolgen umfassen: Microsoft Zeichenfolgen 1250 und 28592 für tschechisch, ungarisch und polnisch, Microsoft Zeichenfolge 1252 für dänisch, holländisch, englisch, finnisch, französisch, deutsch, italienisch, norwegisch, portugiesisch, spanisch und schwedisch, Microsoft Zeichenfolge 28591 für finnisch, französisch, deutsch, norwegisch, portugiesisch und schwedisch, Microsoft Zeichenfolge 28592 für deutsch, ungarisch und polnisch und Microsoft Zeichenfolge 1257 für litauisch. Beispiele für Zeichenfolgen mit mehreren Bytes oder nicht romanische Einzelbytezeichenfolgen umfassen Microsoft Zeichenfolge **936** und **950** für chinesisch, Microsoft Zeichenfolgen **932** und 51932 für japanisch, Microsoft Zeichenfolge **949** für koreanisch und Zeichenfolgen für griechisch und hebräisch.

[0043] In den Schritten **402** bis **405** erweitert die Einrichtung sowohl ein dreidimensionales Modell der primären Sprache der digitalen Beispieldokumentendarstellung wie auch ein eindimensionales Modell der Kombination aus dieser primären Sprache und der Zeichenfolge der digitalen Beispieldokumentendarstellung mit Charakteristika der digitalen Beispieldokumentendarstellung. Das dreidimensionale Modell wird von der Einrichtung verwendet, um eingegebene digitale Dokumentendarstellungen zu erkennen, die dieselbe primäre Sprache wie diese digitale Beispieldokumentendarstellung aufweist, wohingegen das eindimensionale Modell verwendet wird, um eingegebene digitale Dokumentendarstellungen zu erkennen, die dieselbe primäre Sprache und Zeichenfolge wie diese digitale Beispieldokumentendarstellung aufweist. In Schritt **402** erzeugt die Ein-

richtung eine dreidimensionale Charakterisierung der betrachteten digitalen Beispieldokumentendarstellung, wobei die Abbildung, die unten in Tabelle 1 gezeigt ist, und Häufigkeitsverteilungen für Zeichentrigrame, die in der digitalen Beispieldokumentendarstellung auftreten, verwendet werden.

Quellwert(e)	Zielwert
0 × 00–0 × 40, 0 × 5B–0 × 60, 0×7B – 0×7F	0
0 × 80-0XFF	1
0 × 41, 0 × 61	2
0 × 42, 0 × 62	3
0 × 43, 0 × 63	4
...	...
0 × 5A, 0 × 7A	27

Tabelle 1: Abbildung, um eine Sprache mit romanischem SBCS zu erkennen

[0044] Die Details des Schritts **402** zum Erzeugen einer Charakterisierung der digitalen Beispieldokumentendarstellung werden weiter unten genauer in Verbindung mit **Fig. 5** beschrieben. Als Teil des Erzeugungsverfahrens bildet die Einrichtung jeden Bytewert in der digitalen Beispieldokumentendarstellung auf einen Zielwert ab, wobei die in Tabelle 1 gezeigte Abbildung verwendet wird. Basierend auf einer Feststellung, daß Sprachen, die in einer romanischen Einzelbytezeichenfolge erstellt werden, am effektivsten abhängig von den Identitäten der Buchstaben unterschieden werden können, die darin vorkommen, jedoch unabhängig von dem Kasus, bildet die in Tabelle 1 gezeigte Abbildung die Quellwerte, die jedem Buchstaben entsprechen, auf verschiedene Zielwerte ab. Zum Beispiel werden die Quellwerte 0 × 41 und 0 × 61, die den Zeichenglypten "A" und "a" entsprechen, beide auf den Zielwert 2 abgebildet, die Quellwerte 0 × 42 und 0 × 62, die die Zeichenglypte "B" und "b" darstellen, werden auf den Zielwert 3 abgebildet, usw. Bis auf die Tatsache, daß den Zielwerten Buchstaben zugeordnet werden, definiert die in Tabelle 1 gezeigte Abbildung nur zwei weitere Zielwerte: den Zielwert 1 für "ausgedehnte Zeichen", das heißt für Quellwerte die ihre hohe Bitfolge haben, das heißt 0 × 80–0 × FF, und den Zielwert 0 für Quellwerte, die keine hohe Bitfolge aufweisen und keine Buchstaben darstellen, wobei diese Quellwerte im allgemeinen Zahlen, Interpunktionszeichen und anderen Symbolen sowie Steuerzeichen zugewiesen sind. Durch Abbilden von Quellwerten in dieser Weise reduziert die Einrichtung 256 Quellwerte auf nur 28 Zielwerte, wodurch die Speicheranforderungen für Trigramhäufigkeitsverteilungen um über 99% reduziert werden.

Betrachte zum Beispiel den folgenden Beispieltextstring:

Snow, called "POWDER."

[0045] Die untere Tabelle 2 zeigt in ihrer "Bytewert"-Spalte die Folge von Bytewerten, die den Beispielstring in einer digitalen Beispieldokumentendarstellung ausmachen. Die "Zielwert"-Spalte der Tabelle 2 zeigt eine Folge von Zielwerten, auf die die Bytewerte abgebildet werden, wobei die in Tabelle 1 gezeigte Abbildung verwendet wird.

Zeichennummer	Zeichenglypt	Bytewert	Zielwert
1	s	0 × 73	20
2	n	0 × 6E	15
3	o	0 × 6F	16
4	w	0 × 77	24
5	,	0 × 2C	0
6		0 × 20	0
7	c	0 × 63	4
8	a	0 × 61	2
9	l	0 × 6C	13
10	l	0 × 6C	13
11	e	0 × 65	6
12	d	0 × 64	5
13		0 × 20	0
14	"	0 × 22	0
15	P	0 × 50	17
16	O	0 × 4F	16
17	W	0 × 57	24
18	D	0 × 44	5
19	E	0 × 45	6
20	R	0 × 52	19
21	.	0 × 2E	0
22	"	0 × 22	0

Tabelle 2

[0046] Man kann aus Tabelle 2 sehen, daß jeder verschiedene Buchstabe seinen eigenen Zielwert hat. Zum Beispiel sind sowohl der kleine Buchstabe "o" wie auch der große Buchstabe "0" auf den Zielwert **16** abgebildet. Man kann ferner sehen, daß alle Interpunktionszeichenglypten, umfassend Komma, Leerzeichen, Anführungszeichen und Punkt, auf einen einzelnen Zielwert abgebildet sind, den Zielwert 0. Somit sind es die Identitäten der verschiedenen Buchstaben, die verwendet werden, um verschiedene Sprachen in digitalen Dokumentendarstellungen zu unterscheiden, die in romanischen Einzelbytezeichenfolgen erstellt sind. Die in Tabelle 1 gezeigte Abbildung kann ferner verwendet werden, um Modelle und Charakterisierungen von Zeichenfolgen, wie beispielsweise Microsoft Zeichenfolge 50220 für japanisch, zu erzeugen, die, im Gegensatz zu technischen Doppelbytezeichenfolgen, keine Bytewerte zwischen 0 × 80 und 0 × FF verwenden.

[0047] In Schritt **403** addiert die Einrichtung die dreidimensionale Charakterisierung der digitalen Beispieldokumentendarstellung, die in Schritt **402** erzeugt wurde, zu einem dreidimensionalen Modell der Sprache, die als primäre Sprache der digitalen Beispieldokumentendarstellung identifiziert wurde. Dieses Additionsverfahren beinhaltet ein Addieren der Werte eines jedes Eintrages in der dreidimensionalen Charakterisierung der digitalen Beispieldokumentendarstellung in einen entsprechenden Eintrag in dem dreidimensionalen Modell der primären Sprache. Die Schritte **404** und **405** spiegeln die Schritte **402** und **403** in dem Sinne wieder, daß Schritt **404** eine Charakterisierung der digitalen Beispieldokumentendarstellung erzeugt und Schritt **405** diese Charakterisierung zu einem Modell einer Sprache und einer Zeichenfolge addiert. In Schritt **404** erzeugt die Einrichtung eine eindimensionale Charakterisierung der betrachteten digitalen Beispieldokumentendarstellung, wobei die unten in Tabelle 3 gezeigte Abbildung und Häufigkeitsverteilungen für Zeichenunigramme, die in der digitalen Beispieldokumentendarstellung auftreten, verwendet werden.

Quellwert(e)	Zielwert
0–0 × 7F	0
0 × 80	1
0 × 81	2
0 × 82	3
...	...
0 × FF	128

Tabelle 3: Abbildung, um Zeichenfolgen romanischer SBCS zu untersuchen.

[0048] Man kann aus Tabelle 3 sehen, daß die Abbildung zum Erkennen der richtigen Zeichenfolge aus romanischen Einzelbytezeichenfolgen jeden Quellwert mit einer hohen Bitfolge auf verschiedene Zielwerte abbildet. Zum Beispiel wird der Quellwert 0 × 80 auf den Zielwert 1 abgebildet, der Quellwert 0 × 81 wird auf den Zielwert 2 abgebildet usw. Bei Einzelbytezeichenfolgen stellen nur diese "ausgedehnten Zeichenbytewerte" verschiedene Zeichen in verschiedenen Zeichenfolgen dar, wohingegen auf der anderen Seite jeder der Bytewerte 0–0 × 7F üblicherweise dasselbe Zeichen in all diesen Zeichenfolgen darstellt. Es sind daher dieses ausgedehnten Zeichenbytecodes, die dazu dienen, am besten zwischen verschiedenen romanischen Einzelbytezeichenfolgen zu unterscheiden. Zusätzlich zu den 128 Zielwerten, die jeweils einem der Quellwerte zwischen 0 × 80 und 0 × FF zugewiesen sind, definiert die Abbildung einen weiteren Zielwert: der Zielwert 0 ist allen Quellwerten zwischen 0 und 0 × 7F zugewiesen. Durch eine Abbildung der Quellwerte auf diese Weise reduziert die Einrichtung 256 Quellwerte auf nur 129 Zielwerte, wodurch die Speicheranforderungen für Unigrammhäufigkeitsverteilungen um nahezu 50% reduziert sind.

[0049] Im Schritt **404** addiert die Einrichtung die eindimensionale Charakterisierung der digitalen Beispieldokumentendarstellung, die in Schritt **404** erzeugt wurde, zu einem eindimensionalen Modell der Kombination aus der primären Sprache und Zeichenfolge der digitalen Beispieldokumentendarstellung. Nachdem der Schritt **405** vollendet ist, werden diese Schritte beendet.

[0050] In den Schritten **406**–**407** erzeugt die Einrichtung eine einzelne Charakterisierung der digitalen Beispieldokumentendarstellung und addiert diese zu einem einzigen Modell für sowohl die primäre Sprache wie auch der Zeichenfolge der digitalen Beispieldokumentendarstellung.

[0051] In dem Schritt **406** erzeugt die Einrichtung eine zweidimensionale Charakterisierung der digitalen Beispieldokumentendarstellung, wobei die unten in Tabelle 4 gezeigte Abbildung und Häufigkeitsverteilungen für Zeichenbigramme, die in der digitalen Beispieldokumentendarstellung vorkommen, verwendet werden.

Quellwert(e)	Zielwert
0 × 00–0 × 3F, 0 × 5B–0 × 60, 0 × 7B–0 × 7F	0
0 × 41, 0 × 61	1
0 × 42, 0 × 62	2
0 × 43, 0 × 63	3
...	...
0 × 5A, 0 × 7A	26
0 × 80	27
0 × 81	28
0 × 82	29
...	...
0 × FF	154

Tabelle 4: Abbildung, um die Sprache und Zeichenfolge von nichtromanischen SBCS oder DBCS zu untersuchen.

[0052] Man kann sehen, daß die Abbildung in Tabelle 4 im wesentlichen eine Vereinigung der verschiedenen

Zielwerte ist, die durch die beiden Abbildungen definiert sind, die in Tabelle 1 und Tabelle 3 gezeigt sind. Diese Abbildung reduziert 256 Quellwerte auf 155 Zielwerte, wodurch die Speicheranforderungen für Bigrammhäufigkeitsverteilungen um mindestens 63% reduziert werden.

[0053] Im Schritt **407** addiert die Einrichtung die zweidimensionalen Charakterisierungen der digitalen Beispieldokumentendarstellung, die in Schritt **406** erzeugt wurden, zu einem zweidimensionalen Modell der Kombination aus der primären Sprache und Zeichenfolge der digitalen Beispieldokumentendarstellung. Nach dem Schritt **407** werden diese Schritte beendet.

[0054] Fig. 5 ist ein Flußdiagramm, das die Schritte zeigt, die bevorzugt von der Einrichtung ausgeführt werden, um eine Charakterisierung einer digitalen Beispieldokumentendarstellung zu erzeugen. Die in Fig. 5 gezeigten Schritte werden bevorzugt von der Einrichtung als Teil der Schritte **402**, **404** und **406** ausgeführt. Wie in den Schritten **402**, **404** und **406** gezeigt, verwendet eine Charakterisierung, die entsprechend den Schritten der Fig. 5 erzeugt wurde, eine bestimmte reduktive Abbildung und hat eine spezifizierte Anzahl von Dimensionen, die direkt der Länge der n-Gramme entspricht, für die Häufigkeitsverteilungen in der Charakterisierung erhalten bleiben. Zum Beispiel verwendet die Einrichtung, in Übereinstimmung mit dem Schritt **406** für Zeichenfolgen mit mehreren Bytes und romanischen Einzelbytezeichenfolgen, die in Tabelle 3 gezeigte Abbildung und erzeugt eine zweidimensionale Charakterisierung, die die Häufigkeit eines jeden möglichen Bigramms von Zielwerten enthält.

[0055] In dem Schritt **501** verwendet die Einrichtung die reduktive Abbildung, die für die Charakterisierung spezifiziert ist, um jedes Byte der digitalen Dokumentendarstellung von dem Quellwert für das Byte auf einen Zielwert abzubilden. Zum Beispiel würde die Einrichtung für die Zeichen, die in Tabelle 2 gezeigt sind, das Byte oder Quellwerte auf Zielwerte abbilden, die in Tabelle 2 gezeigt sind. Im Schritt **502** konstruiert und initialisiert die Einrichtung eine Matrix, die die Charakterisierung enthalten soll. Die Matrix weist eine Anzahl von Dimensionen auf, die direkt der Länge, in Zeichen, der spezifizierten n-Gramme für diese Kombination aus Sprache und Zeichenfolge entspricht. Da Schritt **406** die Verwendung von Bigrammhäufigkeitsverteilungen spezifiziert, konstruiert die Einrichtung beispielsweise eine zweidimensionale Matrix, wenn der Schritt **502** ausgeführt wird. Bevorzugt hat die Matrix in jeder ihrer Dimensionen eine Position für jeden möglichen Zielwert in der reduktiven Abbildung.

[0056] In den Schritten **503**–**505** durchlauft die Einrichtung eine Schleife für jedes Auftreten eines n-Grammes der spezifizierten Länge in der Folge von Zielwerten, die in Schritt **501** erzeugt wurden. Das heißt, daß wenn die Bigrammhäufigkeitsverteilungen spezifiziert sind, die Einrichtung mit dem Berücksichtigen der ersten und zweiten Zielwerte in den Folgen beginnt, dann die zweiten und dritten Zielwerte in den Folgen berücksichtigt usw., bis sie die zweiten-bis-letzten und letzten Zielwerte in der Folge der Zielwerte berücksichtigt. Für das momentan betrachtete n-Grammauftreten inkrementiert die Einrichtung im Schritt **504** das Element der Matrix, die in Schritt **502** konstruiert wurde, das dem Zielwert-n-Gramm entspricht, das momentan betrachtet wird. Das heißt das Matrixelement, das inkrementiert wird, ist dasjenige, dessen erster Index den ersten Zielwert des n-Grammes aufweist, dessen zweiter Index den zweiten Zielwert des n-Grammes aufweist, usw. Wenn bspw. die Trigrammhäufigkeiten für das erste Trigramm, das in Tabelle 2 gezeigt ist, mit den Zielwerten **20**, **15** und **16** zusammengeführt werden, würde die Einrichtung die Werte der Matrix an den Stellen (**20**, **15**, **16**) inkrementieren. Wenn im Schritt **505** irgendwelche Zielwert-n-Gramm-Vorkommnisse übrigbleiben, fährt die Einrichtung mit dem Schritt **503** fort, das nächste zu berücksichtigen. Andernfalls sind diese Schritte beendet. Beim Beenden dieser Schritte reflektiert die im Schritt **502** konstruierte Matrix, wie oft jedes Zielwert-n-Gramm in der digitalen Dokumentendarstellung vorkommt.

[0057] Fig. 6 ist ein Flußdiagramm, das die Schritte zeigt, die bevorzugt von der Einrichtung ausgeführt werden, um die Zeichenfolge und Sprache einer eingegebenen digitalen Dokumentendarstellung zu identifizieren, für die diese Information nicht bekannt ist. In dem Schritt **601** bestimmt die Einrichtung, ob es wahrscheinlich ist, daß die eingegebene digitale Dokumentendarstellung in einer romanischen Einzelbytezeichenfolge erstellt ist. Wenn der größte Teil der Bytes der eingegebenen digitalen Dokumentendarstellung vor dem Abbilden einen Wert kleiner als 0 × 80 aufweist, d. h. keine hohe Bitfolge hat, bestimmt die Einrichtung, daß die eingegebene digitale Dokumentendarstellung in einer romanischen Einzelbytezeichenfolge erstellt ist und die Einrichtung fährt im Schritt **602** fort, die Schritte **602**–**605** auszuführen, andernfalls bestimmt die Einrichtung, daß die eingegebene digitale Dokumentendarstellung in einer Zeichenfolge mit mehreren Bytes oder einer nicht romanischen Zeichenfolge ausgedrückt ist und fährt im Schritt **606** fort, die Schritte **606** und **607** auszuführen.

[0058] In dem Schritt **602** erzeugt die Einrichtung auf eine Weise, die ähnlich zu der des Schrittes **402** ist, eine dreidimensionale Charakterisierung der eingegebenen digitalen Dokumentendarstellung, wobei die Abbildung,

die in Tabelle 1 gezeigt ist, und Trigramme verwendet werden. In dem Schritt **603** erkennt die Einrichtung die Sprache der eingegebenen digitalen Dokumentendarstellung, wobei die dreidimensionale Charakterisierung der eingegebenen digitalen Dokumentendarstellung, die in **602** erzeugt wurde, und die dreidimensionalen Modelle der Sprachen, die in Schritt **201** der Trainingsphase erzeugt wurden, verwendet werden. Dieses Erkennungsverfahren wird genauer weiter unten in Verbindung mit Fig. 7 beschrieben. In dem Schritt **604**, nachdem die Sprache der eingegebenen digitalen Dokumentendarstellung erkannt wurde, erzeugt die Einrichtung eine eindimensionale Charakterisierung der eingegebenen digitalen Dokumentendarstellung auf eine Weise, die ähnlich zu der des Schrittes **404** ist, wobei die Abbildung, die in Tabelle 3 gezeigt ist, und Unigramme verwendet werden. In dem Schritt **605** erkennt die Einrichtung die Zeichenfolge der eingegebenen digitalen Dokumentendarstellung, wobei die eindimensionale Charakterisierung der eingegebenen digitalen Dokumentendarstellung, die in Schritt **604** erzeugt wurde, und die eindimensionalen Modelle der Kombinationen der Sprachen und Zeichenfolgen, die die erkannte Sprache umfassen, verwendet werden. Dieses Erkennungsverfahren wird genauer weiter unten in Verbindung mit Fig. 7 beschrieben. Nach dem Schritt **605** sind diese Schritte vollendet.

[0059] In dem Schritt **606** erzeugt die Einrichtung auf eine Weise, die ähnlich zu der des Schrittes **406** ist, eine zweidimensionale Charakterisierung der eingegebenen digitalen Dokumentendarstellung, wobei die Abbildung, die in Tabelle 4 gezeigt ist, und Bigramme verwendet werden. In dem Schritt **607** erkennt die Einrichtung sowohl die Sprache wie auch die Zeichenfolge der eingegebenen digitalen Dokumentendarstellung, wobei die zweidimensionale Charakterisierung der eingegebenen digitalen Dokumentendarstellung, die in Schritt **606** erzeugt wurde, und die zweidimensionalen Modelle der Kombinationen aus Sprachen und Zeichenfolgen, die in Schritt **201** der Trainingsphase erzeugt wurde, verwendet werden. Dieses Erkennungsverfahren wird ebenso weiter unten genauer in Verbindung mit Fig. 7 beschrieben. Nach dem Schritt **607** sind diese Schritte vollendet.

[0060] Fig. 7 ist ein Flußdiagramm, das die Schritte zeigt, die bevorzugt von der Einrichtung ausgeführt werden, um die Zeichenfolge und/oder Sprache der eingegebenen digitalen Dokumentendarstellung in Übereinstimmung mit den Schritten **603**, **605** oder **607** zu erkennen. In diesen Schritten vergleicht die Einrichtung die vorliegende Charakterisierung der eingegebenen digitalen Dokumentendarstellung mit jedem Kandidatenmodell. Die Sprache und/oder Zeichenfolge des Kandidatenmodells, die mit der Charakterisierung der eingegebenen digitalen Dokumentendarstellung am ähnlichsten ist, wird als die der eingegebenen digitalen Dokumentendarstellung erkannt. In den Schritten **701**–**703** durchläuft die Einrichtung Schleifen für jedes Kandidatenmodell einer Sprache und/oder Zeichenfolge, wie in den Schritten **603**, **605** oder **607** spezifiziert. In dem Schritt **702** berechnet die Einrichtung das Skalarprodukt der Matrizen, die für das betrachtete Kandidatenmodell und die Charakterisierung der eingegebenen digitalen Dokumentendarstellung stehen. Eine solche Berechnung beinhaltet das Multiplizieren des Wertes einer jeder Stelle in der Kandidatenmodellmatrix mit dem entsprechenden Wert in der Charakterisierungsmatrix und ein Addieren dieser Produkte, um einen einzigen Wert zu erhalten, der "Score" genannt wird und die Ähnlichkeit des Kandidatenmodells mit der Charakterisierung darstellt. In dem Schritt **703**, wenn weitere Kandidatenmodelle zur Verarbeitung übrigbleiben, fährt die Einrichtung in dem Schritt **701** fort, das nächste Kandidatenmodell zu verarbeiten, andernfalls fährt die Einrichtung mit dem Schritt **704** fort. In dem Schritt **704** identifiziert die Einrichtung das Kandidatenmodell, für den der höchste Score im Schritt **702** berechnet wurde. In dem Schritt **705** erkennt die Einrichtung, daß die eingegebene digitale Dokumentendarstellung von der Sprache und/oder Zeichenfolge des Kandidatenmodells ist, die in Schritt **704** erkannt wurde(n). Nach dem Schritt **705** sind diese Schritte beendet.

[0061] Die Schritte, die in Fig. 6 gezeigt sind, erkennen eine einzelne dominante Sprache als die der eingegebenen digitalen Dokumentendarstellung, weil diese die eingegebene digitale Dokumentendarstellung als Ganzes beurteilen. Für eingegebene digitale Dokumentendarstellungen, die Text in mehr als einer Sprache enthalten, hat das Erkennen jeder dieser Sprachen jedoch oft erhebliche Bedeutung. Fig. 8 ist ein Flußdiagramm, das die Schritte zeigt, die bevorzugt von der Einrichtung ausgeführt werden, um mehrere Sprachen zu erkennen, die in einer eingegebenen digitalen Dokumentendarstellung vorkommen, wie auch die Zeichenfolge zu erkennen, in der die eingegebene digitale Dokumentendarstellung erstellt ist. Diese Schritte beinhalten ein Unterteilen der eingegebenen digitalen Dokumentendarstellung in eine Anzahl von kürzeren Segmenten, ein Erkennen der Sprache und Zeichenfolge für jedes Segment und Berichten aller Sprachen, die für die Segmente erkannt wurden, wie auch der Zeichenfolge, die erkannt wurde.

[0062] In Schritt **801** unterteilt die Einrichtung die eingegebene digitale Dokumentendarstellung in Segmente. In einer ersten Ausführungsform haben diese Segmente eine Größe eines herkömmlichen Satzes oder ca. 75 Zeichen. In einer alternativen Ausführungsform haben die Segmente die Größe eines typischen Absatzes oder ca. 400 Zeichen. Diese Segmentgröße ist bevorzugt einstellbar, wodurch es möglich ist, die Einrichtung anzupassen, um einen Text in verschiedenen Sprachen in Segmenten verschiedener Größe zu untersuchen. Ein Verringern der Segmentgröße erhöht die Anzahl von Sprachen, die in einer mehrsprachigen eingegebenen di-

gitalen Dokumentendarstellung erkannt werden können, wohingegen ein Erhöhen der Segmentgröße die Verarbeitungszeit in der Erkennungsphase reduziert.

[0063] In den Schritten **802**–**804** durchläuft die Einrichtung für jedes Segment eine Schleife. In dem Schritt **803** erkennt die Einrichtung die Sprache und Zeichenfolge des betrachteten Segmentes in Übereinstimmung mit den Schritten, die in Fig. 6 gezeigt sind. Wenn weitere Segmente im Schritt **804** für die weitere Verarbeitung übrigbleiben, fährt die Einrichtung im Schritt **802** fort, diese zu verarbeiten, andernfalls fährt die Einrichtung mit Schritt **805** fort. In dem Schritt **805** berichtet die Einrichtung, daß die eingegebene digitale Dokumentendarstellung jede Sprache enthält, die in einer Iteration des Schrittes **803** für eines der Segmente erkannt wurde, in absteigender Reihenfolge der Segmentzahl, in der die Sprache erkannt wurde. In einer Ausführungsform sind die erkannten Sprachen, die im Schritt **805** berichtet werden, auf die begrenzt, die in einer Schwellwertanzahl von Segmenten größer als 1 erkannt wurden. In einer weiter bevorzugten Ausführungsform werden die berichteten, erkannten Sprachen weiter sortiert, um Doppelbytezeichenfolgensprachen oder nicht romanische Sprachen in der Liste oberhalb romanischer Einzelbytezeichenfolgensprachen darzustellen.

[0064] In dem Schritt **806** wählt die Einrichtung eine Zeichenfolge aus den Zeichenfolgen aus, die in den Iterationen des Schrittes **803** erkannt wurden. Der Schritt **806** wird genauer weiter unten in Verbindung mit Fig. 9 beschrieben. In dem Schritt **807** berichtet die Einrichtung die im Schritt **806** ausgewählte Zeichenfolge. Nach dem Schritt **807** werden diese Schritte beendet.

[0065] Fig. 9 ist ein Flußdiagramm, das die Schritte zeigt, die bevorzugt von der Einrichtung ausgeführt werden, um eine Zeichenfolge für die eingegebene digitale Dokumentendarstellung aus den Zeichenfolgen auszuwählen, die für Segmente der eingegebenen digitalen Dokumentendarstellung erkannt wurden. Wenn in dem Schritt **901** alle der dedektierten Zeichenfolgen romanische Einzelbytezeichenfolgen sind, fährt die Einrichtung mit dem Schritt **902** fort, andernfalls fährt die Einrichtung mit dem Schritt **904** fort. In dem Schritt **902** wählt die Einrichtung die Sprache, die als Sprache der meisten Segmente erkannt wurde, als die Sprache der eingegebenen digitalen Dokumentendarstellung aus. In dem Schritt **903** wählt die Einrichtung die Zeichenfolge aus, die mit der Sprache erkannt wurde, die in Schritt **902** für die meisten Segmente ausgewählt wurde. Nach dem Schritt **903** werden diese Schritte beendet. In dem Schritt **904** wählt die Einrichtung die Sprache aus, die mit einer Zeichenfolge, die mehrere Bytes aufweist, oder einer nicht romanischen Einzelbytezeichenfolge, die in den meisten Segmenten erkannt wurde, erkannt wurde. In dem Schritt **905** wählt die Einrichtung die Zeichenfolge, die mehrere Byte aufweist, oder eine nicht romanische Einzelbytezeichenfolge aus, die mit der in Schritt **904** ausgewählten Sprache erkannt wurde, die für die meisten Segmente erkannt wurde. Nach dem Schritt **905** sind diese Schritte vollendet.

[0066] Die obigen Ausführungen beschreiben die Verwendung der Einrichtung zum Erkennen sowohl der Zeichenfolge wie auch der Sprachen einer eingegebenen digitalen Dokumentendarstellung, von der diese Information nicht bekannt ist. Die Einrichtung ist ferner ausgelegt, die Sprachen zu erkennen, die in einer eingegebenen digitalen Dokumentendarstellung verwendet werden, von der bekannt ist, daß diese in einer bestimmten Unicodezeichenfolge erstellt ist. Die Unicodezeichenfolge, wie sie von dem Unicode-Konsortium in San Jose, Californien, USA, in The Unicode Standard, Version 2.0. definiert ist, ist eine große Zeichenfolge, die ausgelegt ist, die meisten Zeichenglypten darzustellen, die in den meisten Sprachen der Welt verwendet werden. Da die Unicodezeichenfolge den Text von so vielen verschiedenen Sprachen darstellen kann, ist es nützlich, fähig zu sein, die Sprachen, die in einer eingegebenen digitalen Dokumentendarstellung vorhanden sind, zu erkennen, von der bekannt ist, das sie in einer Unicodezeichenfolge erstellt ist. Die Unicodezeichenfolge verwendet 16-Bit Zeichenwerte, die 65536 verschiedene Zeichen ermöglichen. Die Unicodezeichenfolge ist in eine Anzahl von "Schriftbereichen" unterteilt, die die Zeichen bündelt, die im allgemeinen in einer Sprache oder in einer Gruppe von verwandten Sprachen verwendet werden. Die untere Tabelle 5 zeigt einige der "Einzelsprachen"-Schriftbereiche, die Zeichen enthalten, die nur in einer Sprache verwendet werden.

Quellwert	Bedeutung	Sprache
0 × 0600–0 × 6FF	arabisch	arabisch
0 × fB50–0 × FDFF	arabische Darstellung Form-A	
0 × FE70–0 × FEFE	arabische Darstellung Form-B	
0 × 3100–0 × 312F	bopomofo	chinesisch
0 × 0370–0 × 03FF	griechisch	griechisch
0 × 1F00–0 × 1FFF	griechisch, ausgedehnt	
0 × 0590–0 × 05FF	hebräisch	hebräisch
0 × 0900–0 × 097F	devanagari	hindi
0 × 3040–0 × 309F	hiragana	japanisch
0 × 30A0–0 × 30FF	katakana	
0 × FF60–0 × FFDF	Formen mit halber und voller Breite	
0 × 1100–0 × 11FF	hangul jamo	koreanisch
0 × 3130–0 × 318F	hangul Kompatiblität jamo	
0 × AC00–0 × D7A3	hangulsilben	
0 × 0400–0 × 04FF	kyrillisch	russisch
0 × 0E00–0 × 0E7F	thailändisch	thailändisch

Tabelle 5: Unicodeschriftbereich für Einzelsprachen

[0067] Man kann aus Tabelle 5 sehen, daß beispielsweise der Schriftbereich von 0 × 0400 bis 0 × 04FF nur in russisch verwendet wird. Die unten gezeigte Tabelle 6 zeigt auf der anderen Seite einige der "mehrsprachigen" Schriftbereiche, die Zeichen enthalten, die jeweils in einer Gruppe von zwei oder mehr Sprachen verwendet werden.

Quellwerte	Bedeutung	Sprachengruppe
0 × 0041–0 × 005A	Basis romanisch A–Z	romanisch
0 × 0061–0 × 007A	Basis romanisch a–z	
0 × 00C0–0 × 017F	Romanisch-1 Zusatz und ausgedehnt-A	
0 × 1E00–0 × 1EFF	Romanisch ausgedehnt zusätzlich	
0 × FF21–0 × FF3A	volle Breite ASCII A–Z	
0 × FF41–0 × FF5A	volle Breite ASCII a–z	
0 × 3000–0 × 303F	CJK Symbole und Interpunktionszeichen Kanbun	CJK
0 × 3190–0 × 319F 0 × 3200–0 × 32FF	eingeschlossene CJK Buchstaben und Monate CJK Kompatibilität	
0 × 3300–0 × 33FF 0 × 4E00–0 × 9FFF	CJK vereinheitlichte Ideogramme	
0 × F900–0 × FAFF	CJK Kompatiblitätsideogramme	
0 × FE30–0 × FE4F	CJK Kompatiblitätsformen	

Tabelle 6: mehrsprachige Unicodeschriftbereiche

[0068] Man kann beispielsweise sehen, daß der Schriftbereich von 0 × 4E00 bis 0 × 9FFF in jeder der "CJK-Gruppe"-Sprachen, chinesisch, japanisch und koreanisch, verwendet wird. In den Fällen, in denen die Zeichen in einem Segment einer eingegebenen digitalen Unicodedokumentendarstellung zum überwiegenden Teil aus einem dieser Einzelsprachenschriftbereiche stammen, erkennt die Einrichtung die Sprache des Segmentes als die einzige Sprache dieses Schriftbereiches. In den Fällen, in denen die Zeichen eines Segmentes zum überwiegenden Teil aus einem Mehrsprachenschriftbereich stammen, bildet die Einrichtung bevorzugt die Quellwerte des Segmentes ab, um eine statistische Charakterisierung des Segmentes zu erzeugen, und vergleicht sodann diese Charakterisierung mit den Modellen der Sprachen in der Sprachengruppe auf eine Weise, die zuvor beschrieben wurde.

[0069] Für Segmente, die zum überwiegenden Teil aus Zeichen eines Schriftbereiches oder mehrerer Schriftbereiche der "romanischen" Sprachengruppe bestehen, die die romanischen Sprachen, wie beispielsweise die oben aufgeführten, umfaßt, verwendet die Einrichtung bevorzugt die Abbildung, die unten in Tabelle 7 gezeigt ist, um die Quellwerte des Segmentes auf dieselben 28 Zielwerte abzubilden, wie die in Tabelle 1 gezeigte Abbildung zum Erkennen der Sprache einer romanischen Einzelbytezeichenfolge.

Quellwert(e)	Zielwert
0 × 0000–0 × 0040, 0 × 005B–0 × 0060, 0 × 007B–0 × 00BF, 0 × 0180–0 × 1DFF, 0 × 1F00–0 × FF20, 0 × FF3B–0 × FF40,	0
0 × FF5B–0 × FFFF	
0 × 00C0–0 × 017F, 0 × 1E00–0 × 1EFF,	1
0 × 0041, 0 × 0061, 0 × FF21, 0 × FF41	2
0 × 0042, 0 × 0062, 0 × FF22, 0 × FF42,	3
0 × 0043, 0 × 0063, 0 × FF23, 0 × FF43,	4
0 × 005A, 0 × 007A, 0 × FF3A, 0 × FFSA	27

Tabelle 7: Abbildung, um Sprachen eines romanischen Unicode zu erkennen

[0070] Wie weiter unten genauer beschrieben werden wird, erkennt die Einrichtung die romanische Sprache innerhalb der romanischen Gruppe durch Vergleichen einer dreidimensionalen Charakterisierung des Segmentes mittels dieser Abbildung mit den dreidimensionalen Modellen der Sprachen in der romanischen Gruppe, die bereits durch die Einrichtung in Übereinstimmung mit Fig. 3 erzeugt wurden.

[0071] Für Segmente, die zum überwiegenden Teil aus Zeichenwerten eines mehrsprachigen Schriftbereiches einer Sprachengruppe bestehen, die eine andere als die romanische Sprachengruppe ist, verwendet die Einrichtung eine anwenderspezifische Abbildung, die von der Einrichtung speziell für diese Sprachengruppe auf eine Weise erzeugt wurde, die ausgelegt ist, um zwischen den verschiedenen Sprachen der Sprachengruppe unterscheiden zu können. Diese anwenderspezifischen Abbildungen, wie beispielsweise die in Tabelle 7 gezeigte Abbildung, bilden von allen 65536 Unicodezeichenwerten auf eine viel kleinere Anzahl von Zielwerten, wie beispielsweise 256 ab. Die anwenderspezifische Abbildung für die Sprachengruppe wird verwendet, um eine eindimensionale Charakterisierung des Segmentes zu konstruieren, die sodann mit eindimensionalen Modellen der verschiedenen Sprachen in der Sprachengruppe verglichen wird, die mittels der gleichen anwenderspezifischen Abbildung erzeugt wurden.

[0072] Fig. 10 ist ein Flußdiagramm, das die Schritte zeigt, die von der Einrichtung bevorzugt in der Trainingsphase ausgeführt werden, um eine anwenderspezifische reduktive Abbildung für eine Unicodesprachengruppe zu erzeugen, die verwendet wird, um zwischen den Sprachen der Sprachengruppe zu unterscheiden. Diese Schritte werden bevorzugt für jede Gruppe von Sprachen ausgeführt, die in einem großen Ausmaß auf dem gleichen Schriftbereich beruhen. Zum Beispiel beruhen die chinesische, japanische und koreanische Sprache sehr auf den Schriftbereichen, die für die "CJK"-Sprachengruppe in Tabelle 6 gezeigt sind. Als ein weiteres Beispiel beruht eine Gruppe von indischen Sprachen ebenso stark auf einem einzelnen Schriftbereich. Diese Schritte werden einmalig für eine jede derartige Gruppe ausgeführt, wodurch eine anwenderspezifische Abbildung für die Gruppe erzeugt wird, die ausgelegt ist, um die Sprachen der Gruppe zu unterscheiden. In einer bevorzugten Ausführungsform hat die anwenderspezifische Abbildung 256 Zielwerte. In alternativen Ausführungsformen bildet die anwenderspezifische Abbildung auf eine Anzahl von Zielwerten ab, die größer oder kleiner als 256 ist.

[0073] In den Schritten **1001**–**1004** durchläuft die Einrichtung Schleifen für jede Sprache in der Sprachengruppe. In dem Schritt **1002** weist die Einrichtung beliebige Quellwerte des betrachteten, exklusiven Schriftbereiches der Sprache einen einzelnen Zielwert zu. In dem Schritt **1003** zählt die Einrichtung die Häufigkeit des Auftretens eines jeden einzelnen Quellwertes. Wenn in dem Schritt **1004** zusätzliche Sprachen zur Verarbeitung übrigbleiben, führt die Einrichtung in dem Schritt **1001** damit fort, die nächste Sprache in der Sprachengruppe zu verarbeiten, andernfalls führt die Einrichtung mit dem Schritt **1005** fort. In den Schritten **1005**–**1008** durchläuft die Einrichtung wiederum Schleifen für jede Sprache in der Sprachengruppe. In dem Schritt **1006** weist die Einrichtung Quellwerten, die nur in dem Korpus der betrachteten Sprache auftreten, einen einzelnen Zielwert zu. In dem Schritt **1007** weist die Einrichtung jeder Gruppe von 100 Quellwerten, die am häufigsten in dem Korpus der betrachteten Sprache auftreten, einen einzelnen Zielwert zu. Falls in dem Schritt **1008** weitere Sprachen zu verarbeiten sind, führt die Einrichtung mit dem Schritt **1005** fort, die nächste Sprache in der Sprachengruppe zu verarbeiten, andernfalls führt die Einrichtung mit dem Schritt **1009** fort. In dem Schritt **1009** weist die Einrichtung den restlichen Quellwerten die restlichen Zielwerte zu.

[0074] Nach dem Schritt **1009** bildet die anwenderspezifische Abbildung jeden Quellwert auf einen der 256

Zielwerte ab und diese Schritte werden beendet.

[0075] Nach dem Erzeugen von anwenderspezifischen reduktiven Abbildungen für jede Unicodesprachengruppe in Übereinstimmung mit Fig. 10, verwendet die Einrichtung bevorzugt jede anwenderspezifische Abbildung, um ein eindimensionales statistisches Modell von jeder der Sprachen in der Sprachengruppe zu erzeugen. Die Einrichtung führt ein solches Erzeugen in Übereinstimmung mit den Schritten aus, die in Fig. 3 gezeigt sind.

[0076] Fig. 11 ist ein Flußdiagramm, das die Schritte zeigt, die bevorzugt von der Einrichtung ausgeführt werden, um die Sprachen zu erkennen, die in einer eingebenen digitalen Dokumentendarstellung auftreten, die in einer Unicodezeichenfolge erstellt ist. In dem Schritt **1101** unterteilt die Einrichtung die eingegebene digitale Dokumentendarstellung in Segmente, wie beispielsweise Segmente von der Länge eines Absatzes mit ca. 400 Zeichen. In den Schritten **1102**–**1104** durchlauft die Einrichtung für jedes Segment eine Schleife. In dem Schritt **1103** erkennt die Einrichtung die Sprache des Segmentes. Dieses Erkennungsverfahren wird weiter unten in Verbindung mit Fig. 12 beschrieben. Falls in dem Schritt **1104** weitere Segmente verarbeitet werden müssen, fährt die Einrichtung mit dem Schritt **1102** fort, das nächste Segment zu verarbeiten, andernfalls fährt die Einrichtung mit dem Schritt **1105** fort. In dem Schritt **1105** berichtet die Einrichtung alle Sprache, die in dem Schritt **1103** erkannt wurden, in absteigender Reihenfolge der Anzahl der Segmente, in der die Sprache erkannt wurde. Nach dem Schritt **1105** sind diese Schritte vollendet.

[0077] Fig. 12 ist ein Flußdiagramm, das die Schritte zeigt, die bevorzugt von der Einrichtung ausgeführt werden, um die Sprache eines Segmentes einer eingegebenen digitalen Dokumentendarstellung zu erkennen, die in der Unicodezeichenfolge erstellt ist. In den Schritten **1201**–**1204** zählt die Einrichtung, wie oft Unicodezeichen in dem Segment auftreten, die sich innerhalb eines jeden Schriftbereiches befinden, die in Tabelle 5 und 6 gezeigt sind. In den Schritten **1201**–**1204** durchläuft die Einrichtung für jedes Auftreten eines Unicodezeichens in dem Segment eine Schleife. Falls in dem Schritt **1202** das Auftreten des Zeichens innerhalb eines Schriftbereiches liegt, fährt die Einrichtung mit dem Schritt **1203** fort, um einen Zähler für den Schriftbereich zu erhöhen, andernfalls fährt die Einrichtung mit dem Schritt **1204** fort. Falls in dem Schritt **1204** aufgetretene Unicodezeichen noch zu verarbeiten sind, fährt die Einrichtung in dem Schritt **1201** fort, das nächste aufgetretene Unicodezeichen in dem Segment zu verarbeiten, andernfalls fährt die Einrichtung mit dem Schritt **1205** fort. Falls in dem Schritt **1205** der Schriftbereich, dessen Zähler den höchsten Wert aufweist, ein Einzelsprachenschriftbereich ist, fährt die Einrichtung mit dem Schritt **1206** fort, andernfalls ist der Schriftbereich, dessen Zähler den höchsten Wert aufweist, ein mehrsprachiger Schriftbereich und die Einrichtung fährt mit dem Schritt **1207** fort. In dem Schritt **1206** erkennt die Einrichtung die Sprache des Segmentes als die Sprache des Einzelsprachenschriftbereiches. Nach dem Schritt **1206** sind diese Schritte vollendet.

[0078] Falls in dem Schritt **1207** die Sprachengruppe des Schriftbereiches, dessen Zähler den höchsten Wert aufweist, die romanische Sprachengruppe ist, fährt die Einrichtung mit dem Schritt **1208** fort, andernfalls fährt die Einrichtung mit dem Schritt **1210** fort. In dem Schritt **1208** erzeugt die Einrichtung eine dreidimensionale Charakterisierung des Segmentes, wobei die Abbildung verwendet wird, die in Tabelle 7 gezeigt ist, die ausgelegt ist, die Sprachen der romanischen Sprachengruppe zu unterscheiden. Der Schritt **1208** wird in Übereinstimmung mit den Schritten, die in Fig. 5 gezeigt sind, ausgeführt, wie zuvor beschrieben. In dem Schritt **1209** erkennt die Einrichtung die Sprache des Segmentes, wobei die erzeugte dreidimensionale Charakterisierung und dreidimensionale Modelle der Sprachen in der romanischen Gruppe in Übereinstimmung mit den Schritten, die in Fig. 7 gezeigt sind, verwendet werden. Nach dem Schritt **1209** sind diese Schritte vollendet.

[0079] In dem Schritt **1210** erzeugt die Einrichtung eine eindimensionale Charakterisierung des Segmentes, wobei eine anwenderspezifische Abbildung der Sprachengruppe verwendet wird, die während der Trainingsphase in Übereinstimmung mit Fig. 10 erzeugt wurde. Der Schritt **1210** wird in Übereinstimmung mit den Schritten, die in Fig. 5 gezeigt sind, ausgeführt, wie zuvor beschrieben. In dem Schritt **1211** erkennt die Einrichtung die Sprache des Segmentes, wobei die eindimensionale Charakterisierung, die in dem Schritt **1210** erzeugt wurde, und eindimensionale Modelle der Sprachen der Sprachengruppe in Übereinstimmung mit den Schritten, die in Fig. 7 gezeigt sind, verwendet werden. Nach dem Schritt **1211** sind diese Schritte vollendet.

[0080] Während diese Erfindung unter Bezug auf bevorzugte Ausführungsformen gezeigt und beschrieben wurde, werden Fachleute verstehen, daß verschiedene Änderungen und Modifikationen in Form und Detail vorgenommen werden können, ohne den Bereich der Erfindung zu verlassen. Zum Beispiel können n-Gramme von größerer Länge, als die oben beschriebene, verwendet werden, um die Genauigkeit der Einrichtung zu erhöhen. Auf der anderen Seite können n-Gramme kleinerer Länge verwendet werden, um die Speicheranforderungen der Einrichtung zu verringern. Während statistische Charakterisierungen der digitalen Dokumenten-

darstellungen und statistische Modelle der Sprachen und/oder Zeichenfolgen auf einfache Weise in der Einrichtung durch Verwenden eines ein- oder mehrdimensionalen Arrays von Werten zwischen 0 und 255 dargestellt werden, kann die Einrichtung auf einfache Weise ausgelegt werden, andere Speicherschemata zu verwenden, wie beispielsweise Arrays mit Elementen von verschiedener Größe, Sparse-Arrays oder Datenstrukturen anderer Arten. Während die Einrichtung in dieser Schrift unter Bezug auf die bestimmten Schriftnatursprachen und Zeichenfolgen beschrieben wurde, kann die Einrichtung ferner auf einfache Weise angewendet werden, um Sprachen und Zeichenfolgen aller Art, umfassend diejenigen, die momentan noch nicht verwendet werden, zu modellieren und zu erkennen.

Patentansprüche

1. Verfahren, um in einem Rechnersystem eine unbekannte Sprache und einen unbekannten Zeichensatz einer Textkette von Datenwerten zu erkennen, welche Text in der unbekannten Sprache darstellt, entsprechend einem unbekannten Zeichensatz, der Zeichenglyphen bestimmten Datenwerten eines Wertebereichs zuordnet, wobei das Verfahren folgende Schritte umfasst:
– Feststellen, ob die Textkette von Datenwerten wahrscheinlich Text in einem unbekannten, auf Lateinbasierenden Ein-Byte-Zeichensatz darstellt, wobei die Feststellung darauf basiert, ob die höherwertigen Bits von mehr als einem vorab festgelegten Teil der Bytewerte der Vielzahl „off" sind;
– Anwenden einer reduzierenden Zuordnung der Textkette von Datenwerten, um eine reduzierend zugeordnete transformierte Kette von Datenwerten zu erzeugen, die die Textkette von Datenwerten kennzeichnet, wobei Datenwerte der reduzierend zugeordneten transformierten Kette von Datenwerten, die die Textkette von Datenwerten kennzeichnet, einen kleineren Wertebereich hat als den Wertebereich der Datenwerte in der Textkette, die Text in der unbekannten Sprache darstellt, und wobei eine andere reduzierende Zuordnung verwendet wird in Abhängigkeit von der Feststellung, ob die Textkette von Datenwerten wahrscheinlich Text in einem unbekannten, auf Lateinbasierenden Ein-Byte-Zeichensatz darstellt;
– Erzeugen einer statistischen Analyse der reduzierend zugeordneten transformierten Kette von Datenwerten, die die Textkette von Datenwerten kennzeichnet;
für jede aus einer Vielzahl von Sprachen und zugehöriger Zeichensätze:
– Abrufen eines reduzierend zugeordneten statistischen Modells, das die Sprache und den zugehörigen Zeichensatz hinsichtlich der statistischen Häufigkeit bestimmter Datenwerte in reduzierend zugeordneten repräsentativen Stichproben des Textes in der Sprache nachbildet, und
– Vergleichen des abgerufenen reduzierend zugeordneten statistischen Modells mit der statistischen Analyse der reduzierend zugeordneten transformierten Kette von Datenwerten, die die Textkette von Datenwerten kennzeichnet; und
– Identifizieren als die unbekannte Sprache und der unbekannte Zeichensatz der Sprache und des zugehörigen Zeichensatzes aus der Vielzahl der Sprachen und zugehörigen Zeichensätze, deren Modell, das die Sprache und den zugehörigen Zeichensatz durch die reduzierende Zuordnung repräsentiert, in dem Vergleich mit der statistischen Analyse der reduzierend zugeordneten transformierten Kette von Datenwerten, die die Textkette von Datenwerten kennzeichnet, am besten abschneidet.

2. Verfahren gemäß Anspruch 1, bei dem die reduzierende Zuordnung erweiterte Zeichen einem einzelnen Wert zuordnet.

3. Maschinenlesbarer Datenträger, dessen Inhalt ein oder mehrere Rechnersystem(e) veranlasst, die Schritte von Anspruch 1 auszuführen.

4. Verfahren gemäß Anspruch 1, bei dem der vorab festgelegte Teil eine Hälfte ist.

5. Verfahren in einem Rechnersystem zum Identifizieren jeder einer Vielzahl von Sprachen, die in einer Darstellung eines digitalen Dokuments vorkommen, welche eine Abfolge von Datenwerten beinhaltet, wobei das Verfahren folgende Schritte umfasst:
– Teilen der Abfolge von Datenwerten, aus denen die Darstellung des digitalen Dokuments besteht, in eine Vielzahl von zusammenhängenden, sich gegenseitig ausschließenden Textketten;
– Identifizieren einer vorherrschenden Sprache, die in jeder Textkette der Darstellung des digitalen Dokuments verwendet wird, mithilfe des Verfahrens von Anspruch 1; und
– Anzeigen, dass die Darstellung des digitalen Dokuments jede der identifizierten Sprachen enthält.

6. Verfahren gemäß Anspruch 1, bei dem die unbekannte Sprache der Textkette eine aus einer Gruppe von auf Lateinbasierenden Sprachen ist, wobei mindestens zwei Sprachen in der Gruppe von auf Lateinbasierenden Sprachen durch verschiedene Ein-Byte-Zeichensätze dargestellt werden, wobei ein oder mehrere glei-

che Bytewert(e) verschiedenen Zeichen in den verschiedenen Ein-Byte-Zeichensätzen zugeordnet sind und ein oder mehr gleiche Bytewert(e) demselben Zeichen in den verschiedenen Ein-Byte 30 Zeichensätzen zugeordnet sind;
wobei die reduzierende Zuordnung beinhaltet, Bytewerte in der Textkette einer kleineren Gruppe von Zeichenwerten zuzuordnen, die ausschließlich Buchstaben des englischen Alphabets unterscheidet, unabhängig von Groß und Kleinschreibung, um eine Kette von transformierten Zeichen zu erzeugen, wobei erweiterte Zeichen einem einzelnen Zeichenwert in der Kette der transformierten Zeichen zugeordnet sind;
wobei die statistische Analyse beinhaltet, die Häufigkeit von NGrammen innerhalb der Kette von transformierten Zeichen zu bestimmen; und
wobei das Identifizieren der unbekannten Sprache beinhaltet, eine statistische Analyse der Häufigkeit von NGrammen innerhalb der Textkette von transformierten Zeichen vorzunehmen, um die Sprache der Textkette aus der Gruppe auf Latein basierender Sprachen zu erkennen.

7. Verfahren gemäß Anspruch 1, wobei die unbekannte Sprache eine auf Latein- basierende Sprache ist;
wobei die reduzierende Zuordnung beinhaltet, eine erste transformierte Version der Textkette zu erzeugen, die die Unterscheidung zwischen erweiterten Zeichen negiert;
wobei die statistische Analyse beinhaltet, eine statistische Analyse der ersten transformierten Version der Textkette, welche die Unterscheidung zwischen erweiterten Zeichen negiert, durchzuführen, um die Sprache der Textkette zu erkennen und eine Vielzahl von in Frage kommenden Zeichensätzen zu identifizieren, die der Sprache der Textkette zugeordnet sind;
wobei das Verfahren ferner umfasst, eine zweite transformierte Version der Textkette zu erzeugen, die die Unterscheidung zwischen nicht erweiterten Zeichen negiert; und eine N-Gramm-Analyse der zweiten transformierten Version der Textkette, die die Unterscheidung zwischen nicht erweiterten Zeichen negiert, durchzuführen, um einen Zeichensatz der Textkette zu erkennen, wobei der Zeichensatz der Textkette aus einer Vielzahl von in Frage kommenden Zeichensätzen ausgewählt wird, die der Sprache der Textkette zugeordnet sind, welche durch die Durchführung der N-Gramm-Analyse der ersten transformierten Version der Textkette, die die Unterscheidung zwischen erweiterten Zeichen negiert, identifiziert wurde.

Es folgen 12 Blatt Zeichnungen

Anhängende Zeichnungen

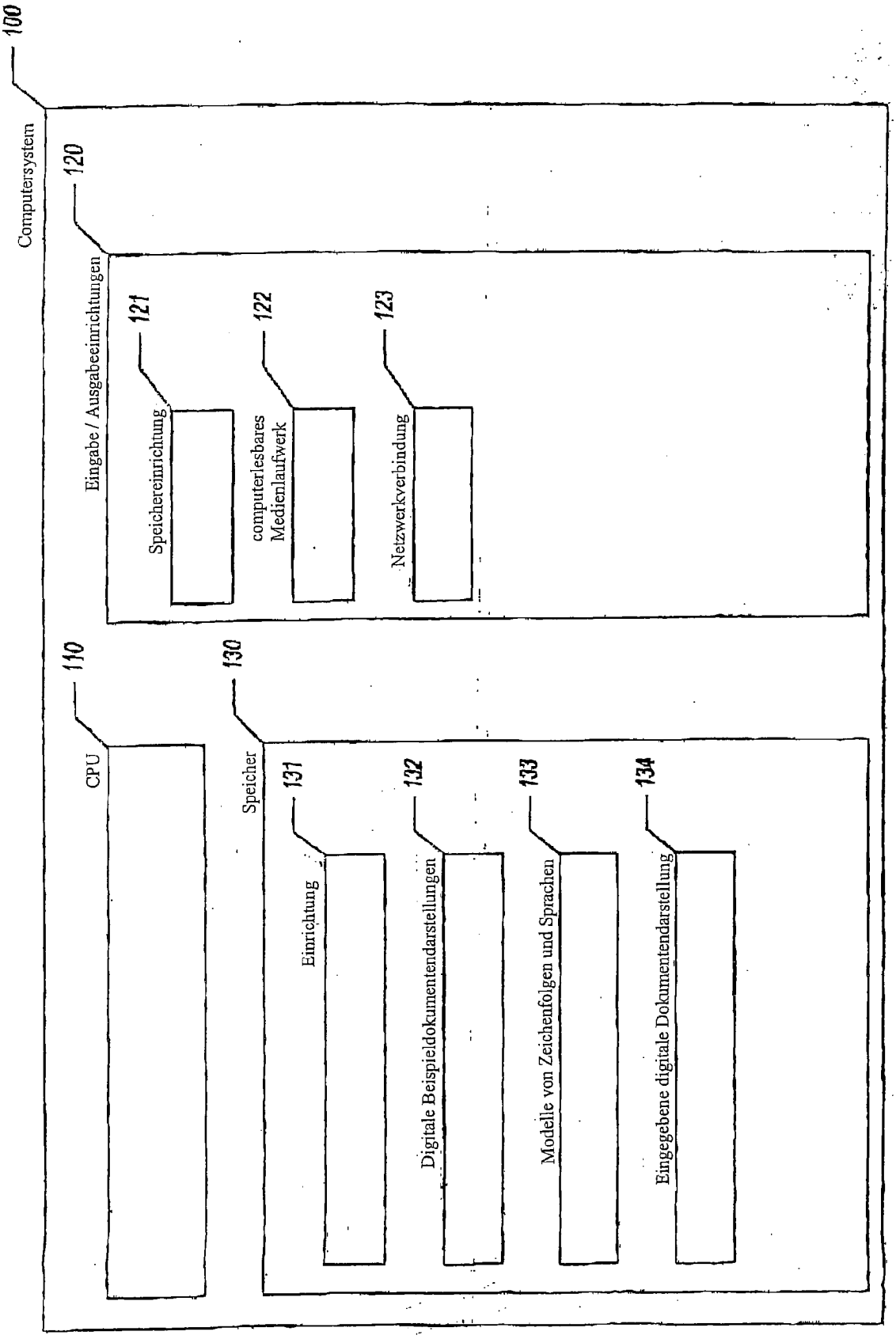

Fig. 1

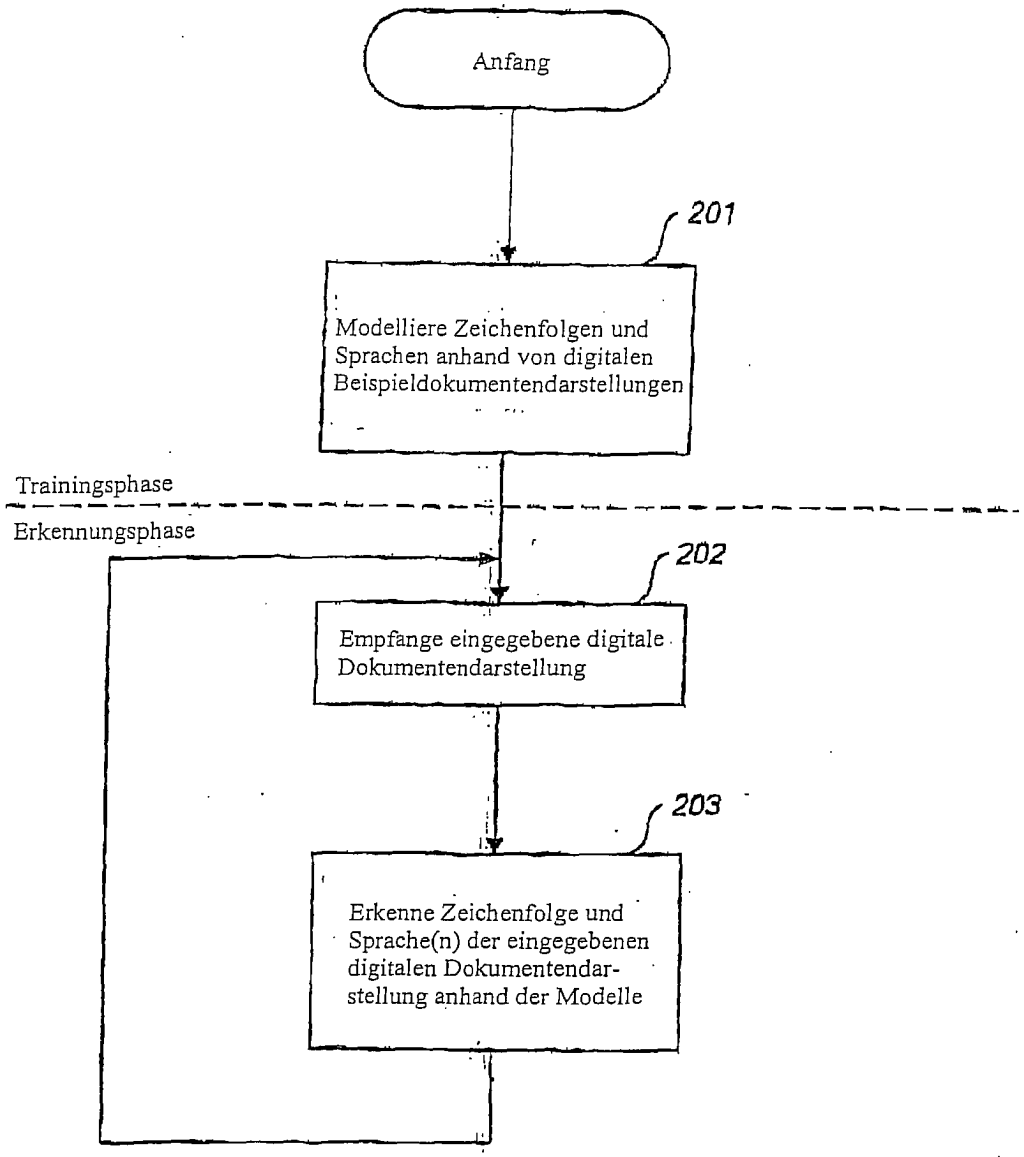
Anfang
201
Modelliere Zeichenfolgen und Sprachen anhand von digitalen Beispieldokumentendarstellungen
Trainingsphase
Erkennungsphase
202
Empfange eingegebene digitale Dokumentendarstellung
203
Erkenne Zeichenfolge und Sprache(n) der eingegebenen digitalen Dokumentendar-stellung anhand der Modelle

Fig. 2

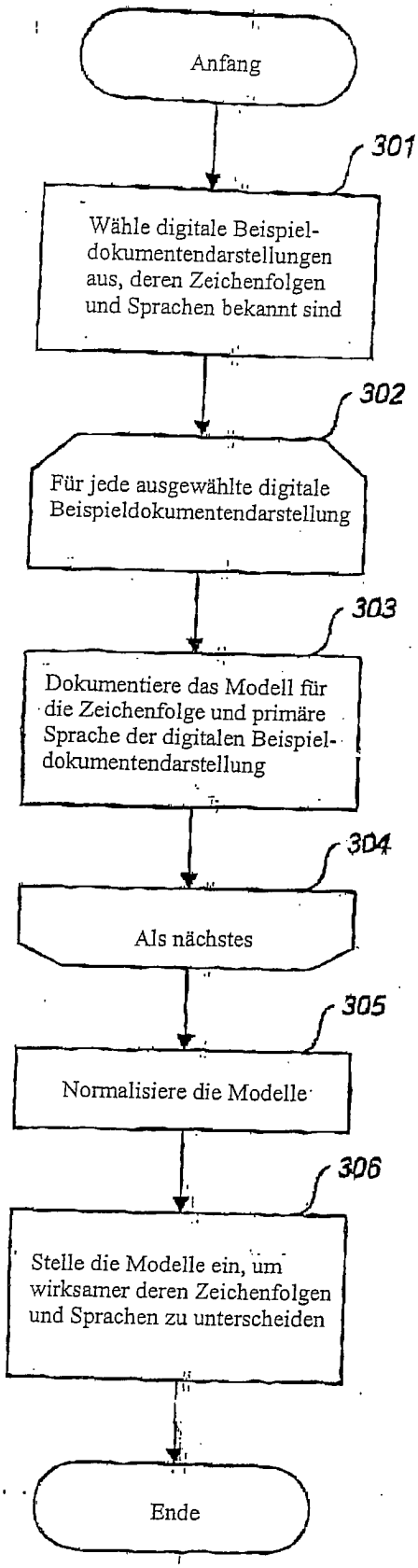
Anfang
301
Wähle digitale Beispiel-dokumentendarstellungen aus, deren Zeichenfolgen und Sprachen bekannt sind
302
Für jede ausgewählte digitale Beispieldokumentendarstellung
303
Dokumentiere das Modell für die Zeichenfolge und primäre Sprache der digitalen Beispiel-dokumentendarstellung
304
Als nächstes
305
Normalisiere die Modelle
306
Stelle die Modelle ein, um wirksamer deren Zeichenfolgen und Sprachen zu unterscheiden
Ende

Fig. 3

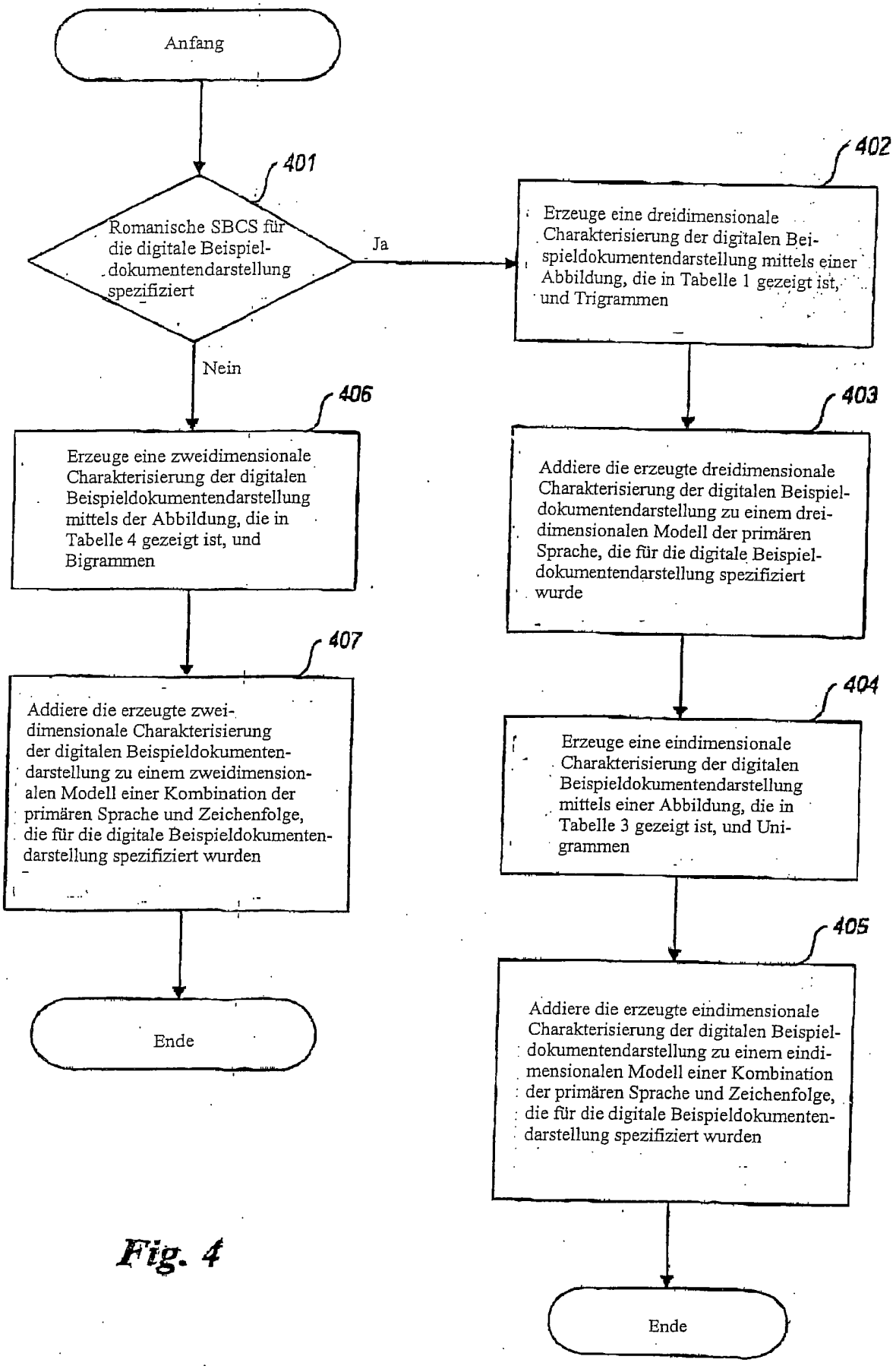
Anfang
401
Romanische SBCS für die digitale Beispieldokumentendarstellung spezifiziert
Ja
Nein
402
Erzeuge eine dreidimensionale Charakterisierung der digitalen Beispieldokumentendarstellung mittels einer Abbildung, die in Tabelle 1 gezeigt ist, und Trigrammen
403
Addiere die erzeugte dreidimensionale Charakterisierung der digitalen Beispieldokumentendarstellung zu einem dreidimensionalen Modell der primären Sprache, die für die digitale Beispieldokumentendarstellung spezifiziert wurde
404
Erzeuge eine eindimensionale Charakterisierung der digitalen Beispieldokumentendarstellung mittels einer Abbildung, die in Tabelle 3 gezeigt ist, und Unigrammen
405
Addiere die erzeugte eindimensionale Charakterisierung der digitalen Beispieldokumentendarstellung zu einem eindimensionalen Modell einer Kombination der primären Sprache und Zeichenfolge, die für die digitale Beispieldokumentendarstellung spezifiziert wurden
Ende
406
Erzeuge eine zweidimensionale Charakterisierung der digitalen Beispieldokumentendarstellung mittels der Abbildung, die in Tabelle 4 gezeigt ist, und Bigrammen
407
Addiere die erzeugte zweidimensionale Charakterisierung der digitalen Beispieldokumentendarstellung zu einem zweidimensionalen Modell einer Kombination der primären Sprache und Zeichenfolge, die für die digitale Beispieldokumentendarstellung spezifiziert wurden
Ende

Fig. 4

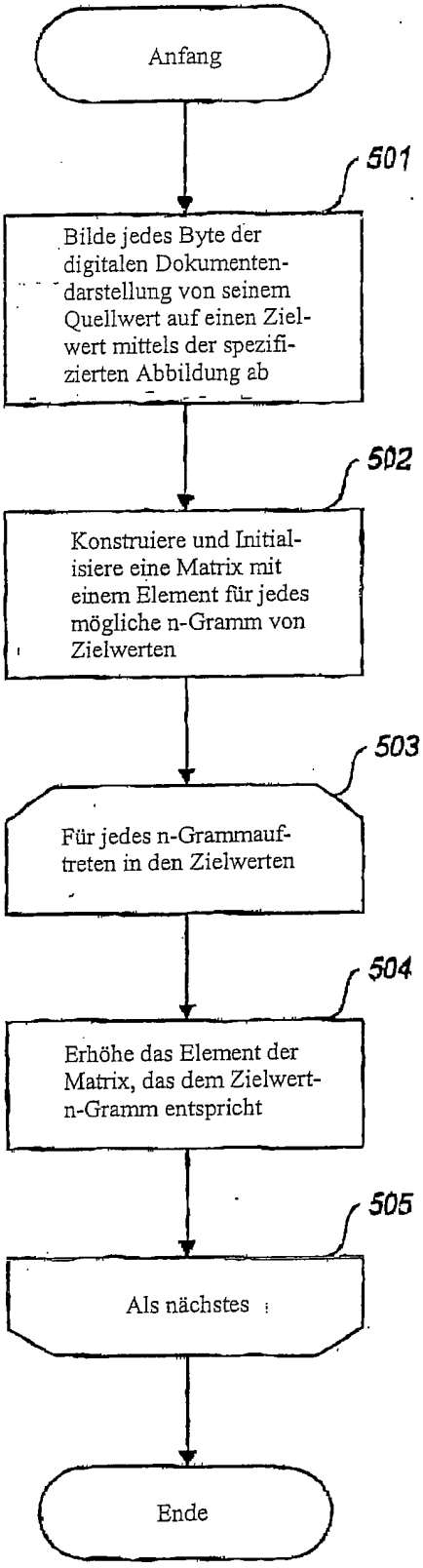
Anfang
501
Bilde jedes Byte der digitalen Dokumenten-darstellung von seinem Quellwert auf einen Ziel-wert mittels der spezifi-zierten Abbildung ab
502
Konstruiere und Initial-isiere eine Matrix mit einem Element für jedes mögliche n-Gramm von Zielwerten
503
Für jedes n-Grammauf-treten in den Zielwerten
504
Erhöhe das Element der Matrix, das dem Zielwert-n-Gramm entspricht
505
Als nächstes
Ende

Fig. 5

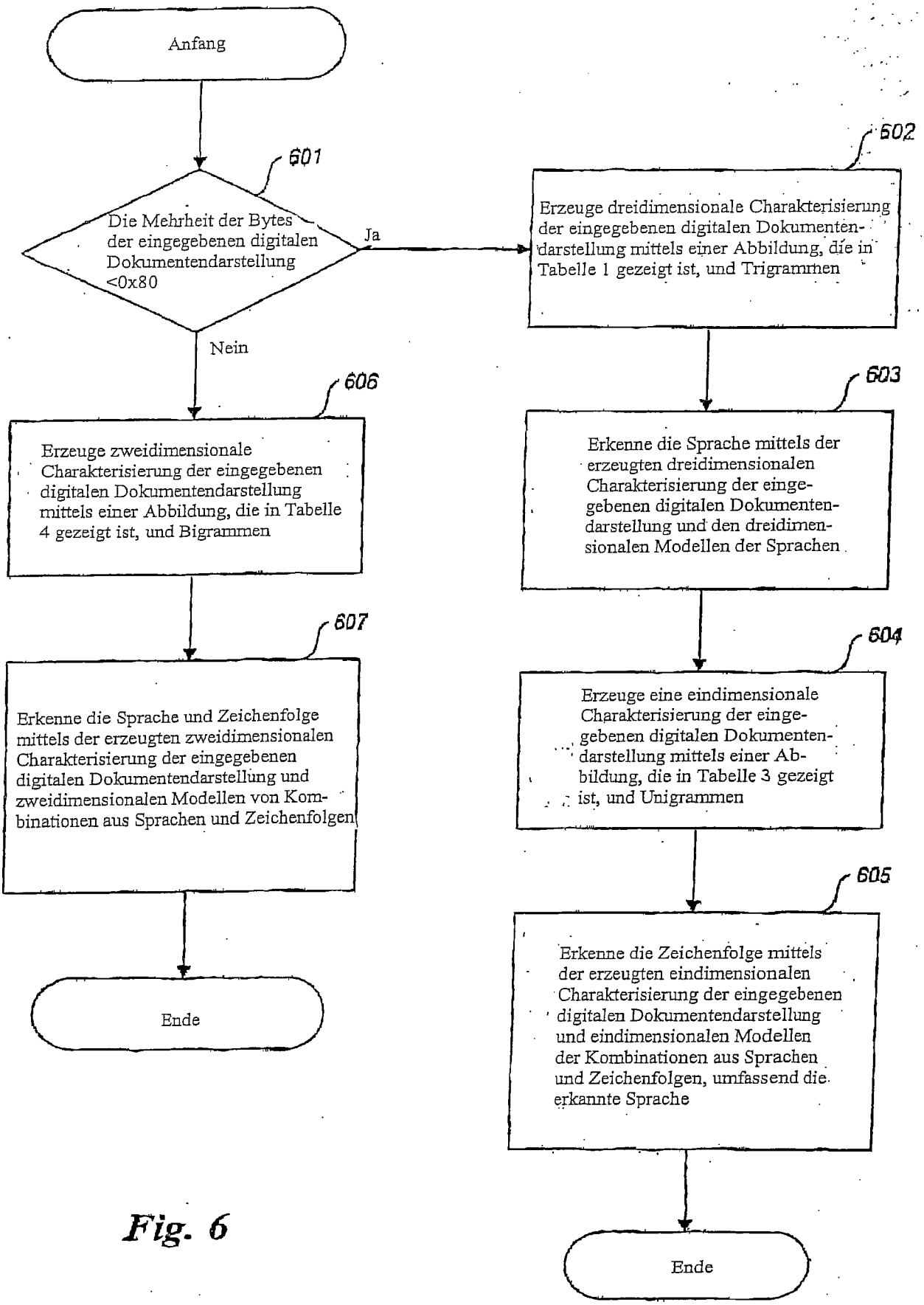
Anfang
601
Die Mehrheit der Bytes der eingegebenen digitalen Dokumentendarstellung <0x80
Ja
Nein
602
Erzeuge dreidimensionale Charakterisierung der eingegebenen digitalen Dokumentendarstellung mittels einer Abbildung, die in Tabelle 1 gezeigt ist, und Trigrammen
603
Erkenne die Sprache mittels der erzeugten dreidimensionalen Charakterisierung der eingegebenen digitalen Dokumentendarstellung und den dreidimensionalen Modellen der Sprachen
604
Erzeuge eine eindimensionale Charakterisierung der eingegebenen digitalen Dokumentendarstellung mittels einer Abbildung, die in Tabelle 3 gezeigt ist, und Unigrammen
605
Erkenne die Zeichenfolge mittels der erzeugten eindimensionalen Charakterisierung der eingegebenen digitalen Dokumentendarstellung und eindimensionalen Modellen der Kombinationen aus Sprachen und Zeichenfolgen, umfassend die erkannte Sprache
Ende
606
Erzeuge zweidimensionale Charakterisierung der eingegebenen digitalen Dokumentendarstellung mittels einer Abbildung, die in Tabelle 4 gezeigt ist, und Bigrammen
607
Erkenne die Sprache und Zeichenfolge mittels der erzeugten zweidimensionalen Charakterisierung der eingegebenen digitalen Dokumentendarstellung und zweidimensionalen Modellen von Kombinationen aus Sprachen und Zeichenfolgen
Ende

Fig. 6

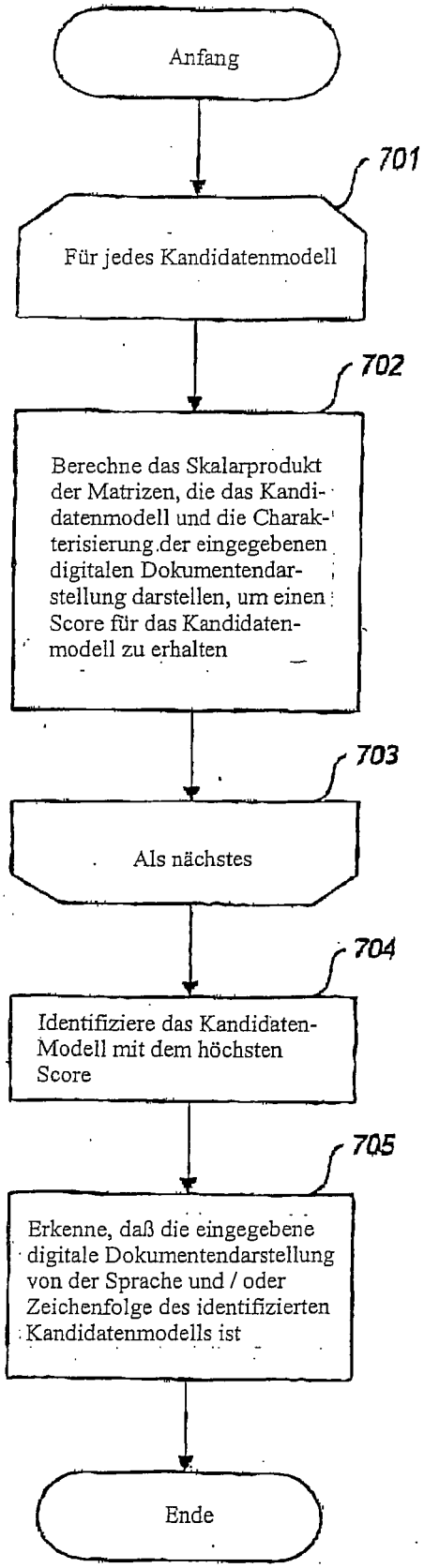
Anfang
701
Für jedes Kandidatenmodell
702
Berechne das Skalarprodukt der Matrizen, die das Kandidatenmodell und die Charakterisierung der eingegebenen digitalen Dokumentendarstellung darstellen, um einen Score für das Kandidatenmodell zu erhalten
703
Als nächstes
704
Identifiziere das Kandidaten-Modell mit dem höchsten Score
705
Erkenne, daß die eingegebene digitale Dokumentendarstellung von der Sprache und / oder Zeichenfolge des identifizierten Kandidatenmodells ist
Ende

Fig. 7

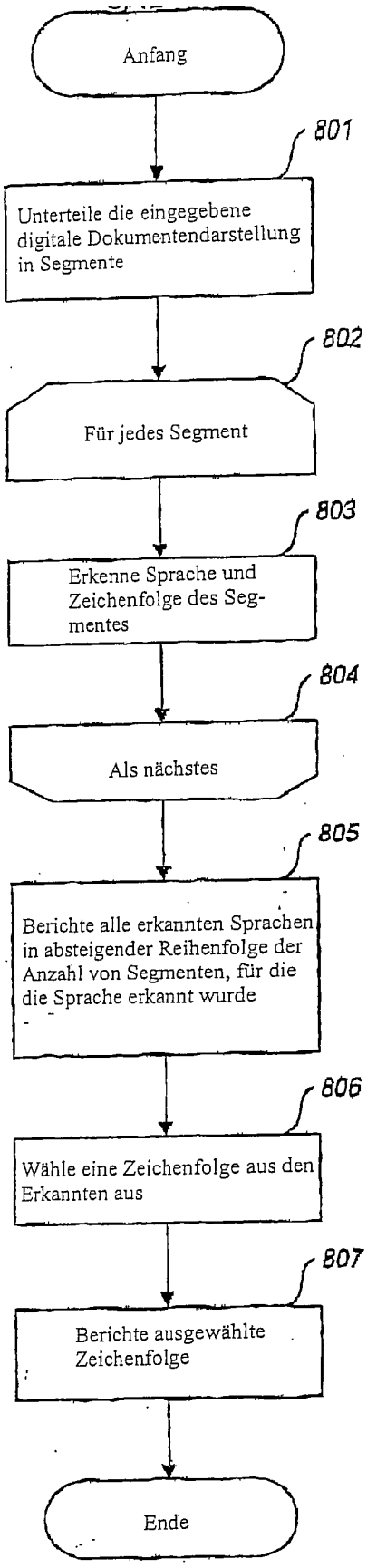
Anfang
801
Unterteile die eingegebene digitale Dokumentendarstellung in Segmente
802
Für jedes Segment
803
Erkenne Sprache und Zeichenfolge des Segmentes
804
Als nächstes
805
Berichte alle erkannten Sprachen in absteigender Reihenfolge der Anzahl von Segmenten, für die die Sprache erkannt wurde
806
Wähle eine Zeichenfolge aus den Erkannten aus
807
Berichte ausgewählte Zeichenfolge
Ende

Fig. 8

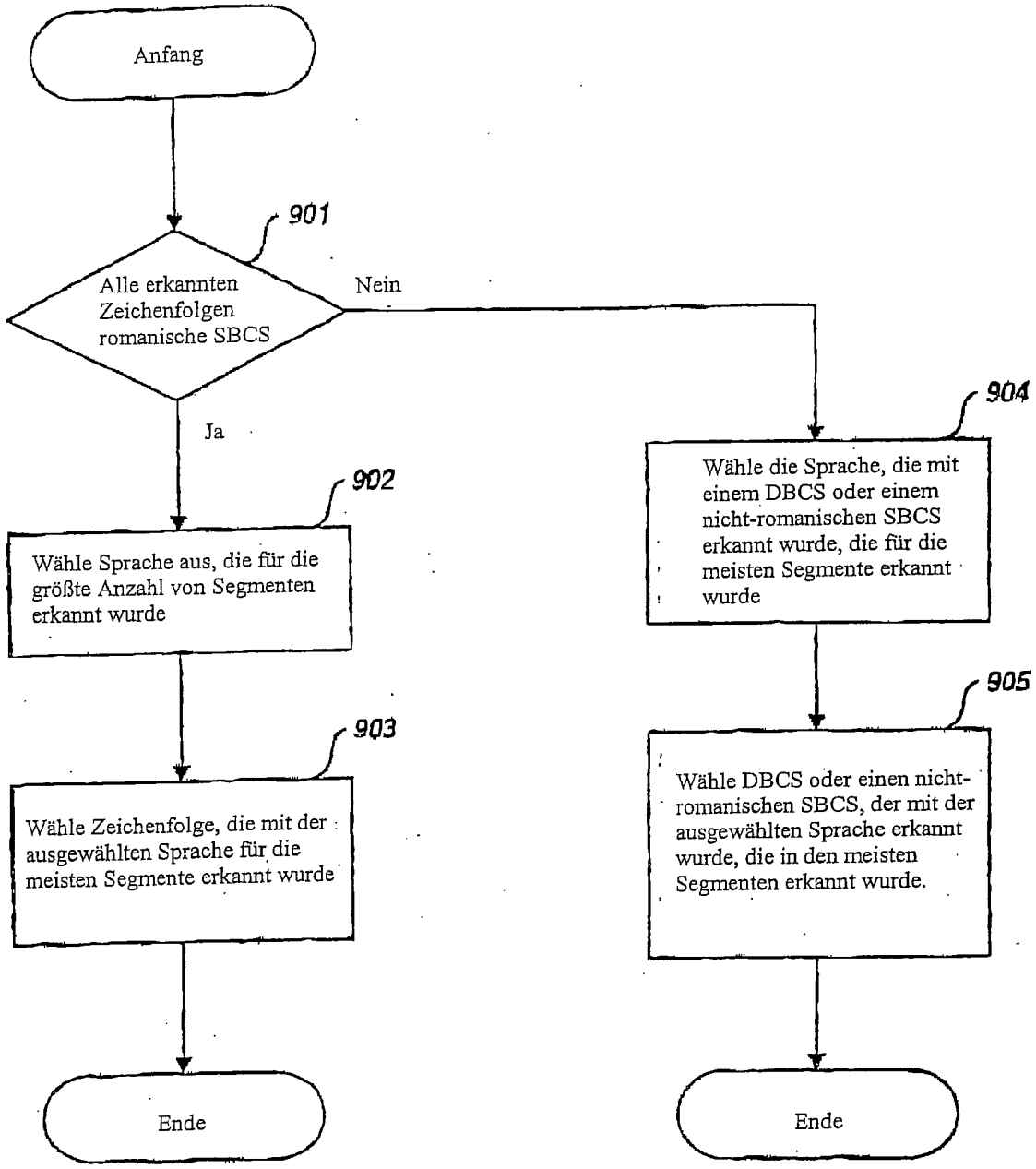
Anfang
901
Alle erkannten Zeichenfolgen romanische SBCS
Nein
Ja
902
Wähle Sprache aus, die für die größte Anzahl von Segmenten erkannt wurde
903
Wähle Zeichenfolge, die mit der ausgewählten Sprache für die meisten Segmente erkannt wurde
Ende
904
Wähle die Sprache, die mit einem DBCS oder einem nicht-romanischen SBCS erkannt wurde, die für die meisten Segmente erkannt wurde
905
Wähle DBCS oder einen nicht-romanischen SBCS, der mit der ausgewählten Sprache erkannt wurde, die in den meisten Segmenten erkannt wurde.
Ende

Fig. 9

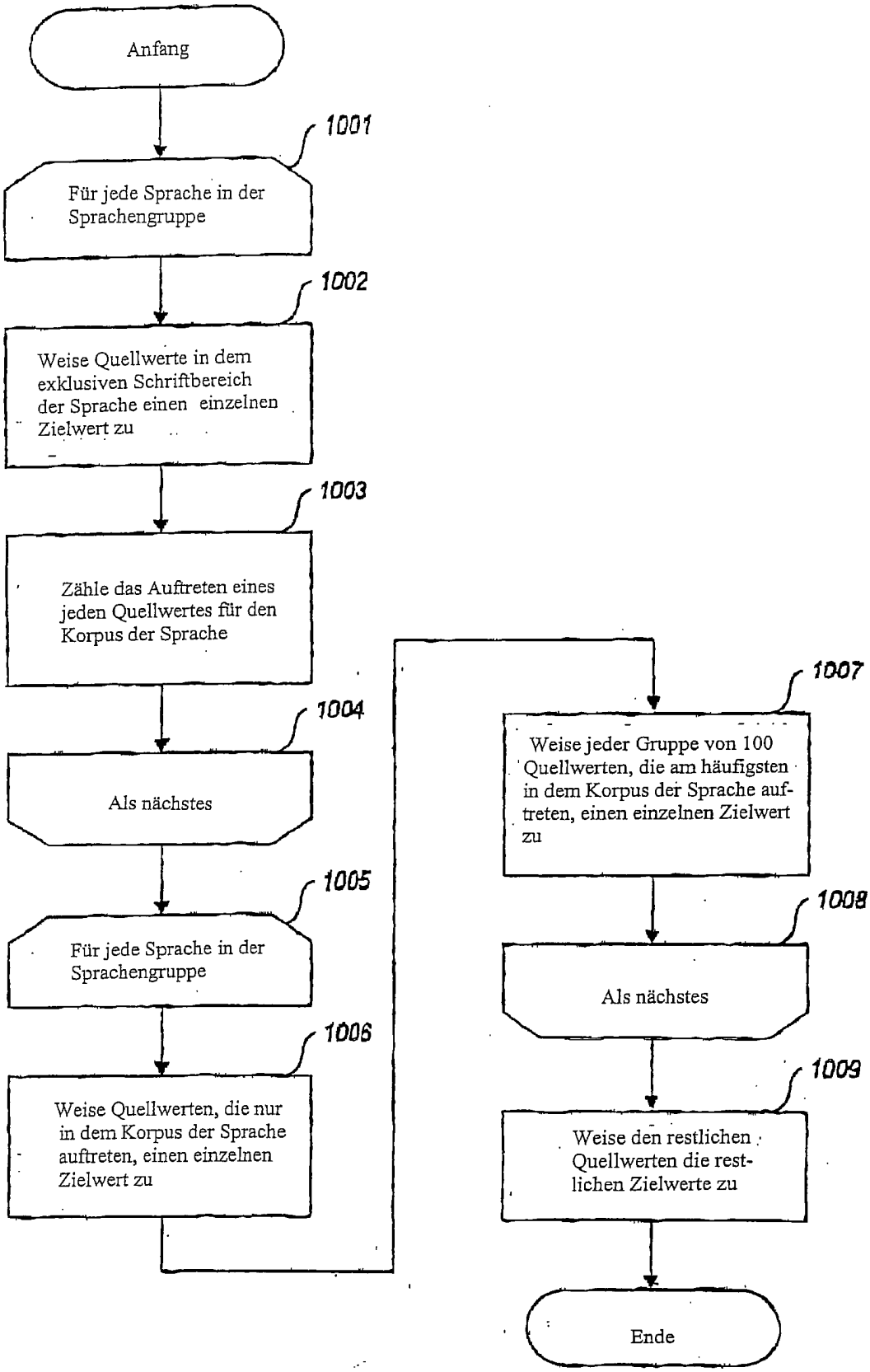
Anfang
1001
Für jede Sprache in der Sprachengruppe
1002
Weise Quellwerte in dem exklusiven Schriftbereich der Sprache einen einzelnen Zielwert zu
1003
Zähle das Auftreten eines jeden Quellwertes für den Korpus der Sprache
1004
Als nächstes
1005
Für jede Sprache in der Sprachengruppe
1006
Weise Quellwerten, die nur in dem Korpus der Sprache auftreten, einen einzelnen Zielwert zu
1007
Weise jeder Gruppe von 100 Quellwerten, die am häufigsten in dem Korpus der Sprache auftreten, einen einzelnen Zielwert zu
1008
Als nächstes
1009
Weise den restlichen Quellwerten die restlichen Zielwerte zu
Ende

Fig. 10

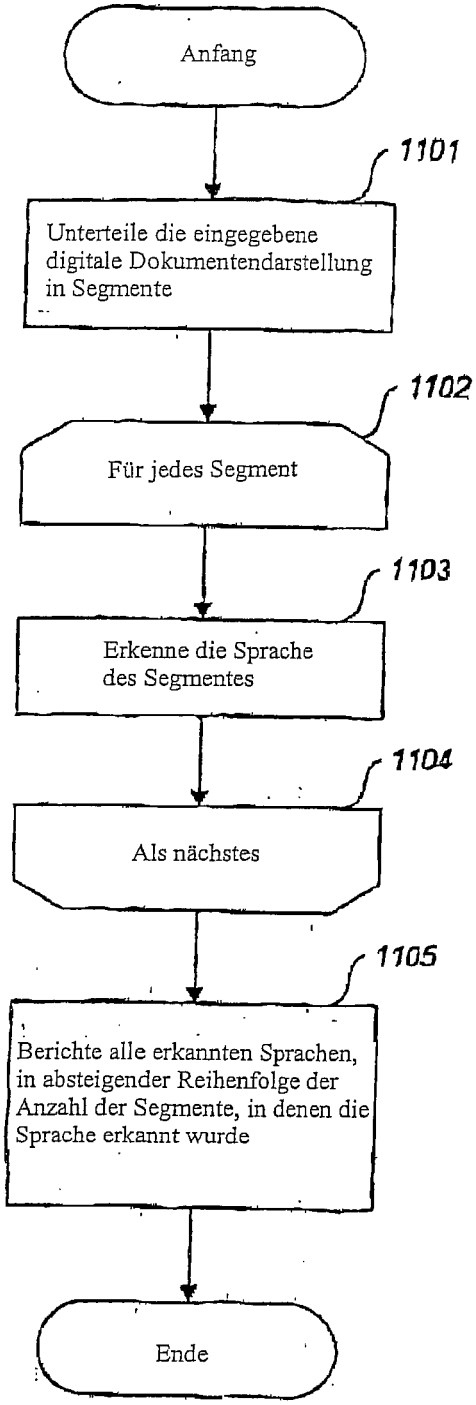
Anfang
1101
Unterteile die eingegebene digitale Dokumentendarstellung in Segmente
1102
Für jedes Segment
1103
Erkenne die Sprache des Segmentes
1104
Als nächstes
1105
Berichte alle erkannten Sprachen, in absteigender Reihenfolge der Anzahl der Segmente, in denen die Sprache erkannt wurde
Ende

Fig. 11

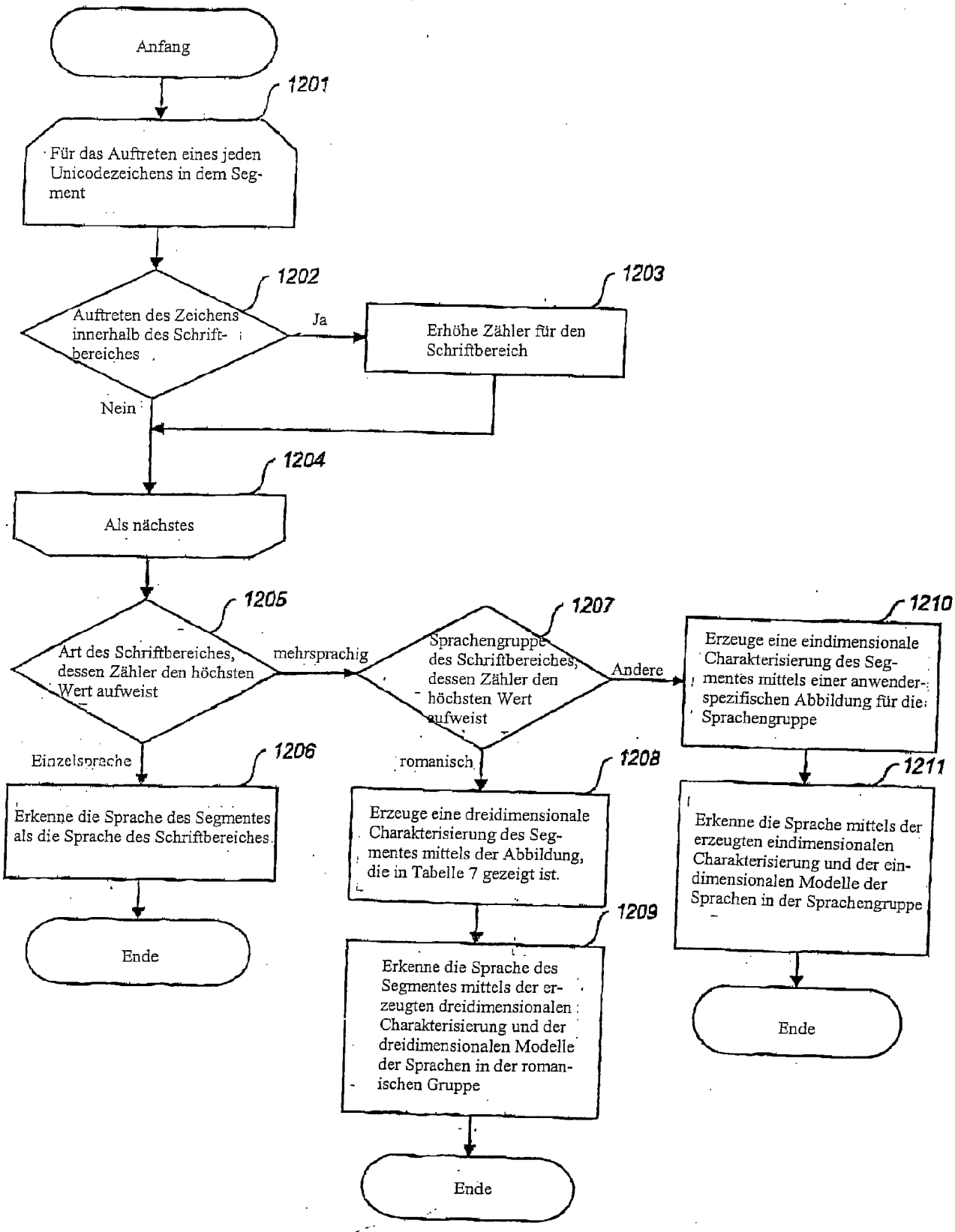
Anfang
1201
Für das Auftreten eines jeden Unicodezeichens in dem Segment
1202
Auftreten des Zeichens innerhalb des Schriftbereiches
Ja
1203
Erhöhe Zähler für den Schriftbereich
Nein
1204
Als nächstes
1205
Art des Schriftbereiches, dessen Zähler den höchsten Wert aufweist
mehrsprachig
1207
Sprachengruppe des Schriftbereiches, dessen Zähler den höchsten Wert aufweist
Andere
1210
Erzeuge eine eindimensionale Charakterisierung des Segmentes mittels einer anwenderspezifischen Abbildung für die Sprachengruppe
Einzelsprache
1206
Erkenne die Sprache des Segmentes als die Sprache des Schriftbereiches
Ende
romanisch
1208
Erzeuge eine dreidimensionale Charakterisierung des Segmentes mittels der Abbildung, die in Tabelle 7 gezeigt ist.
1209
Erkenne die Sprache des Segmentes mittels der erzeugten dreidimensionalen Charakterisierung und der dreidimensionalen Modelle der Sprachen in der romanischen Gruppe
Ende
1211
Erkenne die Sprache mittels der erzeugten eindimensionalen Charakterisierung und der eindimensionalen Modelle der Sprachen in der Sprachengruppe
Ende

Fig. 12