Bewertung von linktopologischen Verfahren als bestimmender Ranking-Faktor bei WWW-Suchmaschinen

DIRK LEWANDOWSKI
Heinrich-Heine-Universität Düsseldorf
Institut für Sprache und Information, Abt. Informationswissenschaft
Universitätsstraße 1, Geb. 23.21
40225 Düsseldorf
dirk.lewandowski@uni-duesseldorf.de

Zusammenfassung
Nutzerstudien haben gezeigt, dass in der Regel nur die erste Seite der von WWW-Suchmaschinen ausgegebenen Trefferlisten Beachtung findet. Dies unterstreicht die Bedeutung des automatischen Rankings durch die Suchmaschinen: Dokumente, die es nicht auf eine Top-Position der Trefferliste schaffen, finden keine oder wenigstens nur eine geringe Beachtung. Alle bedeutenden Suchmaschinen setzen deshalb als einen wesentlichen Faktor des Rankings linktopologische Verfahren ein. Diese bewerten die Qualität von Webseiten anhand ihrer Verlinkungsstruktur, wobei nicht nur die Zahl der eingehenden Links als Votum gewertet wird, sondern auch die Reputation der verweisenden Seite. Die wichtigsten linktopologischen Ansätze werden erläutert. Dabei wird insbesondere auf die Frage eingegangen, ob bestimmte Arten von Webseiten bevorzugt werden bzw. welche das sind.

Abstract
This article discusses link-based ranking algorithms in web search engines. User studies show that users usually only look at the first few hits on the search engine results pages, which underlines the importance of results ranking. Documents which do not appear in the first few results are often omitted by the users and receive only a small attention. To list the “best” documents on top of the results pages, all web search engines use link-based ranking algorithms. These are described in this article. The main question is whether there are factors implied in these algorithms that prefer certain kinds of pages.

1. EINLEITUNG

Der vorliegende Text befasst sich mit linktopologischen Rankingverfahren und deren Bewertung unter Hinblick auf die in diesen Verfahren (teils implizit) enthaltenen Wertungen, was die Eigenschaften eines „guten“ Dokuments sein sollen. Die Untersuchung steht im Kontext des Information Retrieval in Web-Suchmaschinen (Lewandowski 2005), insbesondere der Rankingverfahren, die für die Anordnung der Dokumente in den Trefferlisten sorgen.

1.1 Notwendigkeit des Rankings in Suchmaschinen

Suchmaschinen nutzen Rankingverfahren, um die als Antwort auf eine Suchanfrage gefundenen Dokumente in eine bestimmte Ordnung zu bringen, wobei die Dokumente absteigend nach ihrer Relevanz geordnet werden sollen. Rankingverfahren gab es allerdings auch schon weit vor der Zeit der Web-Suchmaschinen (vgl. Harman 1992), von besonderer Bedeutung für Suchmaschinen sind sie allerdings, da Suchmaschinen erstens

in der Regel sehr große Treffermengen zurückgeben, die vom Nutzer nicht vollständig gesichtet werden können und zweitens die Nutzer oft nicht in der Lage oder willens sind, ihre Suchanfragen so einzuschränken, dass eine überschaubare Treffermenge generiert wird.

Die Größe der Treffermengen resultiert aus der enormen Anzahl der von den Suchmaschinen erfassten Dokumente und der geringen Anzahl der formal erfassten bzw. teils erfassbaren Merkmal der Dokumente. So lassen sich Suchanfragen etwa nur sehr eingeschränkt nach Themen einschränken. Dem Nutzer ist es zwar möglich, seine Suchanfrage mit entsprechenden (Indikator-)Begriffen einzuschränken, eine Einschränkung beispielsweise durch eine Vorauswahl der zu durchsuchenden Quellen ist von sehr allgemeinen Ausnahmen wie der Suche nur nach Nachrichten oder Newsgroup-Beiträgen allerdings nicht gegeben. So ist zu verstehen, dass eine fokussierte Suche im Datenbestand der mit etwa 4,3 Milliarden Dokumenten[1] zur Zeit größten Web-Suchmaschine Google für den professionellen Rechercheur weit problematischer ist als im Datenbestand des Online-Hosts Lexis-Nexis, der mit 4,6 Milliarden Dokumenten (Lexis-Nexis 2004) eine ähnlich große Menge von Dokumenten anbietet.

Auf Seiten der Nutzer besteht eine nur mangelhafte Kenntnis der Suchmaschinen und ihrer Funktionen. Erweiterte Suchfunktionen, die bei Suchmaschinen – wenn auch nicht in gleichem Umfang wie bei Datenbank-Hosts – durchaus in nennenswertem Umfang vorhanden sind (Lewandowski 2004a), werden nur selten genutzt (Machill et al. 2003; Spink u. Jansen 2004), wobei im Falle der Nutzung eine hohe Fehlerquote zu verzeichnen ist. Auch lassen sich kaum Fortschritte der Nutzer hinsichtlich der fortgeschrittenen Nutzung der Suchmaschinen feststellen (Spink u. Jansen 2004, 79).

Den linktopologische Verfahren, wie sie in diesem Aufsatz vorgestellt werden, ist gemeinsam, dass sie das Dilemma der zu großen Treffermengen erkannt haben und zu ihrem Ausgangspunkt machen. Notwendig ist eine Sortierung der Trefferlisten nach Relevanz, wobei die zweite grundlegende Annahme lautet, dass Web-Dokumente von unterschiedlicher Qualität sind und die Dokumentinhalte unzuverlässig bis irreführend sein können.[2] Deshalb müssen Verfahren angewendet werden, die die Qualität der Dokumente bewerten und hochwertige Dokumente auf die vorderen Plätze der Trefferlisten bringen.

[1] So die Angaben auf der Startseite www.google.com [9.11.2004]
[2] Die Problematik der vertrauensunwürdigen Dokumente wird in Mintz (2002) ausführlich behandelt.

2. FORSCHUNGSUMFELD

In diesem Abschnitt wird das Thema des vorliegenden Artikels eingegrenzt und der Rahmen aufgezeigt, in dem dieses Thema steht. Es wird kurz auf die Forschungsergebnisse im Umfeld eingegangen, die die Basis der weiteren Analyse bilden.

2.1 Qualitätsmerkmale von Suchmaschinen

2.1.1 Indexgrößen

Die Qualität von Suchmaschinen lässt sich nicht durch einen einzelnen Faktor bestimmen. Während bei klassischen Online-Datenbanken die Erfassung der Dokumente in der Qualitätsbestimmung keine entscheidende Rolle spielt, da angenommen werden kann, dass die Dokumente im Rahmen der von der Datenbank abgedeckten Inhalte vollständig erfasst werden, ist die Menge erschlossenen Dokumente und deren Vertrauenswürdigkeit in Suchmaschinen von großer Bedeutung. Banal formuliert, heißt das: Dokumente, die eine Suchmaschine nicht kennt, können auch bei einer eventuell passenden Suchanfrage nicht ausgegeben werden.

Die Größe der Suchmaschinen-Datenbanken stellt das einzige direkt vergleichbare objektive Qualitätsmerkmal dar. Die Größe der Indizes wird von den Suchmaschinen (auch zu Werbezwecken) veröffentlicht[3]; diese Daten können nach den Ergebnissen einer von Greg Notess durchgeführten Untersuchung (Notess 2003a, Notess 2003b) weitgehend als zuverlässig gelten.

2.1.2 Präzision bei der Beantwortung von Suchanfragen

Die Qualität von Suchmaschinen (auch die von Information-Retrieval-Systemen allgemein) wird oft mit Hilfe von Retrievaltests gemessen. Dabei werden die Ergebnisse ausgewählter Anfragen auf ihre Relevanz hin beurteilt und die Suchmaschinen nach der Anzahl der ausgegebenen relevanten Dokumente bewertet. Der Recall, in seiner genauen Bestimmung sowieso problembehaftet, kann bei Suchmaschinen in der Regel nicht gemessen werden, so dass sich die Tests auf den Precision-Wert beschränken. Dieser wird in den meisten Fällen nur bis zu einem bestimmten Cut-Off-Wert ermittelt, in der Untersuchung von Griesbaum et al. (2002) beispielsweise werden die ersten 20 Treffer jeder Suchanfrage untersucht.

Bei den Retrievaltests fällt auf, dass es zwar Unterschiede zwischen den untersuchten Systemen gibt, diese jedoch nicht allzu groß ausfallen (vgl. z.B. Griesbaum et al. 2002; Griesbaum 2004).

2.1.3 Abfragesprachen und erweiterte Suchfunktionen

Die Qualität von Suchmaschinen bestimmt sich auch durch ihre Abfragemöglichkeiten. Zwar werden erweiterte Suchformulare und Kommandos nur selten genutzt (Spink u. Jansen 2004), ihre grundsätzliche Nützlichkeit bzw. bei manchen Anfragearten sogar Notwendigkeit ist jedoch unumstritten. Lewandowski (2004a) gibt einen Überblick der Abfragesprachen der wichtigsten Suchmaschinen, ordnet diese in Klassen und zeigt, dass sich hinsichtlich des Vokabulars und der Abfragemöglichkeiten der Suchmaschinen bisher kein umfassender Standard herausgebildet hat. Vielmehr gibt es einige Funktionen, über die ein Konsens zu herrschen scheint, während andere Abfragemöglichkeiten nur von wenigen oder gar nur von einer Suchmaschine angeboten werden.

2.1.4 Zugriff auf unterschiedliche Datenbestände

Suchmaschinen greifen nicht mehr nur auf eine einzige Datenbank zu (Lewandowski 2004b). Die Erweiterung der Datenbestände wurde nötig, um dem Nutzer eine differenzierte Recherche gemäß seinem momentanen Informationsbedürfnis zu ermöglichen. Die wichtigsten einzelnen Datenbestände, die bei den großen Suchmaschinen einzeln abgefragt werden können, sind die Standard-Datenbank der Webseiten, die Verzeichniseinträge und die Nachrichtensuche.

2.2 Search Engine Optimization

Suchmaschinen-Optimierung („Search Engine Optimization“, SEO) beschäftigt sich mit der Platzierung von Web-Angeboten in Suchmaschinen. Dabei werden einerseits die Inhalte der Dokumente so optimiert, dass Suchmaschinen diese leicht auswerten können (Textoptimierung). In diesem Schritt werden potentielle Suchbegriffe in den Dokumenten in einer „natürlichen“ Häufung eingestreut und an exponierter Stelle (z.B. Titel, Überschriften, Zwischenüberschriften) platziert.

Die technische Suchmaschinen-Optimierung macht die Dokumente für die Suchmaschinen *überhaupt* erschließbar, indem sie diese an die technischen

[3] Eine Übersicht findet sich in Sullivan (2003).

Möglichkeiten der Suchmaschinen anpasst. Beispielsweise werden Sites, die komplett in *Flash* erstellt wurden (für Suchmaschinen nicht bzw. nur schwer erschließbar), durch HTML-Seiten ergänzt, die von den Suchmaschinen erfasst werden können.

Im Kontext dieses Aufsatzes von größter Bedeutung ist die dritte Komponente der Suchmaschinen-Optimierung, nämlich die Optimierung der Linkstruktur, um die Seite(n) in den Suchmaschinen auf einem der ersten Plätze der Trefferliste zu platzieren. Dazu werden Linkstrukturen, wie sie auch bei einer Verlinkung durch echte Webmaster zustande kommen würden, künstlich (automatisiert) nachgebildet. Die zu platzierenden Zieldokumente bekommen dadurch einen höheren Wert in den Suchmaschinen und werden bevorzugt gelistet.

Der Grad zwischen ehrlicher Optimierung und Manipulation der Suchmaschinen ist schmal und einzelne Verfahren werden von unterschiedlichen Suchmaschinen-Betreibern auch unterschiedlich bewertet. In der Regel werden bei der Textoptimierung keine Probleme gesehen, solange sich die verwendeten Begriffe im Themenfeld des tatsächlichen Inhalts des Dokuments bewegen. Als problematischer werden manche Möglichkeiten der technischen Optimierung sowie die künstliche Erzeugung von Linkstrukturen gesehen.

2.3 Rankingverfahren

2.3.1 Klassische Rankingverfahren

Suchmaschinen stützen sich auf die aus dem „klassischen“ Information Retrieval bekannten Rankingverfahren (s. Harman 1992). Zu den klassischen Faktoren des Rankings gehören beispielsweise die Worthäufigkeit, die inverse Dokumenthäufigkeit, die Position der Suchbegriffe und deren Nähe zueinander im Dokument. Die Suchmaschinen-Betreiber haben allerdings erkannt, dass eine alleinige Verwendung dieser Faktoren nicht ausreicht, da sie kein Urteil über die Qualität der Dokumente abgeben, was aber in Dokumentkollektionen von variabler Qualität von großer Bedeutung ist.

2.3.2 Nutzungsstatistische Rankingverfahren

Nutzungsstatistische Rankingverfahren gehen vom Nutzer als demjenigen aus, der die Qualität der Dokumente am besten einschätzen kann. Sie messen beispielsweise die Häufigkeit, mit der ein Dokument in der Trefferliste angeklickt wird und nehmen an, dass Dokumente, die häufig angeklickt werden, für den Nutzer interessanter sind, als solche,

die weniger häufig angeklickt werden (Culliss 2000). Die alleinige Wertung auf dieser Basis kann bei den Suchmaschinen als gescheitert gelten, da einerseits die Manipulationsmöglichkeiten zu groß sind und andererseits aufgrund der großen Menge unterschiedlicher Dokumente die statistische Basis für die Bewertung nicht ausreichend ist. Ein „Comeback“ erleben diese Verfahren allerdings als *zusätzliches* Werkzeug zur Qualitätsbewertung; vor allem im Kontext des „personalisierten Rankings“. Dieses basiert entweder auf dem Klickverhalten eines einzelnen Nutzers oder aber auf dem einer bestimmten Nutzergruppe (vgl. Dean et al. 2002).

2.3.3 Linktopologische Rankingverfahren

Die Grundannahme aller linktopologischen Verfahren ist es, dass Dokumente, die stark verlinkt sind, von größerer Bedeutung sind als Dokumente, die wenig oder gar nicht verlinkt sind. Ein Link wird – wie in der wissenschaftlichen Zitationsanalyse auch – grundsätzlich als Empfehlung für das Dokument angesehen, auf das verwiesen wird. Es wird angenommen, dass sich aus der Verlinkungsstruktur des Web (dem sog. *web graph*) Qualitätsurteile über Dokumente ableiten lassen.

Mittlerweile arbeiten alle bedeutenden Suchmaschinen mit linktopologischen Verfahren, wobei Unterschiede in den Verfahren und ihrer Implementierung bestehen. Vorreiter bei der Verwendung dieser Verfahren war die Suchmaschine Google, bei der der auf Basis des Web-Grahen für jede Seite ermittelte *PageRank*-Wert weiterhin eine große Rolle im Ranking der Trefferlisten spielt.

3. PAGE RANK ALS BEISPIEL EINES LINKTOPOLOGISCHEN VERFAHRENS

Das PageRank-Verfahren (Page et al. 1998) ist nach seinem Erfinder Lawrence Page benannt und bildet eine wesentliche Grundlage der Suchmaschine Google (Brin u. Page 1998). PageRank ordnet jedem indexierten Dokument einen statischen PageRank-Wert zu, der also unabhängig von einer gestellten Suchanfrage besteht.

PageRank basiert auf dem Modell eines *random surfer*, also eines angenommenen Web-Nutzers, der das Web abwandert, indem er auf jeder gefundenen Seite wahllos einen Link verfolgt, um zur nächsten Seite zu kommen. Hier folgt er wieder einem Link, usw. Eine Ausnahme bildet die Möglichkeit, dass der Nutzer „sich langweilt“, die Seite verlässt und an einer neuen, zufällig gewählten Stelle des Netzes wieder einsteigt.

Der PageRank-Wert einer Seite soll die Wahrscheinlichkeit angeben, mit der dieser Nutzer auf diese Seite stößt.

3.1 Aufbau des Verfahrens

Im Gegensatz zu einer reinen Zählung der auf eine Seite eingehenden Links, die anfällig für Manipulationen wäre, errechnet sich der PageRank einer Seite durch eine Bewertung der eingehenden Links. Dabei wird angenommen, dass ein Link von einer Seite, die ihrerseits als bedeutend eingeschätzt wird (also einen hohen PageRank hat), bedeutender ist, als ein Link von einer Seite, die als unbedeutend angesehen wird. In einer Gleichung ausgedrückt lautet die Berechnung des PageRank:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

Der PageRank einer Seite PR(A) wird aus den PageRank-Werten der auf diese Seite verweisenden Dokumente berechnet, wobei jede Seite nicht ihren eigenen PageRank-Wert weitergibt, sondern diesen auf die durch sie verlinkten Seiten verteilt. T1 gibt die Anzahl der ausgehenden Links auf einer Seite an. Eine Seite „vererbt“ also ihren eigenen PageRank geteilt durch die Anzahl ihrer ausgehenden Links. Dadurch erfolgt ein Ausgleich zwischen Seiten, die viele ausgehende Links haben und solchen mit wenigen.

Für die Berechnung des PageRank einer Seite werden die PageRanks der verweisenden Seiten zusätzlich jeweils mit einem Dämpfungsfaktor reduziert, der zwischen 0 und 1 liegen kann.[4] Die so gewonnenen PageRanks der auf eine Seite verweisenden Seiten werden addiert, dazugezählt wird noch die Subtraktion von Eins und dem Dämpfungsfaktor. So wird für jedes Dokument ein PageRank-Wert ermittelt, der später im Ranking der Dokumente angewendet wird.

Die Schwierigkeit in der Berechnung der PageRank-Werte liegt in der großen Menge der zu berücksichtigenden Verweise; in die Berechnung müssen schließlich alle Verweise aus allen erfassten Dokumenten eingehen.

Zu Beginn der Berechnung wird jedem Dokument der gleiche Ausgangswert zugeteilt, wobei dieser keinen Einfluss auf die späteren Endwerte hat, die Wahl des Ausgangswerts aber die Performance verbessern kann (Page et al. 1998, 7). Basierend auf den Ausgangswerten werden für jedes Dokument nun in einem iterativen Prozess die PageRank-Werte immer weiter angenähert, wobei sich die Werte von Durchgang zu Durchgang immer weniger verändern, so dass ein Cut-Off-Wert festgelegt werden kann, nach dem die Berechnung abbricht.

[4] In Brin u. Page (1998) ist als regulärer Dämpfungsfaktor 0,85 angegeben.

3.2 Probleme

Der Hauptkritikpunkt an PageRank bezieht sich auf die Zuweisung eines statischen Werts für jede Seite (u.a. Chakrabarti 2003, 212; Haveliwala 2002). Die Werte sind von den gestellten Suchanfragen unabhängig; deshalb kann es zu hohen Trefferplätzen von Seiten führen, die zwar allgemein populär sind, jedoch für die Suchanfrage keine oder nur eine eingeschränkte Relevanz haben, während sie allerdings die Suchbegriffe enthalten. Chakrabarti spricht in diesem Zusammenhang von einer künstlichen Entkoppelung von Relevanz und Qualität (Chakrabarti 2003, 212).

Mandl (2003) kritisiert, dass sich PageRank stets auf eine einzelne Seite anstatt auf eine komplette Site bezieht. Dabei würden mitunter wichtige Dokumente innerhalb einer Site schlecht bewertet, nur weil sie weniger Links auf sich ziehen als andere Dokumente derselben Site. Als Beispiel gibt er eine innerhalb einer Site bestehende Linksammlung an, die weniger Links auf sich ziehen konnte als andere Dokumente derselben Site, deren Wert aber trotzdem unumstritten ist. Denkbar wäre eine Lösung, die solchen Seiten einer Domain einen „Bonus-Wert“ gibt, der auf dem durchschnittlichen Wert der Seiten dieser Domain basiert.

Eine Besonderheit der Suchmaschine Google ist die starke Beachtung des Linktexts der auf eine Seite eingehenden Links und dessen Zurechnung zum Text der Zielseite. Durch diese Berücksichtigung ist es einerseits möglich, Dokumenten, die keine oder nur wenige textuelle Informationen enthalten (z.B. Bilder oder Programm-Dateien), Beschreibungen zuzuordnen. Zweitens enthalten die fremden Beschreibungen von Dokumenten oft treffendere Beschreibungen bzw. Bezeichnungen des Inhalts des Zieldokuments als dieses selbst. So ließe sich die Homepage des „Bundesministeriums für Wirtschaft und Arbeit“ mittels Verfahren, die allein den Inhalt der Zieldokumente auswerten, mit einer Suchanfrage nach dem „Wirtschaftsministerium“ nicht finden. Dies wird aber dadurch möglich, dass externe Seiten dieses Wort in ihren Linktexten enthalten.

Problematisch ist dieses Verfahren insofern, als es ausgenutzt werden kann. Ein bekanntes Beispiel hierfür ist die Suche nach „miserable failure“ („Kläglicher Versager“) bei Google.[5] An erster Stelle erscheint die Biographie des amtierenden US-Präsidenten George W. Bush auf den offiziellen Seiten des Weißen Hauses. Die Begriffe tauchen auf der Seite selbst nicht auf, jedoch im Linktext von Seiten, die auf dieses Dokument

[5] http://www.google.de/search?num=50&hl=de&q=miserable+failure&meta= [5.11.2004]

verweisen.[6] Dabei sind es in diesem Fall nicht die unter diesem Suchbegriff verweisenden Seiten gewesen, die dem Zieldokument einen hohen PageRank-Wert geliefert haben, sondern andere Seiten, die unter seriösen Absichten auf die Präsidenten-Biographie verlinkt haben.

An diesem Beispiel zeigt sich, dass es durchaus problematisch ist, dass der PageRank-Wert statisch und themenunabhängig vergeben wird. Die hohe Wertschätzung kommt aus seriösen Quellen, die zusätzlichen Begriffe stammen von Seiten, die es auf die Verhöhnung von Bush abgesehen haben. Ähnliche Fälle sind in anderen thematischen Zusammenhängen, vor allem aus der Manipulation für Zwecke von Werbung und Verkauf, bekannt.

4. WEITERE LINKTOPOLOGISCHE VERFAHREN

In diesem Abschnitt werden zwei weitere linktopologische Rankingverfahren beschrieben. Allerdings können diese hier nur kurz in ihren Grundzügen dargestellt werden; eine ausführlichere Beschreibung mehrerer linktopologischer Verfahren findet sich beispielsweise in Narsingh u. Gupta (2001).

4.1 HITS

Das neben PageRank wohl bedeutendste linktopologische Rankingverfahren ist HITS („Hyperlink Induced Topic Search“; auch „Kleinberg-Algorithmus“; Kleinberg 1999). Dieses Verfahren versucht, die Einschränkungen einfacher Linkzählungen bzw. die themenunabhängige Bewertungen von Webseiten zu überwinden. Es sollen die wichtigsten Seiten (sog. *Autoritäten*) passend zum Thema der jeweiligen Suchanfrage ermittelt werden. Eine Besonderheit des Verfahrens ist die Unterscheidung von zwei Maßen, die für jedes Web-Dokument vergeben werden: ein *Hub*-Gewicht und ein *Authority*-Gewicht.[7] Ein *hub* ist ein Dokument, welches auf viele als besonders hochwertige Dokumente verweist, während eine *authority* ein solches bedeutendes Dokument ist. Die Bedeutung wird wiederum daran gemessen, wie viele *hubs* auf das Dokument verweisen. Auch diese Werte müssen also – ähnlich wie bei PageRank beschrieben – in einem iterativen Verfahren berechnet werden, da zu Beginn der

[6] Dies wird aus einer Kopie aus dem Google-Cache ersichtlich, in der explizit angegeben wird, dass die Suchbegriffe nur auf externen Seiten vorkommen, jedoch nicht im Dokument selbst (http://www.google.de/search?q=cache:GPN6xA7xUV8J:www.whitehouse.gov/president/gwbbio.html+miserable+failure&hl=de [5.11.2004]).

[7] Diese Unterteilung ist in etwa vergleichbar mit der Unterteilung von Research- und Review-Artikeln im Bereich der wissenschaftlichen Informationen.

Berechnung keine Werte bekannt sind. Die Details der Berechnung finden sich in Kleinberg 1999.

Die Ideen Kleinbergs sind in der Suchmaschine Teoma[8] umgesetzt. Inwieweit das Ranking in dieser Suchmaschine exakt nach dem Kleinberg-Algorithmus abläuft, kann rekursiv nicht überprüft werden; klar ist jedoch, dass sich Teoma die Unterscheidung von Hubs und Authorities zu eigen gemacht hat. Ein frühes Paper über diese Suchmaschine (damals noch unter dem Namen „DiscoWeb“) bezieht sich explizit auf Kleinbergs Text (Davison et al. 1999).

Die Ergebnisse in Teoma werden unterteilt in *results*, *resources* und Vorschläge zur Verbesserung der Suchanfrage. Unter *results* werden die u.a. nach ihrer Autorität sortierten Suchergebnisse angezeigt, die *resources* entsprechen den Kleinberg'schen *hubs*.

4.2 Hilltop

Bharat und Mihaila (2001) stellen mit „Hilltop“ ein Verfahren vor, das die besten Seiten zu populären Themen finden soll. Dabei gehen sie davon aus, dass zu populären Suchanfragen von Suchmaschinen potenziell zu viele Ergebnisse zurückgegeben werden, während doch aus dem Nutzerverhalten bekannt ist, dass die Nutzer nur die ersten zehn bis höchstens 20 Treffer sichten. Das Verfahren ist deshalb darauf angelegt, eine hohe Precision zu erreichen und dabei auf einen hohen Recall zu verzichten. Dazu sollen nur solche Dokumente zurückgegeben werden, die von „unabhängigen Experten“ für gut befunden wurden. Das Verfahren soll Seiten finden, deren Ziel es ist, auf relevante Dokumente zu einem Thema hinzuweisen.

Konzeptionell ist dies den Kleinberg'schen *Hubs* vergleichbar, handelt es sich doch um Seiten, die als wichtigstes Element Links auf *Autoritäten* enthalten. Im Hilltop-Algorithmus werden alle Verweise, die von „expert pages“ ausgehen, gezählt. Je mehr Links von Experten eine Seite auf sich ziehen kann, desto höher steht sie schließlich im Ranking. Bei diesem Verfahren besteht allerdings die Gefahr, dass zu einer Anfrage keine Dokumente gefunden werden, weil schlicht nicht genügend Experten-Seiten zur Verfügung stehen, um ein sinnvolles Ranking zu ermöglichen.

[8] www.teoma.com

5. PROBLEMBEREICHE LINKTOPOLOGOSCHER RANKINGVERFAHREN

In diesem Abschnitt werden einige Problembereiche linktopologischer Verfahren dargestellt, wobei nur grundsätzliche Probleme behandelt werden, die sich aus den *Grundannahmen* dieser Verfahren ergeben. Auf eine kritische Würdigung der Funktionstüchtigkeit einzelner Algorithmen wird bewusst verzichtet.

5.1 Qualitätsmodelle

Die bekannten linktopologischen Verfahren wie PageRank und HITS definieren die Qualität von Dokumenten als deren Autorität bzw. abgestufte Popularität. Dieser Qualitätsbegriff lässt alle weiteren Faktoren außer Acht und beschränkt sich auf die Maßstäbe, die bereits im klassischen Citation Indexing verwendet wurden. Mandl (2003) sieht die Gründe für die Popularität dieser Bewertung in der relativ leicht möglichen Extraktion der Linkstruktur, dem Rückgriff auf etablierte bibliometrische Verfahren und die hohe Plausibilität der Grundidee.

5.2 Motivationen für das Setzen von Links

Linktopologische Verfahren sehen jeden Link als eine „Empfehlung“ für das Dokument an, auf welches verwiesen wird. Allerdings gibt es durchaus auch andere Gründe, auf eine Seite zu verlinken. An erster Stelle ist hier schlicht die Navigation zu nennen. Links werden gesetzt, um eine Website zu erschließen und übersichtlich zu gestalten und damit dem Nutzer die Möglichkeit zu geben, sich in diesem Informationsraum zu bewegen.

Schwerer ins Gewicht fallen bei der Bewertung von Links diejenigen, die zwar inhaltlich vergeben werden, jedoch keine originäre Empfehlung darstellen. Links werden beispielsweise als abschreckendes Beispiel gesetzt, um besonders schlechte Dokumente hervorzuheben oder vor diesen zu warnen. Linktopologische Rankingverfahren können nicht zwischen Empfehlungen und solchen Warnungen unterscheiden.

Weiterhin werden Links aus Gefälligkeit oder aus Gründen der Werbung gesetzt. Dabei ist nur schwer zu entscheiden, wo die Manipulation der Suchmaschinen beginnt und wo es noch in Ordnung ist, der Popularität der eigenen Seite ein wenig nachzuhelfen. Jede Bitte um einen Link könnte in diesem Sinne als eine Manipulation betrachtet werden, umgekehrt wäre es aber auch möglich, den Linkaustausch liberal zu sehen und hier keine oder nur eine sehr weite Grenze zu setzen.

5.3 Wertigkeit einzelner Links

In linktopologischen Verfahren werden alle Links als gleichwertig angesehen. Dies bedeutet einerseits, dass beispielsweise die Position eines Links innerhalb eines Dokuments keine Rolle spielt, obwohl die Position für den Nutzer durchaus eine Rolle spielt und seine Aufmerksamkeit lenkt (Chakrabarti 2003, 219). Links, die an exponierter Stelle eines Dokuments stehen, werden mit einer höheren Wahrscheinlichkeit geklickt als solche, die eher versteckt platziert sind. Dies wird allerdings von den linktopologischen Verfahren nicht berücksichtigt.

5.4 Bevorzugen bestimmter Seiten beim Setzen von Links

Beim Setzen von Links werden diejenigen Seiten bevorzugt, die bereits gut durch Suchmaschinen gefunden werden bzw. die eine hohe Wahrscheinlichkeit haben, überhaupt von einem Nutzer angesehen zu werden. Hier ist an das oben angesprochene Random-Surfer-Modell zu denken. Neue Links werden also nicht gleichmäßig auf alle Seiten verteilt, sondern es liegt ein *preferential attachment* (bevorzugte Anfügung) vor.

In der Untersuchung von Pennock et al. (2002) wird allerdings festgestellt, dass zwar tatsächlich *preferential attachment* vorliegt, allerdings wird dies relativiert, wenn statt des gesamten Web-Graphen nur Teilgraphen, die ein bestimmtes Thema abbilden, betrachtet werden. So wurde bei der Untersuchung von Universitäts- und Unternehmens-Homepages herausgefunden, dass sich dort nicht wie bei vorliegendem *preferential attachment* die meisten Links auf nur wenige Seiten verteilen, sondern eine hohe Anzahl von Seiten existiert, die eine mittlere Anzahl von Links auf sich ziehen kann.

5.5 Bearbeitung unterschiedlicher Anfragetypen

Die Anfragen an Suchmaschinen lassen sich auf verschiedene Weise unterteilen. Broder (2002) schlägt ein einfaches Modell vor, indem er Suchanfragen in navigationsorientierte, informationsorientierte und transaktionsorientierte Anfragen einteilt In zwei Untersuchungen (Nutzerbefragung und Logfile-Analyse) werden die gestellten Anfragen jeweils einer der Klassen zugeordnet. Die Auswertung ergibt, dass auf jede Klasse ein nennenswerter Anteil von Suchanfragen entfällt (vgl. Tabelle 1). Die Ergebnisse werden durch die Logfile-Analysen von Spink u. Jansen bestätigt, die eine zunehmende Anzahl von navigationsorientierten Anfragen verzeichnen (Spink u. Jansen 2004, 77).

Navigationsorientierte Anfragen fragen nach einer bestimmten Webseite, die aufgespürt werden soll, beispielsweise nach der Homepage des Weißen Hauses.

Informationsorientierte Anfragen fragen nach einer Menge von Dokumenten, die zu einem Thema Auskunft gibt. Transaktionsorientierte Anfragen schließlich zielen beispielsweise auf einen Buchungs-, Bestell- oder Downloadvorgang, also auf eine Transaktion im weiteren Sinne, ab.

Ein Ranking mittels linktopologischer Verfahren entfaltet seine Stärken bei den navigationsorientierten Anfragen. In einer Untersuchung wurde gezeigt, dass ein linktopologisches Verfahren nur bei der Suche nach Homepages Vorteile gegenüber anderen Verfahren bringt (Savoy u. Rasolofo 2000).

Tabelle 1 Arten von Suchanfragen und ihre Häufigkeit (Broder 2002)

Type of query	User Survey	Query Log Analysis
Navigational	24,5%	20%
Informational	?? (estimated 39%)	48%
Transactional	>22% (estimated 36%)	30%

5.6 Integration neuer Dokumente in den Index

Während klassische Rankingverfahren neue wie auch alte Dokumente gleich behandeln, ergibt sich bei linktopologischen Verfahren das Problem, dass neue Dokumente oft nur einen Link (nämlich von der eigenen Website) haben. Diese werden dann aufgrund der fehlenden In-Links niedriger gewichtet als bereits durch eine umfangreiche Verlinkung „etablierte“ Dokumente.

Zwar werden von den Suchmaschinen hier Ausgleichsfaktoren angewendet (vgl. Lewandowski 2004c), tendenziell sind jedoch ältere Dokumente trotzdem im Vorteil. Dazu kommt, dass bereits stark verlinkte Dokumente eher gefunden werden und damit die Wahrscheinlichkeit steigt, dass Links auf sie gesetzt werden (s. Abschnitt 5.4).

6. FAZIT

Linktopologische Verfahren sind aus den Suchmaschinen nicht mehr wegzudenken. Sie helfen, wenigstens eine eingeschränkte Qualitätsbewertung aller Dokumente vorzunehmen und haben so dazu beigetragen, die Ergebnisse der Suchmaschinen zu verbessern. Deutlich ist jedoch, dass sie sich besonders für navigationsorientierte Anfragen eignen und von ihrer Anlage her bestimmte Arten von Dokumente bevorzugen.

Auch linktopologische Rankingverfahren sind anfällig für Manipulationen bzw. „Optimierungen“. Insbesondere in dieser Hinsicht sind weitere Verbesserungen nötig, um diese zu identifizieren und im Ranking angemessen bzw. nicht zu berücksichtigen.

LITERATUR

Bharat, K.; Mihaila, G. A.: When Experts Agree: Using Non-Affiliated Experts to Rank Popular Topics. WWW10, May 1-5, 2001, Hong Kong. http://www10.org/cdrom/papers/pdf/p474.pdf [1.4.2004]

Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

Broder, A. (2002): A taxonomy of web search. SIGIR Forum 36(2). http://www.acm.org/sigir/forum/F2002/broder.pdf [12.7.2004]

Chakrabarti, S. (2003): Mining the web: Discovering knowledge from hypertext data. Amsterdam (u.a.): Morgan Kaufmann

Culliss, G. (2000): The Direct Hit Popularity Engine Technology. A White Paper. http://web.archive.org/web/20010619013748/www.directhit.com/about/products/technology_whitepaper.html [10.2.2004]

Davison, B. D.; Gerasoulis, A.; Kleisouris, K.; Lu, Y.; Seo, H.; Wu, B.: DiscoWeb: Applying Link Analysis to Web Search. http://www.cse.lehigh.edu/~brian/pubs/1999/www8/www99.pdf [26.10.2004]

Dean, J. A.; Gomes, B.; Bharat, K.; Harik, G.; Henzinger, M.: Methods and Apparatus for employing Usage Statistics in Document Retrieval / Google Inc. US Patent Application Nr. US2002/0123988 A1 (2002)

Griesbaum, J. (2004): Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. Information Research 9(4) paper 189. http://informationr.net/ir/9-4/paper189.html [3.8.2004]

Griesbaum, J., Rittberger, M., Bekavac, B. (2002): Deutsche Suchmaschinen im Vergleich: AltaVista.de, Fireball.de, Google.de und Lycos.de. In: Hammwöhner, R., Wolff, C., Womser-Hacker, C. (Hrsg.): Information und Mobilität. Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft, 201-223

Harman, D.: Ranking algorithms. – In: Frakes, W. B.; Baeza-Yates, R. (Hrsg.): Information Retrieval. Data Structures & Algorithms. – Upper Saddle River, NJ: Prentice Hall PTR, 363-392 (1992)

Haveliwala, T. H. (2002): Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithms for Web Search. WWW 2002, May 7–11, 2002, Honolulu, Hawaii, USA. http://ranger.uta.edu/~alp/ix/readings/topicSensitivePageRank.pdf [10.11.2004]

Kleinberg, J. (1999): Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46(5), 604-632

Lewandowski,, D. (2004a): Abfragesprachen und erweiterte Funktionen von WWW-Suchmaschinen. IWP - Information: Wissenschaft und Praxis 55(2), 97-102 (2004)

Lewandowski, D. (2004b): Technologie-Trends im Bereich der WWW-Suchmaschinen. Information Professional 2011: 26. Online-Tagung der DGI; Frankfurt am Main 15. bis 17. Juni 2004; Proceedings, 183-195

Lewandowski, D. (2004c): Datumsbeschränkungen bei WWW-Suchanfragen: Eine Untersuchung der Möglichkeiten der zeitlichen Einschränkung von Suchanfragen in den Suchmaschinen Google, Teoma und Yahoo. In: Bekavac, B.; Herget, J.; Rittberger, M.: Information zwischen Kultur und Marktwirtschaft: Proceedings des 9. Internationalen Symposiums für Informationswissenschaft (ISI 2004), Chur, 6.-8. Oktober 2004, S. 301-316

Lewandowski, D. (2005): Web Information Retrieval. IWP - Information: Wissenschaft und Praxis 56(1) [i. Dr.]

Lexis-Nexis (2004): Pressemitteilung vom 5.4.2004. http://www.lexisnexis.de/downloads/040405pressemitteilung.pdf [9.7.2004]

Machill, M.; Welp, C. (Hrsg.) (2003): Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. Gütersloh: Verlag Bertelsmann Stiftung

Mandl, T. (2003): Projekt Automatische Qualitätsabschätzung von Internet Ressourcen (AQUAINT). Arbeitsbericht 3/2003, Universität Hildesheim, Informationswissenschaft. http://www.uni-hildesheim.de/~mandl/Publikationen/Ab_aquaint02.pdf [2.11.2004]

Mintz, A. P. (ed.) (2002): Web of Deception: Misinformation on the Internet. Medford, NJ: Information Today

Narsingh, D.; Gupta, P. (2001): Graph-Theoretic Web Algorithms: An Overview. In: Thomas Böhme, Herwig Unger (Eds.): Innovative Internet Computing Systems, International Workshop IICS 2001, Ilmenau, Germany, June 21-22, 2001, Proceedings. Lecture Notes in Computer Science 2060 Springer, 91-102

Notess, G. (2003a): Search Engine Statistics: Database Total Size Estimates. http://www.searchengineshowdown.com/stats/sizeest.shtml [10.11..2004]

Notess, G. (2003b): Search Engine Statistics: Relative Size Showdown. http://www.searchengineshowdown.com/stats/size.shtml [10.11..2004]

Page, L., Brin, S., Motwani, R., Winograd, T. (1998): The PageRank citation ranking: Bringing order to the Web. http://dbpubs.stanford.edu:8090/pub/1999-66 [26.10.2004]

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., Giles, C. L. (2002): Winners don't take it all: Charecterizing competition for links on the web. Proceedings of the National Academy of Sciences of the United States of America 99(8), 5207-5211

Savoy, J.; Rasolofo, Y. (2000): Report on the TREC-9 Experiment: Link-Based Retrieval and Distributed Collections. http://trec.nist.gov/pubs/trec9/papers/unine9.pdf [6.7.2004]

Spink, A.; Jansen, B. J.: Web Search: Public Searching of the Web. Dordrecht: Kluwer Academic Publishers

Sullivan, D. (2003): Search Engine Sizes. http://searchenginewatch.com/reports/article.php/2156481 [2.7.2004]