SIEMENS AKTIENGESELLSCHAFT

Methode zum Auffinden von Dokumenten

pdffulltext Die Erfindung betrifft das Auffinden von Dokumenten in einem Bestand, in dem die Dokumente Verweise auf andere Dokumente enthalten. Das als World Wide Web (WWW) bekannte System umfaßt eine große Zahl von Dokumenten, die Verweise auf andere Dokumente enthalten, welche wiederum Verweise auf weitere Dokumente usw. enthalten können. Dokumente, die solche Verweise hinter Text- oder Bildobjekten verbergen, werden auch als Hypertext und die Verweise selbst als Hyperlinks bezeichnet. Üblicherweise werden die Hypertext-Dokumente des WWW in der Markierungssprache HTML codiert. Um ein Dokument in diesem größten existierenen Bestand gleich formatierter Dokumente zu finden, sind seit einiger Zeit Suchmaschinen bekannt. Diese verproben in regelmäßigen Abständen die Dokumente und folgen den Hyperlinks. Dabei werden die Dokumente in einen Index eingetragen, der entweder aus den in der HTML angegebenen Indexbegriffen oder aus dem Text extrahierten Wörtern besteht. Ein Benutzer des WWW, der ein Dokument sucht, veranlaßt eine Absuchung eines solchen Index mit von ihm vorgegebenen Suchwörtern.War diese Methode in der Anfangszeit des WWW noch gut brauchbar, so wird die Ergebnismenge nur dann handhabbar klein, wenn sehr spezielle Such- und Stichwörter verwendet werden können. Gerade unerfahrene Benutzer erhalten häufig entweder zu kleine oder zu große Ergebnismengen. Demgemäß werden, ausgehend von den Such- und Stichwörtern, die Dokumente nach Relevanz geordnet angezeigt, wobei die Relevanz auch kommerzielle Bevorzugungenen enthalten kann. Für die Bildung der Relevanz werden in der Regel die Häufigkeiten von Wörtern herangezogen, wie es bereits 1958 in dem Artikel 'The Automatic Creation of Literatur Abstracts' von H.P.Luhn, IBM Journal, P.159-165 vorgeschlagen wurde. Dennoch besteht weiterhin Bedarf an einer verbesserten, auch ungeübten Benutzern zugängliche Methode. In der Patentschrift US 6167398 wird dazu vorgeschlagen, ausgehend von einem Referenzdokument zu jedem untersuchten Dokument einen Abstand mittels einer Abstandmetrik zu berechnen und dann, nach dem Durchsuchen einer vorgegebenen oder anderswie begrenzten Zahl von Dokumenten, diese mit Hilfe der bestimmten Abstände in eine Reihenfolge zu bringen. Dabei sollen mehrere unterschiedliche Abstandsmetriken gemeinsam benutzt werden. Nachteilig an dieser Lösung ist es, daß zunächst eine Menge von Dokumenten bereitgestellt wird und dann jedes der Dokumente bewertet wird. Es ist also weiterhin notwendig, mit einer Stichwort-Suchfrage zunächst eine Teilmenge der Dokumente zu bestimmen. In der Patentschrift US 6,144,973 wird vorgeschlagen, bei der Suche nach Dokumenten im WWW die Verweise in einem Dokument danach zu bewerten, ob ein vorgegebener Grad an Ähnlichkeit mit dem ursprünglichen Dokument gegeben ist. Die Verweise werden entweder verwendet, wenn eine vorgegebene Schwelle überschritten ist; oder sie werden verworfen, wenn die Schwelle unterschritten ist. Parallelarbeit oder eine Anpassung an schon gefundene Dokumente ist nicht vorgesehen; das primäre Mittel zur Begrenzung der Zahl der abgerufenen Dokumente besteht darin, die Suchtiefe zu beschränken. Die Erfindung hingegen beruht auf der Erkenntnis, daß der ermittelte Grad der Ähnlichkeit zweckmäßig dazu verwendet werden kann, die weitere Suche zu steuern und die zu durchsuchenden Verweise zu ordnen. Dazu trägt die Verwendung besserer

Ähnlichkeitsmaße wie dem Vektorraummodell bei. Es handelt sich also um eine Methode zum Durchsuchen einer Dokumentenbasis, in der Dokumente durch Verweise verknüpft sind. Eine Liste von zu bearbeitenden Dokumenten ist nach Priorität sortiert. Das zu dem Eintrag der höchsten Priorität gehörige Dokument wird abgerufen, und von diesem Dokument wird der Abstand zu einer Dokumentenbasis bestimmt. Alle Verweise aus dem Dokument werden in die Liste der zu bearbeitenden Dokumente eingetragen, wobei als Priorität der Abstand des Dokuments zu Dokumentenbasis verwendet wird. In Fig. 1 ist in einem Diagramm der prinzipielle Ablauf gezeigt.Es werden zwei gewichtete Warteschlangen, die Quellschlange SQ und die Ziellschlange TQ, verwendet. Diese werden in üblicher Technik, insbesondere durch Methoden in objektorientierter Programmierung, bereitgestellt. Im folgenden wird angenommen, daß das Gewicht eine Zahl zwischen 0 und 1 ist. Die Quellschlange SQ umfaßt pro Eintrag mindestens ein Feld für das Gewicht, also eine Zahl zwischen 0 und 1, sowie einen Verweis auf das zu berücksichtigende Dokument, vorzugsweise als 'uniform resource locator' (URL, Verweis auf ein Dokument im im WWW). Die Einträge in der Quellschlange sind stets so sortiert, daß in Richtung des Pfeile das Gewicht zunimmt und neue Einträge stets gemäß ihrem Gewicht einsortiert werden. Die Zielschlange TQ ist ähnlich aufgebaut; auch sie enthält pro Eintrag ein Gewicht und einen Verweis auf ein Dokument, das hier als in einem Dokumentenspeicher DS befindlich dargestellt wird, da sich die Verweise stets auf abgerufene Dokumente beziehen. In dieser Zielschlange entsteht das Ergebnis der Methode gemäß der Erfindung. Die Methode geht von einem Anfangsdokument aus, welches das aktuelle Dokument CD wird. Ferner ist eine Vergleichsbasis RD von einem oder mehreren Dokumenten vorhanden. In dem ersten Schritt wird, mit 1a und 1b bezeichnet, sowohl das aktuelle Dokument CD als auch das bzw. die Referenzdokumente RD einem Vergleicher C zugeführt, der beispielsweise nach der Vektorraum-Methode einen Abstand des aktuellen Dokuments CD zu der Referenz RD bestimmt. Beispielsweise durch Bildung des Kehrwerts wird daraus ein Gewicht als Zahl zwischen 0 und 1 erzeugt, wobei ein großer Abstand zu einem kleinen Gewicht und umgekehrt führt.In dem Schritt 2 wird das Gewicht für den Schritt 4 bereitgestellt.Im Schritt 3 werden die Verweise aus dem aktuellen Dokument CD extrahiert und in einer Verweisliste LL gesammelt.Im Schritt 4 wird die Verweiseliste LL mit dem in Schritt 2 bereitgestellten Gewicht in die Quellschlange SQ übertragen. Alle in dem aktuellen Dokument enthaltenen Verweise werden also mit dem Gewicht des Dokuments, in dem sie enthalten sind, in die Quellschlange eingetragen.In dem nächsten Schritt wird das aktuellel Dokument in die Zielliste TQ eingetragen, wobei das ermittelte Gewicht 5a und das aktuelle Dokument bzw. ein Verweis 5b darauf in die Zielliste eingetragen werden. Das aktuelle Dokument selbst wird bevorzugt in einen Dokumentenbestand DS abgelegt.Als Schritt 6 ist dargestellt, daß der Verweis des höchsten Gewichts der Quellschlange SQ von einem Agenten AG entnommen wird und aus dem WWW abgerufen wird, was in Fig. 1 als Schritt 7 dargestellt ist. Ergebnis ist, als Schritt 8 dargestellt, ein Dokument, welches nunmehr das aktuelle Dokument CD wird und so iterativ die Methode angewendet wird. Eine bevorzugt eingesetzte Verbesserung verwendet mehrere Agenten anstelle des in Fig. 1 dargestellten Agenten AG. Denn der Abruf eines Dokuments aus dem WWW kann erhebliche Zeit in Anspruch nehmen. Anstelle der einfachen Übertragung von Schritt 8 tritt dann eine (nicht gezeigte) Pufferschlagen BQ, in der die abgerufenen Dokumente mit dem Gewicht eingeordnet werden, das die zugehörigen Verweise in

der Quellschlange SQ hatten. Ist das jeweils aktuelle Dokument CD bewertet und abgelegt, so wird der Eintrag des dann höchsten Gewichts aus der Pufferschlange BQ als aktuellen Dokument betrachtet. In diesem Fall werden die Dokumente bevorzugt sogleich in den Dokumentenbestand DS eingetragen und in der Pufferschlange BQ nur noch die Verweise geführt.Eine Parallelarbeit mit mehreren aktuellen Dokumenten CD ist möglich, insbesondere bei Maschinen mit mehreren Prozessoren. Um Überläufe der Warteschlangen zu vermeiden, sind mehrere, dem Fachmann im Prinzip bekannte Maßnahmen möglich.Die Pufferschlange BQ kann einfach mit fester Maximallänge gegeben sein. Ein Agent kann immer nur dann tätig werden, wenn ein Platz in der Pufferschlange BQ frei (geworden) ist. Bevorzugt wird die Anzahl der Agenten dynamisch so eingestellt, daß die Pufferschlange BQ stets teilgefüllt ist. Für die Quellschlange SQ besteht eine Möglichkeit gleichfalls darin, auch deren Maximallänge vorzugeben. Ein neuer Eintrag bei voller Schlange wird entweder verworfen, wenn das Gewicht des neuen Eintrags kleiner ist als das des Eintrags mit dem geringsten Gewicht. Andernfalls wird letzterer verworfen und der neue Eintrag einsortiert.Das gleiche Verfahren ist auch für die Zielschlange möglich. Alternativ oder gleichzeitig kann auch direkt nach der Bestimmung des Gewichts des aktuellen Dokuments entschieden werden, daß dessen Eintrag in die Zielschlange sowie seiner Verweise in die Quellschlange unterbleibt, wenn das Gewicht unter einer vorgegebenen Schranke liegt. Bislang wird bei einem sehr großen Bestand wie dem WWW das Verfahren, wenn überhaupt, nur nach sehr langer Zeit zum Stillstand kommen. Einerseits kann die Zielschlange einem Benutzer regelmäßig zur Begutachtung dargestellt werden, so daß dieser das Verfahren abbrechen kann, wenn das Ergebnis seiner Ansicht nach ausreichend ist.Eine andere Möglichkeit besteht darin, einen Mittelwert der Gewichte der in der Zielschlange gespeicherten Dokumente zu brechnen und das Verfahren abzubrechen, wenn dieser Mittelwert sich nach dem Hinzufügen einer vorgegebenen Anzahl von Dokumenten nicht mehr erhöht hat. Hat die Zielschlange TQ eine vorgegebene Maximallänge erreicht und werden, wie oben beschrieben, Dokumente geringen Gewichts verworfen, dann nimmt dieser Mittelwert ohnehin nur zu, so daß eine Stagnation als Abbruchkriterium dienen kann.Sicherlich besteht auch die Möglichkeit, eine vorgegebenen Schranke, wie oben dargestellt, für den Eintrag in die Quellschlange SQ zu verwenden. Dies wird dazu führen, daß die Quellschlange irgendwann leer ist und damit das Verfahren ohnehin beendet ist. Da in der Dokumentenbasis des WWW zyklische Verweise häufig sind, wird bevorzugt eine Liste der bereits behandelten Verweise, in der Regel als Hash-Tabelle, geführt und ein Verweis aus einem Dokument bereits vor seinem Eintrag in die Verweisliste verworfen. Alternativ kann dies von den Agenten oder einem dafür vorgesehenen Prozeß übernommen werden. Als Abstandsmaß wird bevorzugt eines nach dem Vektorraummodell verwendet. Dieses ist z.B. in "Introduction to Modern Information Retrieval" von Gerald Salton, McGraw Hill 1983, S.121-122, beschrieben. Hierbei wird zunächst eine Tabelle der Wörter aus den zu vergleichenden Dokumenten und ihrer Häufigkeit aufgestellt. Aus der Tabelle werden, in der Regel bereits bei deren Erstellung über sogenannte Stoppwort-Listen, die häufigen Wörter geringer Aussagekraft wie Artikel und Konjunktionen gestrichen. Weitere Maßnahmen sind der einschlägigen Literatur zu entnehmen. Die Zahlen der Häufigkeiten bilden für jedes Dokument einen n-dimensionalen Vektor, wobei n die Anzahl der berücksichtigten Wörter ist. Als Abstand zweier Dokumente wird nunmehr ein Skalarprodukt der beiden Vektoren verwendet. Wörter, die in nur einem Dokument

vorkommen, tragen hierzu sinnvollerweise nichts bei und können auch vorab eliminiert werden. Als Skalarprodukt findet bevorzugt das "Cosinus-Maß" Verwendung, wie es in der o.a. Literatur beschrieben ist. Eine Übersicht hierüber findet sich auch in der Diplomarbeit "Visualisierung latent semantischer Hypertext-Strukturen" von Hardy Höfer, Univ. Paderborn, Dezember 1999, im Kapitel 4.3. Die Erfindung wurde an Hand des WWW als Dokumentenbestand beschrieben, bei dem die Dokumente als HTML-Dokumente vorliegen, in denen die Verweise enthalten sind. Die Anwendung auf andere Dokumentenbestände ist ohne weiteres möglich, solange diese im Volltext vorliegen und miteinander verkettet sind. Diese Verkettung kann auch durch nicht im Dokument enthaltene Indices erfolgen; ob die Verweise im Dokument selbst codiert oder in parallel geführten Indices enthalten sind, spielt ersichtlich keine Rolle, solange die Adressierung des Dokuments im Index und umgekehrt eindeutig ist. Falls die Dokumente zwar nicht im Volltext vorliegen, aber einem der bekannten Klarschrift-Leseverfahren zugänglich sind, ist die Anwendung der Erfindung mehr eine Frage der Effizienz als des Prinzips, da die Dokumente automatisch dem Klarschrift-Leser zugeführt und die so erhaltenen Texte verwendet werden können. Dies trifft im übrigen im besonderen auch für Patentschriften zu, bei denen die Verweise auf andere Patentschriften leicht automatisch auffindbar sind, nachdem das Dokument von dem Klarschriftleser in Volltext gewandelt wurde. Abgesehen davon sind die Zitate der Patentschriften untereinander vollständig erfaßt und demgemäß ein Beispiel für den oben genannten externen Index.

简体中文网页