(19) Bundesrepublik Deutschland
Deutsches Patent- und Markenamt

(10) **DE 100 28 624 B4** 2007.07.05

(12) **Patentschrift**

(21) Aktenzeichen: **100 28 624.0**
(22) Anmeldetag: **09.06.2000**
(43) Offenlegungstag: **23.05.2001**
(45) Veröffentlichungstag der Patenterteilung: **05.07.2007**

(51) Int Cl.[8]: ***G06F 17/30*** (2006.01)

Innerhalb von drei Monaten nach Veröffentlichung der Patenterteilung kann nach § 59 Patentgesetz gegen das Patent Einspruch erhoben werden. Der Einspruch ist schriftlich zu erklären und zu begründen. Innerhalb der Einspruchsfrist ist eine Einspruchsgebühr in Höhe von 200 Euro zu entrichten(§ 6 Patentkostengesetz in Verbindung mit der Anlage zu § 2 Abs. 2 Patentkostengesetz).

(30) Unionspriorität:

11-162068	**09.06.1999**	**JP**
11-360369	**20.12.1999**	**JP**

(73) Patentinhaber:
Ricoh Co., Ltd., Tokyo, JP

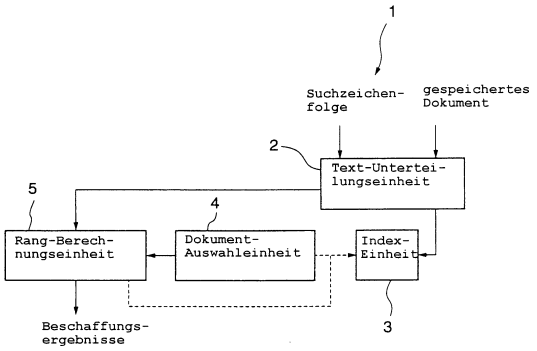
(74) Vertreter:
Schwabe, Sandmair, Marx, 81677 München

(72) Erfinder:
Ogawa, Yasushi, Yokohama, Kanagawa, JP

(56) Für die Beurteilung der Patentfähigkeit in Betracht gezogene Druckschriften:
DE 42 32 507 A1
Development and Evaluation of Full-Document-Based Retrieval System 'Retrieval Express'". In: Proceedings of the Third Annual Meeting of the Association for Natural Language Processing, S.361-364, März 1997;
CHVICHI, KIKUCHI: " A Fast Full-Text Search Method for Japanese Text Database". In: Transactions of the Institute of Electronics, Information and Communication Engeneers, Band J75-D-I, Nr.9, S.836-846, 1992;

(54) Bezeichnung: **Verfahren und Vorrichtung zur Dokumentenbeschaffung**

(57) Hauptanspruch: Verfahren zur Dokumentenbeschaffung, mit folgenden Schritten:
wenigstens eine Suchzeichenfolge wird bereitgestellt;
eine Anzahl Dokumente wird aus einer Vielzahl von gespeicherten Dokumenten ausgewählt;
ein Rangpunkt der wenigstens einen Suchzeichenfolge wird berechnet;
gekennzeichnet durch die folgenden Schritte:
a) die wenigstens eine Suchzeichenfolge wird in partielle Zeichenfolgen unterteilt;
b) jedes der Dokumente der Anzahl Dokumente wird derartig ausgewählt, dass jedes alle partiellen Zeichenfolgen enthält;
c) jeweilige Rangpunkte der partiellen Zeichenfolgen werden für jedes der Anzahl Dokumente berechnet; und
d) der Rangpunkt der jeweiligen Suchzeichenfolge wird von den jeweiligen Rangpunkten der partiellen Zeichenfolgen für jedes der Anzahl Dokumente berechnet.



Beschreibung

[0001] Die vorliegende Erfindung betrifft ein Verfahren zur Dokumentenbeschaffung nach Anspruch 1, eine Vorrichtung zur Dokumentenbeschaffung nach Anspruch 12 sowie ein von einem Computer lesbares Aufzeichnungsmedium nach Anspruch 24.

[0002] Dokumentenbeschaffungstechniken (Dokumenten-„Retrieval"-Techniken) beschaffen Dokumente, die eine Suchzeichenfolge enthalten, von einer Dokumenten-Datenbank. Eine derartige Dokumentenbeschaffungstechnik ist ein Wahrscheinlich-Relevanz-Beschaffungsschema („likelyrelevance retrieval scheme"), das Dokumente beschafft (sucht, lokalisiert und/oder abruft), die Zeichenfolgen enthalten, die einer Suchzeichenfolge ähnlich sind.

[0003] Die Wahrscheinlich-Relevanz-Beschaffungstechnik bzw. die Beschaffungstechnik, die auf einer wahrscheinlichen Relevanz basiert, ist zum Beispiel in der japanischen offengelegten Patentanmeldung Nr. 11-85776 offenbart. Diese Technik berechnet Rangordnungen bzw. Rangordnungspunkte partieller Zeichenfolgen, die Teile einer Suchzeichenfolge sind, und zwar basierend auf der Häufigkeit des Auftretens und sucht nach der Suchzeichenfolge in dem Dokument, in dem die erhaltenen Rangpunkte bzw. Rangordnungspunkte verwendet werden.

[0004] Ein anderes Beispiel der Wahrscheinlich-Relevanz-Beschaffungstechnik findet man in „Development and Evaluation of Full-Document-Based Retrieval System" 'Retrieval Express'", Proceedings of the Third Annual Meeting of the Association for Natural Language Processing, Seiten 361-364, März 1997. Diese Technik erhält die Häufigkeit des Auftretens einer Suchzeichenfolge in einem Dokument, in dem alle Positionen eines derartigen Auftretens in dem Dokument, basierend auf dem Auftreten von partiellen Zeichenfolgen erhalten werden und berechnet eine Rangordnung der Suchzeichenfolgen bezüglich des Dokuments. Die Technik, die in der obigen offengelegten Patentanmeldung offenbart ist, sucht jedoch lediglich nach einer Suchzeichenfolge in einem einzelnen Dokument und kann nicht verwendet werden, um ein Dokument wiederzubeschaffen, das eine Suchzeichenfolge von einer Vielzahl von Dokumenten enthält.

[0005] Weiter werden je länger die Suchzeichenfolge ist, desto größer die Anzahl der partiellen Zeichenfolgen, die bei der Suche zu berücksichtigen sind. Ebenso ist je länger die Suchzeichenfolge ist, desto größer die Anzahl der Dokumentsegmente, die zur Berechnung der Rangpunkte („ranking scores") zu verarbeiten sind. Dies führt zu einer Zunahme in der Beschaffungszeit. Wenn zum Beispiel eine Suchzeichenfolge „ABCDEF" lautet (jeder Großbuchstabe stellt ein einziges japanisches Zeichen zum Zwecke der Erläuterung dar) und partielle Zeichenfolgen, die jeweils aus zwei Zeichen bestehen, als eine Einheit bei der Verarbeitung verwendet werden, kann man fünf partielle Zeichenfolgen, d.h. „AB", „BC", „CD", „DE" und „EF" extrahieren. Wenn im allgemeinen eine Suchzeichenfolge aus m-Zeichen besteht und n-Zeichen eine Verarbeitungseinheit bilden, kann man (m – n + 1) partielle Zeichenfolge extrahieren. Da der Rangpunkt bei jeder Position zu berechnen ist, wo wenigstens eine extrahierte partielle Zeichenfolge erscheint, nimmt die Anzahl der Positionen, für die eine Berechnung erforderlich ist, mit der Anzahl der partiellen Zeichenfolgen zu.

[0006] Ein Rangpunkt partieller Zeichenfolgen in dem Dokument wird basierend auf der Häufigkeit des Auftretens der partiellen Zeichenfolge in dem Dokument berechnet. Für manche der partiellen Zeichenfolgen, die in dem Dokument erscheinen, kann gelten, dass sie nicht durch die Suchzeichenfolge getragen werden. Dennoch werden diese für die Rangpunkte mitgezählt. Dies reduziert die Genauigkeit der Suche. Zum Beispiel kann die Suchzeichenfolge „ABCDEF" nur einmal bei einem gegebenen Dokument erscheinen und eine andere Zeichenfolge „WXYZEF", die eine vollständig unterschiedliche Bedeutung hat, kann mehrere Male in diesem Dokument erscheinen. In einem derartigen Fall erscheint die partielle Suchzeichenfolge „EF" so häufig, wie die Anzahl des Auftretens von „ABCDEF" plus die Anzahl des Auftretens von „WXYZEF". Infolgedessen wird der Rangpunkt der partiellen Zeichenfolge „EF" letztendlich unangemessen hoch, obwohl die Suchzeichenfolge nur selten auftritt, was zu einem unangemessen hohen Rangpunkt für die Suchzeichenfolge führt.

[0007] Ein anderes Problem liegt darin, dass die Suche nicht durchgeführt werden kann, falls die Länge einer Suchzeichenfolge kürzer als eine Verarbeitungseinheit ist. Dies liegt daran, dass die Suchzeichenfolge nicht in partielle Zeichenfolgen unterteilt werden kann, die die Länge der Verarbeitungseinheit haben. Falls zum Beispiel die Suchzeichenfolge „B" ist und zwei Zeichen eine Verarbeitungseinheit bilden, kann die Suche gemäß diesem Verfahren nicht durchgeführt werden, da die Suchzeichenfolge kürzer als die Verarbeitungseinheit ist.

[0008] Die Technik, die in „Development and Evaluation of Full-Document-Based Retrieval System 'Retrieval Express'", Proceedings of the Third Annual Meeting of the Association for Natural Language Processing, Seiten 361-364, März 1997 offenbart ist, hat dasselbe Problem, wie die Technik in der obigen offengelegten Patentanmeldung. Das heißt, der Umfang der Berechnung zum Zählen von Auftreteereignissen einer Suchzeichenfolge in einem Dokument nimmt mit der Länge der Suchzeichenfolge zu, was zu einer Verlängerung einer Verarbeitungszeit für die

Dokumentbeschaffung führt. Je größer die Anzahl der Auftreteereignisse einer Suchzeichenfolge, desto auffälliger ist die Zunahme der Verarbeitungszeit der Dokumentbeschaffung.

Einschub Seite 3a

[0009] Aufgabe der Erfindung ist es, ein Verfahren, eine Vorrichtung und ein von einem Computer lesbares Aufzeichnungsmedium mit einem dem Verfahren entsprechenden Programm bereitzustellen, bei dem ein Beschaffungsschema eingesetzt wird, das ein Dokument mit hoher Geschwindigkeit beschaffen kann.

Einschubseite 3a

[0010] Die DE 42 32 507 A1 offenbart ein Verfahren sowie eine Vorrichtung nach dem Oberbegriff des Anspruchs 1 bzw. des Anspruchs 12. Laut dieser Veröffentlichung wird eine Anzahl von Zeichenfolgen zur Verfügung gestellt, und eine Anzahl von Dokumenten wird nach dem Vorhandensein dieser Zeichenfolgen abgesucht. Anschließend werden Rangpunkte für die Anzahl von Treffern für die jeweiligen Zeichenfolgen in dem Dokument erzeugt und auf der Basis der Rangpunkte wird bewertet, ob das gefundene oder die mehreren gefundenen Dokumente über eines oder mehrere der gesuchten Merkmale verfügen.

[0011] Vorstehende Aufgabe wird durch die Gegenstände der unabhängigen Ansprüche gelöst. Vorteilhafte Weiterbildungen gehen aus den Unteransprüchen hervor.

[0012] Vorteilhaft wird ein Beschaffungsschema eingesetzt, bei dem die Computerlast der Auswahl eines Dokuments und der Berechnung von Rangpunkten reduziert wird, wodurch eine Hochgeschwindigkeitsverarbeitung erzielt wird.

[0013] Vorteilhaft wird ein Beschaffungsschema eingesetzt, das frei von einem Einfluss von anderen Zeichenfolgen ist, die für eine Suchzeichenfolge nicht relevant sind, wodurch die Beschaffungsgenauigkeit verbessert wird.

[0014] Vorteilhaft wird ein Beschaffungsschema eingesetzt, bei dem die Computerlast zum Erzielen bzw. Erfassung von Positionen des Auftretens einer Suchzeichenfolge reduziert werden kann, wodurch eine Dokumentbeschaffung mit hoher Geschwindigkeit erzielt wird.

[0015] Vorteilhaft wird ein Beschaffungsschema eingesetzt, bei dem die Anzahl der Rangpunktsuchvorgänge bzw. Rang relevante Suchvorgänge reduziert werden kann, wodurch die Suchgeschwindigkeit erhöht wird.

[0016] Vorteilhaft wird ein Beschaffungsschema eingesetzt, das ein Dokument selbst dann beschaffen kann, wenn die Länge einer Suchzeichenfolge kürzer als die Verarbeitungseinheit ist.

[0017] Vorteilhaft wird ein Beschaffungsschema eingesetzt, bei dem die Berechnungslast für die Berechnung der Rangpunkt reduziert wird, wodurch eine Beschaffung mit hoher Geschwindigkeit erzielt wird.

[0018] „Ein Dokument oder mehrere Dokumente" werden hierin auch kurz als „Anzahl von Dokumenten" bezeichnet, wobei es sich dabei um ein einzelnes Dokument oder mehrere oder viele Dokumente handeln kann.

[0019] Bei dem oben beschriebenen Verfahren werden das eine Dokument oder die mehreren Dokumente, die partielle Zeichenfolgen enthalten, die der Suchzeichenfolge ähneln, vor der Berechnung der Punkte bzw. der Rangpunkte ausgewählt. Aufgrund dieses Filterungsprozesses wird die Hochgeschwindigkeits-Dokumentbeschaffung erzielt, um ein Dokument aus der Vielzahl von gespeicherten Dokumenten zu beschaffen.

[0020] Vorteilhaft ist ein Verfahren derartig, dass der Schritt des Unterteilens die Suchzeichenfolge in partielle Zeichenfolgen unterteilt, die sich im allgemeinen nicht überlappen und die eine volle Länge der Suchzeichenfolge abdecken.

[0021] Bei dem oben beschriebenen Verfahren kann die Berechnungslast der Auswahl des einen Dokuments oder von mehreren Dokumenten und der Berechnung von Punkten bzw. Rängen reduziert werden, wodurch einen Beschaffung von Dokumenten mit hoher Geschwindigkeit erzielt wird.

[0022] Vorteilhaft ist das beschriebene Verfahren derartig, dass der Schritt der Berechnung jeweiliger Rangpunkte bzw. Punkte für die partiellen Zeichenfölgen die Schritte enthält, wonach ein erster Zählwert erhalten wird, der anzeigt, wie viele der gespeicherten Dokumente eine gegebene Folge der partiellen Zeichenfolgen enthalten, zweite Zählwerte erhalten werden, die jeweils anzeigen, wie viele Male eine ent-sprechende Folge der partiellen Zeichenfolgen bei einem gegebenen Dokument des einen Dokuments oder mehreren Dokumente erscheint, der kleinste der zweiten Zählwerte bzw. Zählungen erhalten wird und ein Rangpunkt bzw. Punkt der gegebe-nen einen Zeichenfolge der partiellen Zeichenfolgen für das gegebene eine Dokument des einen Dokuments oder der mehreren Dokumente von dem ersten Zählwert und dem kleinsten der zweiten Zählwerte derartig erhalten wird, dass der Punkt bzw. Rangpunkt der gegebenen Folge der partiellen Zeichenfolgen mit der Abnahme des ersten Zählwertes

und mit der Zunahme des kleinsten der zweiten Zählwerte zunimmt.

[0023] Vorteilhaft kann der Anschluss eines irrelevanten Auftretens partieller Zeichenfolgen reduziert werden, wenn Ränge bzw. Punkte berechnet werden, wodurch die Beschaffungsgenauigkeit verbessert wird.

[0024] Vorteilhaft ist das oben zuerst beschriebene Verfahren derartig, dass der Schritt der Berechnung jeweiliger Punkte bzw. Rangpunkte der partiellen Zeichenfolgen die Schritte enthält, wonach ein erster Zählwert erhalten wird, der anzeigt, wie viele der gespeicherten Dokumente eine gegebene Zeichenfolge der partiellen Zeichenfolgen enthalten, wonach ein zweiter Zählwert erhalten wird, der anzeigt, wie viele Male die Suchzeichenfolge in einem gegebenen Dokument des einen Dokuments oder der mehreren Dokumente erscheint, und wonach ein Punkt bzw. ein Rangpunkt der gegebenen einen Folge der partiellen Zeichenfolgen für das gegebene eine Dokument des einen Dokuments oder der mehreren Dokumente von dem ersten Zählwert und dem zweiten Zählwert derartig erhalten wird, dass der Punkt bzw. Rangpunkt der gegebenen einen Folge der partiellen Zeichenfolgen mit der Abnahme des ersten Zählwertes und mit der Zunahme des zweiten Zählwertes abnimmt.

[0025] Der Einfluss eines irrelevanten Auftretens partieller Zeichenfolgen innerhalb eines Dokuments kann beseitigt werden, wenn Punkte berechnet werden, wodurch die Beschaffungsgenauigkeit verbessert wird.

[0026] Vorteilhaft ist das oben beschriebene Verfahren derartig, dass der Schritt der Erzielung eines zweiten Zählwerts weiter einen Schritt enthält, wonach eine obere Grenze auf den zweiten Zählwert platziert wird bzw. der zweite Zählwert mit einer oberen Grenze versehen wird.

[0027] Bei dem oben beschriebenen Verfahren kann die Berechnungslast bei der Detektion von Positionen der Suchzeichenfolge reduziert werden, wodurch eine Beschaffung eines Dokuments mit hoher Geschwindigkeit unterstützt wird.

[0028] Vorteilhaft ist das oben zuerst beschriebene Verfahren dergestalt, dass der Schritt der Auswahl eines Dokuments oder mehrere Dokumente das eine Dokument oder mehrere Dokumente auswählt, von denen jedes die Suchzeichenfolge enthält, und der Schritt der Berechnung jeweiliger Rangpunkte der partiellen Zeichenfolgen die Schritte enthält, wonach ein erster Zählwert anzeigt, wie viele der registrierten Dokumente die Suchzeichenfolge enthalten, ein zweiter Zählwert anzeigt, wie viele Male eine gegebene Folge der partiellen Zeichenfolgen in einem gegebenen Dokument des einen Dokuments oder der mehreren Dokumente erscheint und ein Rangpunkt der gegebenen Folge der partiellen Zeichenfolgen für das gegebene eine Dokument des einen Dokuments oder der mehreren Dokumente von dem ersten Zählwert und dem zweiten Zählwert derartig erhalten wird, dass der Rangpunkt bzw. Punkt der gegebenen einen Zeichenfolge der partiellen Zeichenfolgen mit der Abnahme des ersten Zählwertes und der Zunahme des zweiten Zählwertes zunimmt.

[0029] Bei dem oben beschriebenen Verfahren kann der Einfluss von irrelevanten Auftreteereignissen der partiellen Zeichenfolgen über unterschiedliche Dokumente beseitigt werden, wodurch zur verbesserten Genauigkeit der Dokumentbeschaffung beigetragen wird.

[0030] Vorteilhaft ist das oben zuerst beschriebene Verfahren derartig, dass der Schritt der Auswahl eines Dokuments oder mehrerer Dokumente das eine Dokument oder die mehreren Dokumente auswählt, von denen jedes die Suchzeichenfolge enthält, und der Schritt der Berechnung jeweiliger Rangpunkte bzw. Punkte der partiellen Zeichenfolgen die Schritte enthält, wonach ein erster Zählwert erhalten wird, der anzeigt, wie viele der gespeicherten Dokumente die Suchzeichenfolge enthalten, eine Grenze von dem ersten Zählwert berechnet wird, ein zweiter Zählwert erhalten wird, der anzeigt, wie viele Male die Suchzeichenfolge in einem gegebenen Dokument des einen Dokuments oder der mehreren Dokumente erscheint, während ein oberes Ende des zweiten Zählwertes auf die Grenze beschränkt wird, und ein Rangpunkt bzw. ein Punkt einer gegebenen einen Folge der partiellen Zeichenfolgen für das gegebene eine Dokument des einen Dokuments oder der mehreren Dokumente von dem ersten Zählwert und dem zweiten Zählwert derartig erhalten wird, dass der Rangpunkt der gegebenen einen Folge der partiellen Zeichenfolgen mit der Abnahme des ersten Zählwertes und mit der Zunahme des zweiten Zählwertes zunimmt.

[0031] Bei dem oben beschriebenen Verfahren kann der Einfluss irrelevanter Auftreteereignisse der partiellen Zeichenfolgen beseitigt werden und die Berechnungslast von Detektionspositionen der Suchzeichenfolge kann reduziert werden, wodurch dazu beigetragen wird, eine Beschaffung eines Dokuments genau und mit hoher Geschwindigkeit zu erzielen.

[0032] Vorteilhaft beinhaltet ein Verfahren zur Dokumentbeschaffung die Schritte, wonach jeweilige Indizes für Dokumente bereitgestellt werden, wobei jeder der jeweiligen Indizes partielle Zeichenfolgen, die in einem entsprechenden Dokument gefunden wurden, und jeweilige Positionen davon in dem entsprechenden Dokument auflistet, die partiellen Zeichenfolgen ausgewählt werden, die mit einer Zeichenfolge be-

ginnen, die mit einer Suchzeichenfolge identisch ist, ein Dokument oder mehrere Dokumente von den Dokumenten ausgewählt werden, so dass das eine Dokument oder die mehreren Dokumente jeweils wenigstens eine Folge der ausgewählten partiellen Zeichenfolgen enthalten, die jeweiligen Rangpunkte der ausgewählten partiellen Zeichenfolgen für jedes des einen Dokuments oder der mehreren Dokumente berechnet wird und ein Rangpunkt der Suchzeichenfolge von den jeweiligen Rangpunkten der ausgewählten partiellen Zeichenfolgen für jedes Dokument des einen oder der mehreren Dokumente berechnet wird.

[0033] Bei dem oben beschriebenen Verfahren kann eine geeignete Dokumentenbeschaffung selbst dann erzielt werden, wenn die Suchzeichenfolge kürzer als die Länge der partiellen Zeichenfolge ist.

[0034] Bei der folgenden Beschreibung von Ausführungsformen werden weitere Merkmale offenbart, wobei Merkmale unterschiedlicher Ausführungsformen kombiniert werden können.

[0035] Fig. 1 ist ein Blockdiagramm einer Dokumentbeschaffungsvorrichtung gemäß der ersten Ausführungsform der vorliegenden Erfindung;

[0036] Fig. 2A und Fig. 2B sind erläuternde Zeichnungen, die gespeicherte Dokumente zeigen;

[0037] Fig. 3 ist ein Blockdiagramm einer Systemkonfiguration, die die Dokumentbeschaffungsvorrichtung der Fig. 1 realisiert;

[0038] Fig. 4 ist ein Flussdiagramm eines Prozesses zum Berechnung von Rangfolgen für mehrere Dokumente, wobei der Prozess durch die Dokumentbeschaffungsvorrichtung gemäß der ersten Ausführungsform der vorliegenden Erfindung durchgeführt wird;

[0039] Fig. 5 ist ein Flussdiagramm eines Prozesses der Berechnung eines Rangordnungspunktes bzw. einer Rangordnung, die bei einem Schritt S3 der Fig. 4 durchgeführt wird;

[0040] Fig. 6 ist ein Flussdiagramm eines Prozesses der Berechnung eines Rangordnungspunktes bzw. eines Rangpunktes entsprechend der zweiten Ausführungsform der vorliegenden Erfindung;

[0041] Fig. 7 ist ein Flussdiagramm eines Prozesses der Berechnung eins Rangpunktes bzw. eines Rangordnungspunktes gemäß der dritten Ausführungsform der vorliegenden Erfindung;

[0042] Fig. 8 ist ein Flussdiagramm eines Prozesses zum Erzielen eines Auftretungsereignis-Zählwertes einer Suchzeichenfolge mit einer oberen Grenze gemäß der vierten Ausführungsform der vorliegenden Erfindung;

[0043] Fig. 9 ist ein Flussdiagramm eines Prozesses zur Berechnung der Rangordnungspunkte einer Vielzahl von Dokumenten gemäß der fünften Ausführungsform der vorliegenden Erfindung;

[0044] Fig. 10 ist ein Flussdiagramm eines Prozesses der Berechnung von Rangpunkten bzw. Rangordnungspunkten für eine Vielzahl von Dokumenten gemäß der sechsten Ausführungsform der vorliegenden Erfindung;

[0045] Fig. 11A bis Fig. 11C sind erläuternde Zeichnungen, die Beispiele von Dokumenten und ein Beispiel einer entsprechenden Indexeinheit zeigen;

[0046] Fig. 12 ist ein Flussdiagramm eines Prozesses zum Berechnen von Rangpunkten bzw. Rangordnungspunkten für eine Vielzahl von Dokumenten gemäß der siebten Ausführungsform der vorliegenden Erfindung;

[0047] Fig. 13 ist ein Blockdiagramm der Dokumentbeschaffungsvorrichtung gemäß der achten Ausführungsform der vorliegenden Erfindung; und

[0048] Fig. 14 ist ein Flussdiagramm eines Prozesses der Auswahl partieller Zeichenfolgen, die sich nicht überlappen und die eine volle Länge einer Suchzeichenfolge abdecken.

[0049] Im folgenden werden die Ausführungsformen der vorliegenden Erfindung unter Bezugnahme auf die beigefügten Zeichnungen beschrieben.

[0050] Fig. 1 ist ein Blockdiagramm einer Dokumentbeschaffungsvorrichtung **1** gemäß einer ersten Ausführungsform der vorliegenden Erfindung. Die Dokumentbeschaffungsvorrichtung **1** beinhaltet eine Textunterteilungseinheit **2**, eine Indexeinheit **3**, eine Dokumentauswahleinheit **4** und eine Rangberechnungseinheit **5**.

[0051] Die Textunterteilungseinheit **2** unterteilt einen Text in partielle Zeichenfolgen, wo der Text ein gespeichertes Dokument sein kann oder eine Suchzeichenfolge sein kann. Die Indexeinheit **3** speichert darin Information über partielle Zeichenfolgen, die durch Unterteilen eines gespeicherten Dokuments erhalten werden. Die Dokumentauswahleinheit **4** verwendet partielle Zeichenfolgen, die durch Unterteilen einer Suchzeichenfolge erhalten werden, um ein Dokument auszuwählen, für das eine Rangordnung zu berechnen ist. Die Rangberechnungseinheit **5** verwendet partielle Zeichenfolgen, die durch Unterteilen der Suchzeichenfolge erhalten werden, um einen Rangordnungspunkt des Dokuments zu berechnen, das durch die Dokumentauswahleinheit **4** ausgewählt wird. Die Textunterteilungseinheit **2** führt einen

Unterteilungsschritt durch und die Dokumentauswahleinheit **4** führt einen Dokumentauswahlschritt durch. Weiter führt die Rangberechnungseinheit **5** einen Rangberechnungsschritt durch. Details eines jeden Schritts werden später beschrieben.

[0052] Wenn ein Dokument, das zu Speichern bzw. zu Registrieren ist, bereitgestellt wird, unterteilt die Textunterteilungseinheit **2** das Dokument in partielle Zeichenfolgen. Die Information über das Auftreten partieller Zeichenfolgen wird in der Indexeinheit **3** gespeichert.

[0053] Im folgenden wird ein Prozess, der durch die Dokumentbeschaffungsvorrichtung **1** durchgeführt wird, detailliert beschrieben.

[0054] Die Fig. 2A und Fig. 2B sind erläuternde Zeichnungen, die gespeicherte Dokumente zeigen. Jede der Fig. 2A und Fig. 2B zeigt ein gespeichertes Dokument. Bei jedem gespeicherten Dokument zeigen Nummern bzw. Zahlen, die links gezeigt sind, die Zahl der Zeichen an, die von dem Beginn eines Dokuments zu einer Position einer entsprechenden Zeichenfolge gezählt wurden. In einem Dokument der Fig. 2A, beginnt die Zeichenfolge „ABCD" bei dem elften Zeichen von Beginn des Dokuments an und „EF" wird bei dem 20. Zeichen und bei dem 60. Zeichen von Beginn an gefunden. Weiter erscheint die Zeichenfolge „ABCDEF" beim 31. Zeichen von Beginn an. Wenn Zeichenfolgen mit zwei Zeichen als eine Einheit zur Verarbeitung verwendet werden, werden nur zwei Zeichen-Zeichenfolgen von einem Dokument extrahiert und die extrahierten Zeichenfolgen werden in der Indexeinheit **3** zusammen mit ihren Positionen und ihrem Auftreten aufgezeichnet (Zeichenzählung von Beginn des Dokuments an).

[0055] Fig. 2C stellt eine erläuternde Zeichnung dar, die den Inhalt der Indexeinheit **3** zeigt. Zum Beispiel hat das Dokument, das in Fig. 2A gezeigt ist, die Zeichenfolge „AB", die bei dem 11. Zeichen und bei dem 31. Zeichen von Beginn an auftritt, und hat die Zeichenfolge „BC", die bei dem 12. Zeichen und bei dem 32. Zeichen von Anfang an beginnt aufzutreten, so dass diese Auftreteereignisse in dem Dokument in der Indexeinheit **3** aufgezeichnet werden, wie in Fig. 2C gezeigt ist. Die Indexeinheit **3** zeichnet nicht nur die Positionen des Auftretens auf, sondern zeichnet ebenso Dokumentidentifizierer zum Identifizieren von Dokumenten, die relevant für die aufgezeichneten Auftreteereignisse sind, auf. Weiter werden die Anzahl der Auftreteereignisse ebenso aufgezeichnet. Wie in Fig. 2C gezeigt ist, wird die Zeichenfolge „AB" als „{1, 2, (11, 31)}" aufgezeichnet, was anzeigt, dass die Zeichenfolge „AB" zweimal auftritt (die Anzahl der Auftreteereignisse = 2) in dem Dokument der Fig. 2A mit dem Dokumentidentifizierer **1**. Diese Auftreteereignisse werden bei dem 11. Zeichen und bei dem 31. Zeichen von Beginn an gefunden.

[0056] Wenn eine Suchzeichenfolge zum Zweck der Dokumentbeschaffung bereitgestellt wird, unterteilt die Textunterteilungseinheit **2** die Suchzeichenfolge in partielle Zeichenfolgen. Die Dokumentauswahleinheit **4** wählt ein Dokument oder Dokumente aus, für das bzw. für die eine Rangordnung zu berechnen ist, wo eine derartige Auswahl in Hinblick auf die partiellen Zeichenfolgen durchgeführt wird. Die Rangberechnungseinheit **5** berechnet eine Rangordnung für jedes der ausgewählten Dokumente, in dem die partiellen Zeichenfolgen verwendet werden, wobei dadurch Dokumentbeschaffungsergebnisse bereitgestellt werden.

[0057] Die Dokumentauswahleinheit **4** wählt eines oder mehrere Dokumente aus, in dem die Dokumente identifiziert werden, die alle partiellen Zeichenfolgen der Suchzeichenfolge enthalten. Alternativ können Dokumente, die die Suchzeichenfolge selbst enthalten, ausgewählt werden, oder Dokumente, die gewisse geeignete Bedingungen erfüllen, können ausgewählt werden.

[0058] Die Rangberechnungseinheit **5** berechnet einen Rangordnungspunkt (kurz: „Rangpunkt") der Suchzeichenfolge bezüglich eines jeden der ausgewählten Dokumente. Der Rangordnungspunkt der Suchzeichenfolge wird basierend auf Rangordnungspunkte der partiellen Zeichenfolgen erhalten. Hier können die Rangordnungen der partiellen Zeichenfolgen berechnet werden, indem ein Verfahren verwendet wird, das in der Fachwelt als ein tf-Verfahren, ein tf.idf-Verfahren oder dergleichen bekannt ist, die typischerweise bei der Datenbeschaffung verwendet werden. Zum Beispiel wird hierbei auf W.B. Frakes Ed., „Information Retrieval Data Structures & Algorithms", Prentice Hall, 1992 und insbesondere auf Section **14** des Dokuments verwiesen. Um einen Rangordnungspunkt der Suchzeichenfolge von den Rangordnungspunkten der partiellen Zeichenfolgen zu erhalten, kann man eine Summe, einen Mittelwert, ein Maximum usw. der Suchzeichenfolgen der partiellen Zeichenfolgen erhalten.

[0059] Die Berechnung der Rangordnungspunkte wird unter Bezugnahme auf die Indexeinheit **3** beschrieben, die in Fig. 2C gezeigt ist.

[0060] Wenn eine Suchzeichenfolge „ABCDEF" bereitgestellt wird, extrahiert die Textunterteilungseinheit **2** partielle Zeichenfolgen „AB", „BC", „CD", „DE" und „EF". Dann wählt die Dokumentauswahleinheit **4** ein Dokument oder Dokumente aus, das bzw. die alle partiellen Zeichenfolgen von einer Vielzahl von registrierten Dokumenten enthalten. In diesem Beispiel erfüllt nur das Dokument der Fig. 2A die geforderte Bedingung. In dem Stand der Technik werden Dokumente, die wenigstens eines der partiellen Zeichenfolgen enthalten, ausgewählt. Ein derartiges Schema nach dem Stand der Technik würde also das Doku-

ment der Fig. 2B in diesem Beispiel auswählen, wohingegen die vorliegende Erfindung sich nicht für die Auswahl des Dokuments **2B** entscheidet.

[0061] Nach der Auswahl eines Dokuments oder von Dokumenten berechnet die Rangberechnungseinheit **5** ein Rangordnungspunkt bezüglich eines jeden der ausgewählten Dokumente, in dem die partiellen Zeichenfolgen verwendet werden. Im folgenden wird auf die Rangordnungspunkte der partiellen Zeichenfolgen Bezug genommen, als ob sie wie folgt berechnet werden:

$$SCORE(n) = tf(n)\cdot(1 + Log_2(N/df(n)) \quad (1)$$

wobei SCORE(n) ein Rangordnungspunkt der partiellen Zeichenfolge(n) ist und tf(n) die Anzahl der Auftretungsereignisse der partiellen Zeichenfolge n in dem relevanten Dokument ist. Weiter zeigt N die Anzahl der gespeicherten Dokumente an (die in diesem Beispiel zwei beträgt) und df(n) zeigt die Anzahl der gespeicherten Dokumente an, die die partielle Zeichenfolge n enthalten. Im folgenden wird df(n) als Dokumenthäufigkeit bezeichnet. Bei dieser Ausführungsform wird der Rangordnungspunkt für das Dokument als eine Summe der Rangordnungspunkte der partiellen Zeichenfolgen, die in dem Dokument enthalten sind, erzielt. Bezüglich des Dokuments der Fig. 2A wird der Rangordnungspunkt SCORE(AB) der partiellen Zeichenfolge „AB" durch Substituieren von 2 für tf(AB) und von 2 für df(AB) in der Gleichung (1) erhalten. In diesem Fall beträgt SCORE(AB) 2. Weiter werden SCORE(BC) = 4, SCORE(CD) = 4, SCORE(DE) = 1 und SCORE(EF) = 3 erhalten.

[0062] Dementsprechend erhält man SCORE(ABCDEF) wie folgt:

SCORE(ABCDEF) = SCORE(AB) + SCORE(BC) + SCORE(CD) + SCORE(DE) + SCORE(EF) = 14

[0063] Dies ist der Rangordnungspunkt der Rufzeichenfolge „ABCDEF" bezüglich des Dokuments der Fig. 2A.

[0064] Fig. 3 ist ein Blockdiagramm einer Systemkonfiguration, die die Dokumentbeschaffungsvorrichtung **1** realisiert.

[0065] Die Dokumentbeschaffungsvorrichtung **1** beinhaltet eine CPU **11**, ein ROM **12**, ein RAM **13**, einen Bus **14**, eine Festplatte **15**, ein CD-ROM-Laufwerk **16**, eine Ausgabevorrichtung **17**, eine Eingabevorrichtung **18** und eine Kommunikationssteuervorrichtung **20**. Die CPU **11** kümmert sich um verschiedene Ausführungen und die zentrale Steuerung verschiedener Elemente. Das ROM **12** ist ein Nur-Lesespeicher, der darin BIOS-Programme und dergleichen speichert. Das RAM **13** speichert darin Daten und liefert einen Arbeitsbereich für die CPU **11**. Der Bus **14** stellt eine Verbindung zwischen der CPU **11**, dem ROM **12** und dem RAM **13** her. Der Bus **14** ist ebenso über Schnittstellen und/oder Steuerschaltungen (nicht gezeigt) mit der Festplatte **15**, dem CD-ROM-Laufwerk **16**, der Ausgabevorrichtung **17**, wie zum Beispiel eine CRT-Anzeige, eine LCD-Anzeige oder einen Drucker, die Eingabevorrichtung **18**, wie zum Beispiel eine Tastatur und eine Maus und die Kommunikationssteuervorrichtung **20** verbunden, die mit einem Netzwerk **21** verbunden ist.

[0066] Programme, um die Dokumentbeschaffungsvorrichtung **1** zu veranlassen, eine Verarbeitung entsprechend der vorliegenden Erfindung auszuführen, sind in einer CD-ROM **19** gespeichert, die als ein Speichermedium für die vorliegende Erfindung dient. Die CD-ROM **19** wird in das CD-ROM-Laufwerk **16** eingeführt und die Programme werden in die Festplatte **15** geladen und installiert. Mit den Programmen, die in der Festplatte **15** gespeichert sind, ist die Dokumentbeschaffungsvorrichtung **1** bereit, verschiedene Prozesse der vorliegenden Erfindung auszuführen. Es werden nämlich die verschiedenen Einheiten, die in Fig. 1 gezeigt sind, als Prozesse verkörpert, die durch die CPU **11** durchgeführt werden, die die Programme ausführt. Die Indexeinheit **3** ist als eine Datenbank in der Festplatte **15** eingebaut.

[0067] Das Speichermedium der vorliegenden Erfindung ist nicht auf eine CD-ROM beschränkt, sondern es kann sich um jeden Typ von Speichermedium, wie zum Beispiel CD-RW, CD-R, DVD, FD oder MO handeln. Das Programm kann von dem Netzwerk **19**, wie zum Beispiel Internet über die Kommunikationssteuervorrichtung heruntergeladen werden und kann auf der Festplatte **15** installiert werden. In diesem Fall wird eine Speichervorrichtung, die darin die Programme auf der Übertragungsseite des Netzwerkes **19** speichert, als das Speichermedium der vorliegenden Erfindung angesehen. Die Programme können auf einem vorbestimmten Betriebssystem arbeiten.

[0068] Fig. 4 ist ein Flussdiagramm eines Prozesses der Berechnung von Anordnungspunkten für eine Vielzahl von Dokumenten, wo der Prozess durch die Computerbeschaffungsvorrichtung **1** gemäß der ersten Ausführungsform der vorliegenden Erfindung durchgeführt wird. Das Flussdiagramm der Fig. 4 ist unter der Verwendung von Schreibweisen der C-Sprache gezeigt.

[0069] In einem Schritt S1 werden sowohl der Arrayindex m als auch der Dokumentidentifizierer DocId auf 1 gesetzt.

[0070] Bei einem Schritt S2 wird eine Suche nach einem Dokument durchgeführt, das alle partiellen Zeichenfolgen enthält und die kleinste Dokumen-

ten-ID aufweist, die nicht kleiner als DocId ist. Falls ein derartiges Dokument gefunden wird, wird der Dokumentenidentifizierer DocId auf die erhaltenen Dokumenten-ID gesetzt und das Verfahren geht zu einem Schritt S3 über. Ansonsten wird das Verfahren beendet.

[0071] Im Schritt S3 wird ein Rangordnungspunkt für das Dokument mit dem Dokumentidentifizierer DocId berechnet. Der erhaltene Rangordnungspunkt wird in der Struktur, die der C-Sprache ähnlich ist und die den Dokumentenidentifizierer und den Rangpunkt als seine Elemente aufweist, gespeichert.

[0072] In einem Schritt S4 werden der Dokumentenidentifizierer DocId jeweils um eins erhöht. Dann geht das Verfahren zu dem Schritt S2 zurück.

[0073] Fig. 5 ist ein Flussdiagramm eines Prozesses zur Berechnung eines Rangordnungspunktes, die bei dem Schritt S3 der Fig. 4 durchgeführt wird. In einem Schritt S11 wird ein Parameter n zum Anzeigen einer partiellen Zeichenfolge auf eins gesetzt und ein Rangpunkt („score") wird auf null gesetzt.

[0074] In dem Schritt S12 wird die Gleichung (1) unter Bezugnahme auf eine Zeichenfolge str[n] berechnet und zu dem Rangpunkt („score") hinzugefügt.

[0075] In einem Schritt S13 wird eine Überprüfung dahingehend durchgeführt, ob n gleich num ist. Hier ist num die Anzahl aller partieller Zeichenfolgen einer Suchzeichenfolge. Falls n gleich num ist, wird das Verfahren beendet. Ansonsten geht das Verfahren zu einem Schritt S14 über.

[0076] Bei dem Schritt S14 wird n um eins erhöht. Dann kehrt das Verfahren zum Schritt S12 zurück.

[0077] Im folgenden wird eine zweite Ausführungsform der vorliegenden Erfindung beschrieben.

[0078] Bei der zweiten Ausführungsform wird auf dieselben Elemente wie jene der ersten Ausführungsform durch dieselben Bezugszeichen Bezug genommen und eine Beschreibung davon wird weggelassen.

[0079] Die Dokumentbeschaffungsvorrichtung **1** der zweiten Ausführungsform unterscheidet sich von jener der ersten Ausführungsform nur in den Operationen der Rang-Berechnungseinheit **5**. Bei der ersten Ausführungsform berechnet die Rang-Berechnungseinheit **5** einen Rangordnungspunkt einer partiellen Zeichenfolge bezüglich eines ausgewählten Dokuments in Hinblick auf die Anzahl der Auftretensereignisse der partiellen Zeichenfolge in dem Dokument. Diese partielle Zeichenfolge kann in dem Dokument in einem Kontext auftreten, der keine semantische Relevanz hinsichtlich der Suchzeichenfolge aufweist und ein derartiges irrelevantes Auftreten beeinträchtigt den erzielten Rangordnungspunkt, wodurch sich eine weniger genaue Suche ergibt.

[0080] Bei dem Beispiel der Fig. 2A ist es wahrscheinlich, dass die Zeichenfolge „EF", die bei dem 20. Zeichen und bei dem 60. Zeichen auftritt, keine Relevanz hinsichtlich der Suchzeichenfolge hat. Da die Zeichenfolge „EF" dreimal in dem Dokument auftritt, wird jedoch SCORE(EF) einfach als 3 berechnet. Hinsichtlich der Relevanz der Suchzeichenfolge „ABCDEF" kann dieser Rangpunkt zu hoch sein.

[0081] Bei der zweiten Ausführungsform wählt die Dokumentbeschaffungsvorrichtung **1** einen minimalen Auftretungsereignis-Zählwerten von jeweiligen partiellen Zeichenfolgen aus, die in dem Dokument erscheinen, und substituiert den ausgewählten minimalen Auftretungsereignis-Zählwert für die Auftretungsereignis-Zählwerte der partiellen Zeichenfolgen beim Berechnen der Rangordnungspunkte. In dem Beispiel der Fig. 2A beträgt ein minimaler Auftretungsereignis-Zählwert 1, d.h. die Anzahl der Auftretungsereignisse einer partiellen Zeichenfolge „DE", so dass der minimale Auftretungsereignis-Zählwert 1 für die Auftretungsereignis-Zählwerte der anderen partiellen Zeichenfolgen „AB", „BC", „CD" und „EF" zum Zwecke der Berechnung der Rangordnungspunkte substituiert wird. Infolgedessen werden SCORE(AB) = 1, SCORE(BC) = 2, SCORE(CD) = 2, SCORE(DE) = 1 und SCORE(EF) = 1 erhalten, wodurch SCORE(ABCDEF) = 7 erzeugt wird.

[0082] Fig. 6 ist ein Flussdiagramm eines Prozesses zur Berechnung eines Rangordnungspunktes entsprechend der zweiten Ausführungsform der vorliegenden Erfindung. Die Schritte bis auf den Schritt der Berechnung eines Rangordnungspunktes sind dieselben, wie jene der ersten Ausführungsform, wie in Fig. 4 gezeigt ist.

[0083] In dem Schritt S11 der Fig. 6 wird ein Parameter n zum Anzeigen einer partiellen Zeichenfolge auf 1 gesetzt und ein Parameter mintf zum Erzielen eines minimalen Auftretungsereignis-Zählwertes wird auf eine sehr große Zahl bzw. größtmögliche Zahl gesetzt.

[0084] In einem Schritt S12 wird mintf auf den kleinsten Wert von mintf und einem Auftretungsereignis-Zählwert einer Zeichenfolge str[n] gesetzt.

[0085] Bei einem Schritt S13 wird eine Überprüfung durchgeführt, ob n gleich num ist. Wie zuvor ist num die Anzahl aller partieller Zeichenfolgen einer Suchzeichenfolge. Falls n nicht gleich num ist, geht das Verfahren zu einem Schritt S14 über, wo n um 1 erhöht wird, wobei danach eine Prozedur folgt, wo zu dem Schritt S12 zurückgegangen wird. Falls n gleich num ist, dann bedeutet dies, dass mintf gleich den

minimalen Auftretungsereignis-Zählwert ist, so dass das Verfahren zu einem nächsten Schritt übergeht.

[0086] Bei einem Schritt S15 wird n auf 1 gesetzt und ein Zählpunkt wird auf 0 gesetzt.

[0087] Bei einem Schritt S16 wird die Gleichung (1), bei der tf(n) durch mintf ersetzt wird, unter Bezugnahme auf eine Zeichenfolge str[n] berechnet und zu dem Zählpunkt zugefügt.

[0088] Bei einem Schritt S17 wird eine Überprüfung dahingehend durchgeführt, ob n gleich num ist. Falls n gleich num ist, kommt das Verfahren zu einem Ende. Ansonsten geht das Verfahren zu dem Schritt S18 über.

[0089] Bei dem Schritt S18 wird n um 1 erhöht. Dann geht das Verfahren zu dem Schritt S16 zurück.

[0090] Gemäß der zweiten Ausführungsform wird der Einfluss irrelevanter Auftretungsereignisse von partiellen Zeichenfolgen von dem Rangordnungspunkt der Suchzeichenfolge beseitigt, wenn derartige Auftretungsereignisse in einem Kontext bzw. Zusammenhang stattfinden, der keine Relevanz bezüglich der Suchzeichenfolge hat. Dies verbessert die Beschaffungsgenauigkeit.

[0091] Im folgenden wird eine dritte Ausführungsform der vorliegenden Erfindung beschrieben.

[0092] Bei der dritten Ausführungsform werden dieselben Elemente, wie jene der zweiten Ausführungsform durch dieselben Bezugszeichen bezeichnet und eine Beschreibung davon wird weggelassen. Die Dokumentbeschaffungsvorrichtung **1** der dritten Ausführungsform unterscheidet sich von jener der zweiten Ausführungsform dahingehend, dass die dritte Ausführungsform ein Schema verwendet, dass sich von der zweiten Ausführungsform hinsichtlich der Beseitigung des Einflusses von irrelevanten Auftretungsereignissen von partiellen Zeichenfolgen unterscheidet.

[0093] Um den Einfluss irrelevanter Auftretungsereignisse von partiellen Zeichenfolgen zu beseitigen, wird die Anzahl der Auftretungsereignisse der Suchzeichenfolge bezüglich eines ausgewählten Dokuments erhalten und wird dann als ein Ersatz bzw. Substitut für Auftretungs-Zählwerte der partiellen Zeichenfolgen zum Zwecke der Erzielung von Rangordnungspunkten verwendet. Der Auftretungszählwert der Suchzeichenfolge wird erhalten, indem alle Positionen überprüft werden, wo die Suchzeichenfolge bei dem ausgewählten Dokument auftritt.

[0094] Um die Positionen von Erscheinungen der Suchzeichenfolge zu erhalten, kann ein herkömmliches Verfahren verwendet werden das Positionen partieller Zeichenfolge miteinander abgleicht bzw. in Übereinstimmung bringt. Zum Beispiel eine Technik, die auf Seite 839 in Chuichi Kikuchi „A Fast Full-Text Search Method for Japanese Text Database", Transactions of the Institute of Electronics, Information and Communication Engineers, Band J75-D-I, Nr. 9, Seiten 836-846, 1992 offenbart ist, verwendet werden.

[0095] Bezüglich des Beispieles des Dokuments der Fig. 2A kann von dem Index der Fig. 2C gewährleistet werden, dass die Suchzeichenfolge „ABCDEF" nur einmal bei dem von Beginn an 31. Zeichen auftritt. Basierend auf diesem Fund, werden die Auftretungsereignis-Zählwerte der partiellen Zeichenfolgen „AB", „BC", „CD", „DE" und „EF" auf 1 gesetzt, was der Auftretungsereignis-Zählwert der Suchzeichenfolge ist. Der Rangordnungspunkt, der infolge dieser Prozedur erhalten wird, ist derselbe wie jener der zweiten Ausführungsform. Das heißt SCORE(ABCDEF) gleich 7 wird erhalten.

[0096] Fig. 7 ist ein Flussdiagramm eines Prozesses zur Berechnung eines Rangordnungspunktes gemäß der dritten Ausführungsform der vorliegenden Erfindung. Schritte anders als der Schritt zur Berechnung eines Rangordnungspunktes sind dieselben wie jene bei der ersten Ausführungsform, wie in Fig. 4 gezeigt ist.

[0097] In dem Schritt S11 der Fig. 7 wird die Anzahl der Auftretungsereignisse einer Suchzeichenfolge erhalten. Der erhaltene Auftretungsereignis-Zählwert wird auf wordtf festgelegt.

[0098] Bei dem Schritt S12 wird ein Parameter n zum Anzeigen einer partiellen Zeichenfolge auf 1 gesetzt und ein Rangpunkt wird auf 0 gesetzt.

[0099] Bei dem Schritt S13 wird die Gleichung (1), bei der tf(n) durch wordtf ersetzt wird, unter Bezugnahme auf eine Zeichenfolge str[n] berechnet und zu dem Rangpunkt addiert.

[0100] Bei dem Schritt S14 wird eine Überprüfung durchgeführt, ob n gleich num ist. Falls n gleich num ist, kommt das Verfahren zu einem Ende. Ansonsten geht das Verfahren zu einem Schritt S15 über.

[0101] Bei dem Schritt S15 wird n um 1 erhöht. Dann geht das Verfahren zu dem Schritt S13 zurück.

[0102] Gemäß der dritten Ausführungsform wird der Einfluss irrelevanter Auftretungsereignisse partieller Zeichenfolgen von dem Rangordnungspunkt der Suchzeichenfolge beseitigt, wenn derartige Auftretungsereignisse in Kontexten bzw. Zusammenhängen stattfinden, die keine Relevanz hinsichtlich der Suchzeichenfolge haben. Dies verbessert die Beschaffungsgenauigkeit.

[0103] Im folgenden wird eine vierte Ausführungsform der vorliegenden Erfindung beschrieben.

[0104] Bei der vierten Ausführungsform werden dieselben Elemente wie jene der dritten Ausführungsform mit den selben Bezugszeichen bezeichnet und deren Beschreibung wird weggelassen.

[0105] Die Dokumentbeschaffungsvorrichtung **1** der vierten Ausführungsform unterscheidet sich von jener der dritten Ausführungsform in den folgenden Aspekten. Bei der dritten Ausführungsform erzielt die Rangpunkt-Berechnungseinheit **5** den Auftretungsereignis-Zählwert der Suchzeichenfolge beim ausgewählten Dokument durch Überprüfen aller Positionen, wo die Suchzeichenfolge in dem Dokument auftritt.

[0106] Wenn die Suchzeichenfolge lang ist und häufig auftritt, ist jedoch die Berechnungslast zum Erzielen aller Auftretepositionen verbietend hoch, was zu einer länglichen Beschaffungszeit zum Beschaffen eines Dokuments führt.

[0107] Bei der vierten Ausführungsform der vorliegenden Erfindung wird die obere Grenze auf den Auftretungsereignis-Zählwert einer Suchzeichenfolge gesetzt. Falls der Auftretungsereignis-Zählwert einer Suchzeichenfolge in einem Dokument unterhalb der oberen Grenze liegt, wird dieser Zählwert als die Anzahl der Auftretungsereignisse der Suchzeichenfolge verwendet. Falls der Auftretungsereignis-Zählwert einer Suchzeichenfolge die obere Grenze überschreitet, wird die obere Grenze als ein Ersatz bzw. Substitut für den Auftretungsereignis-Zählwert verwendet. In diesem Fall besteht kein Erfordernis, den gesamten Weg zur Überprüfung aller Erscheinungspositionen der Suchzeichenfolge zu gehen, und es genügt, falls das Zählen gestoppt wird, wenn der Zählwert die obere Grenze erreicht.

[0108] Fig. 8 ist ein Flussdiagramm eines Prozesses zum Erzielen des Auftretungsereignis-Zählwerts einer Suchzeichenfolge mit einer oberen Grenze gemäß der vierten Ausführungsform der vorliegenden Erfindung.

[0109] Bei der dritten Ausführungsform wird der Auftretungsereignis-Zählwert einer Suchzeichenfolge einfach durch Zählen aller Auftretungsereignisse der Suchzeichenfolge beim Schritt S11 der Fig. 7 erhalten. Bei der vierten Ausführungsform wird der Auftretungsereigniszählwert wie folgt erhalten.

[0110] In einem Schritt S21 wird wordtf auf 0 gesetzt.

[0111] In einem Schritt S22 wird nach einem neuen Auftretungsereignis der Suchzeichenfolge gesucht. Falls sie gefunden wird, geht das Verfahren zum Schritt S23. Ansonsten kommt das Verfahren zu einem Ende.

[0112] In dem Schritt S23 wird der Auftretungsereignis-Zählwert wordtf um 1 erhöht.

[0113] In dem Schritt S24 wird eine Überprüfung dahingehend durchgeführt, ob wordtf gleich L ist, wo die Anzahl L eine obere Grenze des Auftretungsereignis-Zählwerts festlegt. Falls dies so ist, kommt das Verfahren zu einem Ende. Ansonsten geht das Verfahren zurück zu dem Schritt S22.

[0114] Gemäß der vierten Ausführungsform der vorliegenden Erfindung wird die Berechnungslast zum Überprüfen aller Ereignispositionen einer Suchzeichenfolge im Vergleich zu dem Fall der dritten Ausführungsform reduziert, wodurch eine schnellere Dokumentbeschaffung erzielt wird.

[0115] Im folgenden wird eine fünfte Ausführungsform der vorliegenden Erfindung beschrieben.

[0116] Bei der fünften Ausführungsform werden die gleichen Elemente, wie jene der dritten Ausführungsform durch dieselben Bezugszeichen bezeichnet und eine Beschreibung davon wird weggelassen.

[0117] Die Dokumentbeschaffungsvorrichtung **1** der fünften Ausführungsform unterscheidet sich von jener der dritten Ausführungsform in den folgenden Aspekten. Bei der dritten Ausführungsform erzielt die Rangberechnungseinheit **5** einen Auftretungsereignis-Zählwert einer Suchzeichenfolge bei einem ausgewählten Dokument und verwendet den erzielten Auftretungsereignis-Zählwert als einen Ersatz bzw. ein Substitut für die Auftretungsereignis-Zählwerte partieller Zeichenfolgen, um Rangordnungspunkte zu erzielen. Auf diese Art und Weise kann der Einfluss irrelevanter Auftretungsereignisse von partiellen Zeichenfolgen von dem Rangordnungspunkt der Suchzeichenfolge beseitigt werden, wenn die partiellen Zeichenfolgen außerhalb eines Kontexts für die Suchzeichenfolge erscheinen.

[0118] Bei der dritten Ausführungsform wird der Rangordnungspunkt einer Suchzeichenfolge von den Rangordnungspunkten partieller Zeichenfolgen abgeleitet, die wiederum basierend auf der Anzahl der gespeicherten Dokumente abgeleitet werden, die die partiellen Zeichenfolgen enthalten. Infolgedessen kann das Vorhandensein eines irrelevanten Dokuments den Rangordnungspunkt beeinträchtigen, wenn das irrelevante Dokument eine bestimmte partielle Zeichenfolge aufweist, ohne eine Suchzeichenfolge zu enthalten. Mit anderen Worten wird der Einfluss eines irrelevanten Auftretens von Suchzeichenfolgen nicht vollständig bei der dritten Ausführungsform beseitigt.

[0119] Bei dem Beispiel des Dokuments der Fig. 2A wird SCORE(AB) so berechnet, dass es gleich 1 ist, da die Anzahl der Dokumente, die die Zeichenfolge „AB" enthalten, 2 ist, wenn der Auftretungsereignis-Zählwert der Zeichenfolge „AB" in dem Dokument der Fig. 2A auf die Anzahl der Auftretungsereignisse der Suchzeichenfolge festgelegt wird. Das Dokument der Fig. 2B ist jedoch irrelevant, da dieses Dokument nicht die Suchzeichenfolge „ABCDEF" enthält. In diesem Fall wird deshalb das Vorhandensein eines irrelevanten Dokuments, das ein anderer Typ eines irrelevanten Auftretens von partiellen Zeichenfolgen ist, den Rangordnungspunkt der Suchzeichenfolge beeinträchtigen.

[0120] In Hinblick darauf verwendet die fünfte Ausführungsform der vorliegenden Erfindung die Anzahl der Dokumente mit einer Suchzeichenfolge darin als die Anzahl der Dokumente, die eine gegebene partielle Zeichenfolge enthalten, und zwar für den Zweck der Berechnung eines Rangordnungspunktes der gegebenen partiellen Zeichenfolge. In dem Beispiel des Dokuments der Fig. 2A und Fig. 2B wird die Anzahl der gespeicherten Dokumente, die die Zeichenfolge „AB" enthalten, gleich 1 gesetzt, wobei die Anzahl der Dokumente ist, die die Suchzeichenfolge „ABCDEF" enthalten. Infolgedessen wird SCORE(AB) 2. Weiter wird SCORE(BC) = 2. SCORE(CD) = 2, SCORE(DE) = 2 und SCORE(EF) = 2 erhalten, wodurch SCORE(ABCDEF) = 10 wird.

[0121] Fig. 9 ist ein Flussdiagramm eines Prozesses zur Berechnung von Rangordnungspunkten für eine Vielzahl von Dokumenten gemäß der fünften Ausführungsform der vorliegenden Erfindung.

[0122] Bei einem Schritt S101 der Fig. 9 wird die Anzahl der Dokumente, die die Suchzeichenfolge enthalten, erhalten. Der erhaltene Dokumentenzählwert wird in einem Parameter worddf festgelegt.

[0123] Bei einem Schritt S102 werden ein Arrayindex m und ein Dokumentenidentifizierer DocId auf 1 gesetzt.

[0124] Bei einem Schritt S103 wird eine Suche nach einem Dokument durchgeführt, das die Suchzeichenfolge enthält und das die kleinste Dokument-ID nicht kleiner als DocId enthält. Falls ein derartiges Dokument gefunden wird, wird der Dokumentenidentifizierer DocId auf die erzielte Dokumenten-ID gesetzt und das Verfahren geht zu dem Schritt S104 über. Ansonsten kommt das Verfahren zu einem Ende.

[0125] Bei dem Schritt S104 wird ein Rangordnungspunkt, das den Dokumentidentifizierer DocId aufweist, berechnet. Hier wird bei der Gleichung (1), die zur Berechnung eines Rangordnungspunktes bei dem Schritt S104 verwendet wird, df(str[n]) durch worddf ersetzt. Der erzielte Rangpunkt wird in der Struktur, die der C-Sprache ähnlich ist und die den Dokumentidentifizierer und den Rangpunkt als Elemente aufweist, gespeichert.

[0126] Bei dem Schritt S105 werden der Arrayindex m und der Dokumentidentifizierer DocId jeweils um 1 erhöht. Dann geht das Verfahren zu dem Schritt S103 zurück.

[0127] Gemäß der fünften Ausführungsform der vorliegenden Erfindung wird der Einfluss irrelevanter Auftretungsereignisse partieller Zeichenfolgen im wesentlichen von dem Rangordnungspunkt einer Rufzeichenfolge bezüglich eines ausgewählten Dokuments beseitigt, wenn die partiellen Zeichenfolgen außerhalb eines Kontexts in dem ausgewählten Dokument oder sogar außerhalb eines Kontexts in anderen gespeicherten Dokumenten erscheinen. Dies verbessert die Beschaffungsgenauigkeit.

[0128] Bei dieser Ausführungsform sind die Rangordnungspunkte der partiellen Zeichenfolgen jeweils gleich einem Rangordnungspunkt, der unter Verwendung der Anzahl der Dokumente erhalten wurde, die die Suchzeichenfolge darin aufweisen, und der Anzahl der Auftretungsereignisse der Suchzeichenfolge. Deswegen besteht kein Bedürfnis danach, alle Rangordnungspunkte der partiellen Zeichenfolgen zu berechnen und sie mit dem Ziel zu kombinieren, den Rangordnungspunkt der Suchzeichenfolge zu erzeugen. Alternativ wird der Rangordnungspunkt der Suchzeichenfolge direkt von der Anzahl der Dokumente abgeleitet, die die Suchzeichenfolge darin und die Anzahl der Auftretungsereignisse der Suchzeichenfolge aufweisen. Dies ermöglicht die Reduktion bei der Berechnungslast, wodurch eine Wahrscheinlich-Relevanz-Dokumentsbeschaffung mit hoher Geschwindigkeit erzielt wird.

[0129] Im folgenden wird eine sechste Ausführungsform der vorliegenden Erfindung beschrieben.

[0130] Bei der sechsten Ausführungsform wird auf die gleichen Elemente wie bei der vierten Ausführungsform durch dieselben Bezugszeichen Bezug genommen und eine Beschreibung davon wird weggelassen.

[0131] Die Dokumentbeschaffungsvorrichtung **1** der sechsten Ausführungsform unterscheidet sich von jener der vierten Ausführungsform in den folgenden Aspekten. Bei der vierten Ausführungsform wird eine obere Grenze für die Anzahl der Auftretungsereignisse einer Suchzeichenfolge mit dem Ziel der Beschleunigung der Dokumentbeschaffungsgeschwindigkeit gesetzt.

[0132] Durch die Deckelung bzw. durch das Setzen einer Obergrenze kann jedoch die Beschaffungsgenauigkeit verschlechtert werden, da ein derartiger

Deckel zu einem anderen Rangordnungspunkt führt, als jener, der ohne eine obere Grenze erzielt werden würde. Das Ausmaß, mit dem sich der Rangordnungspunkt aufgrund des Platzierens einer oberen Grenze ändert, hängt von der Anzahl der Dokumente ab, die die Suchzeichenfolge enthalten. Wenn die Gleichung (1) zur Berechnung einer Rangordnungsfolge verwendet wird, so ist der Unterschied des Rangordnungspunktes, der durch das Platzieren einer oberen Grenze verursacht wird, um so größer, je kleiner die Anzahl der Dokumente ist, die die Suchzeichenfolge enthalten. Im Hinblick darauf ist es vorzuziehen, dass die obere Grenze dynamisch entsprechend der Anzahl der Dokumente geändert wird, die die Suchzeichenfolge darin aufweisen, und zwar mit dem Ziel, den Einfluss der Platzierung eines Deckels bzw. Setzens einer Obergrenze zu reduzieren. Zum Beispiel kann eine obere Grenze L_x zur Verwendung in dem Fall der Anzahl von Dokumenten mit einer Suchzeichenfolge darin, die x ist (x 1) wie folgt berechnet werden:

$$L_x = L_1(1 + \log_2(N/x))/(1 + \log_2 N) \quad (2)$$

wobei L_1 eine obere Grenze ist, die in dem Fall verwendet wird, dass die Anzahl der Dokumente, die eine Suchzeichenfolge aufweisen, 1 ist. Gemäß der Gleichung (2) ist die obere Grenze um so größer, je kleiner die Anzahl der Dokumente ist, die die Suchzeichenfolge darin aufweisen. Das heißt, je kleiner der Unterschied des Rangordnungspunktes ist, der durch das Setzen der oberen Grenze verursacht wird, desto geringer ist die Anzahl der Dokumente, die die Suchzeichenfolge darin aufweisen. Mit anderen Worten wird die Reduktion bei der Beschaffungsgenauigkeit aufgrund des Einführens einer oberen Grenze verbessert. Bemerkenswert ist, dass die Gleichung (2) nur ein Beispiel ist und jede Formel verwendet werden kann, solange eine obere Grenze mit einer Zunahme in der Anzahl der Dokumente zunimmt, die eine Suchzeichenfolge darin aufweisen.

[0133] Fig. 10 ist ein Flussdiagramm eines Prozesses zur Berechnung von Rangordnungspunkten für eine Vielzahl von Dokumenten gemäß der sechsten Ausführungsform der vorliegenden Erfindung.

[0134] Mit einem Schritt S111 der Fig. 10 wird die Anzahl der Dokumente, die die Suchzeichenfolge enthalten, erhalten. Der erhaltene Dokumentzählwert wird in einem Parameter worddt festgelegt.

[0135] In einem Schritt S112 wird die obere Grenze L_x erhalten, indem die Gleichung (2) verwendet wird, bei der worddf für x substituiert werden.

[0136] Bei einem Schritt S113 werden ein Arrayindex m und ein Dokumentidentifizierer DocId beide auf 1 gesetzt.

[0137] Beim Schritt S114 wird eine Suche nach einem Dokument durchgeführt, das die Suchzeichenfolge enthält und die kleinste Dokument-ID nicht kleiner als DocId aufweist. Falls ein derartiges Dokument gefunden wird, wird der Dokumentenidentifizierer DocId auf die erzielte Dokumenten-ID gesetzt und das Verfahren geht zu dem Schritt S115 über. Ansonsten kommt das Verfahren zu einem Ende.

[0138] Bei dem Schritt S115 wird ein Rangordnungspunkt für das Dokument, das den Dokumentenidentifizierer DocId aufweist, berechnet. Hier werden die Schritte S11 bis S15 der Fig. 7 bei dem Schritt S115 ausgeführt, wobei bei der Gleichung des Schrittes S13 df(str[n]) durch worddf ersetzt worden ist und wordtf durch die obere Grenze L_x begrenzt worden ist.

[0139] Bei dem Schritt S116 werden der Arrayindex m und der Dokumentenidentifizierer DocId jeweils um 1 erhöht. Dann kehrt das Verfahren zu dem Schritt S103 zurück.

[0140] Im folgenden wird eine siebte Ausführungsform der vorliegenden Erfindung beschrieben.

[0141] Bei der siebten Ausführungsform werden dieselben Elemente wie jene der ersten Ausführungsform durch dieselben Bezugszeichen bezeichnet und eine Beschreibung davon wird weggelassen.

[0142] Die Dokumentenbeschaffungsvorrichtung **1** der siebten Ausführungsform unterscheidet sich von jener der ersten Ausführungsform in den folgenden Aspekten. Wenn bei der ersten Ausführungsform eine Suchzeichenfolge kürzer als die Länge (d.h. die Anzahl der Zeichen) einer partiellen Zeichenfolge ist, die als eine Verarbeitungseinheit dient, kann die Textunterteilungseinheit **2** die Suchzeichenfolge nicht in partielle Zeichenfolgen unterteilen, was zu einer Situation führt, bei der eine Dokumentbeschaffung scheitert. Wenn zum Beispiel eine Suchzeichenfolge „A" ist und zwei Zeichen eine Verarbeitungseinheit bilden, kann keine Beschaffung durchgeführt werden, da die Suchzeichenfolge kürzer als die Verarbeitungseinheit ist.

[0143] Wenn die Suchzeichenfolge kürzer als eine Verarbeitungseinheit ist, wird ein Verfahren bei der siebten Ausführungsform der vorliegenden Erfindung wie folgt verwendet.

(1) Die Textunterteilungseinheit **2** extrahiert alle partiellen Zeichenfolgen von der Indexeinheit **3**, so dass diese partiellen Zeichenfolgen mit dem selben Zeichen bzw. mit den selben Zeichen beginnen, wie dies bei der Suchzeichenfolge der Fall ist.

(2) Die Dokumentauswahleinheit **4** identifiziert ein Dokument oder mehrere Dokumente, die wenigstens eine der partiellen Suchzeichenfolgen enthal-

ten, die durch die Textunterteilungseinheit **2** extrahiert werden.

(3) Die Rangberechnungseinheit **5** berechnet Rangordnungspunkte der Dokumente, die durch die Dokumentauswahleinheit **4** ausgewählt werden, indem die partiellen Zeichenfolgen verwendet werden, die durch die Textunterteilungseinheit **2** extrahiert werden.

[0144] Fig. 11A bis Fig. 11C sind erläuternde Zeichnungen, die Beispiele von Dokumenten und ein Beispiel einer entsprechenden Indexeinheit zeigen. Das obige Verfahren wird weiter im Detail unter Bezugnahme auf die Fig. 11A bis Fig. 11C beschrieben.

[0145] Wenn eine Verarbeitungseinheit aus zwei Zeichen besteht, wird die Indexeinheit **3** Inhalte aufweisen, wie in Fig. 11C gezeigt ist und zwar bezüglich des Dokuments der Fig. 11A und des Dokuments der Fig. 11B. Hier ist das Datenformat der Indexeinheit **3**, die in Fig. 11C gezeigt ist, dasselbe, wie jenes der Indexeinheit **3**, das in Fig. 2C gezeigt ist. Im folgenden wird das Verfahren des Beschaffungsprozesses unter Bezugnahme auf ein Beispiel beschrieben, bei dem „Y" als eine Suchzeichenfolge gegeben ist.

[0146] Die Textunterteilungseinheit **2** extrahiert drei partielle Zeichenfolgen „YI", „YK" und „YB" als partielle Zeichenfolgen, die dasselbe Zeichen am Beginn davon als die Suchzeichenfolge aufweisen. Die Dokumentauswahleinheit **4** wählt das Dokument der Fig. 11A und das Dokument der Fig. 11B aus, da sie wenigstens eines der extrahierten partiellen Zeichenfolgen enthalten. Dann berechnet die Rangberechnungseinheit **5** Rangordnungspunkte der ausgewählten Dokumente, in dem die partiellen Zeichenfolgen verwendet werden, die durch die Textunterteilungseinheit **2** extrahiert wurden.

[0147] Wenn die Rangordnungspunkte berechnet werden, leitet die Rangberechnungseinheit **5** einen Rangordnungspunkt der Suchzeichenfolge bezüglich eines ausgewählten Dokuments von Rangordnungspunkten der partiellen Zeichenfolgen innerhalb des ausgewählten Dokuments ab. Dies wird zum Beispiel erzielt, indem die Summe der Rangordnungspunkte der partiellen Zeichenfolgen berechnet wird. Wenn die Rangordnungspunkte der partiellen Zeichenfolgen basierend auf der Gleichung (1) berechnet wird, wird man SCORE(YI) = 0, SCORE(YK) = 2 und SCORE(YB) = 2 erhalten, wodurch SCORE(Y) = 4 erzielt wird.

[0148] Fig. 12 ist ein Flussdiagramm eines Prozesses zur Berechnung von Rangordnungspunkten für eine Vielzahl von Dokumenten gemäß der siebten Ausführungsform der vorliegenden Erfindung.

[0149] Im Schritt S121 werden sowohl ein Arrayindex m als auch ein Dokumentenidentifizierer DocId auf 1 gesetzt.

[0150] Im Schritt S122 wird eine Suche nach einem Dokument durchgeführt, das wenigstens eine partielle Zeichenfolge enthält und dessen kleinste Dokument-ID nicht kleiner als DocId ist. Hier werden die partiellen Zeichenfolgen als jene festgelegt, die mit dem bzw. den selben Zeichen beginnen, wie dies bei der Suchzeichenfolge der Fall ist. Falls ein derartiges Dokument gefunden wird, wird der Dokumentenidentifizierer DocId auf die erzielte Dokumenten-ID festgelegt und das Verfahren geht zu einem Schritt S122 über. Ansonsten kommt das Verfahren zu einem Ende.

[0151] In dem Schritt S123 wird ein Rangordnungspunkt für das Dokument berechnet, das den Dokumentenidentifizierer DocId aufweist. Der erzielte Rangpunkt wird mit der Struktur, die eine der C-Sprache ähnliche Struktur aufweist und die den Dokumentenidentifizierer und den Rangpunkt als ihre Elemente aufweist. In einem Schritt S124 werden der Arrayindex m und der Dokumentenidentifizierer DocId jeweils um 1 erhöht. Dann kehrt das Verfahren zu dem Schritt S122 zurück.

[0152] Indem dem oben beschriebenen Verfahren gefolgt wird, kann man ein Dokument selbst dann beschaffen, wenn eine Suchzeichenfolge eine kürzere Länge als eine Verarbeitungseinheit aufweist.

[0153] Bei der Berechnung von Rangordnungspunkten kann ein Dokumenten-Zählwert, der die Anzahl der Dokumente, die eine Suchzeichenfolge aufweisen, in derselben Art und Weise wie bei der dritten Ausführungsform der vorliegenden Erfindung verwendet werden. Bei dem Beispiel der Fig. 11A und Fig. 11B gibt es zwei Dokumente, die die Zeichenfolge „Y" aufweisen. Bezüglich des Dokumentes **11A** ist deshalb SCORE(YI) = 0, SCORE(YK) = 1 und SCORE(YB) = 1, was dazu führt, dass SCORE(Y) 2 ist. Dieser Berechnungsprozess kann die Beschaffungsgenauigkeit verbessern, da er die Anzahl der Dokumente, die die Suchzeichenfolge aufweisen, zum Zwecke der Berechnung der Rangordnungspunkte verwendet. Weiter kann die Rangberechnungseinheit **5** die Anzahl der Auftretungsereignisse der Suchzeichenfolge in dem Dokument zum Zwecke der Berechnung der Rangordnungspunkte verwenden. In dem Dokument der Fig. 11A erscheint die Suchzeichenfolge „Y" zweimal, wie durch Aufaddieren der Auftretungsereignis-Zählwerte von „YI", „YK" und „YB", die in **Fig.** 11 C gezeigt sind, berechnet wird. Da die Anzahl der Dokumente, die „Y" enthalten, 2 ist, wird SCORE(Y) als 2 berechnet. Bei diesem Verfahren wird die Gleichung (1) für eine geringere Anzahl von Malen bzw. mit einer geringeren Häufigkeit berechnet als bei den vorhergehenden Verfahren, wodurch die Dokumentbeschaffung be-

schleunigt wird.

[0154] Im folgenden wird die achte Ausführungsform der vorliegenden Erfindung beschrieben.

[0155] Fig. 13 ist ein Blockdiagramm der Dokumentbeschaffungsvorrichtung **1A** entsprechend einer achten Ausführungsform der vorliegenden Erfindung.

[0156] Die Textunterteilungseinheit **2** unterteilt einen Text in partielle Zeichenstrings, wobei der Text ein gespeichertes Dokument oder eine Suchzeichenfolge sein kann. Die Indexeinheit **3** speichert Information über partielle Zeichenfolgen, die durch Unterteilen eines gespeicherten Dokuments erhalten werden. Eine Auswahleinheit **6** für partielle Zeichenfolgen wählt partielle Zeichenfolgen aus, die für Dokumentbeschaffungszwecke zu verwenden sind, wo eine derartige Auswahl aus den partiellen Zeichenfolgen, die durch Unterteilen der Suchzeichenfolge erhalten wurden, getroffen wird. Die Dokumentauswahleinheit **4** verwendet partielle Zeichenfolgen, die durch eine Auswahleinheit **6** für partielle Zeichenfolgen ausgewählt wurden, um ein Dokument auszuwählen, für das ein Rangordnungspunkt zu berechnen ist. Die Rangberechnungseinheit **5** verwendet die partiellen Zeichenfolgen, die durch die Auswahleinheit **6** für partielle Zeichenfolgen ausgewählt wurden, um einen Rangordnungspunkt des Dokuments zu berechnen, der durch die Dokumentauswahleinheit **4** ausgewählt wurde. Die Textunterteilungseinheit **2** führt einen Unterteilungsschritt aus und die Dokumentauswahleinheit **4** führt einen Dokumentauswahlschritt aus. Weiter führt die Rangberechnungseinheit **5** einen Rangberechnungsschritt durch und die Auswahleinheit **6** für partielle Zeichenfolgen führt einen Auswahlschritt für partielle Zeichenfolgen aus.

[0157] Die Speicherung der Dokumente ist dieselbe, wie die bei der ersten Ausführungsform.

[0158] Wenn eine Suchzeichenfolge für den Zweck der Dokumentbeschaffung bereitgestellt wird, unterteilt die Textunterteilungseinheit **2** die Suchzeichenfolge in partielle Zeichenfolgen. Die Auswahleinheit **6** für partielle Zeichenfolgen wählt partielle Zeichenfolgen aus allen partiellen Zeichenfolgen aus, die durch Unterteilen der Suchzeichenfolge erhalten wurden, so dass die ausgewählten partiellen Zeichenfolgen für die Dokumentbeschaffungszwecke zu verwenden sind. Die Dokumentauswahleinheit **4** wählt ein Dokument oder Dokumente aus, für die ein Rangordnungspunkt zu berechnen ist, wobei eine derartige Auswahl in Hinblick auf die ausgewählten partiellen Zeichenfolgen getroffen wird. Die Rangberechnungseinheit **5** berechnet einen Rangordnungspunkt für jedes ausgewählte Dokument, in dem die ausgewählten partiellen Zeichenfolgen verwendet werden, wodurch Dokumentbeschaffungsergebnisse bereitgestellt werden.

[0159] Die Textunterteilungseinheit **2**, die Dokumentauswahleinheit **4** und die Rangberechnungseinheit **5** funktionieren im wesentlichen genauso, wie bei der ersten Ausführungsform. Die Auswahleinheit **6** für partielle Zeichenfolgen wählt partielle Zeichenfolgen so wenig wie möglich, jedoch ausreichend, um die volle Länge der Suchzeichenfolge abzudecken aus, wobei eine derartige Auswahl von allen partiellen Zeichenfolgen, die durch die Textunterteilungseinheit **2** unterteilt wurden, die die Suchzeichenfolge unterteilt, durchgeführt wird. Um die partiellen Zeichenfolgen nicht mehr als notwendig auszuwählen, um die volle Länge der Suchzeichenfolge abzudecken, wählt die Auswahleinheit **6** für die partielle Zeichenfolge partielle Zeichenfolgen eine nach der anderen von Anfang der Suchzeichenfolge aus, so dass sie sich nicht miteinander überlappen. Falls die partiellen Zeichenfolgen, die ausgewählt sind, so dass sie sich nicht miteinander überlappen, nicht die volle Länge der Suchzeichenfolge abdecken können, wird zusätzlich eine partielle Zeichenfolge, die einem Endabschnitt der Suchzeichenfolge entspricht, ausgewählt.

[0160] Wenn eine Suchzeichenfolge „ABCDEF" zum Beispiel bereitgestellt wird, extrahiert die Textunterteilungseinheit **2** fünf partielle Suchzeichenfolgen „AB", „BC", „CD", „DE" und „EF". In diesem Fall wählt die Auswahleinheit **6** für die partielle Suchzeichenfolge drei der fünf Folgen „AB", „CD" und „EF" aus, während die anderen beiden Zeichenfolgen „BC" und „DE" ausgelassen werden. Wenn eine Suchzeichenfolge „BCDEF" auf der anderen Seite lautet, werden zwei partielle Zeichenfolgen „BC" und „DE" zuerst gewählt. Keine weiteren partiellen Zeichenfolgen können ohne Überlapp ausgewählt werden, jedoch sind die ausgewählten zwei partiellen Zeichenfolgen nicht dazu in der Lage, die volle Länge der Suchzeichenfolge abzudecken (d.h. sie können nicht das letzte Zeichen der Suchzeichenfolge abdecken). In diesem Fall wird eine andere partielle Suchzeichenfolge „EF" zusätzlich ausgewählt. Infolgedessen werden drei partielle Zeichenfolgen „BC", „DE" und „EF" von der Suchzeichenfolge „BCDEF" ausgewählt.

[0161] Ausgewählte partielle Zeichenfolgen sind immer weniger als alle partiellen Zeichenfolgen. Wenn eine Suchzeichenfolge aus m-Zeichen besteht, ist die Anzahl der ausgewählten partiellen Zeichenfolgen gleich der kleinsten ganzen Zahl, nicht kleiner als m/n. Falls nämlich n gleich 2 und m gleich 3 ist, ist die kleinste ganze Zahl, die nicht kleiner als 3/2 ist, 2. Falls n gleich 2 ist und m gleich 4 ist, ist die kleinste ganze Zahl, die nicht kleiner als 4/2 ist, 2. Falls n gleich 2 und m gleich 5 ist, ist die kleinste ganze Zahl, die nicht kleiner als 5/2 ist, 3. Diese Zahl der ausgewählten partiellen Zeichenfolgen ist kleiner als (m – n + 1), das ist die Zahl der partiellen Zeichenfolgen, die bei der japanischen offengelegten Patentanmeldung

Nr. 11/85776 verwendet wird. In dieser Art und Weise kann die achte Ausführungsform der vorliegenden Erfindung die Berechnungslast zum Auswählen von Dokumenten und Berechnen von Rangordnungspunkten reduzieren, wodurch eine Dokumentenbeschaffung mit hoher Geschwindigkeit erzielt wird.

[0162] Fig. 14 ist ein Flussdiagramm eines Prozesses zum Auswählen partieller Zeichenfolgen, der nicht die volle Länge einer Suchzeichenfolge überlappt und abdeckt.

[0163] Bei einem Schritt S201 wird ein Parameter s auf 1 gesetzt. Dieser Parameter zeigt eine Startposition einer partiellen Zeichenfolge an.

[0164] Bei einem Schritt S202 wird s plus sublen berechnet. Die sich ergebende sum (Summe) wird in einem Parameter e festgelegt. Hier ist sublen eine Länge von partiellen Zeichenfolgen, d.h. eine Länge einer Verarbeitungseinheit. Der Parameter e zeigt die Position an, die als nächstes einer Endposition der partiellen Zeichenfolge folgt, und zwar beginnend bei der Position s.

[0165] Bei einem Schritt S203 wird eine Überprüfung dahingehend durchgeführt, ob e größer als len plus 1 ist, wobei len die Länge einer Suchzeichenfolge ist. Falls dies nicht so ist, geht das Verfahren zu einem Schritt S204 über.

[0166] In dem Schritt S204 wird eine partielle Zeichenfolge mit der Startposition s als eine der partiellen Zeichenfolge für den Beschaffungszweck ausgewählt.

[0167] In dem Schritt S205 wird eine Überprüfung dahingehend durchgeführt, ob e gleich len plus 1 ist. Falls dies so ist, kommt das Verfahren zu einem Ende.

[0168] In dem Schritt S206 wird die Startposition s auf e festgelegt. Dann geht das Verfahren zu dem Schritt S202 zurück.

[0169] Falls die Überprüfung bei dem Schritt S203 findet, dass e größer ist als len plus 1 ist, geht das Verfahren zu einem Schritt S207 über.

[0170] In dem Schritt S207 wird eine partielle Zeichenfolge mit einer Startposition (len – sublen + 1) als eine der partiellen Zeichenfolgen für ein Beschaffungszweck ausgewählt. Dann kommt das Verfahren zu einem Ende.

[0171] Bemerkenswert ist, dass die achte Ausführungsform der vorliegenden Erfindung in derselben Art und Weise modifiziert werden kann, wie die erste Ausführungsform modifiziert worden ist, um die zweite bis siebte Ausführungsform, wie zuvor beschrieben wurde, bereitzustellen.

[0172] Ein Verfahren zur Dokumentenbeschaffung beinhaltet die Schritte der Unterteilung einer Suchzeichenfolge in partielle Zeichenfolgen, das Auswählen eines Dokuments oder mehrerer Dokumente von einer Vielzahl gespeicherter Dokumenten derartig, dass das eine Dokument oder die mehren Dokumente jeweils alle partiellen Zeichenfolgen enthalten, die Berechnung jeweiliger Ränge der partiellen Zeichenfolgen für jedes der Anzahl Dokumente und die Berechnung eines Rangs der Suchzeichenfolge von den jeweiligen Rängen der partiellen Zeichenfolgen für jedes der Anzahldokumente.

Patentansprüche

1. Verfahren zur Dokumentenbeschaffung, mit folgenden Schritten:
wenigstens eine Suchzeichenfolge wird bereitgestellt;
eine Anzahl Dokumente wird aus einer Vielzahl von gespeicherten Dokumenten ausgewählt;
ein Rangpunkt der wenigstens einen Suchzeichenfolge wird berechnet;
gekennzeichnet durch die folgenden Schritte:
a) die wenigstens eine Suchzeichenfolge wird in partielle Zeichenfolgen unterteilt;
b) jedes der Dokumente der Anzahl Dokumente wird derartig ausgewählt, dass jedes alle partiellen Zeichenfolgen enthält;
c) jeweilige Rangpunkte der partiellen Zeichenfolgen werden für jedes der Anzahl Dokumente berechnet; und
d) der Rangpunkt der jeweiligen Suchzeichenfolge wird von den jeweiligen Rangpunkten der partiellen Zeichenfolgen für jedes der Anzahl Dokumente berechnet.

2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, dass der Schritt zum Unterteilen die Suchzeichenfolge in partielle Zeichenfolgen unterteilt, die sich überlappen.

3. Verfahren nach Anspruch 1, dadurch gekennzeichnet, dass der Schritt der Unterteilung die Suchzeichenfolge in partielle Zeichenfolgen unterteilt, die sich im allgemeinen nicht überlappen und die eine volle Länge der Suchzeichenfolge abdecken.

4. Verfahren nach Anspruch 1, dadurch gekennzeichnet, dass der Schritt der Berechnung jeweiliger Rangpunkte der partiellen Zeichenfolgen die folgenden Schritte umfasst:
ein erster Zählerwert wird erhalten, der anzeigt, wie viele der gespeicherten Dokumente eine gegebene der partiellen Zeichenfolgen beinhaltet;
ein zweiter Zählerwert wird erhalten, der anzeigt, wie viele Male die gegebene eine Zeichenfolge der partiellen Zeichenfolgen in einem gegebenen Dokument

der Anzahl von Dokumenten erscheint; und
ein Rangpunkt der gegebenen einen Folge der partiellen Zeichenfolgen wird für das gegebene eine Dokument der Anzahl von Dokumenten von dem ersten Zählwert und dem zweiten Zählwert derartig erhalten, dass der Rangpunkt der gegebenen einen Folge der partiellen Zeichenfolgen mit der Abnahme des ersten Zählwertes und der Zunahme des zweiten Zählwertes zunimmt.

5. Verfahren nach Anspruch 1, dadurch gekennzeichnet, dass der Schritt der Berechnung jeweiliger Rangpunkte der partiellen Zeichenfolgen die folgenden Schritte umfasst:
ein erster Zählwert wird erhalten, der anzeigt, wie viele der gespeicherten Dokumente eine gegebene Zeichenfolge der partiellen Zeichenfolgen enthalten;
zweite Zählwerte werden erhalten, die jeweils anzeigen, wie häufig eine entsprechende Folge der partiellen Zeichenfolgen in einem gegebenen Dokument der Anzahl Dokumente erscheint;
ein kleinster Zählwert der zweiten Zählwerte wird erhalten; und
ein Rangpunkt der gegebenen einen Folge der partiellen Zeichenfolgen wird für das gegebene eine Dokument der Anzahl Dokumente von dem ersten Zählwert und dem kleinsten der zweiten Zählwerte erhalten, so dass der Rangpunkt der gegebenen einen Zeichenfolge der partiellen Zeichenfolgen zunimmt, wenn der erste Zählerwert abnimmt und wenn der kleinste Zählwert der zweiten Zählwerte zunimmt.

6. Verfahren nach Anspruch 1, dadurch gekennzeichnet, dass der Schritt der Berechnung jeweiliger Rangpunkte der partiellen Zeichenfolgen die folgenden Schritte enthält:
ein erster Zählwert wird erhalten, der anzeigt, wie viele der gespeicherten Dokumente eine gegebene Folge der partiellen Zeichenfolgen enthalten;
ein zweiter Zählwert wird erhalten, der anzeigt, wie häufig die Suchzeichenfolge in einem gegebenen Dokument der Anzahl Dokumente erscheint; und
ein Rangpunkt der gegebenen einen Folge der partiellen Zeichenfolgen wird für das gegebene eine Dokument der Anzahl Dokumente von dem ersten Zählwert und dem zweiten Zählwert erhalten, so dass der Rangpunkt der gegebenen einen Folge der partiellen Zeichenfolgen zunimmt, wenn der erste Zählwert abnimmt und wenn der zweite Zählwert zunimmt.

7. Verfahren nach Anspruch 6, dadurch gekennzeichnet, dass der Schritt der Erzielung eines zweiten Zählwerts weiter einen Schritt beinhaltet, eine obere Grenze für den zweiten Zählwert festzulegen.

8. Verfahren nach Anspruch 1, dadurch gekennzeichnet, dass der Schritt des Auswählens einer Anzahl von Dokumenten die Anzahl von Dokumenten auswählt, wobei jedes die Suchzeichenfolge enthält, und der Schritt der Berechnung jeweiliger Rangpunkte der partiellen Zeichenfolgen die folgenden Schritte enthält:
ein erster Zählwert wird erhalten, der anzeigt, wie viele der gespeicherten Dokumente die Suchzeichenfolge enthalten;
ein zweiter Zählwert wird erhalten, der anzeigt, wie häufig eine Folge der partiellen Zeichenfolgen in einem gegebenen Dokument der Anzahl Dokumente erscheint; und
ein Rangpunkt der gegebenen Folge der partiellen Zeichenfolgen wird für das gegebene Dokument der Anzahl Dokumente von dem ersten Zählwert und dem zweiten Zählwert erhalten, so dass der Rangpunkt der gegebenen einen Folge der partiellen Zeichenfolgen zunimmt, wenn der erste Zählwert abnimmt und wenn der zweite Zählwert zunimmt.

9. Verfahren nach Anspruch 1, dadurch gekennzeichnet, dass der Schritt des Auswählens einer Anzahl von Dokumenten die Anzahl von Dokumenten auswählt, von denen jedes die Suchzeichenfolge enthält, und der Schritt der Berechnung jeweiliger Rangpunkt der partiellen Suchzeichenfolgen die folgenden Schritte enthält:
ein erster Zählwert wird erhalten, der anzeigt, wie viele der gespeicherten Dokumente die Suchzeichenfolge enthalten;
von dem ersten Zählwert wird eine Grenze berechnet;
ein zweiter Zählwert wird erhalten, der anzeigt, wie häufig die Suchzeichenfolge in einem gegebenen Dokument der Anzahl von Dokumenten erscheint, während ein oberes Ende des zweiten Zählwerts auf die Grenze beschränkt wird; und
ein Rangpunkt eines gegebenen der partiellen Zeichenfolgen wird für das gegebene eine Dokument der Anzahl Dokumente von dem ersten Zählwert und dem zweiten Zählwert derartig erhalten, dass der Rangpunkt der gegebenen einen Folge der partiellen Zeichenfolge zunimmt, wenn der erste Zählwert abnimmt und wenn der zweite Zählwert zunimmt.

10. Verfahren zur Dokumentenbeschaffung nach Anspruch 1, gekennzeichnet durch die folgenden Schritte:
jeweilige Indizes werden für Dokumente bereitgestellt, wobei jeder der jeweiligen Indizes eine der partiellen Zeichenfolgen, die in einem entsprechenden Dokument gefunden wurden und jeweilige Positionen dafür in dem entsprechenden Dokument auflistet;
die partiellen Zeichenfolgen werden ausgewählt, die mit einer Zeichenfolge starten, die identisch mit einer Suchzeichenfolge ist;
die Anzahl Dokumente wird aus den Dokumenten derartig ausgewählt, dass die Anzahl Dokumente jeweils wenigstens eine Folge der ausgewählten partiellen Zeichenfolgen enthalten;
die jeweiligen Rangpunkte der partiellen Zeichenfolgen werden aufgrund der ausgewählten partiellen Zeichenfolgen für jedes Dokument der Anzahl Doku-

mente berechnet; und der Rangpunkt der Suchzeichenfolge wird basierend auf den jeweiligen Rangpunkten der ausgewählten partiellen Zeichenfolgen für jedes Dokument der Anzahl Dokumente berechnet.

11. Verfahren nach Anspruch 10, dadurch gekennzeichnet, dass der Schritt der Berechnung jeweiliger Rangpunkte der ausgewählten partiellen Zeichenfolgen die folgenden Schritte enthält:
ein erster Zählwert wird erhalten, der anzeigt, wie viele der gespeicherten Dokumente eine gegebene Folge der ausgewählten partiellen Zeichenfolgen anzeigt;
ein zweiter Zählwert wird erhalten, der anzeigt, wie häufig die gegebene Zeichenfolge der ausgewählten Zeichenfolgen in einem gegebenen Dokument der Anzahl Dokumente erscheint; und
ein Rangpunkt der gegebenen Folge der ausgewählten partiellen Zeichenfolgen wird für das gegebene Dokument der Anzahl Dokumente von dem ersten Zählwert und dem zweiten Zählwert derartig erhalten, dass der Rangpunkt der gegebenen Zeichenfolge der ausgewählten partiellen Zeichenfolgen zunimmt, wenn der erste Zählwert abnimmt und wenn der zweite Zählwert zunimmt.

12. Vorrichtung zur Dokumentenbeschaffung, mit folgenden Merkmalen, wobei
wenigstens eine Suchzeichenfolge bereitgestellt wird,
einer Dokumentenauswahleinheit (**4**), die ausgebildet ist, um eine Anzahl Dokumente aus einer Vielzahl von gespeicherten Dokumenten auszuwählen;
eine Rangberechnungseinheit (**5**), die ausgebildet ist, um einen Rangpunkt der wenigstens einen Suchzeichenfolge zu berechnen;
gekennzeichnet durch die folgenden Merkmale:
a) einer Unterteilungseinheit (**2**), die ausgebildet ist, um die wenigstens eine Suchzeichenfolge in partielle Zeichenfolgen zu unterteilen;
b) die Dokumentenauswahleinheit (**4**) ist ferner ausgebildet, um jedes der Dokumente der Anzahl Dokumente derartig auszuwählen, dass jedes alle partiellen Zeichenfolgen enthält;
c) die Rangberechnungseinheit (**5**) ist ferner ausgebildet, um jeweilige Rangpunkte der partiellen Zeichenfolgen für jedes der Anzahl Dokumente zu berechnen, und
d) die Rangberechnungseinheit (**5**) ist ferner ausgebildet, um den Rangpunkt der jeweiligen Suchzeichenfolge von den jeweiligen Rangpunkten der partiellen Zeichenfolgen für jedes der Anzahl Dokumente zu berechnen.

13. Vorrichtung nach Anspruch 12, dadurch gekennzeichnet, dass die Unterteilungseinheit (**2**) die Suchzeichenfolge in partielle Zeichenfolgen unterteilt, die einander überlappen.

14. Vorrichtung nach Anspruch 13, die weiter eine Auswahleinheit (**6**) für partielle Zeichenfolgen umfasst, die die partiellen Zeichenfolgen auswählt, die im allgemeinen nicht überlappen und die die volle Länge der Suchzeichenfolge abdecken, wobei die ausgewählten partiellen Zeichenfolgen aufeinander folgend berechnet werden, um die jeweiligen Rangpunkte der ausgewählten partiellen Zeichenfolgen zu berechnen.

15. Vorrichtung nach Anspruch 12, dadurch gekennzeichnet, dass die Rangberechnungseinheit (**5**) folgendes enthält:
eine erste Einrichtung, die einen ersten Zählwert erhält, der anzeigt, wie viele der gespeicherten Dokumente eine gegebene Folge der partiellen Zeichenfolgen enthalten;
eine zweite Einrichtung, die einen zweiten Zählwert erhält, der anzeigt, wie häufig die gegebene Folge der partiellen Zeichenfolgen in einem gegebenen Dokument der Anzahl Dokumente erscheint; und
eine Rangeinrichtung, die einen Rangpunkt der gegebenen Folge der partiellen Zeichenfolgen für das gegebene Dokument der Anzahl Dokumente von dem ersten Zählwert und dem zweiten Zählwert derartig erhält, dass der Rangpunkt der gegebenen Folge der partiellen Zeichenfolgen zunimmt, wenn der erste Zählwert abnimmt und wenn der zweite Zählwert zunimmt.

16. Vorrichtung nach Anspruch 12, dadurch gekennzeichnet, dass die Rangberechnungseinheit (**5**) folgendes enthält:
eine erste Einrichtung, die einen ersten Zählwert erhält, der anzeigt, wie viele Dokumente der gespeicherten Dokumente eine gegebene Folge der partiellen Zeichenfolgen enthalten;
eine zweite Einrichtung, die zweite Zählwerte erhält, die anzeigen, wie häufig eine entsprechende Folge der partiellen Zeichenfolgen in einem gegebenen Dokument der Anzahl Dokumente erscheint;
eine Minimaleinrichtung, um den kleinsten der zweiten Zählwerte zu erhalten; und
eine Rangeinrichtung, die einen Rang der gegebenen Folge der partiellen Zeichenfolgen für das gegebene Dokument der Anzahl Dokumente von dem ersten Zählwert und dem kleinsten der zweiten Zählwerte derartig erhält, dass der Rangpunkt der gegebenen Folge der partiellen Zeichenfolgen zunimmt, wenn der erste Zählwert abnimmt und wenn der kleinste der zweiten Zählwerte zunimmt.

17. Vorrichtung nach Anspruch 12, dadurch gekennzeichnet, dass die Rangberechnungseinheit (**5**) folgendes enthält:
eine erste Einrichtung, die einen ersten Zählwert erhält, der anzeigt, wie viele der gespeicherten Dokumente eine gegebene Folge der partiellen Zeichenfolgen enthalten;
eine zweite Einrichtung, die einen zweiten Zählwert

erhält, der anzeigt, wie häufig die Suchzeichenfolge in einem gegebenen Dokument der Anzahl Dokumente erscheint; und
eine Einrichtung, die einen Rangpunkt der gegebenen Zeichenfolge der partiellen Zeichenfolgen für das gegebene Dokument der Anzahl Dokumente von dem ersten Zählwert und dem zweiten Zählwert derartig erhält, dass der Rangpunkt der gegebenen Zeichenfolge der partiellen Zeichenfolgen zunimmt, wenn der erste Zählwert abnimmt und wenn der zweite Zählwert zunimmt.

18. Vorrichtung nach Anspruch 17, dadurch gekennzeichnet, dass die zweite Einrichtung zum Erzielen eines zweiten Zählwertes weiter eine Festlegeinrichtung zum Festlegen einer oberen Grenze des zweiten Zählwerts enthält.

19. Vorrichtung nach Anspruch 12, dadurch gekennzeichnet, dass die Dokumentauswahleinheit (**4**) die Anzahl Dokumente auswählt, von denen jedes die Suchzeichenfolge enthält und wobei die Rangberechnungseinheit (**5**) folgendes enthält:
eine erste Einrichtung, die einen ersten Zählwert erhält, der anzeigt, wie viele der gespeicherten Dokumente die Suchzeichenfolge enthalten;
eine zweite Einrichtung, die einen zweiten Zählwert erhält, der anzeigt, wie häufig eine gegebene Folge der partiellen Zeichenfolgen in einem gegebenen Dokument der Anzahl Dokumente erscheint; und
eine Einrichtung, die einen Rangpunkt der gegebenen Folge der partiellen Zeichenfolgen für das gegebene Dokument der Anzahl Dokumente von dem ersten Zählwert und dem zweiten Zählwert derartig erhält, dass der Rangpunkt der gegebenen Folge der gegebenen Zeichenfolgen zunimmt, wenn der erste Zählwert abnimmt und der zweite Zählwert zunimmt.

20. Vorrichtung nach Anspruch 12, dadurch gekennzeichnet, dass die Dokumentauswahleinheit (**4**) die Anzahl Dokumente auswählt, von denen jedes die Suchzeichenfolge enthält, und die Rangberechnungseinheit (**5**) folgendes enthält:
eine erste Einrichtung, die einen ersten Zählwert erhält, der anzeigt, wie viele der gespeicherten Dokumente die Suchzeichenfolge enthalten;
eine Berechnungseinrichtung, die eine Grenze von dem ersten Zählwert berechnet;
eine zweite Einrichtung, die einen zweiten Zählwert erhält, der anzeigt, wie häufig die Suchzeichenfolge in einem gegebenen Dokument der Anzahl Dokumente erscheint, während ein oberes Ende des zweiten Zählwerts auf die Grenze beschränkt wird; und
eine Rangeinrichtung, die einen Rangpunkt einer gegebenen Folge der partiellen Zeichenfolgen für das gegebene Dokument der Anzahl Dokumente von dem ersten Zählwert und dem zweiten Zählwert derartig erhält, dass der Rangpunkt der gegebenen Folge der partiellen Zeichenfolgen zunimmt, wenn der erste Zählwert abnimmt und wenn der zweite Zählwert zunimmt.

21. Vorrichtung zur Dokumentenbeschaffung nach Anspruch 12, die folgendes umfasst:
eine Textunterteilungseinheit (**2**), die ausgebildet ist, um jeweilige Indizes für Dokumente bereitzustellen,
wobei jeder der jeweiligen Indizes eine der partiellen Zeichenfolgen, die in einem entsprechenden Dokument gefunden wurden und jeweilige Positionen dafür in dem entsprechenden Dokument auflistet;
wobei die Dokumentauswahleinheit (**4**) ferner ausgebildet ist, um die partiellen Zeichenfolgen auszuwählen, die mit einer Zeichenfolge starten, die identisch mit einer Suchzeichenfolge ist und um
die eine Anzahl Dokumente von den Dokumenten derartig auswählt, dass die Anzahl Dokumente jeweils wenigstens eine Folge der ausgewählten partiellen Zeichenfolge enthalten; und
wobei die Rangberechnungseinheit (**5**) ferner ausgebildet ist, um die jeweiligen Rangpunkte der partiellen Zeichenfolgen aufgrund der ausgewählten partiellen Zeichenfolgen für jedes Dokument der Anzahl Dokumente zu berechnen; und um
den Rangpunkt der Suchzeichenfolge basierend auf den jeweiligen Rangpunkten der ausgewählten partiellen Zeichenfolgen für jedes Dokument der Anzahl Dokumente zu berechnen.

22. Vorrichtung nach Anspruch 21, dadurch gekennzeichnet, dass die Rangberechnungseinheit (**5**) folgendes enthält:
eine erste Einrichtung, die einen ersten Zählwert erhält, der anzeigt, wie viele der gespeicherten Dokumente eine gegebene Folge der ausgewählten partiellen Zeichenfolgen enthalten;
eine zweite Einrichtung, die einen zweiten Zählwert erhält, der anzeigt, wie häufig die gegebene Folge der ausgewählten partiellen Zeichenfolgen in einem gegebenen Dokument der Anzahl Dokumente erscheint; und
eine Rangeinrichtung, die einen Rang für die gegebene Folge der ausgewählten partiellen Zeichenfolgen für das gegebene Dokument der Anzahl Dokumente von dem ersten Zählwert und dem zweiten Zählwert derartig erhält, dass der Rangpunkt der gegebenen Folge der ausgewählten partiellen Zeichenfolgen zunimmt, wenn der erste Zählwert abnimmt und wenn der zweite Zählwert zunimmt.

23. Vorrichtung nach Anspruch 22, dadurch gekennzeichnet, dass die Rangberechnungseinheit (**5**) folgendes enthält:
eine erste Einrichtung, die einen ersten Zählwert erhält, der anzeigt, wie viele der gespeicherten Dokumente eine gegebene Folge der ausgewählten partiellen Zeichenfolgen enthält;
eine zweite Einrichtung, die eine zweite Folge erhält, die anzeigt, wie häufig die Suchzeichenfolge in einem gegebenen Dokument der Anzahl Dokumente erscheint; und

eine Rangeinrichtung, die einen Rangpunkt für die gegebene Folge der ausgewählten partiellen Zeichenfolgen für das gegebene Dokument der Anzahl Dokumente von dem ersten Zählwert und dem zweiten Zählwert derartig erhält, dass der Rangpunkt der gegebenen Folge der ausgewählten partiellen Zeichenfolgen zunimmt, wenn der erste Zählwert abnimmt und wenn der zweite Zählwert zunimmt.

24. Computer lesbares Aufzeichnungsmedium, das ein Programm enthält, dessen Schritte das Verfahren nach einem der Ansprüche 1 bis 11 auf einem Computer ausführen.

Es folgen 14 Blatt Zeichnungen

Anhängende Zeichnungen

FIG. 1

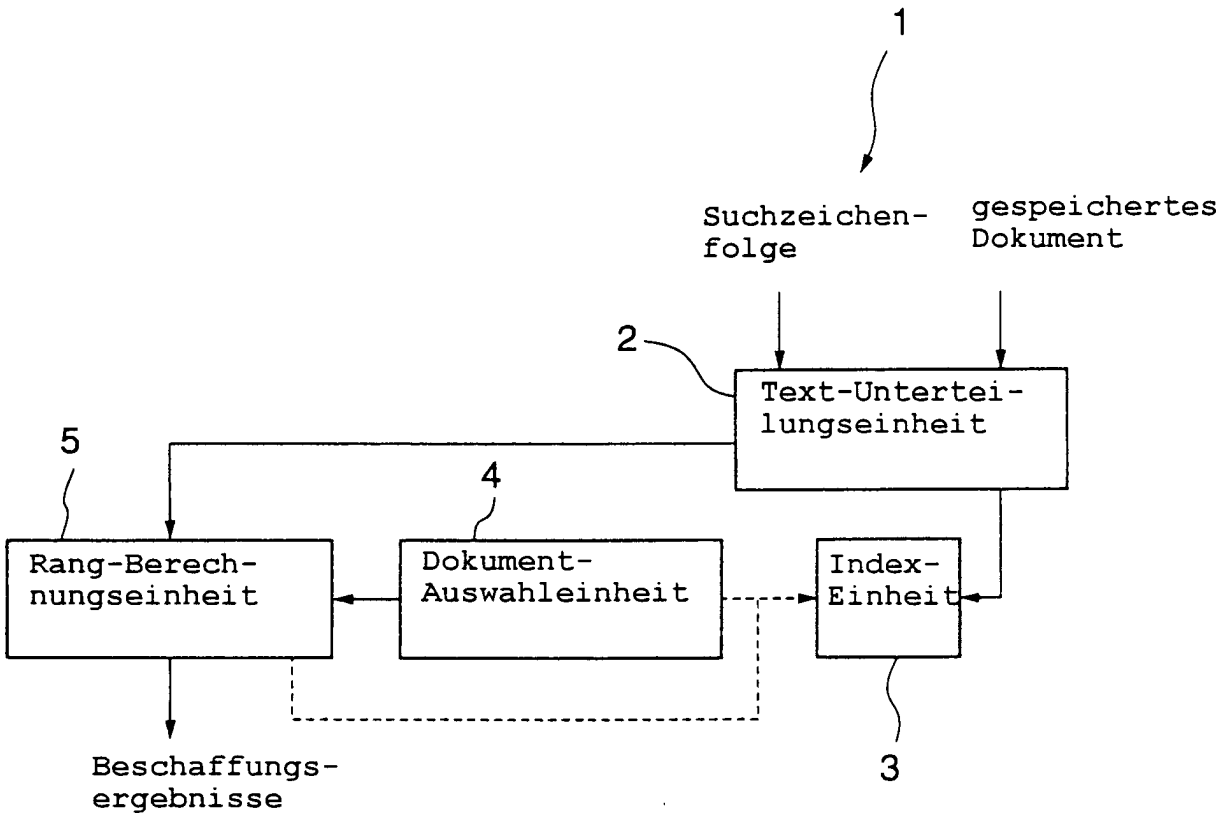


FIG. 2A

Dokument 1

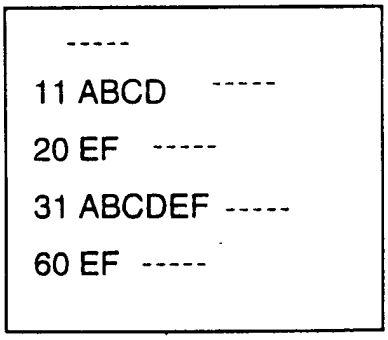


FIG. 2B

Dokument 2

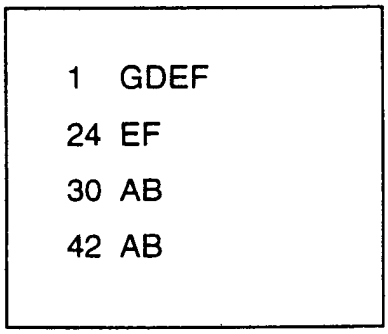


FIG. 2C

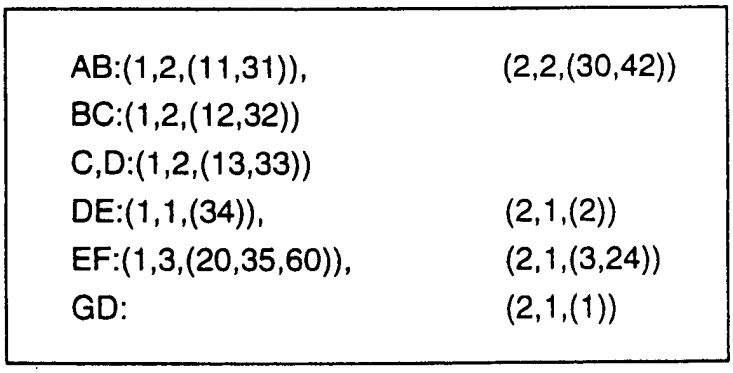


FIG. 3

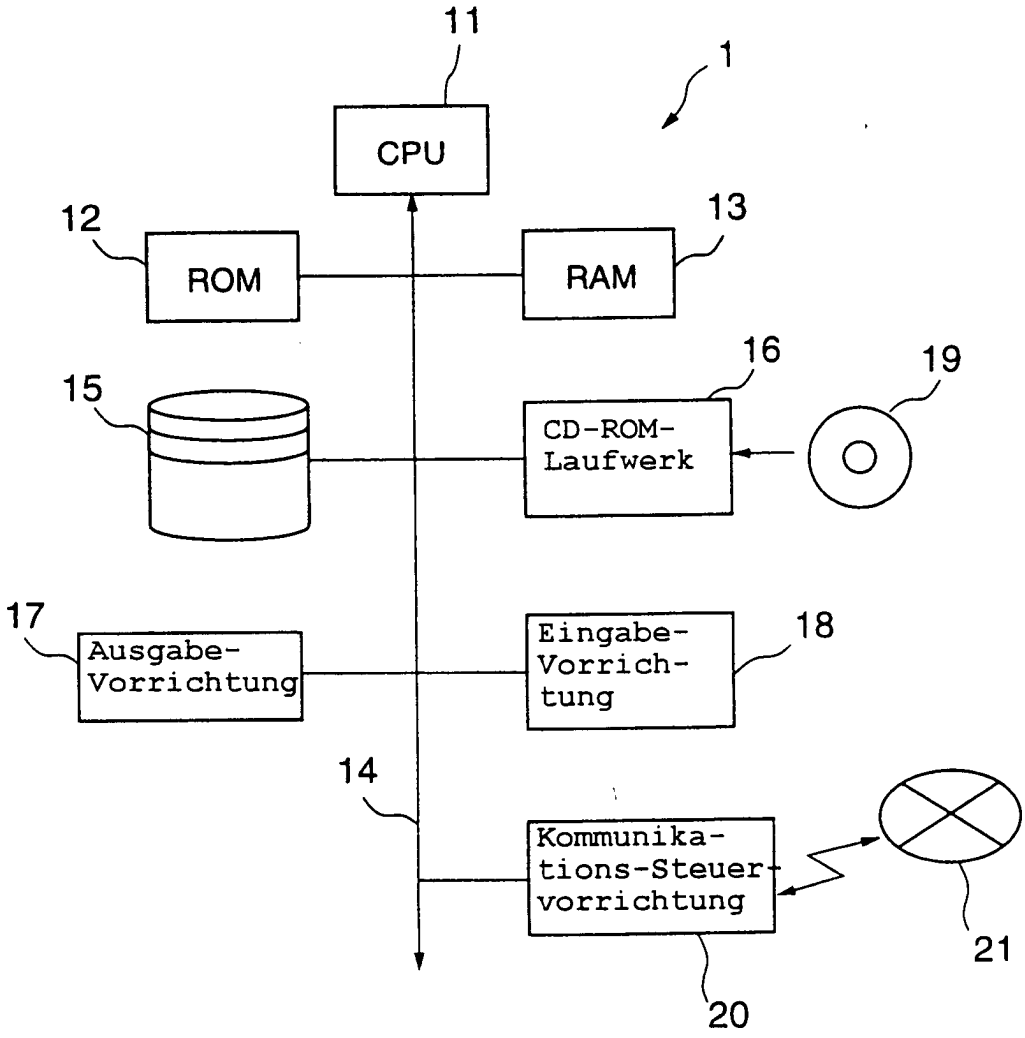

11
1
CPU
12
ROM
13
RAM
16
19
15
CD-ROM-
Laufwerk
17
Ausgabe-
Vorrichtung
Eingabe-
Vorrich-
tung
18
14
Kommunika-
tions-Steuer-
vorrichtung
21
20

FIG. 4

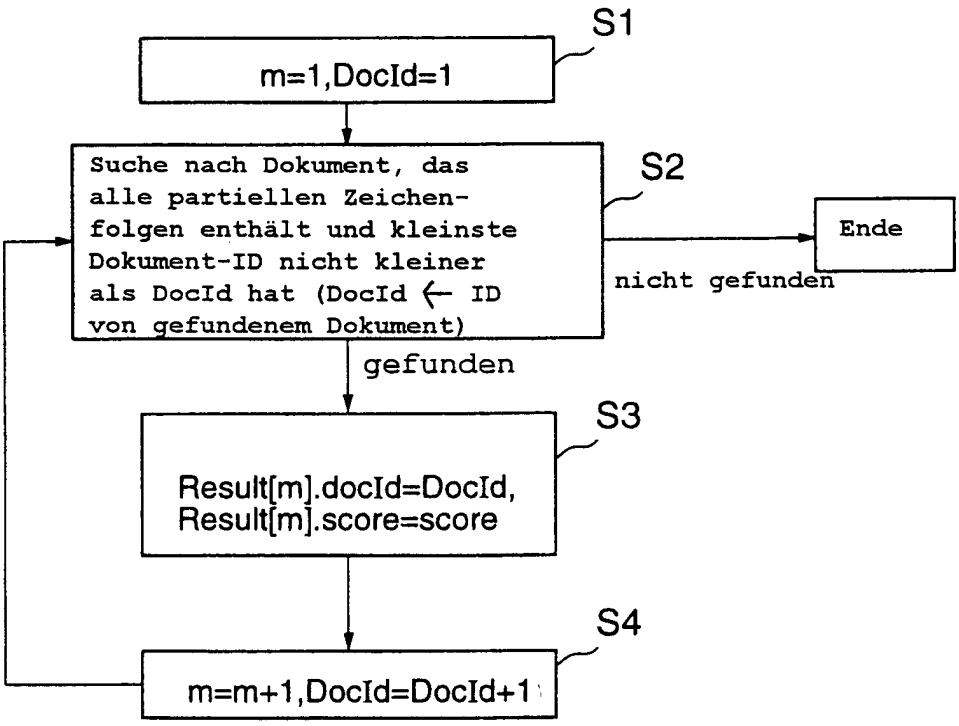

S1
m=1,DocId=1
Suche nach Dokument, das alle partiellen Zeichen-folgen enthält und kleinste Dokument-ID nicht kleiner als DocId hat (DocId ← ID von gefundenem Dokument)
S2
Ende
nicht gefunden
gefunden
S3
Result[m].docId=DocId,
Result[m].score=score
S4
m=m+1,DocId=DocId+1

FIG. 5

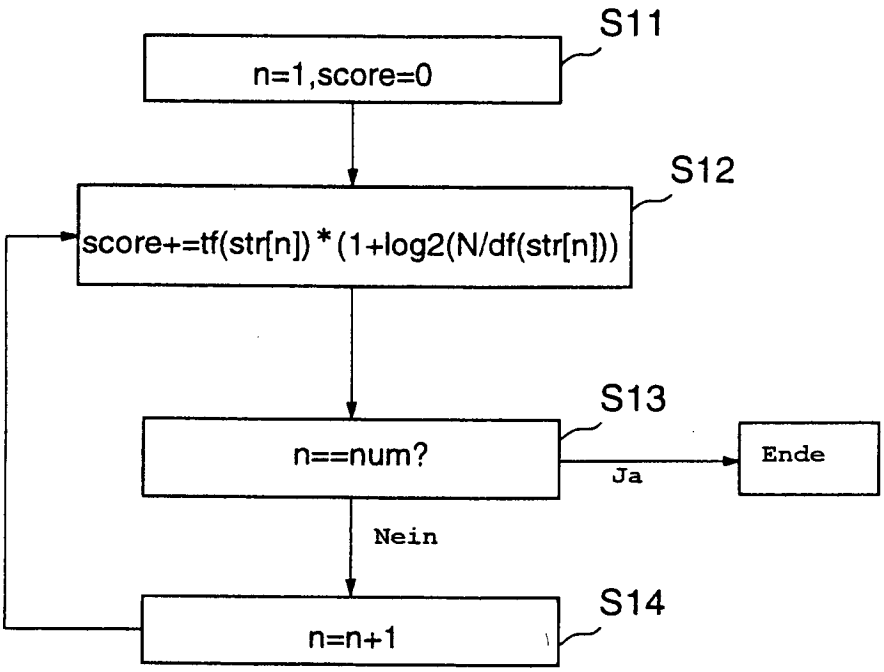

S11
n=1,score=0
S12
score+=tf(str[n])*(1+log2(N/df(str[n])))
S13
n==num?
Ja
Ende
Nein
S14
n=n+1

FIG. 6

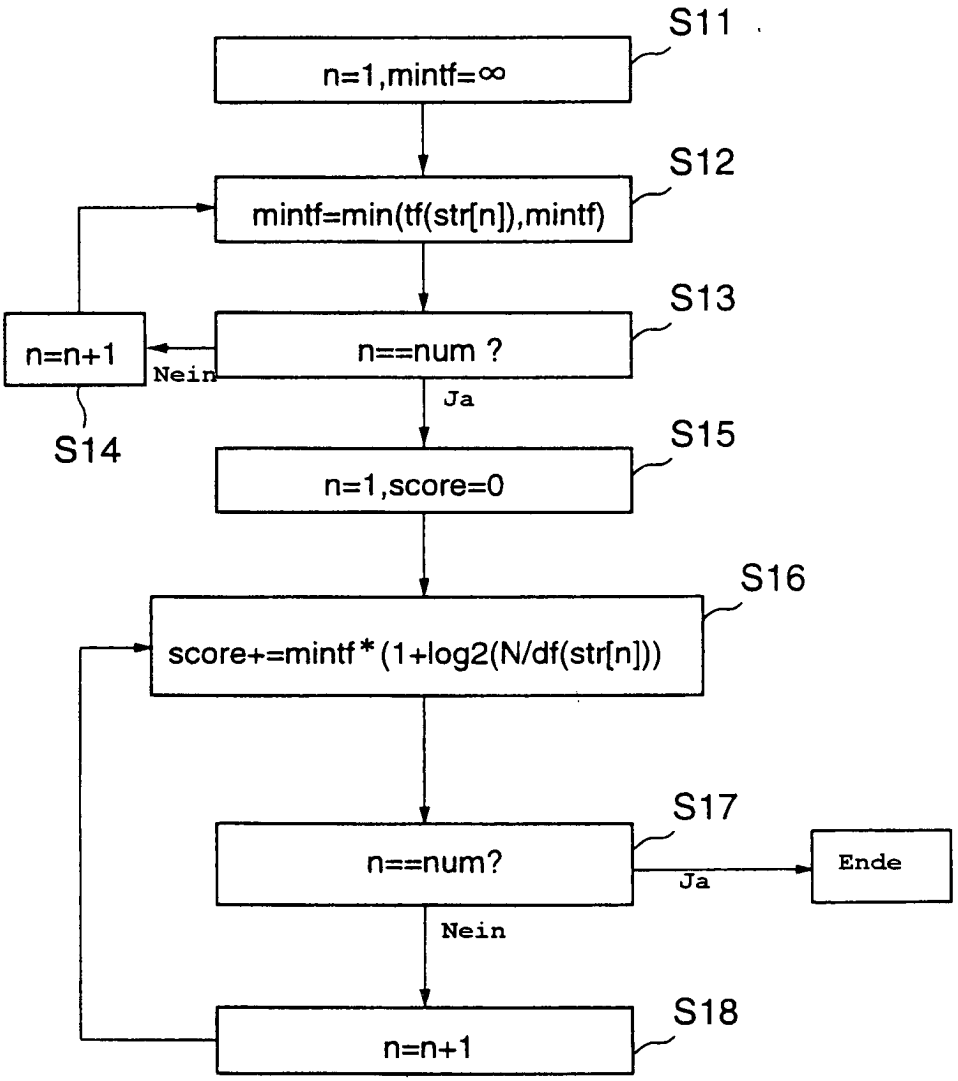

S11
n=1,mintf=∞
S12
mintf=min(tf(str[n]),mintf)
S13
n==num ?
Nein
n=n+1
S14
Ja
S15
n=1,score=0
S16
score+=mintf*(1+log2(N/df(str[n])))
S17
n==num?
Ja
Ende
Nein
S18
n=n+1

FIG. 7

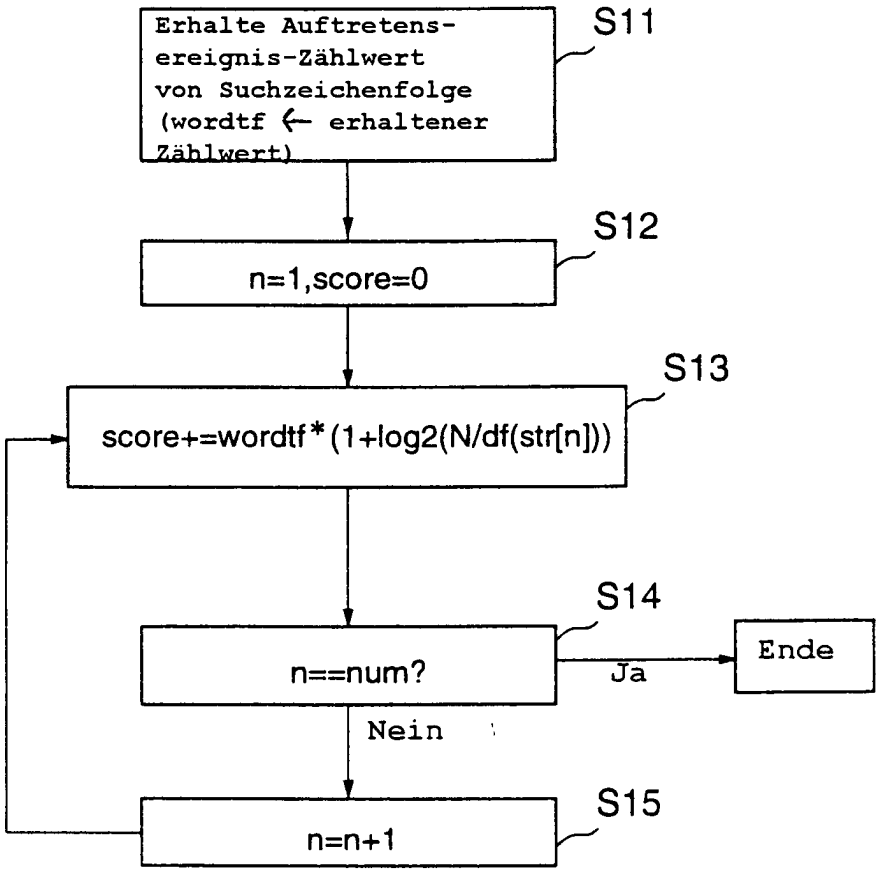

Erhalte Auftretens-ereignis-Zählwert von Suchzeichenfolge (wordtf ← erhaltener Zählwert)
S11
n=1,score=0
S12
score+=wordtf*(1+log2(N/df(str[n]))
S13
n==num?
S14
Ja
Ende
Nein
n=n+1
S15

FIG. 8

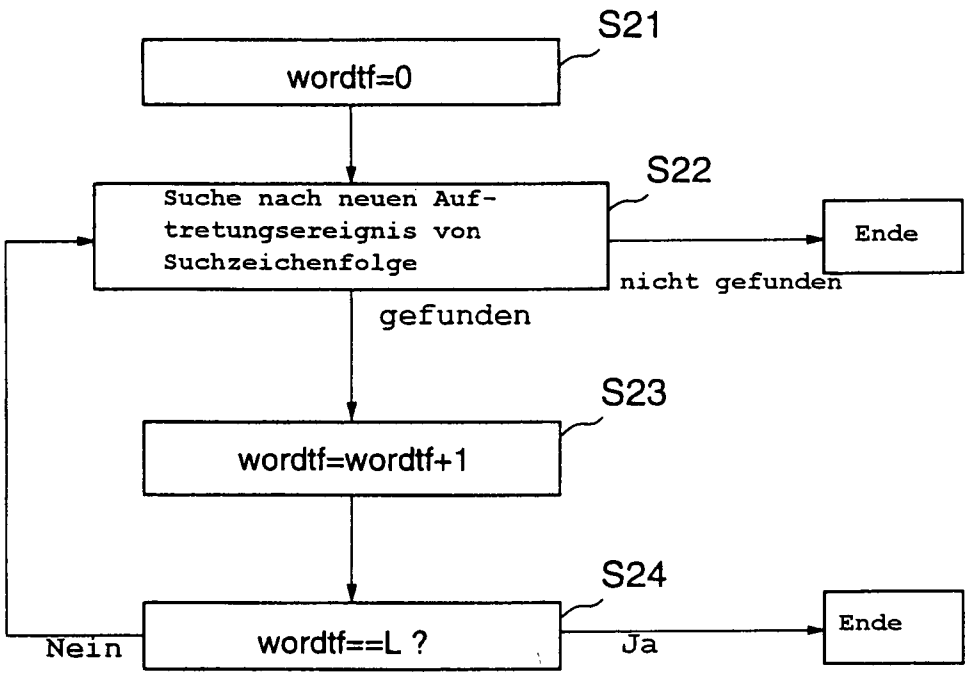

S21
wordtf=0
S22
Suche nach neuen Auf-
tretungsereignis von
Suchzeichenfolge
Ende
nicht gefunden
gefunden
S23
wordtf=wordtf+1
S24
wordtf==L ?
Nein
Ja
Ende

FIG. 9

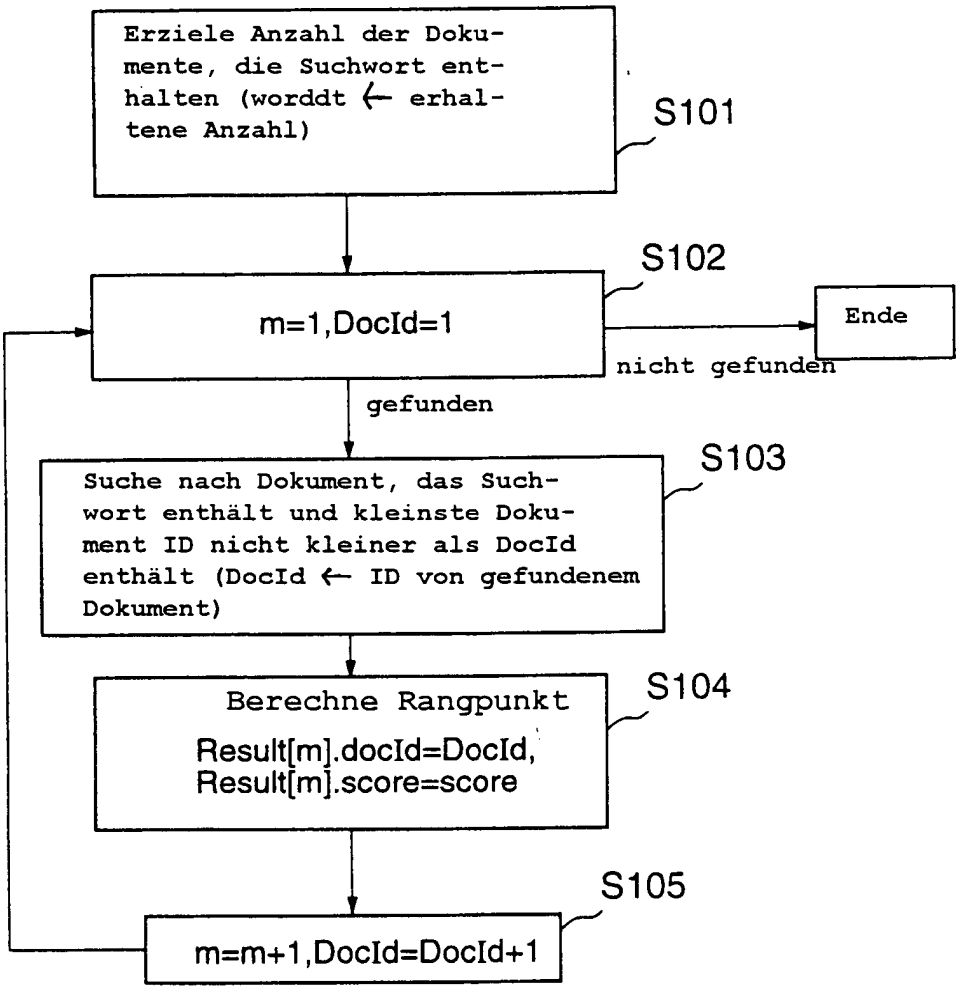

Erziele Anzahl der Dokumente, die Suchwort enthalten (worddt ← erhaltene Anzahl)
S101
S102
m=1,DocId=1
Ende
nicht gefunden
gefunden
S103
Suche nach Dokument, das Suchwort enthält und kleinste Dokument ID nicht kleiner als DocId enthält (DocId ← ID von gefundenem Dokument)
S104
Berechne Rangpunkt
Result[m].docId=DocId,
Result[m].score=score
S105
m=m+1,DocId=DocId+1

FIG. 10

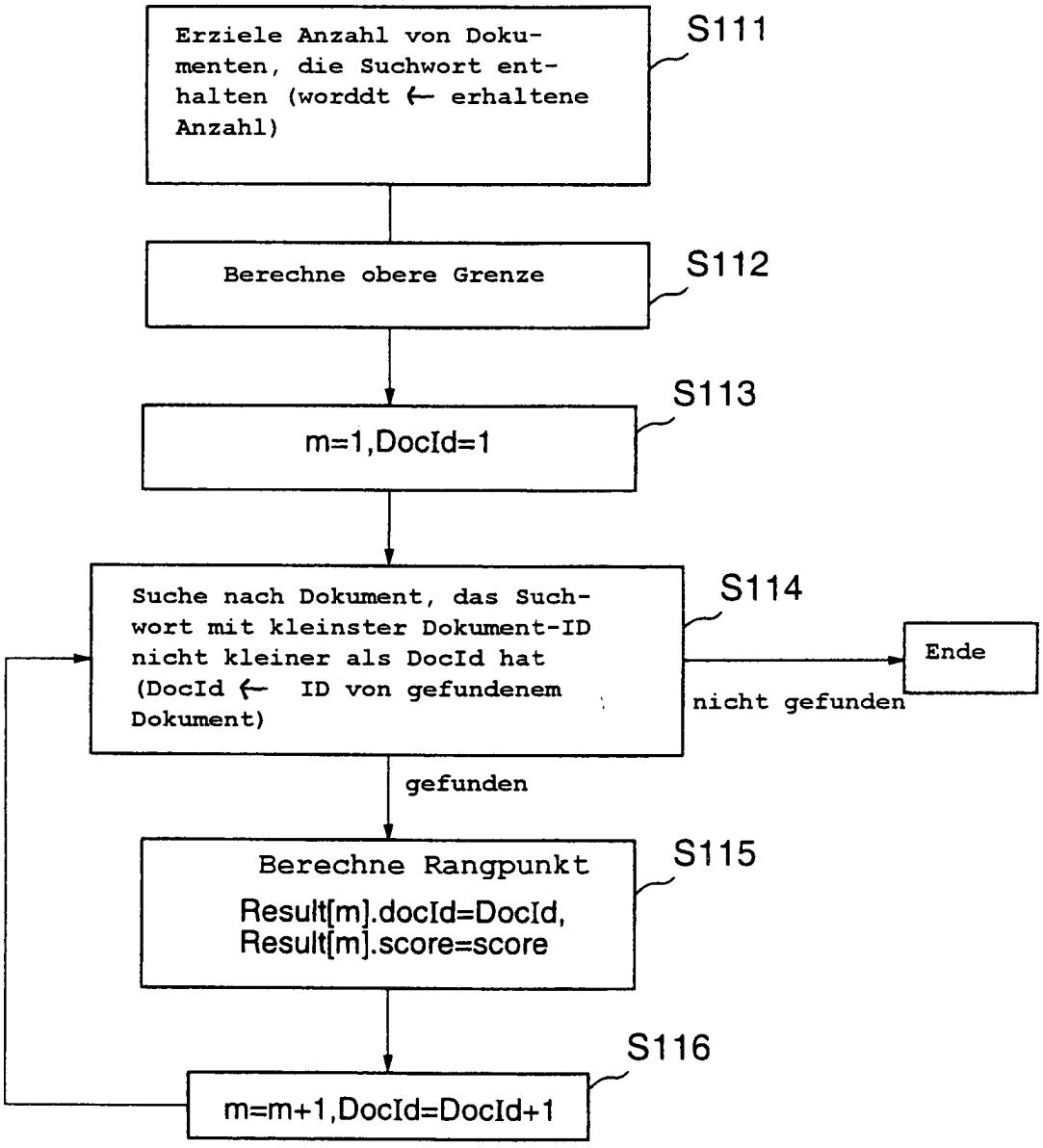

Erziele Anzahl von Dokumenten, die Suchwort enthalten (worddt ← erhaltene Anzahl)
S111
Berechne obere Grenze
S112
m=1,DocId=1
S113
Suche nach Dokument, das Suchwort mit kleinster Dokument-ID nicht kleiner als DocId hat (DocId ← ID von gefundenem Dokument)
S114
Ende
nicht gefunden
gefunden
Berechne Rangpunkt
Result[m].docId=DocId,
Result[m].score=score
S115
m=m+1,DocId=DocId+1
S116

FIG. 11A

Dokument 1

TKYBYKT。

FIG. 11B

Dokument 2

SNMWYIT。

FIG. 11C

IT : (2,1,(6))

YI: (2,1,(5))
YK: (1,1,(5))
YB: (1,1,(3))

FIG. 12

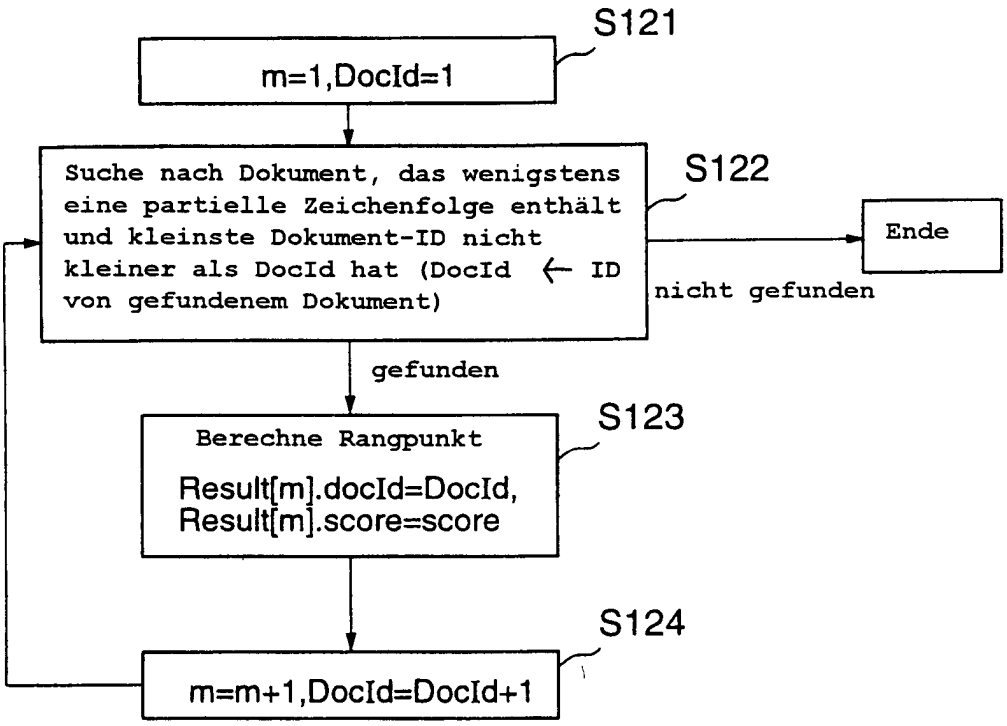

S121
m=1,DocId=1
Suche nach Dokument, das wenigstens eine partielle Zeichenfolge enthält und kleinste Dokument-ID nicht kleiner als DocId hat (DocId ← ID von gefundenem Dokument)
S122
Ende
nicht gefunden
gefunden
Berechne Rangpunkt
Result[m].docId=DocId,
Result[m].score=score
S123
S124
m=m+1,DocId=DocId+1

FIG. 13

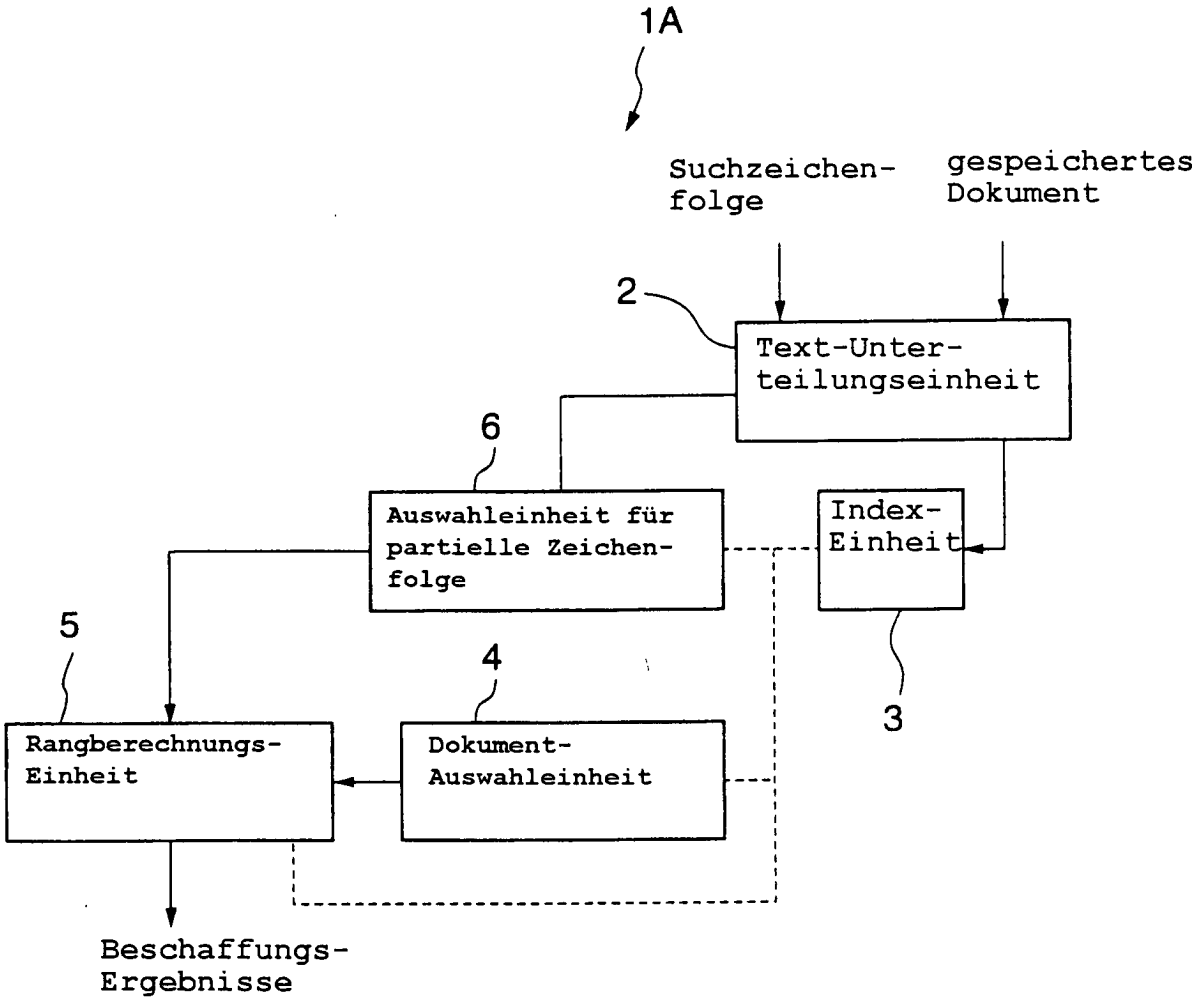
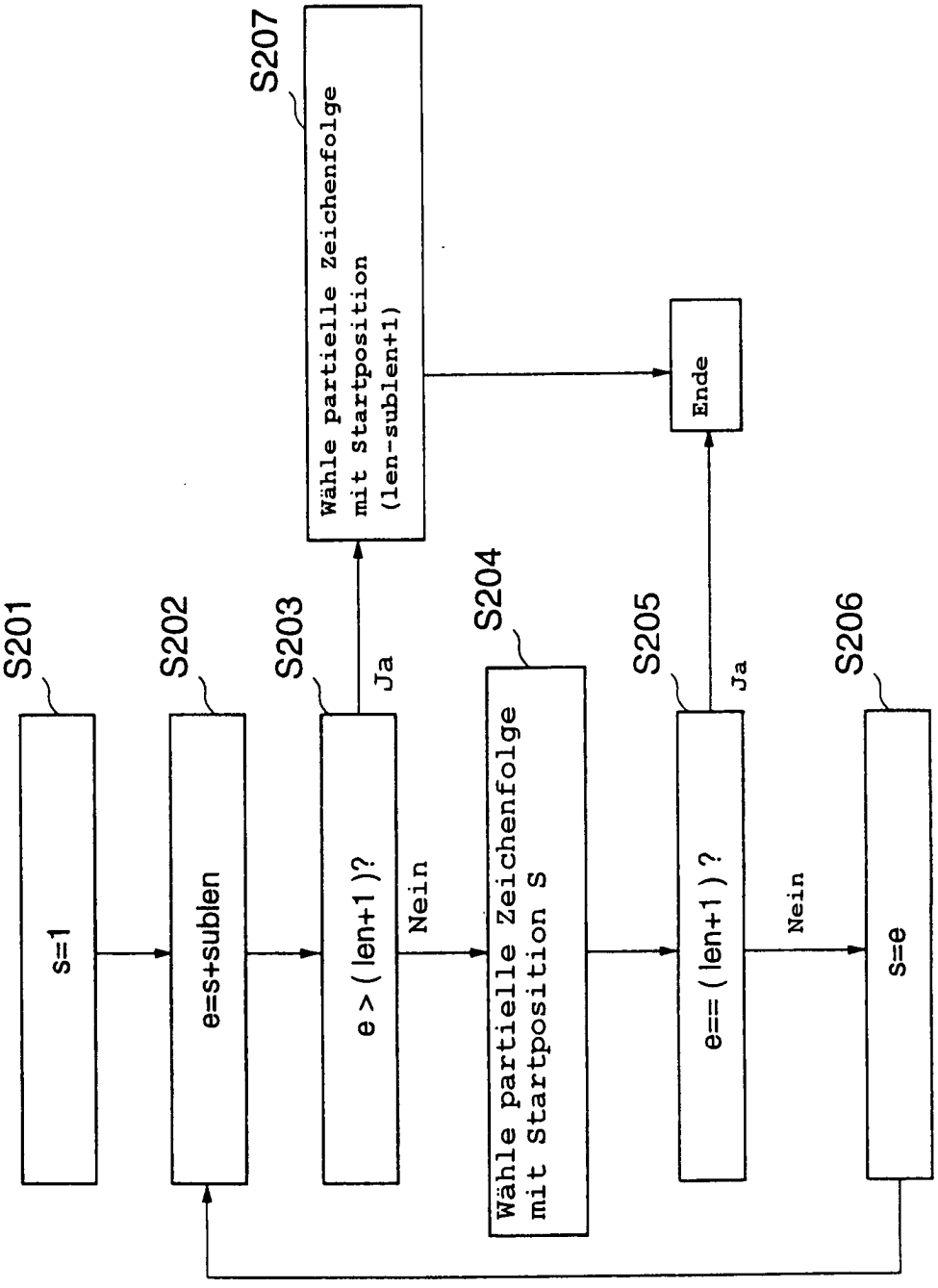

1A
Suchzeichen-folge
gespeichertes Dokument
2
Text-Unter-teilungseinheit
6
Auswahleinheit für partielle Zeichen-folge
Index-Einheit
5
4
3
Rangberechnungs-Einheit
Dokument-Auswahleinheit
Beschaffungs-Ergebnisse

FIG. 14


S201
s=1
S202
e=s+sublen
S203
e > (len+1)?
Ja
Nein
S204
Wähle partielle Zeichenfolge mit Startposition S
S205
e== (len+1) ?
Ja
Nein
S206
S=e
S207
Wähle partielle Zeichenfolge mit Startposition (len-sublen+1)
Ende