

Linkanalyse und alternative Verfahren zur Qualitätsbewertung im Web Information Retrieval

Linkanalyseverfahren wie etwa der Page-Rank-Algorithmus haben sich in den letzten Jahren als zusätzlicher Faktor im Ranking von Internet-Suchmaschinen etabliert und zu zahlreichen Forschungsarbeiten geführt. Diese Algorithmen bewerten letztendlich die Qualität von Internetseiten auf der Basis eines einzelnen Parameters. Für die Qualitätsbewertung existieren bereits experimentelle Ansätze, welche mehrere Parameter in Betracht ziehen.

Dieser Artikel gibt einen Überblick über den Stand der Forschung zur Linkanalyse und geht auf die Anwendungen und Nachteile derartiger Ansätze ein. Zusätzlich werden die experimentellen Verfahren aus unterschiedlichen Fachrichtungen in den Forschungsüberblick integriert.

1 Web Information Retrieval

Web-Suchmaschinen setzen Technologie aus dem Information Retrieval für das Internet um und haben sich als wichtigstes Werkzeug für den Zugang zum Wissen im Internet etabliert. Suchmaschinen sind aus den folgenden Modulen aufgebaut [Arasu et al. 2001]:

- Der *Crawler* sammelt Seiten im Internet, indem er nach einem Kontrollalgorithmus die Links auf bekannten Seiten auswertet und weiterverfolgt.
- Der *Indexer* baut eine Repräsentation der vom Crawler gefundenen Seiten auf, die deren Inhalt wiedergibt. Dazu werden die Texte mit Standardverfahren des Information Retrieval wie linguistische Vorverarbeitung und Gewichtungungsverfahren analysiert.
- Die *Benutzungsoberfläche* (meist ein Web-Client) erlaubt dem Benutzer das Stellen von Anfragen, präsentiert die Ergebnisse und sollte Suchstrategien unterstützen.
- Die *Anfragenverarbeitung* analysiert die Anfragen und vergleicht sie mit den im Index repräsentierten Seiten. Abhängig von der Ähnlichkeit zwischen Anfrage und Dokumenten wird ein Ranking erstellt.

Das Information Retrieval steht im Web vor einigen neuen Herausforderungen. Dazu zählen nicht nur die Dynamik und die enorme Anzahl der Dokumente, die zu Kompromissen zwischen Effizienz und Effektivität führen. Die Qualität der Retrieval-Ergebnisse wird stark von der Heterogenität der Dokumente hinsichtlich Datenformaten, Länge, Sprache und nicht zuletzt ihrer Qualität beeinflusst.

Die schlechte Qualität von Internetseiten führt häufig zu unerwünschten Ergebnissen in Web-Suchmaschinen, obwohl eine inhaltliche Ähnlichkeit zwischen Anfrage und Dokument besteht. Unter anderem wertet die Forschungsleiterin von Google die unterschiedliche Qualität im Web als eine der wichtigsten Herausforderungen für zukünftige Suchmaschinen [Henzinger et al. 2002]. Auch die Vortäuschung von Inhalten, die gar nicht geboten werden, also Spam, lässt sich als Problem der Qualität betrachten.

Als Reaktion darauf versuchen Suchmaschinen-Betreiber automatisch die Qualität von Seiten im Internet abzuschätzen. Neben das Ranking der Dokumente nach dem Inhalt der Anfrage tritt ein Ranking nach der Qualität. Die endgültige Reihenfolge der Ergebnisdokumente wird aus beiden Maßzahlen bestimmt.

Was kann aber Qualität von Dokumenten und der enthaltenen Information bedeuten? Die unterschiedlichsten Definitionen lassen sich finden. Einen Rahmen hierfür liefert [Marchand 1990], der fünf typische Ansätze für die Definition von Informationsqualität nennt:

- **Transzendent:** Diese Definitionen setzen eine objektive und absolute Qualität voraus, die universell gültig ist.
- **Benutzerorientiert:** Dieser Ansatz betont die Subjektivität der Qualität und stellt sie in den Kontext der jeweiligen Situation des Benutzers.
- **Produktorientiert:** Das Informationsprodukt und seine Eigenschaften stehen im Fokus dieser Ansätze, die da-

von ausgehen, dass Qualität weitgehend messbar und quantifizierbar ist.

- **Produktionsorientiert:** Im Mittelpunkt steht hier der Prozess der Erstellung des Produkts. Qualität besteht in der Abdeckung des vorab festgelegten Pflichtenhefts.
- **Wertorientiert:** Neben die positiven Eigenschaften treten hier die Kosten des Informationsprodukts. Qualität ist die Suche nach einer optimalen Balance zwischen den beiden Forderungen.

Obwohl zahlreiche Autoren versuchen, transzendente Definitionen zu geben, ist die Annahme einer objektiven Qualität von Internetdokumenten letztlich wohl nicht angebracht. Vielmehr sollten die Benutzer mit ihrer Subjektivität und die wahrgenommene Qualität im Zentrum stehen. Die meisten Autoren im Bereich der Linkanalyse gehen dagegen nach wie vor von einer objektiven und absoluten Qualität aus, welche durch geeignete Algorithmen der Linkanalyse gemessen werden kann.

2 Linkanalyse

Die Linkanalyse untersucht die Verbindungen zwischen Webseiten und extrahiert daraus neues Wissen. Linkanalyse ist damit ein Teilbereich von Web-Mining und insbesondere Web-Structure Mining [Rahm 2002].

Wer eine Internetseite sehr positiv bewertet, verweist möglicherweise in seinen eigenen Seiten mit einem Link darauf. Die Grundannahme der Linkanalyse besteht darin, dass Links auf eine bestimmte Seite ein Qualitätsurteil darstellen, welches zur Bewertung dieser Seite ausgenutzt werden kann: »A simple means to measure the quality of a Web page ... is to count the number of pages with pointers to the page« [Kobayashi & Takeda 2000, S. 161]. Der bedeutendste Algorithmus ist PageRank, der von einem der Gründer von Google formuliert und zum Patent angemeldet wurde. PageRank ist ein Qualitätsmaß, das für jede Seite in

einem formalen Graphen definiert ist und aus der Linkstruktur bestimmt wird. Zentraler Faktor ist die Anzahl der Links, welche auf eine Seite verweisen. Ferner gewichtet der Algorithmus den Einfluss jedes Links mit dem PageRank der Ausgangsseite.

Die Ideen der Linkanalyse sind nicht neu, sondern werden seit langem von der Bibliometrie und Szientometrie eingesetzt. Dort entsprechen den Links die Zitate, mit welchen wissenschaftliche Publikationen aufeinander verweisen. In der Regel wird damit z.B. der so genannte *impact factor* einer Zeitschrift bestimmt.

2.1 PageRank und seine Varianten

PageRank wurde vor allem durch seine Integration in der Suchmaschine Google bekannt. Der Algorithmus weist jeder Seite in einem formalen Graphen einen Wert zu. Das Internet lässt sich als ein Graph interpretieren, in dem die Seiten die Knoten und die Links gerichtete Verbindungen darstellen. Der PageRank einer Seite p ergibt sich aus einer Konstante für den jeweiligen Graphen und dem Mittelwert aus den PageRank-Werten aller Seiten, welche eine Verbindung hin zu p besitzen [Henzinger 2000, S. 2 f.]:

$$R(p) = \frac{\epsilon}{n} + (1 - \epsilon) \cdot \sum_{(q,p)} \frac{R(q)}{\text{outlinks}(q)}$$

$R(p)$ PageRank von Seite p

ϵ Parameter (zwischen 0,1 und 0,2)

n Zahl der Links im Graph

(q, p) Seiten q mit Link zu Seite p

$\text{outlinks}(q)$ Zahl der out - Links von Seite q

In der ursprünglichen Form ist der Parameter vor der Summe noch nicht abhängig von der Anzahl der Seiten im untersuchten Graphen [Page et al. 1998, S. 3]. Die Berechnung kann auch als Funktion der Verbindungsmatrix des Internets (bzw. des untersuchten Ausschnitts) betrachtet werden. Im iterativen Ablauf wird dann bei jedem Schritt der PageRank-Vektor neu aus dem vorherigen PageRank-Vektor R sowie der Verbindungsmatrix \mathfrak{R} berechnet.

$$\vec{R} = f(\mathfrak{R})$$

$$\vec{R}^i = f(\vec{R}^{i-1}, \mathfrak{R})$$

Der Algorithmus konvergiert nach einer Anzahl von Schritten [Page et al. 1998],

d.h., die Autoritätswerte verändern sich dann kaum mehr. In der Praxis werden zwischen fünf und 50 Iterationen berechnet.

Der PageRank-Vektor R kann also auch durch wiederholtes Multiplizieren mit der Verbindungsmatrix berechnet werden. Der PageRank-Vektor löst also folgende Gleichung und stellt damit den Eigenvektor der Verbindungsmatrix \mathfrak{R} dar [Haveliwala 2002, S. 3]:

$$\vec{R} = \vec{R} \times \mathfrak{R}$$

Mittlerweile existieren zahlreiche Varianten von PageRank. Eine Schwäche des ursprünglichen Algorithmus liegt in der fehlenden thematischen Fokussierung. Zum einen stellt der globale Ansatz natürlich einen Vorteil dar. Die Werte aller Seiten können vorab und ohne Berücksichtigung des Kontexts berechnet werden. Gleichwohl stellt der fehlende thematische Bezug von PageRank ein von vielen Autoren bemängeltes Problem dar.

Sehr gute und hochspezialisierte Seiten erreichen auch bei relevanten Anfragen nicht immer ein hohes Ranking, da ihre Relevanzwerte von fachlich weniger spezifischen Seiten mit sehr hohem PageRank überlagert werden können. Würde aber für jedes Themengebiet ein eigener PageRank-Vektor berechnet, dann könnten sich hochspezialisierte Seiten eher durchsetzen. Ein derartiges System würde auch berücksichtigen, dass z.B. ein Internet-Verzeichnisdienst für manche Themen sehr hohe Qualität liefert, für andere Themen dagegen nicht.

[Haveliwala 2002] stellt eine Variante des PageRank-Algorithmus vor, die eine solche thematische Fokussierung vornimmt. In dem Ansatz werden als Themen die sechzehn obersten Kategorien des Internet-Verzeichnisdienstes *Open Directory Project* (<http://dmoz.org>) gewählt. Für jede dieser Kategorien werden die Seiten unterhalb der Kategorie und die darin enthaltenen Terme extrahiert. Der so bestimmte Termvektor repräsentiert das jeweilige Thema. Für jede Seite wird die Ähnlichkeit des Termvektors der Seite zu den Termvektoren aller Themen bestimmt. Damit wird eine Seite nicht nur einem Thema zugeschlagen, sondern erhält für jedes Thema ein Gewicht, das diese Ähnlichkeit widerspiegelt. Zudem erhält jede Seite einen PageRank für alle Themen. Der endgültige PageRank ergibt

sich dann als lineare Kombination der einzelnen PageRank-Werte $R(p)$, die mit dem Themengewicht der Seite multipliziert werden [Haveliwala 2002].

$$R(p) = \sum_{topic} R_{topic}(p) \text{sim}(topic, p)$$

Eine Voraussetzung bildet die Berechnung der themenspezifischen PageRank-Werte $R_{topic}(p)$. Sie lässt sich als eine Modifikation der Verbindungsmatrix \mathfrak{R} zu \mathfrak{R}^* vor der Bestimmung des Eigenvektors interpretieren. In der Verbindungsmatrix \mathfrak{R}^* zur Bestimmung des themenspezifischen PageRanks besitzen die Seiten zu diesem Thema »mehr« eingehende Links [Haveliwala 2002, S. 4]:

$$\vec{R}_{topic} = \vec{R}_{topic} \times \mathfrak{R}^*$$

$$\vec{R}_{topic} = (1 - \alpha) \mathfrak{R} \times \vec{R}_{topic} + \alpha \vec{t}$$

Dieses Vorgehen zur Erhöhung der Wahrscheinlichkeit, dass die Seiten zu einem Thema erreicht werden, lässt sich im Rahmen des so genannten *Random-Surfer-Modells* anschaulich darstellen. Der PageRank einer Seite entspricht der relativen Wahrscheinlichkeit, dass ein Benutzer, der für unbegrenzte Zeit zufällig Links verfolgt, auf diese Seite trifft. Da es im Web viele Sackgassen gibt, springt dieser Benutzer im *Random-Surfer-Modell* nach einer Reihe von Schritten unabhängig von einem Link auf eine zufällig ausgewählte Seite. Um dies zu simulieren, addiert man in der Linkmatrix zu allen Zellen eine sehr kleine positive Zahl (Teleportations-Parameter). Damit besteht eine Übergangswahrscheinlichkeit zwischen allen Seiten im betrachteten Graphen.

Bei der Berechnung der themenspezifischen PageRank-Vektoren wird dieser kleine Wert nicht mehr über alle Seiten gleich verteilt, sondern Seiten zu einem Thema werden höher gewichtet, und diese gewinnen sodann auf die Berechnung des PageRanks höheren Einfluss. Dieses Verfahren kann ebenso auf individuell ausgewählte Seiten begrenzt werden und dient dann der Personalisierung des PageRanks.

Einen personalisierten PageRank beschreiben [Jeh & Widom 2003]. Dabei stellt das System nicht real für jeden Benutzer ein eigenes Qualitätsranking, dennoch kann bei diesem Verfahren durchaus jeder Benutzer ein anderes Ergebnis er-

halten. Das Interesse des Benutzers ermitteln [Page et al. 1998] sowie [Jeh & Widom 2003] aus einer Sammlung von relevanten Seiten. Dieses Verfahren hat den Vorteil, dass jede beliebige Menge von Webseiten als Ausgangspunkt dienen kann. Meist verwenden die Systeme die *Bookmarks* des Benutzers, die zwar leicht zu extrahieren sind, jedoch nur ein sehr eingeschränktes Benutzermodell darstellen. Die Menge von Seiten wirkt dann während der Berechnung von PageRank als *Bias*.

Allerdings erscheint diese Art der Personalisierung fragwürdig. Die Links mit stärkeren Gewichten sind Verbindungen, welche der Benutzer von den ihm bekannten *Bookmarks* aus ohnehin durch Navigieren erreichen kann. Bei der Suche sind aber häufig völlig neue und bisher unbekannte Seiten gefragt. Das *Intelligent-Surfer-Modell* kommt ohne die explizite Angabe von Seiten aus und bestimmt die interessanten Seiten aus einer Anfrage. Es gewichtet bei der PageRank-Berechnung Links stärker, bei denen sowohl die Ausgangsseite als auch die Zielseite einen Anfrageterm enthält [Richardson & Domingos 2002].

Die vom Benutzer bevorzugten Seiten könnten am besten durch die Integration weiterer Wissensquellen ermittelt werden. Durch Integration von realen Nutzungsdaten lassen sich bei der PageRank-Berechnung die Links stärker gewichten, die häufiger benutzt werden. [Oztekin et al. 2003] stellt ein entsprechendes *Usage Aware PageRank* vor, in dem häufig verfolgte Links den PageRank der Zielseiten stärker erhöhen.

Ausgangspunkt eines sehr ähnlichen Verfahrens ist die empirisch nicht weiter belegte Aussage, Linkanalyse liefere gute Ergebnisse bei Suchen im gesamten Web und schlechtere Ergebnisse bei Suchen in kleineren Mengen von Internetseiten wie etwa *Sites* [Xue et al. 2003]. Innerhalb dieser Menge greifen [Xue et al. 2003] dann ebenfalls auf Log-Daten der Benutzeraktionen zu. Für die PageRank-Berechnung benutzen die Autoren nicht die ursprüngliche Linkmatrix, sondern ersetzen diese durch eine Matrix so genannter impliziter Verknüpfungen, welche eher den Charakter positiver Empfehlungen tragen. Diese Links bestehen aus Paaren von Seiten, die häufig gemeinsam in Benutzerpfaden vorkommen. Die Evaluierung weist auf eine Verbesserung der Re-

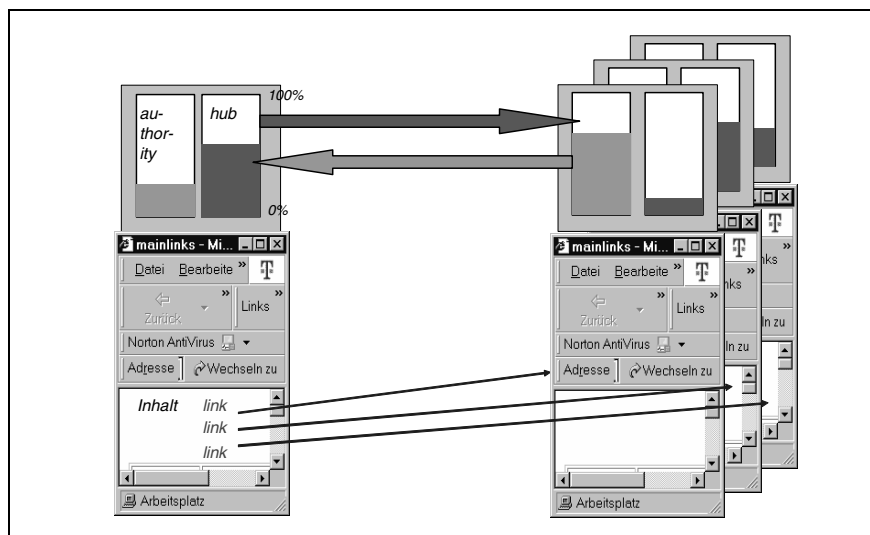


Abb. 1: Der HITS-Algorithmus als gegenseitige Verstärkung von Hub- und Authority-Gewicht

trieval-Ergebnisse hin, beruht jedoch nur auf 30 Anfragen [Xue et al. 2003].

2.2 Maße mit zwei Rollen

Der HITS- (*Hyperlink Induced Topic Search*) oder Kleinberg-Algorithmus führt zwei Rollen ein, um die Autorität zu bewerten [Kleinberg 1998]. Er weist jeder Webseite ein Gewicht für die Rollen *Hubs* und *Authority* zu. Ein *Hub* entspricht einem Mittelpunkt oder Verteiler, dessen Aufgabe im Wesentlichen in der Bereitstellung von Verbindungen zu anderen Seiten besteht. Dahinter verbirgt sich die Vorstellung eines Clearinghouses oder in der Wissenschaft der eines guten Überblicksartikels. Ein hoher *Hub*-Wert kennzeichnet also einen guten Informationsvermittler. Die *Authorities* dagegen enthalten die eigentliche Information in unterschiedlicher Qualität. HITS findet nur Anwendung auf eine Menge von ca. 5000 bis 10.000 Seiten, die aus einer Suchanfrage ermittelt werden. Diese werden analysiert und die enthaltenen Verbindungen extrahiert. Die Seiten, auf welche Links aus dem Suchergebnis verweisen, gelangen ebenfalls in die Ausgangsmenge. *Hub* und *Authority* verstärken sich wechselseitig. Die *Authority* einer Seite steigt mit der Anzahl der ankommenden Verbindungen. Diese Zahl wird aber mit dem *Hub*-Gewicht der Ausgangsseite relativiert. Nur die Links von guten *Hubs* wirken sich somit stark auf die Autorität einer Seite und damit auf das *Authority*-Gewicht aus. Ebenso unterliegt das *Hub*-Gewicht einer Verände-

rung, die von der Autorität der Zielseiten abhängt. Je besser die Seiten sind, auf die der *Hub* verweist, desto höher wird er selbst bewertet und desto stärker steigt sein *Hub*-Gewicht [Henzinger 2000, S. 4 f.]:

$$A(p) = \sum_{(q,p)} H(q)$$

$$H(p) = \sum_{(q,p)} A(q)$$

$H(p)$ *hub* – Wert von Seite p

$A(p)$ *authority* – Wert von Seite p

Die Trennung von *Hub*- und *Authority*-Werten wirkt sehr plausibel, jedoch besitzt der HITS-Algorithmus einige Schwächen wie die Gefahr der thematischen Entfernung durch die Integration weiterer Seiten neben dem eigentlichen Suchergebnis (*topic-drift*).

Ein weiteres Problem von HITS besteht darin, dass auch sehr schlechte Seiten immer noch einen positiven Beitrag leisten und somit in einem gewissen Maße Quantität mehr zählt als Qualität. Verweisen etwa zwei *Hubs* A und B auf zehn sehr gute *Authorities* und *Hubs* B noch zusätzlich auf zwei sehr schwache *Authority*-Seiten, so gilt intuitiv B als der schlechtere *Hub*, weil er zusätzlichen *Noise* einführt und nicht ausschließlich auf beste Seiten verweist wie der *Hub* A. HITS bewertet aber B als den besseren *Hub*. Dieses kontra-intuitive Ergebnis wird vermieden durch die Bildung des Durchschnitts aller *Authority*-Werte der Seiten, auf welche ein *Hub* verweist

[Borodin et al. 2001]. Ein analoger Verbesserungsvorschlag zielt darauf ab, zu verhindern, dass eine Seite hohe *Authority*-Werte erhält, obwohl nur viele schlechte *Hubs* auf sie verweisen. Dazu berücksichtigt das System lediglich die *Hubs*, die einen bestimmten Schwellenwert überschreiten. Dieser liegt mindestens beim Durchschnitt aller *Hub*-Werte der Seiten, die auf die aktuelle Seite verweisen. Der Algorithmus berücksichtigt nur diese *Hub*-Werte für die Berechnung der *Authority*. Analog zum *Hub-Threshold* ist also auch ein *Authority-Threshold* nötig. Sodann tragen nur die *Authorities*, welche über dem Durchschnitt liegen, zur Berechnung des *Hub*-Wertes einer Seite bei [Borodin et al. 2001].

2.3 Nachteile der Linkanalyse

Die Grundannahme der Linkanalyse besteht darin, dass der Autor einer Internetseite seine Links eher auf qualitativ gute Seiten setzt. Demnach müsste der Autor vorab eine Qualitätsüberprüfung vornehmen. Dies ist jedoch eher selten der Fall. Es ist sogar völlig unrealistisch, dass jeder Web-Autor das Ziel seiner Links ständig intensiv überprüft. Zum einen verändern sich viele Seiten häufig [Fetterly et al. 2003] und zum anderen sind besonders populäre Angebote oft sehr groß. So wird oft auf den Verzeichnisdienst Yahoo verlinkt, obwohl sicher kaum jemand vorher den gesamten hierarchischen Baum des Verzeichnisdienstes betrachtet.

Ferner herrschen in sozialen Netzwerken wie dem Internet Gesetze, die dazu führen, dass Seiten mit vielen Inlinks mit größerer Wahrscheinlichkeit wieder das Ziel von Links werden als weniger populäre Seiten [Barabási 2002]. Eine Simulation konnte zeigen, dass lediglich 10% der Wahrscheinlichkeit für das Erhalten eines Inlinks gleichmäßig auf alle Seiten verteilt ist und 90% von der Anzahl der bereits erhaltenen Links abhängt [Pennock et al. 2002]. Unter Web-Autoren bereits bekannte und populäre Seiten wachsen sehr viel stärker in ihrer Popularität bzw. in ihrem PageRank als andere Seiten (*the rich get richer* bzw. Matthäus-Effekt). Demnach ist ein Inlink nicht nur das Ergebnis der hohen Qualität einer Seite, sondern auch das Resultat eines dynamischen Prozesses beim Wachsen eines Netzes.

Teilweise bedeuten Links sogar ein explizit negatives Urteil. Links innerhalb von Newsgroups bringen meist eine negative Einschätzung zum Ausdruck. In der sozialen Struktur einer Onlinediskussion besteht eine starke Tendenz, dann auf einen Beitrag zu antworten, wenn man nicht mit ihm übereinstimmt [Agrawal et al. 2003].

Diese Argumente gegen die Linkanalyse werden auch von den bekannten Evaluierungsergebnissen gestützt. Im Rahmen des *Web Track* von TREC (<http://trec.nist.gov>) zeigte sich, dass kein Linkanalyseverfahren zu einer Verbesserung der Retrieval-Ergebnisse führte [Mandl 2003, Craswell & Hawking 2002].

3 Anwendungen von Linkanalyse

Die wichtigste Anwendung der Linkanalyse besteht im Information Retrieval. Die Qualität tritt als weiteres Merkmal für die Bestimmung des Rankings hinzu. Daneben spielen Netzwerkanalysen und Crawling eine Rolle.

3.1 Ranking

Die Benutzung mehrerer Evidenzen zur Berechnung eines endgültigen Rankings besitzt im Information Retrieval bereits Tradition [Ingwersen 1994, Mandl & Womser-Hacker 2000]. Die Fusion muss nun auch Werte aus der Linkanalyse mit der inhaltlichen Ähnlichkeit zwischen Anfrage und Dokumenten verbinden. Für die Integration der Werte muss ein Algorithmus gewählt werden. Dabei muss die Verteilung der PageRank-Werte beachtet werden, die dem Power Law folgt. Der ursprüngliche Entwurf von PageRank thematisiert den Aspekt der Kombination mit den unmittelbaren Retrieval-Ergebnissen noch nicht näher [Page et al. 1998].

Häufig wird der Linkwert mit dem Retrieval-Wert multipliziert [Zhu & Gauch 2000, Kraaij & Westerveld 2000]. Der HITS-Algorithmus ordnet die Dokumente der anfrageabhängigen Menge nach dem *Authority*-Wert [Kleinberg 1998]. Andere Implementierungen summieren etwa *Authority*- und *Hub*-Wert sowie den Retrieval *Status Value*. Unterschiedliche Verfahren wurden evaluiert, jedoch hat sich wie bei der Fusion im Standard-Retrieval noch kein optimales Verfahren herauskristallisiert [Silva et al.

2000, Plachouras & Ounis 2002, Richardson & Domingos 2002]. Eine sicher häufig genutzte heuristische und effiziente Variante ist ein Zweischrittverfahren. Dabei bildet das Ergebnis des inhaltlichen Retrieval die Basis, und die besten Treffer werden anhand eines zweiten Maßes wie des PageRanks neu geordnet (z.B. [Fagin et al. 2003]).

3.2 Netzwerkanalyse

Die Linkanalyse dient wie die Bibliometrie auch zur Analyse sozialer Strukturen wie etwa der Erkennung von zusammengehörigen Gruppen (*Web Communities*). Diese Gemeinschaften stellen thematisch oder anderweitig zusammengehörende Angebote oder Seiten dar, die sehr häufig aufeinander Bezug nehmen.

Communities lassen sich auch mit Hilfe des HITS-Algorithmus ableiten [Gibson et al. 1998]. Eine Studie wählte jeweils die zehn Seiten mit dem höchsten *Hub*-Wert und mit dem höchsten *Authority*-Wert und untersuchte Robustheit und Stabilität der Themen.

Den umgekehrten Weg geht das System *Topic* (<http://www.cs.toronto.edu/db/topic>) [Mendelson & Rafiei 2000]. Nach Eingabe einer Seite liefert es die Themen, für die diese Seite bekannt ist. *Topic* kombiniert dazu Link- und Inhaltsanalysen. Zunächst werden mit Hilfe einer Suchmaschine alle Seiten ermittelt, die auf die Seite verweisen. Aus der Kurzfassung der Seiten in der Suchmaschine (*snippet*) extrahiert das System dann die am häufigsten vorkommenden Schlagwörter und liefert diese als die Themen zurück. Damit kann sich der Benutzer einen Überblick über die Themen der Seite verschaffen, ohne sich auf die Selbstbeschreibung der Autoren verlassen zu müssen.

[Matsumura et al. 2001] untersuchen, ob sich auch Außenseiter für die innerhalb einer Community diskutierten Themen interessieren, und werten dies als Maß für die Verbreitung des Themas. [Bun & Ishizuka 2001] interessieren sich für die Änderungen innerhalb einer Gruppe von thematisch zusammengehörigen Web-Angeboten und analysieren in diesem Korpus die wichtigsten Sätze, die neu entstehende Themen am besten repräsentieren.

3.3 Crawling

Das Sammeln von Internetseiten für den Aufbau einer Suchmaschine ist keine triviale Aufgabe. Viele Betreiber wünschen einen hohen Abdeckungsgrad und damit Vollständigkeit, während Ansätze des *Focused Crawling* nach thematisch ähnlichen Seiten suchen. Dabei verwendet der Suchalgorithmus eine *Best-first*-Strategie und die Bewertungsfunktion belohnt thematisch ähnliche Seiten. Ebenso kann sich ein Crawler einer Qualitätsdefinition bedienen, um verstärkt auf qualitativ hochwertige Seiten zu stoßen. Die bisher publizierten Ergebnisse zeigen aber, dass solche Strategien momentan noch nicht ausgereift sind.

So zeigt sich bei einem Download von 500 Millionen Seiten, dass eine *Breadth-first*-Strategie zum schnellen Erreichen von Seiten mit hohen PageRank-Werten führt. Während des Crawls sank der Durchschnitt der PageRank-Werte der Seiten stetig. An den ersten drei Tagen lag der Durchschnitt der PageRank-Werte über eins und ab dann darunter. Am ersten Tag war der Wert mit 7,04 noch mehr als dreimal so hoch wie mit 2,07 am zweiten Tag. Eine *Best-first*-Suche mit PageRank als Bewertungsfunktion verstärkt diesen Effekt zwar, jedoch rechtfertigt der hohe Aufwand der PageRank-Berechnung dieses Vorgehen nicht [Najork & Wiener 2001].

Ein weiteres umfangreiches Experiment testet drei *Crawling*-Strategien, die alle eine *Best-first*-Suche realisieren und aus Perspektive des zu erwartenden Information-Retrieval-Erfolgs evaluiert werden. Eine Orientierung an inhaltlicher Ähnlichkeit führte zu weit besseren Ergebnissen als die Ausrichtung am PageRank [Menczer et al. 2001].

4 Alternative Verfahren zur automatischen Qualitätsabschätzung

Mehrere experimentelle Verfahren versuchen, die Schwächen der Linkanalyse zu überwinden und zu besseren Qualitätsabschätzungen zu gelangen. Gemeinsam ist den alternativen Verfahren, dass sie zusätzliche Wissensquellen integrieren. Meist sind dies formale Eigenschaften der Internetseiten. Dies mag überraschend erscheinen, da formale Eigenschaften zunächst wenig über den Inhalt

aussagen. Jedoch fand eine Untersuchung selbst bei Forschungsanträgen einen Zusammenhang zwischen Formalia und Erfolgsquote [Berleant 2000]. Im Internet bestimmt die Gestaltung einer Seite die Qualität der Interaktion, und das Aussehen einer Seite besitzt großen Einfluss auf die wahrgenommene Qualität. Die Ergebnisse der im Folgenden kurz vorgestellten vier Projekte zeigen, dass sich die Qualitätsdefinitionen von Benutzern durchaus auf der Basis formaler Kriterien nachbilden lassen.

4.1 Information-Retrieval-Filter

Der Ansatz von Zhu & Gauch integriert einen Ansatz zur Bewertung von Qualität als Filter in ein Information-Retrieval-System. Das Ziel wird wie folgt formuliert: »Present an approach that combines similarity-based ranking with quality ranking in distributed search environments« [Zhu & Gauch 2000]. Dieser Ansatz ist einer der wenigen, die eine komplexe Definition von Qualität realisieren. Die Autoren schlagen sechs Kriterien für Qualität vor: »*Currency, availability, information-to-noise ratio, authority, popularity, and cohesiveness*« [Zhu & Gauch 2000]. Sie stellen konkrete formale Umsetzungen dieser Aspekte vor: Die Aktualität (*currency*) wird mit Hilfe der letzten Änderung bestimmt, die dem Änderungsdatum der Datei (*timestamp*) entnommen wird. Verfügbarkeit (*availability*) wird anhand der nicht mehr erreichbaren (*dead*) Links bestimmt. Die Größe *Information-to-Noise Ratio* lässt sich am besten mit Informationsgehalt ausdrücken. Die Umsetzung berücksichtigt die Anzahl der Wörter im Text und setzt sie in Verhältnis zu der Dateigröße. Popularität (*popularity*) nennen die Autoren die Anzahl der Verbindungen, die zu einer Seite führen. Autorität (*authority*) messen die Autoren anhand intellektueller Bewertungen im Rahmen eines Internetdienstes von Yahoo (Yahoo Internet Life: <http://www.zdnet.com/yil/>), der auf einer Skala von zwei bis vier liegt. Die Kohäsion (*cohesiveness*) von Internetseiten bezieht sich auf die enthaltenen Texte und misst den inhaltlichen Zusammenhang einer Seite oder eines gesamten Angebots. Dazu suchen die Autoren die dominantesten Themen der Objekte und messen deren semantischen Abstand. Je größer der Abstand, desto geringer ist die Qualität der

Objekte. Realisiert wird die Messung der thematischen Abstände über die Ontologie eines hierarchisch gegliederten Internetkatalogs. Die meist ca. 20 Web-Angebote einer Hierarchie werden zusammengefasst und indiziert. Der entstehende Gewichtsvektor definiert eine Art Cluster-Zentroid, ein exemplarisches Dokument, das diese jeweilige Kategorie vertritt. Jede betrachtete Internetseite wird ebenfalls indiziert und mittels des Kosinus-Ähnlichkeitsmaßes werden die dazu 20 ähnlichsten Konzepte identifiziert. Der Abstand zwischen den 20 passendsten Konzepten gilt als Maßstab für die Kohäsion der Seite. Dazu wird der Abstand über die Länge des zu durchschreitenden Pfades zwischen den Konzepten gemessen und mit dem Maß der Ähnlichkeit zwischen Seite und Konzept relativiert [Zhu & Gauch 2000].

Ausgehend von diesen Definitionen steht ein Modell für verteiltes Information Retrieval im Mittelpunkt, das den Fusionsaspekt betont.

- Verteilte Informationssuche: Die Berechnung des Retrieval *Status Value* beinhaltet die Qualitätsmerkmale, die jeweils mit einem Gewicht multipliziert werden, das die Wichtigkeit des entsprechenden Merkmals wiedergibt. Die mittels des Produkts aus Termfrequenz und inverser Dokumentfrequenz berechnete Ähnlichkeit zwischen Anfrage und Dokument wird mit dem Endergebnis der Qualitätsanalyse multipliziert. Das Gewicht für die Wichtigkeit der Ähnlichkeitsmerkmale wurde folgendermaßen ermittelt: Jedes Qualitätsmerkmal wurde einzeln mit dem Ergebnis des Standard-Retrieval kombiniert. Aus den Ergebnissen wurde die Verbesserung gegenüber einem Versuch ohne Einfluss von Qualitätsaspekten gemessen.
- Site-Auswahl: Das gleiche Experiment wurde auf der Ebene der gesamten Site durchgeführt. Die Qualitätsmerkmale werden für alle Seiten innerhalb der Site berechnet und daraus durch Mittelwertbildung die Qualität der gesamten Site bestimmt.
- Fusion von Retrieval-Ergebnissen: Im vorliegenden Ansatz stellt jede Website eine Evidenzquelle dar, die mit einem entsprechenden Gewicht für ihre Güte gewichtet wird. Diese

Güte entspricht der Qualität nach den oben angeführten Kriterien. Die Gewichte der einzelnen Qualitätsaspekte der auch in diesem Fall linearen Kombination ergeben sich nach dem gleichen Prinzip wie in den vorhergehenden Experimenten.

Die Datenbasis ist nicht sehr umfangreich, für jedes der fünf Anwendungsgebiete werden vier Internetangebote aus dem kommentierten und bewerteten *Yahoo Internet Life* ausgewählt, wobei die Autoren angeben, dass die Qualität variierte. Die Anfragen stammen aus einem Logfile und spiegeln reale Benutzerbedürfnisse wider.

Der Aufbau des Experiments offenbart einige Schwächen. Das Bedürfnis, das umfassende System auch zu implementieren, führt zu einigen heuristischen Annahmen, die nur schwer zu begründen sind. Besonders die Definition von Kohäsion ist problematisch, da sie etwa eine beliebige Ontologie als alleinige Wissensquelle wählt und die Anzahl der Themen nicht begründet wird. Da zwischen den Seiten aus der Ontologie und der Testmenge für das Retrieval explizit keine Doppelung ausgeschlossen ist, kann es hierbei zu Problemen kommen. Dokumente, die den Maßstab für Kohäsion mit festlegen, können mit diesen Kriterien bewertet werden und erhalten so notwendigerweise hohe Qualitätswerte.

Die Autoritätsdefinition von Zhu & Gauch basiert auf intellektuellen Urteilen, die sich eher auf globale Qualität beziehen. Dieser Aspekt ist unter den andern fünf der einzige, hinter dem sich eine intellektuelle Einschätzung verbirgt. Alle anderen Faktoren lassen sich vollautomatisch für jede Seite erfassen, so dass die so definierte Autorität eine ernsthafte Einschränkung für die Menge der zu verarbeitenden Seiten darstellt.

Die Ergebnisse der Experimente wurden mit Standard-Evaluationsmaßen aus dem Information Retrieval bewertet. Die zurückgelieferten Seiten wurden von menschlichen Evaluatoren betrachtet und als relevant oder nicht relevant eingeordnet. Daraus wurde die durchschnittliche Precision berechnet. Alle drei Experimente liefen zunächst ohne Qualitätsmerkmale und danach mit allen Qualitätsmerkmalen einzeln ab. Fast immer ergab sich eine Verbesserung der durchschnittlichen Precision, die dann als

Gewicht der Wichtigkeit des Merkmals diente. Durch die Kombination mehrerer Merkmale ergaben sich teilweise bessere Werte als bei einzelnen Qualitätsmerkmalen, aber in keinem Fall basierte das beste Ergebnis auf allen Merkmalen. In allen drei Experimenten kam es zu unterschiedlichen Resultaten:

- Im ersten Experiment ergaben die Qualitätsmerkmale Verbesserungen von 5% bis zu 15%, wobei die vier besten Merkmale Informationsgehalt, Kohäsion, Erreichbarkeit und Aktualität waren. Die Kombination aller Merkmale konnte die durchschnittliche Precision um 20% erhöhen, während bei der Kombination der vier genannten besten Merkmale die Verbesserung 25% betrug. Interessant daran ist vor allem, dass die am häufigsten benutzten Merkmale Autorität und Popularität am schlechtesten abschneiden. Die Ergebnisse sind statistisch signifikant.
- Das Experiment mit Qualitätswerten für gesamte Sites ergab eine Erhöhung der durchschnittlichen Precision um 25% bei Berücksichtigung der Kohäsion. Dahinter lagen fast gleichauf Popularität, eine Kombination aus Erreichbarkeit, Informationsgehalt und Popularität, eine Kombination aller Merkmale, Erreichbarkeit und Informationsgehalt. Dieses Resultat weist darauf hin, dass die Definition von Kohäsion sich für gesamte Sites gut eignet.
- Im Fusionsexperiment dagegen führte die Kohäsion zu einer Verschlechterung um 10%. Nur die Popularität erreichte hier eine Verbesserung von 5%, die statistisch signifikant ist.

Unklar bleibt, ob die Ergebnisseiten auch tatsächlich von höherer Qualität waren. Die Seiten wurden offensichtlich nur auf binäre Relevanz überprüft und nicht daraufhin, ob nun bessere Seiten nachgewiesen wurden. Trotz dieser Schwächen verweisen die Ergebnisse auf interessante Tendenzen. Eine Analyse mehrerer Merkmale und ihrer Kombinationen zahlt sich demnach aus. Teilweise können auch sehr einfache Merkmale eine gute Annäherung von Qualität erreichen.

4.2 WebTango

Einen Ansatz mit wesentlich mehr Kriterien liefert das System WebTango [Ivory

& Hearst 2002]. Darin werden aus dem Blickwinkel des Web-Designs bzw. der Gebrauchstauglichkeit 157 einzelne Maße für Seiten und Sites untersucht. Ziel ist es, statistische Zusammenhänge zwischen Qualitäturteilen und den untersuchten Kriterien zu finden und die Diskrepanzen in Vorschläge für Modifikationen umzusetzen, um die entsprechenden Seiten zu verbessern. Da der Fokus auf der Gebrauchstauglichkeit der Internetseiten liegt, werden keine inhaltlichen Maße wie semantische Kohäsion untersucht. Zwar umfasst die Studie Eigenschaften von Textelementen, jedoch geht es vorwiegend um die Rezipierbarkeit und nicht den Inhalt. Dementsprechend erfassen [Ivory & Hearst 2002] z.B. die Menge an Text, die Größe der Schrift, die Komplexität des Textaufbaus sowie Ergebnisse des Syntaxprüfers Weblint.

Die Datengrundlage stammt von einem Internetpreis für populäre Seiten [Ivory & Hearst 2002]. Die mit diesem *Webby-Award* ausgezeichneten Seiten werten die Autoren als qualitativ sehr hochstehend. Insgesamt wurden ca. 5400 Seiten aus 639 Sites ausgewählt. Davon fielen jeweils ungefähr ein Drittel in die Kategorien *good*, *average* und *poor*, und für diese Zuordnung wurde ein Klassifizierer trainiert. Ein Klassifikations- und Regressionsbaum mit 14 Regeln konnte 94% der Seiten der Testmenge korrekt zuordnen.

Innerhalb der einzelnen Klassen zeigte eine statistische Analyse einige Eigenschaften der Cluster auf. Gute Seiten enthielten zum Beispiel weniger Farbanweisungen, mehr Links, längere Link-Label, mehr Interaktionselemente und verstießen häufiger gegen Standards. Ferner ergab *k-means*-Clustering drei Cluster innerhalb der Kategorie *good*. Zwei der Cluster unterschieden sich vorwiegend in der Menge an Text und der dritte Cluster stach durch die hohe Anzahl von HTML-Tabellen hervor, die meist dem Layout dienten. Zur Analyse der Sites wurden die Seiten einer *Site* zusammengefasst. Für *Sites* erzielte ein Klassifizierer eine Trefferquote von 81%. Eine statistische Analyse innerhalb der Kategorien zeigte, dass gut bewertete *Sites* in ihrer Struktur breiter angelegt waren, während schlechtere *Sites* tiefe Ebenen beinhalteten [Ivory & Hearst 2002].

Für schlechte Seiten und *Sites* versuchten die Autoren abschließend, aus

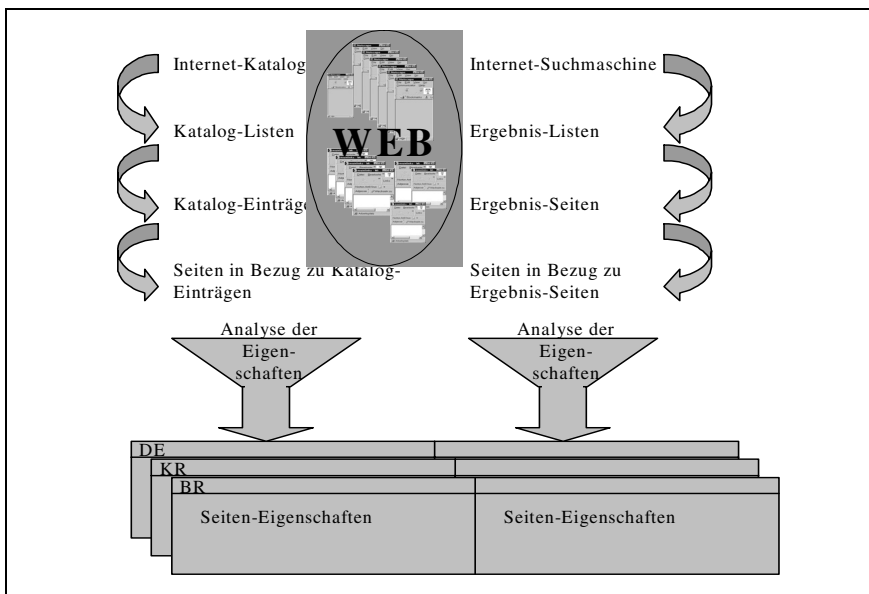


Abb. 2: Erstellung der Vergleichsdaten

den Regeln des Klassifizierers Verbesserungsvorschläge abzuleiten. WebTango ist ein gutes Beispiel für die sich etablierende empirische Forschung zum Web-Design. Aus dem Blickwinkel der Anwendung von Qualitätsfiltern im Information Retrieval wirken die gewählten Qualitätsurteile insgesamt zu positiv. Das Filtern zielt vorwiegend darauf ab, Seiten mit besonders negativer Qualität zu entfernen. Deshalb sollten Seiten, die überhaupt nicht für einen Preis wie den *Webby-Award* in Erwägung gezogen wurden, ebenfalls untersucht werden. Zudem wurden insgesamt nicht sehr viele Seiten betrachtet. Gleichwohl geht WebTango konsequent den Weg der automatischen Qualitätsbewertung anhand von maschinell extrahierten Eigenschaften und deren Korrelationen mit Expertenurteilen.

4.3 AQUAINT

Das Projekt AQUAINT (Automatic Quality Assessment for Internet Resources) geht noch einen Schritt weiter als die vorgestellten Projekte. AQUAINT stellt Information Retrieval als Anwendungsfall in den Mittelpunkt und will vor allem die subjektive Qualitätswahrnehmung untersuchen und die Suchergebnisse qualitativ verbessern. Das vorrangige Ziel besteht in der Erstellung eines Modells für Qualität, das auf menschlichen Urteilen beruht und diese weitgehend wiedergibt. Ein derartiges Modell muss mehrere Aspekte von Qualität integrieren und zu-

mindest sowohl auf die Autorität als auch die Gebrauchstauglichkeit abzielen. Anders als in anderen Projekten stehen sowohl als Ausgangsdaten als auch in der Evaluierung subjektive, menschliche Qualitätsurteile im Zentrum [Mandl 2004, Mandl 2005].

Zunächst mussten dazu Qualitätsentscheidungen erfasst werden. Besonders ergiebig hierfür sind von Redakteuren erstellte Internetkataloge und *Clearinghouses*. Die Aufnahme in einen oder mehrere solche Dienste spiegelt ein Qualitätsurteil wider. Als Vergleichsdaten dienen beliebige Seiten, die mit Hilfe einer Suchmaschine gesucht werden. In der Vergleichsmenge können natürlich auch qualitativ gute Seiten enthalten sein, die den Redakteuren aber nicht bekannt sind. Die Evaluierung darf sich deshalb nicht in der Approximation der Aufnahmeentscheidungen erschöpfen.

Wie in anderen Projekten erfolgt eine formale Analyse der Seiten nach insgesamt 110 unterschiedlichsten Kriterien, die sich automatisch erkennen lassen. Dabei wurden sowohl aus der Literatur bekannte Kriterien mit einbezogen als auch eigene, komplexe Kriterien entwickelt. Eine geringe Rolle spielte der Inhalt einer Seite, während die Analyse des HTML-Quellcodes im Zentrum stand. Diese Fokussierung hatte mehrere Gründe:

- Inhalt und Darstellung sind im Internet sehr eng verbunden. Die Bewertung des Inhalts kann daher selten

von dessen Darstellung getrennt werden.

- Die subjektive Bewertung von Internetseiten durch den Benutzer hängt in hohem Maße von visuellen Eindrücken ab. Diese lassen sich aus der Struktur der Seite ableiten.
- Der gleiche Inhalt kann bei unterschiedlicher Darstellung und Präsentation stark unterschiedlich gut benutzbar sein. Die Gebrauchstauglichkeit oder Benutzbarkeit stellt einen wichtigen Faktor der Qualität dar, und zu dessen automatischer Bewertung liegen erste Ansätze vor, die oben erläutert wurden. Die Benutzbarkeit offenbart sich zu einem Teil in der Präsentation. Die Anteile grafischer Inhalte, die Ausgewogenheit sowie die Überladenheit oder Klarheit und Einfachheit einer Seite lassen sich an dem HTML-Quellcode ablesen. Zu einem Teil gelingt dies auch automatisch.
- Die Linkanalyse bewertet die Qualität ebenfalls ohne Berücksichtigung des Inhalts. Sie hat sich in der Praxis etabliert.

Die Erstellung der Vergleichsdaten für die Qualitätsmodelle läuft in mehreren Schritten ab, die Abbildung 2 veranschaulicht. Dabei werden zunächst die Seiten eines Internetkatalogs analysiert und die darin enthaltenen Links auf externe und positiv bewertete Seiten extrahiert. Diese Seiten werden anschließend auf ihre Eigenschaften hin untersucht. Parallel extrahiert das System Seiten ohne Qualitätsurteile, indem es häufig in den Katalogseiten vorkommende Begriffe als Suchanfrage an Suchmaschinen sendet. Aus den Ergebnislisten werden die Treffer extrahiert und ebenfalls auf ihre Eigenschaften hin untersucht. Maschinelles Lernen leistet die Abbildung von den erfassten Kriterien der Seiten auf die Qualitätsurteile. Bisher wurde in AQUAINT ein umfangreiches Modell anhand des Unterpunktes *Gesundheit* des Internet-Verzeichnisdienstes Yahoo erstellt. Dabei wurde mit einem linearen Lernverfahren (*Naïve Bayes*) eine Trefferquote von ca. 70% erzielt (*ten fold cross validation*).

Die Evaluierung erfolgt im Kontext der Anwendung der Qualitätsabschätzung im Rahmen der Suche im Internet. Das Qualitätsmodell wurde in eine prototypische Suchmaschine integriert, welche

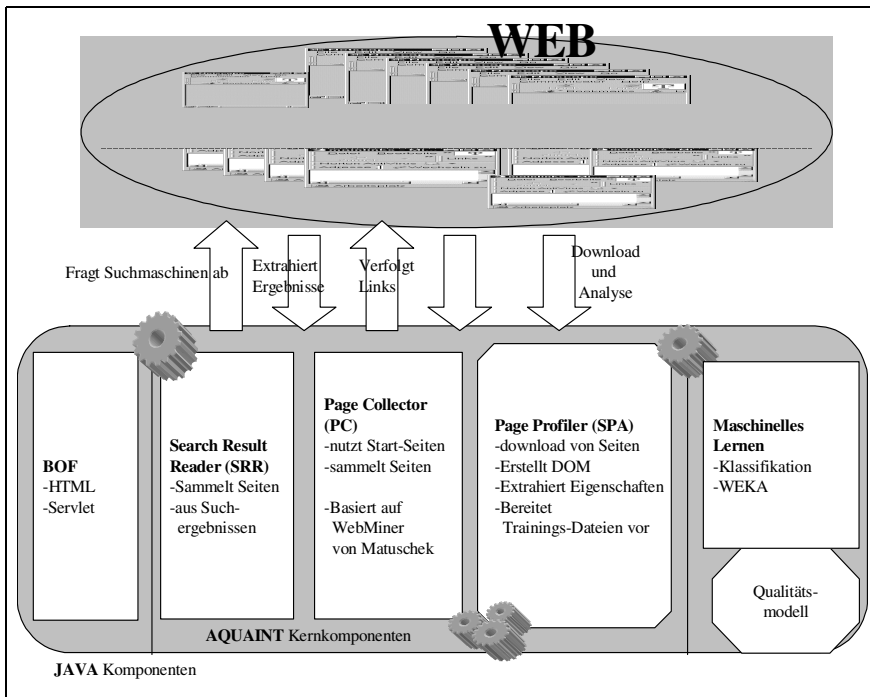


Abb. 3: Aufbau des AQUAINT-Suchsystems

die Ergebnisse verschiedener Internet-Suchmaschinen abfragt und die Ergebnisseiten mit dem Qualitätsmodell bewertet und in veränderter Reihenfolge wieder präsentiert (siehe Abb. 3). In einem Benutzertest mit 20 Testpersonen wurde die Qualitäts- und Relevanzbewertung der Ergebnisse erhoben. Dabei wurden verschiedene Suchdienste und drei unterschiedliche Re-Ranking-Strategien verwendet [Mandl 2004, Mandl 2005]. Die Auswertung steht noch aus. Anhand der Ergebnisse wird das Qualitätsmodell verfeinert.

4.4 Bloodhound

Die Popularität von Internetseiten gilt als ein wichtiger Indikator für ihre Qualität, sie ist jedoch im Vergleich zu anderen Sites schwer zu erfassen. Beschränkt sich die Analyse auf eine Site, so ergeben sich zusätzliche Möglichkeiten. Das Projekt *Bloodhound* wagt methodisch diesen Schritt und geht damit weiter als die vorher diskutierten Systeme. *Bloodhound* untersucht die Navigationsstruktur in Zusammenhang mit dem Inhalt und den Log-Dateien und kombiniert so *Usage*-, *Structure*- und *Content-Mining* [Chi et al. 2003].

Die Autoren gehen von einer Theorie für das Informationsverhalten aus, die auf

der Nahrungssuche von Lebewesen aufbaut. Demnach sind Menschen Informationsverarbeiter, die abschätzen, welche Informationsquellen bei möglichst geringem Aufwand einen hohen Ertrag bringen (*information foraging*) [Chi et al. 2000]. Dazu evaluieren sie ständig Anhaltspunkte für den Ertrag bzw. Inhalt von Information. Im Internet bewerten Benutzer vor allem Links anhand des Textes und schätzen den Wert der Zielseiten ab. *Bloodhound* bewertet nun vor allem die Übereinstimmung von zusammengehörenden Link- und Seitentexten, da Linktexte die wichtigsten Anhaltspunkte für den Informationssucher im Internet sind. Informationsspuren in und um Links bezeichnen die Autoren als *Information Scent* [Chi et al. 2000].

Ausgehend von beispielhaften Benutzeranforderungen in Form einer Anfrage analysiert das System, inwieweit der Benutzer beim Verfolgen von Links, deren Linktext oder deren Umfeld seiner Anfrage ähnelt, tatsächlich zu Seiten gelangt, die seinem Problem am ähnlichsten sind. *Bloodhound* simuliert Logfiles anhand von typischen Informationsbedürfnissen, die der Evaluator als Menge von Suchtermen vorgibt. Das System analysiert die Linkstruktur der Site und berechnet die Ähnlichkeit aller Seiten und Links zu den Anfragen. Dieser Wert spiegelt laut den

Autoren ungefähr wider, wie viele Benutzer auf ihren Suchpfaden zum Erfolg gelangen. *Bloodhound* bietet damit einen Report zur Gebrauchstauglichkeit einer Site.

In einem Benutzertest mit 240 Benutzern wurde überprüft, inwieweit die plausiblen Annahmen mit realem Benutzerverhalten übereinstimmen. Der Test umfasste acht Aufgaben für vier Sites. Die Benutzer browsen auf der Site und erreichten die verschiedenen Seiten mit einer bestimmten Frequenz. Über alle Seiten der Site ergaben die Tests für den Zugriff der Benutzer eine Häufigkeitsverteilung. Als Maßstab wurde diese Verteilung mit der von *Bloodhound* berechneten verglichen. Die Korrelation war für alle Aufgaben und Sites höher als 0,4 und in einem Drittel der Fälle über 0,8 [Chi et al. 2003].

Zwar sind diese Ergebnisse sehr positiv, jedoch berücksichtigt das Experiment nicht die Zufriedenheit der Benutzer oder deren Informationserfolg. Gleichwohl ist *Bloodhound* ein sehr viel versprechender Ansatz, der jedoch wie bereits erwähnt aufgrund der Integration von *Usage*-Daten auf eine Site beschränkt bleibt. Zudem gelten die Ergebnisse nur für die Suchergebnisse bei bestimmten Termen.

5 Fazit

Die Qualitätsbestimmung wird in zukünftigen Suchdiensten im Internet eine noch größere Rolle spielen. Die Linkanalyse wird ihren Platz finden, die Algorithmen werden sich jedoch für weitere Kriterien öffnen. Unterschiedliche Ansätze für die Bewertung der Qualität haben mit zahlreichen formalen Kriterien experimentiert. Eine wissenschaftliche Aufarbeitung bietet [Mandl 2005]. Zukünftige Verfahren zur Qualitätsbewertung müssen vor allem die hohe Dynamik im Internet beachten und für einzelne Fachgebiete oder Seitentypen optimiert werden.

Danksagung

Der Autor wurde im Projekt AQUAINT von der Deutschen Forschungsgemeinschaft (DFG) unter dem Kennzeichen MA 2411/3-1 gefördert.

6 Literatur

- [Agrawal et al. 2003] *Agrawal, Rakesh; Rajagopalan, Sridhar; Ramakrishnan, Srikant; Xu, Yirong*: Mining Newsgroups Using Networks Arising From Social Behavior. In: Proceedings of the Twelfth International World Wide Web Conference (WWW 2003), Budapest. 20.-24. Mai 2003, S. 529-535; <http://www2003.org/cdrom/papers/refereed/p688/688-agrawal/index.html>.
- [Arasu et al. 2001] *Arasu, Arvind; Cho, Jung-hoo; Garcia-Molina, Hector; Paepcke, Andreas; Raghavan, Sriram*: Searching the Web. In: ACM Transactions on Internet Technology, 1(1), 2001, S. 2-43.
- [Barabási 2002] *Barabási, Albert-László*: Linked: The New Science of Networks. Perseus, Cambridge, 2002.
- [Berleant 2000] *Berleant, Daniel*: Does Typography Affect Proposal Assessment? In: Communications of the ACM, Vol. 43, Nr. 8, 2000, S. 24-25; <http://tc.eserver.org/19908.html> (Zugriff am 15.06.2004).
- [Borodin et al. 2001] *Borodin, Allan; Roberts, Gareth; Rosendahl, Jeffrey; Tsapares, Panyiotis*: Finding Authorities and Hubs from Link Structure on the World Wide Web. In: Proceedings of the Tenth International World Wide Web Conference (WWW 10), 2001; <http://www.www10.org/cdrom/papers/314>.
- [Bun & Ishizuka 2001] *Bun, Khyou; Ishizuka, Mitsuru*: Emerging Topic Tracking System. In: Web Intelligence: Research and Development. Proceedings First Asia-Pacific Conference (WI 2001) Maebashi City, Japan, Oktober 2001. LNCS 2198. Springer-Verlag, Berlin et al., 2001, S. 125-130.
- [Chi et al. 2000] *Chi, Ed H.; Pirolli, Peter; Pitkow, James*: The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage, and Usability of a Web Site. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '00), Amsterdam, April 2000.
- [Chi et al. 2003] *Chi, E.; Rosien, A.; Supattanasiri, G.; Williams, A.; Royer, C.; Chow, C.; Robles, E.; Dalal, B.; Chen, J.; Cousins, S.*: The Bloodhound Project: Usability Issues Using the InfoScent™ Simulator. Proc ACM Conference on Human Factors in Computing Systems (CHI '03), Ft. Lauderdale, USA, 2003, S. 505-512.
- [Craswell & Hawking 2002] *Craswell, Nick; Hawking, David*: Overview of the TREC-2002 Web Track. In: Voorhees, Ellen; Buckland, Lori (Hrsg.): The Eleventh Text Retrieval Conference (TREC 2002). NIST Special Publication 500-251. National Institute of Standards and Technology. Gaithersburg, Maryland, November 2002; http://trec.nist.gov/pubs/trec11/t11_proceedings.html.
- [Fagin et al. 2003] *Fagin, Ronald; Kumar, Ravi; McCurley, Kevin; Novak, Jasmine; Sivakumar, D.; Tomlin, John; Williamson, David*: Searching the Workplace Web. In: Proceedings of the Twelfth International World Wide Web Conference (WWW 2003), Budapest. 20.-24. Mai 2003, S. 366-375; <http://www2003.org/cdrom/papers/refereed/p641/xhtml/p641-mccurley.html>.
- [Fetterly et al. 2003] *Fetterly, Dennis; Manasse, Mark; Najork, Marc; Wiener, Janet*: A Large-Scale Study of the Evolution of Web Pages. In: Proceedings of the Twelfth International World Wide Web Conference (WWW 2003), Budapest. 20.-24. Mai 2003, S. 669-678; <http://www2003.org/cdrom/papers/refereed/p097/P97%20sources/p97-fetterly.html>.
- [Gibson et al. 1998] *Gibson, David; Kleinberg, Jon; Raghavan, Prabhakar*: Inferring Web Communities from Link Topology. In: Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia, 1998; <http://cite.seer.nj.nec.com/gibson98inferring.html> (Zugriff am 22.04.2004).
- [Haveliwala 2002] *Haveliwala, Taher*: Topic-Sensitive PageRank. In: Proceedings of the Eleventh International World Wide Web Conference 2002 (WWW 2002), Honolulu, Hawaii. 7.-11. Mai 2002; <http://www2002.org/CDROM/refereed/127/> (Zugriff am 15. 06. 2004).
- [Henzinger 2000] *Henzinger, Monika*: Link Analysis in Web Information Retrieval. In: Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 23, Nr. 3, 2000, S. 3-8.
- [Henzinger et al. 2002] *Henzinger, Monika; Motwani, R.; Silverstein, C.*: Challenges in Web Search Engines. In: SIGIR Forum. 36 (2), 2002, S. 11-22.
- [Ingwersen 1994] *Ingwersen, Peter*: Polyrepresentation of Information Needs and Semantic Entities. Elements of a Cognitive Theory for Information Retrieval Interaction. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, 1994, S. 101-110.
- [Ivory & Hearst 2002] *Ivory, Melody; Hearst, Marti*: Statistical Profiles of Highly-Rated Sites. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2002), Minneapolis, USA, 20.-25. April 2002, S. 367-374.
- [Ivory & Hearst 2002] *Ivory, Melody; Hearst, Marti*: Statistical Profiles of Highly-Rated Sites. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2002), Minneapolis, USA, 20.-25. April 2002, S. 367-374.
- [Jeh & Widom 2003] *Jeh, Glen; Widom, Jennifer*: Scaling Personalized Web Search. In: Proceedings of the Twelfth International World Wide Web Conference (WWW 2003), Budapest. 20.-24. Mai 2003, S. 271-279; <http://www2003.org/cdrom/papers/refereed/p185/html/p185-jeh.html>.
- [Kleinberg 1998] *Kleinberg, Jon*: Authoritative Sources in a Hyperlinked Environment. In: Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithms, San Francisco, USA, Jan 1998, S. 668-677; <http://citeseer.ist.psu.edu/kleinberg99authoritative.html> (Zugriff am 05.05.2004).
- [Kobayashi & Takeda 2000] *Kobayashi, Mei; Takeda, Koichi*: Information retrieval on the web. In: ACM Computing Surveys (CSUR) 32 (2), 2000, S. 144-173.
- [Kraaij & Westerveld 2000] *Kraaij, Wessel; Westerveld, Thijs*: TNO/UT at TREC-9: How Different are Web Documents? In: Voorhees, E.; Harman, D.: The Ninth Text Retrieval Conference (TREC-9). NIST Special Publication 500-249. National Institute of Standards and Technology. Gaithersburg, Maryland, 2000, S. 665-671; <http://trec.nist.gov/pubs/trec9/papers/tno-ut.pdf>.
- [Mandl 2003] *Mandl, Thomas*: Neuere Entwicklungen bei der Evaluierung von Information Retrieval Systemen: Web- und Multimedia-Dokumente. In: Information – Wissenschaft und Praxis, Vol. 54 (4), 2003, S. 203-210.
- [Mandl 2004] *Mandl, Thomas*: AQUAINT: Automatische Qualitätsabschätzung für Internet-Ressourcen. Vortrag auf dem 3. Hildesheimer Evaluierungs- und Retrieval (HIER) Workshop, 21.7.2004; <http://www.uni-hildesheim.de/~mandl/hier3.html>.
- [Mandl 2005] *Mandl, Thomas*: Automatische Bewertung der Qualität von Web-Seiten. Habilitationsschrift. erscheint.
- [Mandl & Womser-Hacker 2000] *Mandl, Thomas; Womser-Hacker, Christa*: Ein adaptives Information Retrieval Modell für Digitale Bibliotheken. In: Knorz, Gerhard; Kuhlen, Rainer (Hrsg.): Informationskompetenz – Basiskompetenz in der Informationsgesellschaft. Proceedings 7. Intl. Symposium für Informationswissenschaft. (ISI 2000). 8.-10.11.2000, Darmstadt. Schriften zur Informationswissenschaft Bd. 38. Universitätsverlag, Konstanz, 2000, S. 1-16.
- [Marchand 1990] *Marchand, Donald*: Managing Information Quality. In: Wormell, Irene (Hrsg.): Information Quality. Definitions and Dimensions. Proceedings of a NORD-INFO Seminar. Copenhagen. Taylor Graham, Los Angeles, USA, 1990, S. 7-17.
- [Matsumura et al. 2001] *Matsumura, Naohiro; Ohsawa, Yukio; Ishizuka, Mitsuru*: Discovery of Emerging Topics between Communities on WWW. In: Web Intelligence: Research and Development. Proceedings First Asia-Pacific Conference (WI 2001) Maebashi City, Japan, Oktober 2001. LNCS 2198. Springer-Verlag, Berlin et al., 2001, S. 473-482.
- [Menczer et al. 2001] *Menczer, Filippo; Pant, Gautam; Srinivasan, Padmini; Ruiz, Miguel*: Evaluating Topic-Driven Web Crawlers. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 01), New Orleans, Louisiana, USA, 2001, S. 241-249; <http://informatics.buffalo.edu/faculty/ruiz/publications/p241-menczer.pdf>.
- [Mendelzon & Rafiei 2000] *Mendelzon, Alberto; Rafiei, Davood*: What Do the Neighbours Think? Computing Web Page Reputations. In: IEEE Data Engineering Bulletin, Vol. 23, Nr. 3, 2000, S. 9-16; <http://www.cs.ualberta.ca/~drafiel/papers/bull00.pdf> (Zugriff am 16. 06. 2004).
- [Najork & Wiener 2001] *Najork, Marc; Wiener, Janet*: Breadth-First Search Crawling Yields High-Quality Pages. In: Proceedings of the Tenth International Conference on World Wide Web (WWW 10), Hongkong, 2001, S. 114-118; <http://www10.org/cdrom/papers/208/> (Zugriff am 16.06.2004).

- [Oztekin et al. 2003] *Oztekin, B. Uygur; Ertoz, Levent; Kumar, Vipin*: Usage Aware PageRank. In: Poster Proceedings of the Twelfth International World Wide Web Conference (WWW 2003), Budapest, 20.-24. Mai 2003; <http://www2003.org/cdrom/papers/poster/p219/p219-oztekin.html> (Zugriff am 16.06.2004).
- [Page et al. 1998] *Page, Larry; Brin, Sergey; Motwani, R.; Winograd, T.*: The PageRank Citation Ranking: Bringing Order to the Web, 1998; <http://citeseer.nj.nec.com/page98/pagerank.html>.
- [Pennock et al. 2002] *Pennock, David; Flake, Gary; Lawrence, Steve; Glover, Eric; Giles, Lee*: Winners Don't Take All: Characterizing the Competition for Links on the Web. In: Proceedings of the National Academy of Sciences, April, 2002. Vol. 99, Nr. 8. S. 5207-5211; <http://modelingtheweb.com/modelingtheweb.pdf>.
- [Plachouras & Ounis 2002] *Plachouras, Vassilis; Ounis, Iadh*: Query-Based Combination of Evidence on the Web. In: Workshop on Mathematical/Formal Methods in Information Retrieval, ACM SIGIR Conference, Tampere, Finland, 2002; <http://ir.dcs.gla.ac.uk/terrier/publications/query-scope.pdf>.
- [Rahm 2002] *Rahm, Erhard*: Kurz erklärt: Web Usage Mining. In: Datenbank-Spektrum: Zeitschrift für Datenbanktechnologie Vol. 2, 2002; S. 75-76.
- [Richardson & Domingos 2002] *Richardson, Matthew; Domingos, Pedro*: The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In: Advances in Neural Information Processing Systems 14. MIT Press, Cambridge, MA, 2002, S. 1441-1448.
- [Silva et al. 2000] *Silva, Ilmério; Ribeiro-Neto, Berthier; Calado, Pável; Moura, Edleno; Ziviani, Nívio*: Link-Based and Content-Based Evidential Information in a Belief Network Model. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athen, 2000, S. 96-103.
- [Xue et al. 2003] *Xue, Gui-Rong; Zeng, Hua-Jun; Chen, Zheng; Ma, Wei-Ying; Zhang, Hong-Jiang; Lu, Chao-Jun*: Implicit Link Analysis for Small Web Search. In: Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 03), Toronto, Canada. 28. Juli-1. August 2003, S. 56-63.
- [Zhu & Gauch 2000] *Zhu, Xiaolan; Gauch, Susan*: Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web. In: Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 2000), Athen, Griechenland, 2000, S. 288-295.



Thomas Mandl ist seit 1998 wissenschaftlicher Mitarbeiter an der Universität Hildesheim und lehrt im Studiengang Internationales Informationsmanagement.

Er studierte Informationswissenschaft an der Universität Regensburg und war von 1995 bis 1998 am Informationszentrum Sozialwissenschaften in Bonn als Projektmitarbeiter tätig. Im Jahr 2000 promovierte er an der Universität Hildesheim über neuronale Netze im Information Retrieval und arbeitet seitdem an einer Habilitation. Seit drei Jahren leitet er die Teilnahme der Universität Hildesheim am Cross Language Evaluation Forum (CLEF).

Dr. Thomas Mandl
 Universität Hildesheim
 Informationswissenschaft
 Marienburger Platz 22
 31141 Hildesheim
 mandl@uni-hildesheim.de
<http://www.uni-hildesheim.de>