Eine der wichtigsten Entscheidung bei der Erstellung eines Indexes ist die Definition der Indexierungsstrategie und damit verbunden die Bestimmung geeigneter Gewichtungsmodelle (s. Kap. 4). D.h. ob und wie Dokumente sowie deren Deskriptoren gewichtet und mittels welcher Retrieval-Funktion geeignete Dokumente identifiziert und bewertet werden können.

Indexe stellen den Zugang zur Datenbasis für die Anfragen via Querry Processor, der Suchkomponente einer Suchmaschine, dar. Entsprechend der definierten Suchstrategien (z.B. einfache Keywordsuche, Suche mittels Boolescher Operatoren, Volltextsuche, Expertensuche, etc.) und deren Retrieval-Funktionen müssen die Indexe so konzipiert sein, dass sie hierzu über alle erforderlichen Daten und Gewichtungsinformationen verfügen und diese effizient organisieren.

Im nachfolgenden Kapitel beschäftigen wir uns ausgiebig mit den verschiedenen Gewichtungsmodellen, auch Indexierungsmodelle genannt, die bei den Suchmaschinen zum Einsatz kommen.

Links

Index-sequential files
- [www.dcs.gla.ac.uk/~iain/keith/data/pages/72.htm]

B-Trees
- [www.dcs.gla.ac.uk/~iain/keith/data/pages/83.htm]

Sortierung der Indexdatei
- [www.linguistik.uni-erlangen.de/tree/html/corsica/zierl97/node67.html]

File Structures
- [www.dcs.gla.ac.uk/Keith/Chapter.4/Ch.4.html]

4 Relevanz und Gewichtungsmodelle

Das wesentlichste Unterscheidungsmerkmal von Information Retrieval Systemen im Vergleich zu klassischen Tabellen orientierten Datenbanksystemen ist deren Funktionalität, Suchergebnisse entsprechend ihrer *Relevanz* zu einer Suche differenzieren zu können. Relevanz kann auch im Sinne von *Ähnlichkeit* gedeutet werden. Information Retrieval Systeme sind so konzipiert, dass es möglich ist, aus der Gesamtheit aller im Datenbestand vorhandenen Dokumente, genau diejenigen Textdokumente zu finden, die zu einer Suchanfrage ein Mindestmaß an Ähnlichkeit besitzen. Dies erfordert von den Suchmaschinen relevante Dokumente von irrelevanten Dokumenten unterscheiden und relevante Dokumente hinlänglich ihres Ähnlichkeitsgrads zur Suchanfrage sortieren zu können.

Um eine Unterscheidung vornehmen zu können, welche Dokumente eines Datenbestandes inhaltlich über einen Bezug zu einer Suchanfrage verfügen und zu welchem Grad eine Ähnlichkeit besteht, müssen Gewichtungsmodelle eingesetzt werden, die Dokumente hinlänglich ihrer Relevanz zu einer Suchanfrage unterscheiden können. Nachfolgend werden die wichtigsten Gewichtungsmodelle dargestellt, die von Suchmaschinen im Internet verwendet werden. Sie lassen sich grob in *Vektorraum basierte Gewichtungsmodelle* und *Hypermedia basierte Gewichtungsmodelle* unterteilen. Da die Bestimmung von Relevanz auf mathematischen Modellen zur Berechnung der Ähnlichkeit basiert, muss bei der Beschreibung der einzelnen Modelle hierauf Bezug genommen werden. Die Erklärung der einzelnen mathematischen Modelle ist jedoch allgemein gut verständlich.

Was unter dem Begriff *Relevanz* zu verstehen ist und warum er ein wichtiges Merkmal bei der Verarbeitung und Suche von Dokumenten in einem Datenbestand darstellt, wird nachfolgend ausführlich erklärt. Ein grundlegendes Verständnis über die Bedeutung der Relevanz und damit verbunden, eine Kenntnis über die einzelnen Gewichtungsmodelle ist Voraussetzung, um eine Website für Suchmaschinen optimieren zu können.

4.1 Zusammenhang von Relevanz und Rangbildung

Betrachten wir bekannte Datenbanksysteme wie z.B. SQL basierte Datenbanken, so basieren sie auf einem binären Entscheidungskriterium, das nur dann Datensätze als Ergebnis berücksichtigt, wenn sie einer Suchanfrage zu 100 Prozent entsprechen. Faktisch erfolgt dies durch einen Vergleich des Suchstrings mit dem

Inhalt aller Felder einer oder mehrerer Kolumnen. Die Suche in einer Adressdatei nach dem Namen *Meier* kann nur dann zu einem Ergebnis führen, wenn mindestens ein Datensatz in der Datenbank existiert, der die Zeichenfolge *Meier* unter der gestellten Suchbedingung in dem betreffenden Feld exakt erfüllt.

Bei Information Retrieval Systemen ist dies anders. Dokumente werden auch dann Teil eines Suchergebnisses, wenn sie eine Suchanfrage nur teilweise erfüllen; d.h. sie werden auch dann als Ergebnis angezeigt, wenn der berechnete Relevanzgrad nicht 100 Prozent sondern beispielsweise nur 95 Prozent, 75 Prozent oder auch weniger aufweist. Die Erfordernis Daten auch bei einem geringeren Relevanzgrad als 100 Prozent als geeignetes Suchergebnis zu berücksichtigen, liegt in der Besonderheit von Textdokumenten sowie der Systematik begründet, wie Inhalte von Dokumenten über Keywords erschlossen werden.

Analysiert man verschiedene Texte die von unterschiedlichen Verfassern zu einem vorgegebenen Thema erstellt wurden wird deutlich, dass die einzelnen Texte je nach Fachwissen, Schreibstil, persönlicher Zielsetzung, anvisiertem Publikum sowie aufgewendeter Zeit und Motivation inhaltlich sehr unterschiedlich ausfallen können. In der Terminologie des Information Retrieval kann auch gesagt werden, die einzelnen Texte sind hinlänglich eines eindeutig vorgegebenen Themas *inhaltlich* abweichend und bezogen auf das Thema unterschiedlich *relevant*. Die Ähnlichkeit der einzelnen Dokumente zueinander als auch zum vorgegebenen Thema ist verschieden.

Überträgt man die Erkenntnis der *unterschiedlichen Relevanz* von Dokumenten, die ein und das selbe Thema behandeln auf das Information Retrieval wird deutlich, dass die Suchmaschinen über automatisierte Verfahren verfügen müssen, die dieser Realität entsprechend gerecht werden und unter dieser Vorbedingung dennoch präzise Suchergebnisse liefern. Das *Thema* zu dem ein Anwender Dokumente sucht bestimmt er durch seine Suchworte. Das Ergebnis bilden all diejenigen Dokumente, die aufgrund der einzelnen Verfahren der Retrieval Systeme und der eingegebenen Suchworte als *ähnlich* zur Suchanfrage gelten.

Zur Realisation von Suchmethodiken, die die Ähnlichkeit von Textdokumenten zu einer Suchanfrage berücksichtigen, sind spezielle Datenstrukturen erforderlich. Weiter müssen Methoden angewendet werden, die aufgrund der Inhalte von Dokumenten automatisiert eine Bestimmung vornehmen können, welche Dokumente zu einem bestimmten Thema relevant sind und in welcher Intensität diese Ähnlichkeit ist.

Die zur Realisation erforderlichen Datenstrukturen sind das invertierte Dateisystem, das bereits im vorherigen Kapitel dargestellt wurde. Zur Umsetzung der Bestimmung des Grades der Ähnlichkeit von Dokumenten sind Gewichtungsmodelle notwendig, die aufgrund von verschiedenen Parametern bestimmen, zu welchen Themen ein Dokument relevant ist und wie stark diese Relevanz hinlänglich jeden Themas ist.

Betrachtet man alle Elemente des Hypermedia so können die für Gewichtungsverfahren erforderlichen Parameter aus einem Dokument selbst, als auch auf Basis der Systematik des Hypertext und HTTP-Protokolls gewonnen werden.

In den nachfolgenden Abschnitten wird auf die wichtigsten Gewichtungsmodelle eingegangen, die eine Bewertung der Ähnlichkeit von Dokumenten zu Suchanfragen ermöglichen. Es ist zu beachten, dass die verschiedenen Gewichtungsmodelle jedoch nur die Parameter festlegen, mittels derer Werte eine Berechnung der Relevanz erfolgen kann. Die konkrete mathematische Berechnung eines Ähnlichkeitsgrads erfolgt über eine Retrievalfunktion.

Eine Retrievalfunktion ist ein Algorithmus der auf Basis verschiedener Gewichtungsmodelle den Relevanzgrad eines Dokuments, bezogen auf eine Suchanfrage, berechnet. Zwei Punkte sind hierbei zu beachten.

Eine Retrievalfunktion kann einen oder mehrere Parameter verschiedener Gewichtungsmodelle berücksichtigen. Weiter können die verschiedenen Gewichtungswerte der einzelnen Modelle in ihrer Wirkungsstärke von der Retrievalfunktion unterschiedlich stark bei ihrer Berechnung berücksichtigt werden. In der praktischen Anwendung kann dies beispielsweise bedeuten, dass eine Retrievalfunktion die als Gewichtungsmodelle *Term Frequency* und *Inverse Term Frequency* zur Relevanzberechnung einsetzt, die Intensität der Wirkung die sie den Werten der einzelnen Gewichtungsmodelle zuweist, unterschiedlich stark berücksichtigt. So kann beispielsweise der Algorithmus einer Suchmaschine der Term Frequency einen Anteil von 80 Prozent und der Inverse Term Frequency einen Anteil von 20 Prozent bei der Berechnung der Relevanz zuordnen. In Abhängigkeit der Verteilung der verschiedenen Gewichtungsmodelle und Bewertungskriterien von Schlüsselwörtern ergibt sich dann eine unterschiedliche Relevanz der Dokumente zur Suchanfrage.

Während die Bestimmung der einzusetzenden Gewichtungsmodelle eine Grundsatzentscheidung für eine Suchmaschine ist, die dazu führt, dass die für die jeweiligen Gewichtungsmodelle erforderlichen Parameter bei der Dokumentenanalyse erfasst und im invertierten Dateisystem entsprechend berücksichtigt werden, stellt die Verteilung des Wirkungsgrades der einzelnen Gewichtungsmodelle eine flexible Einstellungsoption dar, die zur Feineinstellung der *Precision* dient. Sie hat insofern direkte Auswirkung auf die Präzision von Suchanfragen und dient, basierend auf den eingesetzten Gewichtungsmodellen, zur Verbesserung der Qualität von Suchergebnissen.

Auch wenn grundsätzlich alle Gewichtungsmodelle des Information Retrieval bekannt sind, kann keine wirklich eindeutige rekursive Ermittlung des Algorithmus einer Suchmaschine erfolgen, da die Kombinationsformen, die individuelle Berechnung der Stärke einzelner Parameter als auch die Verteilung der Wirkungsstärke der Modelle zueinander nicht exakt nachvollzogen werden kann. Die rekursive Analyse eines Suchmaschinen-Algorithmus kann folglich immer nur ein approximatives Ergebnis zu den einzelnen Parametern, deren Zusammensetzung und Intensität liefern.

Das Ergebnis einer Suchanfrage von Suchmaschinen die gewichtete Verfahren einsetzen, ist eine Ergebnisliste die nach bestimmten Kriterien sortiert wird. Während noch vor wenigen Jahren die Suchergebnislisten zum Teil optional nach Al-

phabet sortiert werden konnten, hat sich eine Sortierung nach Relevanz, d.h. nach Ähnlichkeitsgrad der gefundenen Dokumente zur Suchanfrage durchgesetzt. Die *Rangbildung*, auch *Ranking* genannt, orientiert sich bei den Suchmaschinen heute nach dem Grad der Ähnlichkeit. Je eher ein Dokument einer Suchanfrage entspricht, desto weiter oben auf der Ergebnisliste erscheint es. Das Dokument das auf der ersten Seite und auf der ersten Position einer Suchergebnisliste angezeigt wird, entspricht der gestellten Suchanfrage, gemäß der Berechnungsmethodik der Suchmaschine, am genauesten. Die Rangposition entspricht folglich dem Ähnlichkeitsgrad eines Dokuments zur Suche.

Links

Retrieval Evaluation
- [www.dcs.gla.ac.uk/~iain/keith/data/pages/144.htm]

Technology To Ensure Most Relevant Results of Any Search Engine Today
- [www.northernlight.com/docs/press_company_pr99_1025.html]

Search Strategies
- [www.dcs.gla.ac.uk/Keith/Chapter.5/Ch.5.html]

4.2 Effektivität von Suchmaschinen

Die Qualität einer Suchmaschine ist sowohl für den suchenden Anwender als auch für einen Content-Anbieter gleichermaßen bedeutsam. Eine qualitative Systembewertung von Retrieval-Systemen kann anhand der *Systemeffizienz* und der *Systemeffektivität* erfolgen.

Mit der Systemeffizienz werden die Kosten und die Zeit gemessen, die zur Ausführung bestimmter Systemoperationen erforderlich sind. Faktoren die die Systemeffizienz beeinflussen sind u.a. die Art der Datenstrukturen, Organisation der Speichermedien und Methoden der Query-Abfrage. In Hinblick auf die Intention dieses Buches, Methodiken zur Website-Optimierung zu identifizieren, ist eine Kenntnis über die Systemeffektivität und deren Kennzahlen sowie Einflussfaktoren wichtiger, als eine Analyse der Systemeffizienz bestimmenden Parameter. Aus diesem Grund soll die Systemeffizienz nicht weiter vertieft werden.

Unter Systemeffektivität versteht man im Allgemeinen die Fähigkeit eines Information Retrieval Systems, einem Nutzer genau die Informationen nachzuweisen die er sucht. Im Hinblick auf Systemeffektivität sind zwei Maßgrößen relevant um die Leistung der Retrieval Systeme von Suchmaschinen zu definieren. Die Maßgröße *Recall* beschreibt die Fähigkeit eines Retrievalsystems *alle* für eine bestimmte Suchanfrage relevanten Dokumente nachzuweisen. Die Precision beurteilt hingegen die *Exaktheit* mit der relevante Dokumente nachgewiesen werden.

Welche Maßzahl konkret für Nutzer von Retrieval-Systemen relevanter ist, lässt sich nicht allgemein beantworten. Die Praxis zeigt, dass von Anwender zu Anwender unterschiedliche Informationsbedürfnisse bestehen. So existieren Nutzer die einen hohen Recall wünschen, also ein Ergebnis bevorzugen, das möglichst viele Dokumente generiert. Auf der anderen Seite existieren Nutzer, die möglichst alles Irrelevante zu vermeiden suchen und insofern eine möglichst hohe *Precision* bevorzugen.

Trennt man die Menge der *nachgewiesenen* Dokumente von den *nicht nachgewiesenen* Dokumenten einer Datenbank und trennt man die *relevanten* Dokumente von den *irrelevanten* Dokumenten, so kann der Recall R und die Precision P wie folgt definiert werden:

Recall R

$$\frac{\text{Anzahl der nachgewiesenen relevanten Dokumente}}{\text{Anzahl aller relevanten Dokumente in der Datenbank}} \quad (4.1)$$

Precision P

$$\frac{\text{Anzahl der nachgewiesenen relevanten Dokumente}}{\text{Anzahl aller nachgewiesenen Dokumente}} \quad (4.2)$$

Die Werte für Recall und Precision liegen jeweils zwischen 0 und 1; je näher an 1, desto besser.

- **Recall = 1** bedeutet, dass alle relevanten Dokumente im Datenbestand gefunden wurden.
- **Precision = 1** bedeutet, dass alle gefundenen Dokumente auch relevant sind.

Der Recall bezieht sich folglich auf die Fähigkeit eines Systems verwertbare Dokumente nachzuweisen, während die Precision hingegen die Fähigkeit eines Systems misst, ungenaue Dokumente vom Suchergebnis auszuschließen.

Eine hochgradig umfangreiche Indexierung, d.h. es werden möglichst viele Begriffe aus dem Text eines Dokuments indexiert, führt zu einem Nachweis vieler potentiell relevanter Dokumente. Gleichzeitig leidet aber die Precision, da auch nur wenig bedingte Dokumente Teil des Suchergebnisses werden. Werden hingegen überwiegend hochspezifische Deskriptoren zur Dokumentenrepräsentanz berücksichtigt, d.h. mittels bestimmter Verfahren die repräsentativsten Wörter identifiziert, steigt zwar die Precision der gefunden Dokumente, es sinkt jedoch gleichzeitig auch die Anzahl der möglich relevanten Dokumente. Diese Überlegungen haben direkte Auswirkung auf die Einstellungen des Keyword-Relevanzfilters (s. Kap. 3.2.7) sowie die Höhe der Gewichtung von Schlüsselwörtern.

Einer messbaren Effektivität eines Retrievalsystems mittels oben dargestellter Maßzahlen, steht jedoch immer auch die *subjektive* Erwartung bzw. Bewertung eines Nutzers gegenüber, inwieweit ein Retrievalsystem in der Lage ist, seine Anfrage zufriedenstellend zu beantworten. Ein großes Hindernis stellt dabei die Problematik dar, dass verschiedene Nutzer das Gleiche mittels unterschiedlicher Suchworte und Suchmethoden bzw. Unterschiedliches mittels gleicher Wortwahl suchen.

Natürliche Sprachen ermöglichen aufgrund von Synonymen die Verwendung verschiedener Wörter für ein und den selben Begriff. Dieses Problem versuchen z.B. klassische Retrieval Systeme mittels Thesaurusregeln zu lösen. Suchmaschinen im Internet besitzen für diese Problematik hingegen keine Lösung, denn Suchanfragen mit Synonymen werden aufgrund der Keyword orientierten Suchbeantwortung mit einem anderen Ergebnis beantwortet.

Ein anderes Problem stellt die Mehrdeutigkeit von Worten dar. So kann beispielsweise mit dem Wort „Golf" entweder eine Sportart, ein bestimmtes Fahrzeugmodell oder auch eine geographische Definition für eine Meeresküste gemeint sein. Das Problem der Wortmehrdeutigkeit versuchen Information Retrieval Systeme zu lösen, indem Dokumente nicht nur ausschließlich auf Basis einzelner Deskriptoren indexiert werden. Sondern es werden zusätzlich Dokumente mittels bestimmter Verfahren, wie z.B. dem Cluster-Verfahren, inhaltlich analysiert und dann einem bestimmten Themen-Cluster zugeordnet.

Relevanz von Informationen ist aber auch immer subjektiv in Abhängigkeit des eigenen Wissens und muss in Bezug auf die Information beurteilt werden, die benötigt wird und nicht diejenige, die angefordert wird. Das bedeutet, dass Relevanz bzw. Irrelevanz eines gefundenen Dokuments konkret von dem aktuellen eigenen Wissenstand eines Anwenders zu einem bestimmten Thema abhängt.

Bereits diese wenigen Überlegungen machen klar, dass es zwar Maßzahlen gibt die versuchen die Qualität einer Suchmaschine objektiv zu beschreiben, aber eine Vielzahl von subjektiven Faktoren der Anwender zu berücksichtigen sind. Das bedeutet für den Entwurf einer Website, dass eine Website aufgrund ihrer inhaltlichen Ausgestaltung nicht alle Personen erreichen kann, die Interesse an dem Thema einer Website haben. Eine Website kann deshalb auch nur hinlänglich einer vom Content-Anbieter definierten Zielgruppe technisch optimal entworfen werden, die mittels einer gedanklich antizipierten Suchmethodik und vermuteten Suchworten versucht, Wissenstand und Erwartung der Zielgruppe abzuschätzen.

Links

Information Retrieval-Precision und Recall
- [www.uni-duisburg.de/FB3/CL/ses/docs/ir.html]

Introduction Course Outline Introduction to Information Retrieval
- [http://ir.iit.edu/~dagr/cs529/files/handouts/01Introduction-6per.PDF]

Introduction to Information Retrieval Evaluation
- [www.sims.berkeley.edu/courses/is202/f98/Lecture14/]

Information Retrieval
- [www.issco.unige.ch/ewg95/node214.html]

Introduction in Information Retrieval
- [www.dcs.gla.ac.uk/Keith/Chapter.1/Ch.1.html]

4.3 Statistische Gewichtungsmodelle

Das traditionelle Information Retrieval setzt schon sehr lange statistische Gewichtungsmodelle zur Bestimmung der Relevanz eines Textdokumentes ein. Sie basieren im Allgemeinen auf der Häufigkeit des Vorkommens eines Begriffs oder eines Parameters im Dokument. Da eine inhaltliche Erschließung von Textdokumenten überwiegend auf der Identifikation von repräsentativen Deskriptoren beruht, beziehen sich die statistischen Gewichtungsmodelle zur Diskriminierung von Dokumenten, in vielen Fällen auf der Häufigkeit des Vorkommens eines Begriffs im Textdokument. Neben bibliothekarischen Information Retrieval Systemen setzten die bekannten Suchmaschinen neben anderen Bewertungsverfahren statistische Gewichtungsmodelle zur Relevanzbewertung ein. Auch wenn beispielsweise Google dem PageRank-Verfahren erhebliche Dominanz bei der Relevanzbewertung zuordnet, kommen dennoch statistische Gewichtungsmodelle ergänzend zum Einsatz.

4.3.1 Das Vektorraummodell

Gegenwärtig basieren verschiedene Retrieval-Algorithmen der Suchmaschinen auf dem *Vektorraummodell*. Es ist ein einfaches und benutzerfreundliches Modell, das unmittelbar auf neue Datenbestände angewendet werden kann und in Abhängigkeit der gewählten Retrievalfunktion eine relativ gute Retrievalqualität bietet.

Beim Vektorraummodell wird jedes Dokument durch einen Vektor von *n*-Deskriptoren repräsentiert. D.h. für jedes Dokument existiert *ein* Vektor, dessen Vektorenraum durch *n*-Schlüsselwörter des betreffenden Dokuments gebildet wird. Konkret bedeutet das, dass jedes gefundene Schlüsselwort eines Dokumentes im Vektor *eine* Dimension bildet und der Vektor eines Dokuments somit n-dimensional ist. Werden beispielsweise zwanzig Keywords für ein Dokument bestimmt, besitzt der Vektor des betreffenden Dokuments zwanzig Dimensionen (n = 20).

Eine Suchanfrage wird ihrerseits als m-dimensionaler Vektor dargestellt. In Analogie zum Vektor eines Dokuments bestimmt sich der Vektorraum einer Suchanfrage aus der Anzahl der Suchworte. Besteht eine Suche aus vier Suchworten, hat der Suchvektor vier Dimensionen (m = 4).

Während beim simplen Booleschen-Modell, das überwiegend bei Tabellen orientierten Datenbanksystemen zum Einsatz kommt, lediglich binär überprüft wird, ob ein Begriff in einem Datensatz vorkommt oder nicht, ist es Ziel von vektorbasierten Retrieval-Systemen, über Gewichtungsverfahren Dokumente in Bezug auf ihre Ähnlichkeit zur Suchanfrage zu identifizieren und in eine gewichtete Rangfolge zu bringen. Das Vektorraummodell evaluiert aus diesem Grund den Grad der Ähnlichkeit der Dokumentenvektoren in Bezug auf Suchanfragevektoren als Korrelation zwischen dem Vektor der Suchanfrage und dem Vektor eines Dokuments. Es existieren zwei grundsätzliche Ansätze des Vektorraummodells.

Das binäre Vektorraummodell

Im *binären Vektorraummodell* wird lediglich die Existenz eines Begriffs im Dokument binär mit *1* dargestellt sofern der Begriff vorkommt bzw. mit *0* wenn er nicht vorkommt. Diese binäre Abbildung des reinen Auftretens eines Keywords in einem Dokument ermöglicht jedoch keine Differenzierung von Dokumenten hinlänglich ihrer *Ähnlichkeit* zueinander bzw. eine Berechnung der Ähnlichkeit hinlänglich einer Suchanfrage.

Das gewichtete Vektorraummodell

Im *gewichteten Vektorraummodell*, das nachfolgend detaillierter besprochen wird, werden Gewichtungsmodelle eingesetzt die einem Begriff einen bestimmten Wert in Hinblick auf seine Relevanz zuordnet.

Folgende Gegenüberstellung macht den Unterschied zwischen gewichteten und ungewichteten Modellen deutlich.

Schlüsselwort	Computer	Prozessor	Netzkarte
Ungewichteter Vektor	{ 1	0	1 }
Gewichteter Vektor	{ 2.3	3.5	1.6 }

Abb. 4.1. Vergleich binärer Vektor – gewichteter Vektor

Der Ansatz des gewichteten Vektorraummodells, nämlich ein Dokument durch seine Keywords in Form eines gewichteten Vektors darzustellen, eröffnet eine einfache Methode, Dokumente und deren Gewichtung physikalisch abzubilden sowie mathematisch zu verarbeiten. Jeder Deskriptor kann als *eine* Dimension im Vektor dargestellt werden. Ein Dokument mit n-Deskriptoren wird somit über einen n-dimensionalen Vektor dargestellt. Im obigen Beispiel verfügt der Vektor über drei Dimensionen, die durch die Schlüsselwörter „Computer", „Prozessor" und „Netzwerkkarte" bestimmt werden. Wie beschrieben, erfolgt eine automatisierte Identifikation von Schlüsselwörtern durch den Einsatz eines Keyword-Relevanzfilter. Alle Keywords die der Filter als relevant für ein Dokument erachtet, werden im invertierten Dateisystem berücksichtigt. Das Dokument wird im System als n-dimensionaler Vektor abgebildet. Die Anzahl der gefundenen Schlüsselwörter bestimmen dabei die Anzahl der Dimensionen eines Dokumentenvektor.

Die *Länge* eines Vektors spiegelt den Wert im gewichteten Vektorraummodell wider, der einem Schlüsselwort zugerechnet wird. Einem Deskriptor kann neben Null ein positiver oder auch ein negativer Wert zugeordnet werden. Betrachten wir obiges Beispiel, bei dem das Dokument mit n = 3 und den Deskriptorengewichten {t1 = 2,3; t2 = 3,5; t3 = 1,6} abgebildet wird, ergibt sich nachfolgend dargestellter dreidimensionaler Vektor.

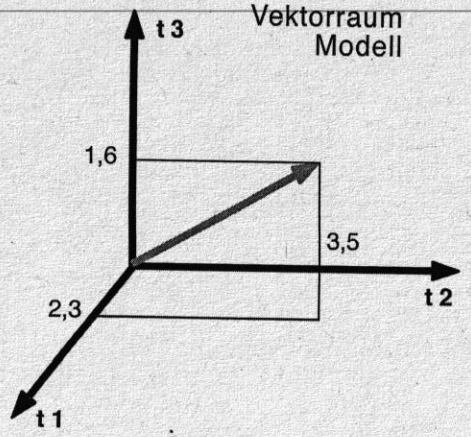

Abb. 4.2. Vektorrepräsentation im Vektorraummodell

Zur Berechnung der Ähnlichkeit von Anfrage und Dokument wird die Anfrage ebenfalls als Vektor mit einem vorbestimmten Wert definiert. Um ein Dokument als relevant auszuweisen, wird nun nicht mehr auf einer völligen Übereinstimmung zwischen Anfrage- und Dokumentenvektor bestanden (wie es im binären Booleschen-Modell erforderlich ist), sondern es wird festgelegt, dass der Nachweis eines Dokuments von dem *Ähnlichkeitswert* zwischen der *Suchanfrage* und dem *Dokument* abhängt. Die Ähnlichkeit wird zwischen einem bestimmten Dokumentenvektor und einem Suchanfragevektor als Funktion, in Abhängigkeit von z.B. der Anzahl der übereinstimmenden Suchbegriffe, bestimmt. Hierzu werden von den Suchmaschinen unterschiedliche Retrieval-Funktionen eingesetzt.

Nahezu alle Suchmaschinen im Internet basieren auf dem gewichteten Vektorraummodell zur Berechnung von Deskriptorengewichten und der Relevanz von Dokumenten. Insbesondere seine Einfachheit und Schnelligkeit beim Retrieval sind für seine weite Verbreitung ursächlich. Das Vektorraummodell macht jedoch keine Vorgaben wie die Dokumentenbeschreibung, Gewichtung und Ähnlichkeitsberechnung zu erfolgen hat. Es bildet sowohl ein Dokument als auch eine Suchanfrage lediglich als mathematischen Wert ab.

Mittlerweile existieren viele Gewichtungsmodelle und Kombinationsformen zur Berechnung der Dokumentenrelevanz, die wiederum auf der Berücksichtigung unterschiedlichster Parameter beruhen. Aus dieser Vielfalt werden nachfolgend die wichtigsten Gewichtungsmodelle beschrieben. Es ist zu beachten, dass die einzelnen Gewichtungsmodelle die Grundlage zur Optimierung von Websites darstellen.

Links

Das Vektorraummodell
- [www.bui.fh-hamburg.de/pers/ulrike.spree/vektor/vektor5.htm]

Das Vektorraummodell als Alternative zum Booleschen Modell
- [www-db.informatik.uni-tuebingen.de/~becker/ir/kap5a-ein.ps]

Das Vektorraummodell
- [www.ifi.unizh.ch/CL/Glossar/Vektorraummodell.html]

Vector Space Model (VSM)
- [www2002.org/CDROM/refereed/643/node5.html]

Information Retrieval Models
- [www.cs.rpi.edu/~sibel/mmdb/lectures/ir_models.pdf]

4.3.2 Die relative Worthäufigkeit (TF-Algorithmus)

Der *Term Frequency Algorithmus (TF)*, auch *Algorithmus der Worthäufigkeit* genannt, beruht auf der Erkenntnis, dass es für den Verfasser beim Erstellen eines Textes grundsätzlich leichter ist, immer den gleichen Begriff für ein und den selben Sachverhalt zu verwenden, als ständig wechselnde Begriffe. Neben dieser Erkenntnis, die auch als *Zipf'sches Gesetz* bzw. als *Gesetz des geringsten Widerstandes* bekannt ist, ist anzumerken, dass für bestimmte Worte schlicht weg keine Synonyme verwendet werden können, da keine existieren.

Wird beispielsweise auf einer Website ein spezieller *Grappa* zum Verkauf angeboten, so gibt es für den *Grappa di Tignanello* aus dem Hause *Antinori* keine Synonyme. Das Produkt kann nur entsprechend seiner Bezeichnung im Text erscheinen und vom Verfasser identisch wiederholt werden.

Aus der Erkenntnis des Zipf'schen Gesetz lässt sich ableiten, dass mit *steigender Häufigkeit* eines Wortes in einem Text seine Bedeutung für den Inhalt an Relevanz zunimmt. Die einfachste Form einen Wert mittels Term Frequency Algorithmus (TF) zu bestimmen, ist die Summe der Häufigkeit eines auftretenden Keywords im Text. Erscheint z.B. ein Wort zwanzig mal im Text wäre entsprechend dieser Kalkulation der TF-Wert = 20.

Diese einfache Berechnungsform führt jedoch dazu, dass bei langen Texten in den ein Begriff nur deshalb häufiger vorkommt weil der Text länger ist, einen höheren Wert zugewiesen bekommt als kürze Dokumente. Zur Vermeidung einer Gewichtung mittels *absoluten Werten* wird die Worthäufigkeit ins Verhältnis zu allen im Dokument vorkommenden Worten gesetzt. Hierdurch wird vermieden, dass ein Dokument nur deshalb bezüglich eines bestimmten Wortes als relevanter bewertet wird als ein anderes, weil die *absolute Worthäufigkeit* höher ist. Viel aussagekräftiger ist folglich die *relative Worthäufigkeit*, da sie eine Bewertung hinlänglich der Wichtigkeit eines bestimmten Wortes zu dem im Text behandelten Thema ermöglicht.

Es lässt sich somit festhalten, dass die relative Worthäufigkeit eines Wortes Auswirkungen auf die Gewichtung eines Dokumentes hat. Der Worttyp der von Suchmaschinen als Keyword bestimmt wird, ist immer ein Substantiv. Nur durch Substantive kann eine Bestimmung über Inhalte und Themen eines Textdokuments erfolgen. Die relative Worthäufigkeit bezieht sich folglich auf die relative Häufigkeit von Substantiven im Text eines Dokuments.

Links

Term Frequency Considerations
- [http://dent.ii.fmph.uniba.sk/~kravcik/IR/AutoIndx/SngTrmIT/TrmFrqCn.html]

Concept of Term Frequency
- [www.dcs.gla.ac.uk/~iain/keith/data/pages/25.htm]

How does the search engine work?
- [www.magportal.com/help/user/search.html]

Information Retrieval and Search
- [www.cs.sfu.ca/~cameron/Teaching/D-Lib/IR.html]

How Search Engines Work
- [www.monash.com/spidap4.html]

Information retrieval models
- [www.cs.rpi.edu/~sibel/mmdb/lectures/ir_models.pdf]

Result Merging in Distributed Indexing
- [www.w3.org/Search/9605-Indexing-Workshop/Papers/Schuetze@Xerox.html]

4.3.3 Die inverse Dokumentenhäufigkeit (ITF-Algorithmus)

Bei der Bestimmung relevanter Dokumente hat ein Keyword zwei wesentliche Aufgaben. Zum einen muss es ein Dokument *inhaltlich* repräsentieren, d.h. das Schlüsselwort muss Aufschluss über den Inhalt eines Textes bzw. das Dokumententhema geben. Zum anderen muss es tauglich sein, ein Dokument gegenüber *anderen* Dokumenten im Datenbestand zu diskriminieren. Anders ausgedrückt, ein Keyword muss es ermöglichen Unterschiede zwischen verschiedenen Dokumenten sichtbar zu machen, um hierdurch bei der Informationssuche die relevanten von den nicht relevanten Dokumenten im Datenbestand unterscheiden zu

können. Diese Anforderung an ein Schlüsselwort entspricht der Precision-Funktion von Deskriptoren. So kann ein Deskriptor wie beispielsweise „Computer" aufgrund seiner relativen Worthäufigkeit in einem bestimmten Dokument durchaus als geeigneter Deskriptor für das *betreffende Dokument* gelten, da er in Bezug auf andere vorkommende Worte den Inhalt des Textes am genauesten repräsentiert.

Kommt jedoch der Begriff „Computer" in der Gesamtheit aller erfassten Dokumente und somit im gesamten Datenbestand sehr häufig vor, eignet er sich nicht die einzelnen Dokumente *zueinander* zu unterscheiden. Dies führt bei der Deskriptorengewichtung zu der Überlegung, Schlüsselworte auch in Bezug auf ihre Unterscheidungsfähigkeit zu den einzelnen Dokumenten zu bewerten. Dabei wächst die Bedeutung eines Begriffs mit der Häufigkeit innerhalb eines Dokuments, ist jedoch umgekehrt proportional zur Gesamtzahl der Dokumente in denen er vorkommt. Dieses Konzept das durch den Inverse Document Frequency Algorithmus (ITF) bzw. die inverse Dokumentenhäufigkeit ausgedrückt wird, bewertet ein Schlüsselwort folglich um so höher, je seltener es in anderen Dokumenten vorkommt, bzw. umso niedriger, je häufiger es in anderen Dokumenten auftritt.

Um den IDF-Inverse Document Frequency Algorithmus in einem sich dynamisch anpassenden System wie dem der Suchmaschinen zu implementieren, wird in der *Word List* die Häufigkeit eines jeden Begriffs gespeichert. Der Faktor der inversen Dokumentenhäufigkeit IDF kann dann zum Zeitpunkt des Dokumenten-Retrieval (Erfassung und Analyse) errechnet werden. Die hierzu erforderlichen Informationen können sehr einfach über die betreffende invertierte Datei kalkuliert werden, in Ergänzung zu einer Variablen die immer dynamisch die Gesamtanzahl aller Dokumente berechnet.

Links

Tf Idf Ranking
- [http://phpwiki.sourceforge.net/phpwiki/TfIdfRanking]

Extracting Document Representations
- [http://agents.www.media.mit.edu/groups/agents/publications/newt-thesis/subsection 2_6_2 _1.html]

CSM06 Information Retrieval
- [www.computing.surrey.ac.uk/personal/pg/A.Salway/csm06/CSM06%20LECTURE%204.ppt]

Ranking Algorithmus
- [www.csie.ncu.edu.tw/~chia/Course/IR/IR1999/QueryOperation.ppt]

4.3.4 Bedeutung der Lage eines Keywords

Gewichtungsverfahren die die *Lage eines Keywords* im Dokument berücksichtigen basieren auf der Überlegung, dass ein Verfasser ein für den Inhalt sehr wichtiges Schlüsselwort eher am Dokumentenanfang als am Ende eines Texts positioniert. Es kann bei Verfahren, die die Position eines Keywords im Text zur Dokumentengewichtung berücksichtigen, zwischen zwei Methoden unterschieden werden. Gewichtungsverfahren die sich auf die *absolute Position* eines Keywords im Dokument beziehen sowie einem Verfahren, das die *Nähe von Schlüsselwörtern zueinander* berücksichtigt. Letzteres Verfahren wird auch *Proximity-Verfahren* genannt.

Zur Bestimmung der Position eines Wortes setzen die Information Retrieval Systeme besondere Parser ein, die genau bestimmen an welcher Stelle sich ein Wort im Dokument befindet. Hierdurch ist es möglich eine Differenzierung der Gewichtung, bezogen auf die konkrete Position eines Wortes, vorzunehmen.

Betrachten wir die Struktur von HTML-Dokumenten so kann sie grob in einen *Dokumentenkopf* und einen *Dokumentenkörper* unterschieden werden. Innerhalb des Dokumentenkopfs befindet sich der Dokumententitel sowie Metaangaben über das Dokument, die in Formen von Meta-Tags dargestellt werden. Eine besondere Bedeutung kommt den Informationen innerhalb des Dokumentenkopfs zu. Information Retrieval Systeme bewerten den Inhalt des Dokumententitels besonders hoch, da davon auszugehen ist, dass der Verfasser eines Textes den Titel dazu verwendet, um den Inhalt möglichst prägnant zu beschreiben. Bei vielen Suchmaschinen erhalten Worte die sich im Dokumentenkopf befinden mit die höchste Gewichtung. Verfeinert kann dieses Verfahren durch eine differenzierte Berücksichtigung der genauen Position eines Wortes im Titel werden. Dabei wird dem ersten Wort im Titel ein höheres Gewicht zugeordnet als dem zweiten Wort, und das zweite Wort wird wiederum höher bewertet als das dritte Wort. Entsprechend ihrer Position können also Begriffe eine unterschiedlich starke Gewichtung erfahren.

Der zweite interessante Bereich innerhalb des Dokumentenkopfs sind für die Suchmaschinen die Angaben innerhalb der Meta-Tags. Über spezielle HTML-Parser werden die einzelnen Meta-Tags exakt erkannt und deren Inhalt entsprechend verarbeitet und bewertet. Analysiert man genau welche Meta-Tags bei der Dokumentenanalyse relevant sind, verbleiben aus der Vielzahl aller möglichen Meta-Tags lediglich das Meta-Tag DESCRIPTION und das Meta-Tag KEYWORDS, die von der Suchmaschine zur Gewichtung berücksichtigt werden. Schlüsselwörter, die in einem der beiden Meta-Tags vorkommen, erhalten Suchmaschinen individuell eine höhere Gewichtung, als ein Keyword das im Dokumentenkörper erscheint. Die Stärke der Gewichtung die ein Keyword erfährt das sich in einem der beiden Meta-Tags befindet, ist jedoch von Suchmaschine zu Suchmaschine unterschiedlich.

Im Dokumentenkörper befindet sich der eigentliche Text eines HTML-Dokuments und stellt für die Erfassung und Auswertung eines Themas den wichtigsten Bereich dar. Bei Systemen die eine differenzierte Gewichtung von Worten in Abhängigkeit ihrer Position im Text vornehmen, wird jedes einzelne Wort exakt mit seiner

Position innerhalb des Textes erfasst. Dabei wird jedes Wort mit genauer Positionsangabe im invertierten Dateisystem abgespeichert. Grundsätzlich gilt bei dieser Methode, je weiter am Dokumentenanfang ein Keyword vorkommt, desto höher ist die Bewertung. Zur Vereinfachung der Bewertungssystematik werden hierfür teilweise Klassen gebildet. Im Zuge der Klassenbildung erhalten beispielsweise Keywords, die sich innerhalb der ersten 50 Worte befinden eine höhere Bewertung, als Schlüsselworte, die sich innerhalb der Sektion von 51 bis 100 Worten befinden. Je nach eingesetzter Systematik kann nur eine begrenzte Anzahl von Sektionen indexiert werden oder auch eine vollständige Erfassung des gesamten Dokuments erfolgen.

Das Hypermedia ermöglicht Keywords auch außerhalb des eigentlichen HTML-Dokuments zu positionieren. Betrachten wir die Struktur und Aufbau der Adresse eines Dokuments im WWW so wird deutlich, dass ein Schlüsselwort auch innerhalb des URL als *Domain-Name*, als *Verzeichnisname* oder auch als *Dokumentenname* vorkommen kann. Ein Beispiel verdeutlicht dies. Möchte man das Keyword „Ferienwohnungen" in einem URL positionieren so ergibt sich unter Ausnutzung aller Möglichkeiten folgender URL:

www.ferienwohnungen.de/ferienwohnungen/ferienwohnungen.html

Bis zum technischen Relaunch Mitte 2002 gewichtete Fireball Schlüsselwörter, die sich innerhalb des URL befanden, besonders stark. Eine Analyse des URL ermöglicht sehr einfach festzustellen, ob ein Schlüsselwort als Domainname, als Verzeichnisname oder als Dokumentenname eingesetzt ist. Je nach Methodik kann eine differenzierte Gewichtung in Abhängigkeit der Lage des Keywords in dem URL erfolgen.

Das Proximity-Verfahren kommt bei Suchanfragen zum Einsatz, die aus mindestens zwei Suchworten bestehen. Die Grundüberlegung auf der das Verfahren beruht ist die Einschätzung, dass zwei Worte, die in einem Text näher zueinander vorkommen, einen Text inhaltlich eher repräsentieren als Worte, die weiter voneinander entfernt sind. Diese Einschätzung basiert auf empirischen Untersuchungen die zeigten, dass ein Verfasser Worte die in einem thematischen Zusammenhang stehen eher nahe beieinander im Text aufführt, als weiter voneinander entfernt. In der konkreten Umsetzung führt dies dazu, dass Suchmaschinen Dokumente differenziert bewerten, wenn Schlüsselwörter die in Kombination gesucht werden, in den jeweiligen Dokumenten unterschiedlich weit von einander entfernt erscheinen.

4.4 Hypermedia basierte Gewichtungsmodelle

Durch die Möglichkeiten des Hypertext im Internet sind zu den bisherigen Gewichtungsverfahren des klassischen Information Retrieval von Textdokumenten neue Techniken hinzu gekommen. Die Systematik des Hypermedia als eine weltweite gegenseitige Verflechtung von Dokumenten mittels Hyperlinks sowie die Möglichkeiten des Anwendungsprotokolls HTTP, führten zu zwei innovativen Methoden.

Die Gründer und Entwickler der Suchmaschine Google *Sergey Brin* und *Larry Page* entwickelten das hoch effiziente *PageRank-Verfahren,* das bei der Relevanzbewertung von Dokumenten explizit die Hyperlink-Verweise von Dokumenten zueinander analysiert und die *Anzahl* und *Qualität* der Hyperlink-Verweise als relevantes Gewichtungskriterium einsetzt. Im Zuge der technischen Entwicklung begannen auch andere Suchmaschinen ein ähnliches Verfahren einzusetzen das allgemein als *Link Popularity* bezeichnet wird. Es findet neben Google mittlerweile auch bei den Suchmaschinen Altavista, Inktomi, Alltheweb und Lycos in Kombination mit anderen Gewichtungsmethoden Anwendung. Verständlicherweise arbeitet jeder eingesetzte Link Popularity-Algorithmus etwas anders und bewirkt in Verbindung mit weiteren Gewichtungsverfahren unterschiedliche Auswirkungen auf das Ranking eines Dokuments.

Neben dem Link Popularity-Verfahren stellt die *Click Popularity-Technik* die zweite wesentliche Hypermedia basierte Innovation für das Information Retrieval dar. Das 1998 von *Gary Culiss* und *Mike Cassidy* entwickelte Verfahren ist ein Gewichtungsverfahren, das bei der Relevanzberechnung von Dokumenten die *Häufigkeit* berücksichtigt, die ein Dokument von Nutzern über die Suchergebnisliste aufgerufen wird und wie lange er darauf verweilt (*Verweildauer*). Es wurde erstmals mit der Suchmaschine DirectHit.com eingesetzt, die jedoch mittlerweile nicht mehr als eigenständige Suchmaschine existiert. Das Click Popularity-Verfahren wurde in den letzen Jahren von verschiedenen Suchmaschinen wie MSN, Lycos, Hotbot und Fireball als auch von Webkatalogen wie Yahoo eingesetzt. Im Gegensatz zur Link Popularity konnte es sich jedoch nie wirklich durchsetzen. Da es aber immer wieder bei verschiedenen Suchmaschinen zum Einsatz kommt, soll es nachfolgend auch erläutert werden.

Obwohl beide Verfahren, gleichwohl wie die Modelle des Vektorraums, statistische Häufigkeitswerte als Maß einsetzen, stellen sie dennoch eine eigene Klasse dar. Vektorraummodelle beziehen sich ausschließlich auf ein Dokument oder eine Sammlung von Dokumenten, bei denen die Dokumente als zweidimensionales Konstrukt definiert werden können. Durch die Einbeziehung des gesamten Hypermedia sind Dokumente im Internet nun als dreidimensionales, interdependentes Konstrukt zu sehen, das eine neue Dimension für das Information Retrieval eröffnet. Es ist wichtig zu beachten, dass die verschiedenen Verfahren nicht exklusiv, sondern in Kombinationen mit anderen Gewichtungsverfahren eingesetzt werden.

4.4.1 PageRank von Google

Aufgrund seiner erheblichen Bedeutung für das Ranking bei verschiedenen Suchmaschinen soll die Funktionsweise des *Link Popularity & Analysis-Verfahren* ausführlich dargestellt werden. Da diese Methodik erstmals mit der Entwicklung von Google eingeführt wurde und es dort nach wie vor einer der zentralen Gewichtungsmethoden darstellt, wird nachfolgend beispielhaft das PageRank-Verfahren von Google beschrieben. Die gegenwärtigen Link Popularity & Analysis-Verfahren anderer Suchmaschinen basieren prinzipiell auf ähnlichen Methoden.

Gegenwärtig setzen Link Popularity & Analysis-Verfahren neben Google u.a. auch Altavista, Lycos, Alltheweb sowie Fireball ein. Während am Anfang der Einführung der Technik die Link Popularity überwiegend in Form der Addierung von ausgehenden Hyperlink Verweisen auf eine andere Web Site realisiert wurde, hat mittlerweile die *Qualität* von Link-Verweisen erheblich an Wichtigkeit und Bewertungsstärke gewonnen.

Der theoretische Ansatz des PageRank-Verfahrens beruht auf den Überlegungen, dass ähnlich wie bei wissenschaftlichen Veröffentlichungen, diejenigen Texte für ein Thema am relevantesten sind, die von anderen Autoren häufig zitiert werden. In der Fortführung dieses theoretischen Ansatzes verlässt sich Google auf die demokratische Struktur des Hypermedia mit seiner Hyperlink-Struktur. Google setzt Dokumentenverweise als sein relevantestes Kriterium der Gewichtung, unter Berücksichtigung der *Anzahl* und *Qualität* von verweisenden Hyperlinks ein. Da die Anzahl an Hyperlink-Verweisen durch Content-Anbieter relativ einfach beeinflusst werden kann, basiert der qualitative Ansatz auf einer Bewertung der verweisenden Seite und kann weniger einfach manipuliert werden.

PageRank ist also eine Methode zur Kalkulation der Relevanz eines Dokuments auf Basis der *Anzahl* und *Qualität* von Link-Verweisen anderer Dokumente und drückt sich als Ergebnis in einem numerischen Wert aus. Je höher der Wert und somit der PageRank ist, desto wichtiger ist ein Dokument. Das PageRank-Verfahren wird zur Diskriminierung der Dokumente in Kombination mit dem Einsatz anderer Verfahren, wie beispielsweise dem Term Frequency Algorithmus oder der differenzierten Bewertung der Position von Schlüsselwörtern, angewendet. Es stellt jedoch bei Google die dominierende Methode zur Bestimmung der Bedeutung eines Dokuments dar.

Verkürzt dargestellt wird beim PageRank-Verfahren ein Dokument in vier Schritten gefunden und bewertet. Die Vorgehensweise kann wie folgt dargestellt werden:

1. Finde alle Dokumente die das Suchwort als Deskriptor beinhalten.
2. Wende Keyword spezifische Verfahren wie z.B. TF oder ITF an und berechne einen initialen Wert für alle Dokumente.
3. Berechne den Wert aller ausgehenden Verweise eines Dokuments.
4. Berechne den PageRank-Wert für jedes Dokument und führe den iterativen Berechnungsprozess n-mal aus.

Zur Berechnung des PageRank setzt Google auf die Anzahl und Qualität der verweisenden Hyperlinks eines Dokuments, wobei die Links als Empfehlung *für* ein Dokument interpretiert werden. Bei Google bleibt jedoch der semantische Inhalt eines verweisenden Links bei der Bewertung unberücksichtigt. D.h. ob ein Link einen sinnvollen Textinhalt in Bezug auf die verweisende Seite aufweist oder nicht, ist bei Google laut Eigenauskunft unerheblich.

Wenn eine Seite A mit einem Link auf eine Seite B verweist (nachfolgend als *eingehender Verweis* für B bezeichnet), bedeutet dies i.S. der Methodik von PageRank, dass die Seite A die Seite B für wichtig erachtet. Der Seite B wird aufgrund eines *eingehenden Verweises* ein Wert zugeschrieben, der den PageRank Wert von B erhöht.

Google berücksichtigt aber auch die Anzahl der Verweise, die von einem Dokument auf andere Dokumente zeigen. Verfügt die Seite B über ausgehende Verweise auf die Dokumente C und D, so wird der Wert von B auf C und D anteilig aufgeteilt. Wie sich nachfolgend noch zeigt, differenziert PageRank seine Kalkulation zusätzlich noch qualitativ nach der Herkunft der eingehenden Verweise.

Wenn eine Seite eine Vielzahl von eingehenden Verweisen hat, wird hierdurch grundsätzlich Ihre Wichtigkeit über das PageRank-Verfahren gesteigert, was sich durch eine erhöhte Gewichtung ausdrückt und zu einer verbesserten Relevanz führt. Verweisen hingegen keine oder zu wenige Seiten auf ein Dokument kann dies zur Folge haben, dass die Seite von Google als irrelevant erachtet und gegebenenfalls wieder aus dem Index gelöscht wird. Die Erfassung der vielfältigen Link-Strukturen aller indexierten Dokumente zueinander wird durch den Einsatz einer URL-Datenbank realisiert. Über eine spezielle URL-Analyse kann dabei die gesamte Link-Struktur erkannt und ausgewertet werden. Hierdurch wird eindeutig erkennbar, welche Dokumente auf andere Dokumente verweisen. Mittels einer ausgefeilten Analyse der Link-Strukturen werden auch Zirkelbezüge von Link-Verweisen identifiziert und bei der Berechnung des PageRank nicht berücksichtigt. Gleichfalls werden Verweise, die von Dokumenten unterhalb der gleichen Domain zueinander erfolgen, von Google bei der Berechnung nicht berücksichtigt.

Die Methodik des PageRank-Verfahren kann sehr einfach an Hand des 1997 veröffentlichten PageRank-Algorithmus näher erklärt werden. Es ist jedoch anzumerken, dass der heute eingesetzte Algorithmus an die fortschreitende Entwicklung im WWW angepasst wurde. Der PageRank Algorithmus wurde von seinen Erfindern *Brian* und *Page* wie folgt definiert:

PageRank

$$\mathbf{PR(A) = (1-d) + d(PR(T1)/C(T1) + ...PR(Tn)/C(Tn))} \qquad (4.3)$$

PR(A) = der PageRank Wert von A berechnet aus allen eingehenden Verweisen.
A = das Dokument für den der PageRank Wert ermittelt wird.
d = ein Dämpfungsfaktor zwischen 0 und 1 (oftmals ~ 0,85).
PR(T1) = der PageRank Wert des Dokuments T1 das auf A verweist.
C(T1) = die Gesamtanzahl aller ausgehenden Verweise von T1.

PR(Tn)/C(Tn) bedeutet, dass der Verweiswert für jede Seite die auf **A** zeigt, aus dem PageRank der Seite **n**, unter Berücksichtigung der Anzahl aller ausgehenden Verweise von **Tn** berechnet wird. Durch die Formel wird deutlich, dass es sich um eine iterative Berechnung des Wertes **PR(A)** handelt, da zur Berechnung des PageRank für Dokument **A** zuerst alle PageRank-Werte **PR(n)** derjenigen Dokumente erforderlich sind, die auf **A** verweisen. Das bedeutet konkret, dass ein neu indexier-

tes Dokument faktisch zunächst keinen PageRank-Wert besitzt. Um dieses Dilemma zu umgehen, wird einem neu erfassten Dokument ein initialer Wert zugeordnet. Dieser Wert kann sich aus der Berechnung der Gewichtung ergeben, die auf dem Term Frequency Algorithmus oder einer differenzierten Bewertung der Position von Schlüsselwörtern beruht. Verfügt ein Dokument hingegen über einen Verweis von einer Website die Google für besonders wichtig erachtet, wie z.B. Yahoo oder den Katalog des Open Directory Projects, erfolgt sofort eine wesentlich höhere anfängliche PageRank-Bewertung des Dokuments. Der Dämpfungsfaktor d stellt innerhalb des Algorithmus eine Individualisierungsvariable dar und reflektiert den Faktor, den eine Seite einer anderen Seite von dem eigenen Wert zuweisen kann. Sie dient zur Feineinstellung der Berechnungsmethode und bedeutet, dass eine Seite einer anderen Seite durch einen ausgehenden Verweis nicht ihren eigenen vollen Wert zuweisen kann.

Das iterative Verfahren der PageRank-Berechnung kann schematisch an einem Beispiel mit den vier Dokumenten A, B, C, D verdeutlicht werden, die unterschiedlich aufeinander verweisen. Zum Start wird der Einfachheit halber jedem Dokument ein Wert von 1 zugewiesen. Dieser Wert kann durch ergänzende Gewichtungsverfahren wie z.B. durch die Berücksichtigung von Worthäufigkeiten berechnet werden. Die Pfeile stellen die Richtung der ausgehenden Verweise dar.

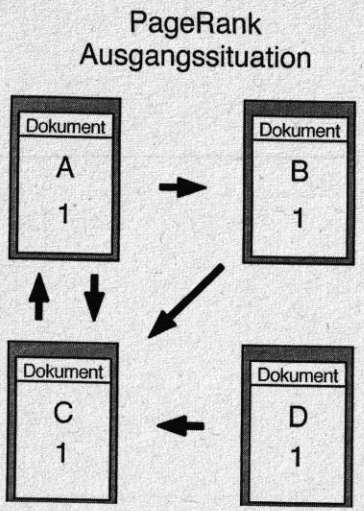

Abb. 4.3. Dokumentengewichtung bei Page Rank-Ausgangssituation

Zuerst wenden wir den Dämpfungsfaktor d mit einem Wert von 0,85 an. Der Dämpfungsfaktor impliziert, dass ein Verweis auf eine andere Seite dieser nicht den gleich hohen Wert zuweisen kann, den die verweisende Seite selbst besitzt.

Kalkulation PageRank-Wert Seite A

Beginnen wir mit der Kalkulation bei Seite A. Der um den Dämpfungswert bereinigte Wert für Verweise von A ist d * PR(TA) =1* 0,85 = 0,85. Da zwei ausgehende Verweise von A weggehen, ist d (PR(TA)/C(TA)) = 0,85 / 2 = 0,425; d.h. am Ende des iterativen Prozesses werden der Seite B und C zu ihrem bisherigen Wert der Wert 0,425 zugewiesen.

Kalkulation PageRank-Wert Seite B

Seite B hat nur einen ausgehenden Verweis, weshalb der Seite C der Wert 1 x 0,85 = 0,85 am Ende des iterativen Prozesses zugewiesen wird.

Kalkulation PageRank-Wert Seite C und D

Da Seite C auch nur einen ausgehenden Verweis auf A besitzt, ist der Wert des Verweises auf A gleichfalls 0,85. Der gleiche Wert ergibt sich für den Verweis Seite D auf C.

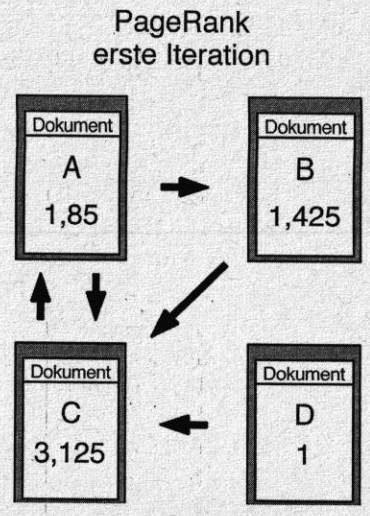

Abb. 4.4. Dokumentengewichtung PageRank – erste Kalkulation

Wendet man die Berechnung in einem ersten (n = 1) iterativen Prozess an, ergeben sich die oben dargestellten Dokumentenwerte (Abb. 4.4). Der Kern der PageRank-Theorie ist jedoch, dass besser verlinkte Dokumente auch einen höheren Wert zugewiesen bekommen was dadurch realisiert wird, dass der iterative Prozess mindestens ein zweites Mal ausgeführt wird. Die erneute Anwendung des obigen Verfahrens führt zu veränderten PageRank-Werten:

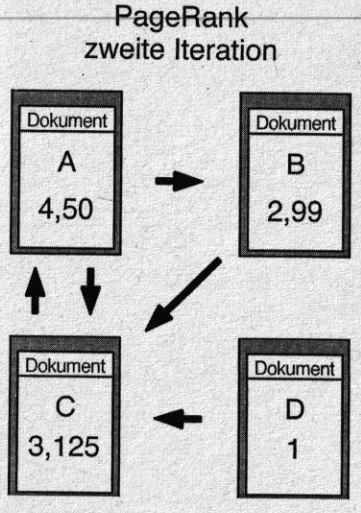

Abb. 4.5. Dokumentengewichtung PageRank – zweite Kalkulation

Der besseren Übersichtlichkeit werden bei der zweiten Iteration nur zwei Dezimalstellen abgebildet. Das Ergebnis zeigt deutlich, dass die Seiten auf die am häufigsten verwiesen wird, den höchsten PageRank Wert erhalten. Dokument D auf das von keinem Dokument verwiesen wird, weist auch bei mehrmaligen Durchgängen immer nur den Wert aus, den es anfänglich durch ein initiales Gewichtungsverfahren erhalten hat. Die genaue Anzahl wie oft der oben beschriebene iterative Prozess zur Bestimmung des PageRank ausgeführt wird, ist nicht bekannt. Von Insidern wird jedoch eine Wiederholungshäufigkeit von 20 bis 100 Wiederholungen genannt.

Neben der Anzahl von eingehenden Verweisen ist für die Berechnung des PageRank der Wert der jeweilig verweisenden Seite besonders relevant, da dieser die Berechnungsbasis für den Verweis darstellt und somit direkt die Gewichtung der empfangenden Seite beeinflusst. Google berücksichtigt neben der Anzahl der eingehenden Verweise insbesondere auch die Qualität der verweisenden Seite, die sich letztendlich durch ihre Wertigkeit ausdrückt. Die Qualität einer Seite kann sich z.B. durch ihre besondere Bedeutung im Web oder durch eine thematische Ähnlichkeit zum Verweis ausdrücken. Besondere qualitative Bedeutung für Google haben in diesem Zusammenhang intellektuell bewertete Webkataloge wie z.B. Yahoo oder der Katalog des Open Directory Project, die manuell einen besonders hohen PageRank-Wert zugeordnet bekommen haben. Identifiziert Google beispielsweise einen Verweis von Yahoo auf ein Dokument, wird diesem Verweis ad hoc ein wesentlich höherer Wert zugeschrieben, als einem Verweis von einer anderen Seite. Positiv für die Höhe des PageRank wirken sich auch Verweise von thematisch ähnlichen Seiten aus. Erhält ein Dokument einen Verweis von einer Seite die ein ähnliches Thema zum Inhalt hat, wird dieser Verweis höher bewertet, als ein Verweis einer zwar von der PageRank-Wertigkeit gleichen, aber inhaltlich unterschiedlichen Seite. Eine Berechnung der Ähnlichkeit zwischen den einzelnen Dokumenten wird z.B. durch den Einsatz von Cluster-Verfahren erreicht.

Links

The Anatomy of a Large-Scale Hypertextual Web Search Engine
- [www7.scu.edu.au/programme/fullpapers/1921/com1921.htm]

Erklärungen zu PageRank
- [www.google.de/intl/de/why_use.html]

PageRank, HITS and a Unified Framework for Link Analysis.
- [www.nersc.gov/research/SCG/cding/papers_ps/sigpage6b.ps]

Link Analysis: Hubs and Authorities on the World Wide Web
- [www.nersc.gov/research/SCG/cding/papers_ps/hits3.ps]

Google Introduces Date Range Search
- [http://searchenginewatch.com/searchday/01/sd0716-realsearch.html]

Autom. Resource Compilation by Analyzing Hyperlink Structure and associated text
- [http://decweb.ethz.ch/WWW7/1898/com1898.htm]

Improved Algorithms for Topic Distillation in a Hyperlinked Environment.
- [ftp://ftp.digital.com/pub/DEC/SRC/publications/monika/sigir98.pdf]

Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect
- [www9.org/w9cdrom/175/175.html]

When Experts Agree: Using Non-Affiliated Experts to Rank Popular Topics
- [www10.org/cdrom/papers/474/]

Notes on Kleinberg's Algorithm and PageRank
- [www.eecs.harvard.edu/~michaelm/E126/klein.pdf]

Efficient Computation of PageRank
- [http://net.cs.pku.edu.cn/~webg/refpaper/papers/taher-efficient.pdf]

PageRank Explained
- [www.ececs.uc.edu/~annexste/Courses/cs690/PageRank.pdf]

PageRank von Google
- [www.suchmaschinentricks.de/ranking/link_popularity.php3]

4.4.2 Systematik der Click Popularity

Die Technologie der *Click Popularity* wurde erstmals mit der 1998 entwickelten Suchmaschine DirectHit.com eingesetzt. Das Konzept der Click Popularity beruht auf der Überlegung, dass diejenigen Seiten die von Nutzern entsprechend einer bestimmten Suche aus der Suchergebnisliste heraus häufiger angeklickt werden,

relevanter sein müssen, als solche Verweise der Ergebnisliste, die von den Anwendern seltener aufgerufen werden.

Beim Relevanzverfahren der Click Popularity werden Dokumente entsprechend den dargestellten Indexierungsmethoden in den Datenbestand aufgenommen und jedem Keyword ein Wert mittels der bekannten Gewichtungsmodelle zugeordnet. Bei neu indexierten Dokumenten ist dieser Wert jedoch im Allgemeinen niemals so hoch, dass er im Ranking zu einer der vordersten Positionen führt.

Eine Verbesserung der Rangposition wird durch die Anzahl der von Anwendern ausgeführten Klicks auf den URL des Verweises erreicht. Hierzu werden alle Klicks die auf einen Verweis der Suchergebnisliste vorgenommen werden gezählt, in einer Datenbank gespeichert und dem betreffenden URL zugeordnet. Die Berechnung des Gewichtungswerts basiert folglich auf der Häufigkeit der Klicks. Das Click Popularity-Maß ist somit ein Wert, der sich über die Anzahl aller erfolgten Klicks auf einen Verweis berechnet.

Der Wert der Click Popularity stellt keinen absoluten Wert dar. Das bedeutet, dass die Anzahl der erfolgten Klicks ins Verhältnis zur Dauer des Dokuments im Datenbestand gesetzt wird. Dadurch wird vermieden, dass Dokumente, die bereits schon lange im Datenbestand geführt werden, einen sehr hohen absoluten Wert erreichen, den neue Dokumente aufgrund ihrer kurzen Zugehörigkeit zum Bestand nur schwer einholen können.

Mit der Anzahl der erfolgten Klicks werden gleichzeitig alle IP-Adressen der ausführenden Clients registriert. Durch die Speicherung der IP-Adresse soll verhindert werden, dass das Ranking künstlich durch den Eigentümer einer Webseite mittels oftmaligem Klicken oder eingesetzten automatisierten Verfahren beeinflusst wird. Um diese potentielle Manipulationsmethode auszuschließen, werden weiter Klicks von gleichen Netzadressen innerhalb einer kurzen Zeitspanne nur einmal gezählt.

Ergänzend können die Suchmaschinen auch *Cookies* einsetzten, die beim ersten Besuch der Suchmaschine auf den Rechner des Users geladen werden. Ein Cookie ist eine kleine Textdatei, die vom Server an den Client übertragen wird, um den Client beim nächsten Besuch eindeutig identifizieren zu können. Sofern das Cookie vom Nutzer nicht abgelehnt wird, kann hierdurch zukünftig eine Erkennung des Rechners und damit verbunden, Manipulationsversuche durch einen Content-Anbieter erkannt werden.

Bis Mitte des Jahres 2002 setzte Fireball die Click Popularity als zusätzliche Gewichtungsmethode ein. Mit Anklicken eines Verweis wurde ein Zählbefehl an die URL-Datenbank in Form von

http://count.fireball.de/.../URL

übermittelt. Nachfolgende Abbildung macht dies deutlich.

Abb. 4.6. Klickzählung bei Fireball.de – Stand 1.2.2002

Die auf der Fast Technology beruhende Suchmaschine Alltheweb setzt die Click Popularity hingegen weiter ein (Stand 11/2002).

Abb. 4.7. Klickzählung bei Alltheweb.com – Stand 1.11.2002

Da verschiedene andere Suchmaschinen und Portale wie beispielsweise Lycos, Tiscali oder das Portal von T-Online auf der Technologie von Fast beruhen, ist es durchaus möglich, dass die Click Popularity zukünftig auch dort zum Einsatz kommt. Wird bei Alltheweb ein Verweis aufgerufen, erfolgt ein Zählbefehl unter Angabe des URL an die Datenbank. Obige Abbildung zeigt dies sehr deutlich.

Die technische Realisierung der Click Popularity erfordert eine Erweiterung der Systematik des Information Retrieval Systems um eine URL basierte Datenbank, die die Klickhäufigkeit permanent erfasst und ad hoc als Wert verfügbar macht.

Das große Problem für Content-Anbieter beim Verfahren der Click Popularity ist dessen nur sehr schwierige Beeinflussung zur Verbesserung der Rangposition. Da automatisierte Verfahren zur Erhöhung der Klickrate weitestgehend von der Suchmaschine unterbunden werden können, bleibt als einzige Möglichkeit die Bildung eines optimalen Dokumententitels und einer treffenden Meta-Tag DESCRIPTION-Angabe. Diese beiden Dokumenten bezogenen Informationen erscheinen in der Suchergebnisliste und dienen dazu, eine Zielgruppe möglichst geschickt anzusprechen. Gelingt dies, erfolgen verstärkt Klicks auf den Verweis und die Bewertung der betreffenden Seite steigt, was konsequenterweise eine Verbesserung der Rangposition mit sich bringt.

Links

Improve Search Engine Ranking with Click Popularity
- [www.apromotionguide.com/click_popularity.html]

The Ins and Outs of Click Popularity and Stickiness
- [www.searchengines.com/directhit.html]

Click Popularity
- [www.metamend.com/click-popularity.html]

Click Popularitty vs. Link Popularity
- [www.searchenginetutorial.com/link-popularity.html]

Improving Click Through
- [www.searchengineethics.com/clickthrough.htm]

Link and Click Popularity
- [www.thewritemarket.com/archives/2-4.htm]

Fast Search Technology
- [www.fastsearch.com]

4.5 Cluster-Verfahren

Eine von den bisher dargestellten Gewichtungsmodellen unterschiedliche Methode zur Bewertung eines Dokuments, ist die Klassifikation von Massendaten mittels Cluster-Verfahren. Cluster-Verfahren haben zum Ziel, aus einer Gesamtheit von Dokumenten Gruppen von Dokumenten zu bilden, die *zueinander ähnlich* sind. Also eine Ähnlichkeitsberechnung vorzunehmen, die zunächst nicht auf einer Suchanfrage beruht, sondern auf den Inhalten und bestimmten Parametern der einzelnen Dokumente zueinander.

Über verschiedene Verfahren der Cluster-Bildung wird, ausgehend von vordefinierten oder sich automatisch selbst generierenden Vorgaben der einzelnen Gruppen, alle Dokumente überprüft, inwieweit sie mit den Definitionen eines bestimmten Clusters übereinstimmen. Die Zuordnung eines Dokuments zu einem Cluster erfolgt u.a. über Berechnungsmethoden die auf statistischen Gewichtungsverfahren beruhen. Die Ergebnisse der Berechnungen von Ähnlichkeiten der einzelnen Dokumente zueinander, bzw. ihre Zugehörigkeit zu bestimmten Clustern, wird im Zuge der Indexierung vorgenommen und im invertierten Dateisystem mit einem numerischen Vermerk auf den jeweiligen Cluster berücksichtigt. Die Klassifikation kann dazu dienen, nicht nur Dokumente bei einer Suche zu berücksichtigen die einer konkreten Suchanfrage optimal entsprechen, sondern auch solche Dokumente Element eines Suchergebnisses werden zu lassen, die eine hohe Ähnlichkeit zu den Dokumenten aufweisen, die als relevant zur Suchanfrage bestimmt wurden.

Google ist einer der wenigen Suchmaschinen die ein Cluster-Verfahren einsetzen. Über die Funktion *Ähnliche Seiten* können aus der Suchergebnisliste weitere Dokumente ausgewählt werden, die eine Ähnlichkeit zu einem bestimmten Dokument aus der Suchergebnisliste besitzen. Nachfolgende Abbildung einer Suchergebnisliste von Google zeigt die Auswahloption „Ähnliche Seiten".

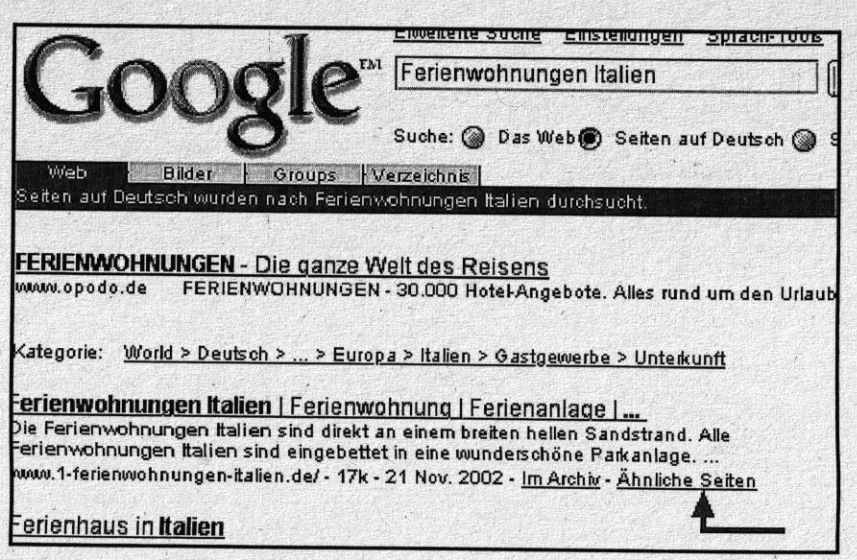

Abb. 4.8. Cluster-Suche bei Google

Klickt man das Link *Ähnliche Seiten* an erscheinen alle Dokumente, die aufgrund des von Google eingesetzten Cluster Verfahrens zu dem betreffenden Verweis als ähnlich definiert wurden. Google basiert seine Objekt bezogenen Cluster-Verfahren auf *verweisende Hyperlinks*. Das verweisende Dokument als auch die ausgewählten Dokumente sind dabei Elemente des gleichen Clusters. Dokumente gehören nicht exklusiv einem einzigen Cluster an, sondern können gleichzeitig unterschiedlichen Gruppen zugeordnet sein. Das auf Verweisen beruhende Cluster-Verfahren von Google ist jedoch nicht sehr effizient, da es eine Gruppenbildung ausschließlich auf verweisende Hyperlinks basiert. Hierdurch werden alle Dokumente Element eines Clusters, wenn sie einen Verweis auf ein Dokument richten oder von einem Dokument erhalten. Eine thematische Differenzierung der Verweise wird nicht vorgenommen. Ziel einer Cluster-Bildung ist es jedoch Dokumente *inhaltlich* zu gruppieren, um hierdurch eine Verbesserung der Precision zu erreichen. Mit einer reinen Verweis-bezogenen Gruppenbildung ist dies nicht zu erreichen

Neben Google setzt auch Teoma, eine sehr gute und präzise Suchmaschine aus den USA, Objekt bezogene Cluster ein. Über das Link *Related Pages* werden all diejenigen Dokumente angezeigt, die bezogen auf das betreffende Dokument eine

hohe Ähnlichkeit besitzen. Im Gegensatz zu Google basiert die Cluster-Bildung nicht auf Basis einer Link-Struktur, sondern erfolgt durch die Berechnung von Ähnlichkeitswerten, die auf den Inhalten bzw. den Keywords sowie deren Position im Dokument beruhen. Nachfolgende Abbildung der Suchergebnisliste von Teoma zeigt die Auswahloption „Related Pages".

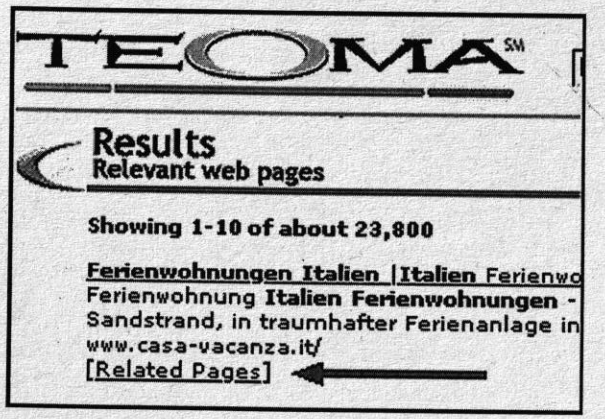

Abb. 4.9. Cluster-Suche bei Teoma

Cluster-Modelle lassen sich in *Word Cluster* zur Erzeugung von automatisierten *Thesauri* und in *Objekt-Cluster* zur Erzeugung von *Dokumenten-Clustern* unterteilen.

Ein Thesaurus stellt eine geordnete Zusammenstellung von Begriffen mit ihren natürlichsprachigen Beziehungen dar. Ziel ist es, durch eine terminologische Kontrolle eine Reduktion von Mehrdeutigkeiten und Unschärfen der Sprache zu erreichen. Durch Gruppenbildung und Äquivalenzrelationen werden automatisiert Synonymgruppen gebildet. Ein Thesaurus bestimmt also einerseits welche Begriffe vor der Speicherung eines Dokuments zur Inhaltsbeschreibung vergeben werden und andererseits, welche Begriffe bei der Suche nach relevanten Dokumenten benutzt werden sollen. Wird beispielsweise ein Word Cluster für den Begriff „Haus" gebildet, muss der betreffende Cluster alle Begriffe beinhalten, die ein Synonym zu dem Begriff „Haus" darstellen. Aber auch Worte ausweisen, die eine hohe thematische Ähnlichkeit zu dem Begriff besitzen. Der Cluster „Haus" kann beispielsweise folgende Worte beinhalten:

[Cluster: Haus]

- Wohnhaus
- Einfamilienhaus
- Mehrfamilienhaus
- Reihenhaus
- Bürohaus

Während der Word Cluster in Form eines Thesauri bei der Indexierung der Information Retrieval Systeme nicht zum Einsatz kommt, findet er gelegentlich Verwendung bei der automatisierten Bildung von Webkatalogen. Mit dem Dokumenten-Cluster der u.a. von Google eingesetzt wird, soll hingegen eine Struktur von ähnlichen Dokumenten aufgebaut werden, mit der Zielsetzung auch Dokumente zu finden, die ähnlich zueinander aber nicht *direkt* ähnlich zur Suchanfrage sind. Durch die Auswahl eines bestimmten Dokuments aus der Suchergebnisliste werden Dokumente geliefert, die ähnlicher zu dem betreffenden Dokument als zu der initialen Suchanfrage sind.

Bei automatisierter Cluster-Bildung wird an Hand bestimmter Parameter eine Ausgangskonfiguration von groben Clustern gebildet, die dann schrittweise verfeinert wird. Dieses Verfahren eignet sich sehr gut zum Einfügen von neuen oder unbekannten Objekten in eine bereits bestehende Cluster-Struktur. Bei Dokumenten-Clustern kann als kleinste Einheit ein einzelnes Dokument eingesetzt werden, das den Ausgangspunkt des Clusters darstellt. Ist keine Struktur als Startpunkt eines Verfahrens vorhanden, werden einfach verschiedene Objekteigenschaften benutzt um eine Startkonfiguration bestimmen zu können. Besondere Beachtung finden die nachfolgend genannten Parameter:

- Begriffe innerhalb des Title-Tags,
- Begriffe innerhalb des URL,
- TLD-Domainsuffix innerhalb des URL,
- Begriffe innerhalb des DESCRIPTION-Tags,
- Begriffe innerhalb des KEYWORDS-Tags,
- Anzahl der Begriffe im Dokument,
- Hyperlink-Verweise von / auf Dokumente.

Im Zuge der Cluster-Bildung werden dann aus einer kleinen Teilmenge von Dokumenten sogenannte Kern-Cluster gebildet. Diese dienen dazu, eine Ähnlichkeitsberechnung mit anderen Dokumenten auszuführen. Die bei der Bildung von Kern-Clustern nicht berücksichtigten Dokumente werden dann schrittweise in diese Cluster-Struktur überführt.

Liegt eine Ausgangskonfiguration vor, werden in einem zweiten Schritt Cluster-Repräsentanten für die bestehenden Cluster berechnet. Die Gruppierung der einzelnen Objekte wird dann wiederum mit Hilfe von Ähnlichkeitskoeffizienten zwischen den einzelnen Objekten und Cluster-Zentroiden (Mittelpunkte) durchgeführt. Die weitere Verfeinerung der groben Ausgangsstruktur wird mit folgenden Schritten erreicht:

(1) Vergleiche jedes Dokument mit allen Cluster-Zentroiden und berechne für alle Cluster einen Ähnlichkeitskoeffizienten.

(2) Bestimme für jedes Dokument das Cluster mit dem maximalen Ähnlichkeitskoeffizienten und integriere das Dokument in das entsprechende Cluster. Berücksichtige den Schwellwert zur Aufnahme in den Cluster. Wenn Überlappungen bei den Clustern gewünscht sind, weise ein Dokument mehreren Clustern zu.

(3) Berechne alle Cluster-Zentroiden nach der Dokumentenzuweisung neu und beginne bei Schritt eins.

(4) Beende das Verfahren nach Durchlauf einer n-Anzahl an Wiederholungen.

Neben den oben beschriebenen inhaltsbezogenen Dokumenten-Clustern können unterschiedliche Cluster parallel bzw. ergänzend entwickelt werden, die Dokumente nach weiteren Kriterien zu Gruppen bzw. Metagruppen zusammenfassen. Oftmals orientieren sich diese Cluster an nur einem oder wenigen statistischen Parametern, die zusätzlich differenzierte Cluster-Bildungen ermöglichen. So stellt z.B. die Einschränkung der Suche auf einen bestimmten Top Level Domain-Bereich bei Suchmaschinen eine häufig verwendete Cluster-Suche dar, die über eine einfache binäre Matrix realisiert wird.

Links

Google-Cluster
- [www.google.de/intl/de/help/refinesearch.html]

Automatic Classification
- [www.dcs.gla.ac.uk/~iain/keith/data/pages/36.htm]

Cluster-based retrieval
- [www.dcs.gla.ac.uk/~iain/keith/data/pages/103.htm]

Search Strategies
- [www.dcs.gla.ac.uk/Keith/Chapter.5/Ch.5.html]

Automatic Classification
- [www.dcs.gla.ac.uk/Keith/Chapter.3/Ch.3.html]

Subject Classification and Indexing
- [www.ctr.columbia.edu/~jrsmith/html/pubs/webseek/node3.html]

5 Suchprozess und Suchformen

Suchanfragen werden über den Query Processor der Suchmaschine ausgeführt, der für den Anwender die Schnittstelle zum Datenbestand der Suchmaschine bildet. Aufgabe des Query Processors ist es, die vom IR-System gefundenen Dokumente an Hand der Gewichtungsinformationen mittels einer Retrieval-Funktion in eine Reihenfolge zu bringen. Die Reihenfolge entspricht der Relevanz der jeweiligen Dokumente zur Suchanfrage. Der Algorithmus (Retrieval-Funktion) des Query Processors, die Gewichtungsmodelle sowie die Datenstrukturen sind folglich nicht unabhängig von einander, sondern stehen in funktionalem Zusammenhang.

Zur Vornahme von Suchanfragen stellen die Suchmaschinen den Anwendern verschiedene Möglichkeiten zur Verfügung, Suchanfragen nicht nur über einen einzelnen Suchbegriff sondern auch über Wortkombinationen oder auch den Ausschluss von Begriffen auszuführen. Suchanfragen können dabei mittels verschiedener Parameter eingegrenzt werden. Eine Eingrenzung des Suchraums kann sich auf bestimmte Bereiche des Dokuments wie z.B. den Titel oder die Meta-Tags erstrecken. Je nach Suchmaschine besteht weiter die Möglichkeit Suchbegriffe nicht nur im Dokument selbst zu suchen, sondern sie auch in der Domain oder im URL zu identifizieren. Ergänzend besteht die Systematik, Suchanfragen nur auf eine bestimmte Domain, einen ausgewählten Host oder einen Top Level Domain-Bereich auszuführen.

Die Kenntnis welche Suchmethoden einem Anwender zur Verfügung stehen und welche Formen der Suchraumeingrenzung bzw. Verfeinerung von Suchergebnissen möglich sind, ist bei der inhaltlichen Entwicklung von Websites von großem Vorteil. Da Anwender immer kompetenter Suchanfragen stellen und sie die zur Verfügung stehenden Methoden zur Verbesserung von Suchergebnissen zielstrebig einsetzen, müssen sie in Hinblick auf die Website Optimierung berücksichtigt werden. Eine wichtige Rolle spielt in diesem Zusammenhang auch die Zusammensetzung der Suchergebnisliste und die Informationen die zur Darstellung von Verweisen eingesetzt werden.

5.1 Der Query Processor – Suchtool der Suchmaschine

An diesem Punkt ist es zum besseren Verständnis der Funktion eines Query Processors sinnvoll, noch einmal kurz die beiden Systemkomponenten *Webrobot-System* und *Information Retrieval System* zu betrachten. Das Webrobot-System ist