Handbuch Internet-Suchmaschinen 3
D. Lewandowski (Hrsg.)
© 2013, AKA Verlag Berlin

Suchmaschinenindices

Dirk LEWANDOWSKI
Hochschule für Angewandte Wissenschaften Hamburg,
Department Information
Finkenau 35, 22081 Hamburg
dirk.lewandowski@haw-hamburg.de

Abstract. Suchmaschinen können danach unterschieden werden, ob sie einen eigenen Index, also eine eigene Datenbank der Webinhalte, aufbauen, oder nicht. Die Erstellung eines solchen Indexes verlangt massive technische und finanzielle Ressourcen, da potenziell alle Inhalte des Webs vollständig und aktuell erfasst werden sollen. In diesem Kapitel wird der Aufbau der Indices beschrieben und auf die damit verbundenen Probleme eingegangen. Es wird gezeigt, dass durch die Vermarktung von Suchmaschinenindices an andere Suchmaschinen letztlich die Vielfalt auf dem Suchmaschinenmarkt reduziert wurde, es aus wirtschaftlicher Sicht allerdings sinnvoll ist, eine Suchmaschine ohne eigenen Index zu betreiben. Dies führt zu der Forderung, im Sinne einer gewünschten Informationsvielfalt die Erstellung eines Suchmaschinenindexes, der zu fairen Konditionen für alle Anbieter zugänglich ist, öffentlich zu fördern.

Keywords. Suchmaschinen, Index, Indexierung, Informationszugang

1. Einleitung

Jede (Web-) Suchmaschine beruht auf einem Index, d. h. einer Datenbank, die die im Web aufgefundenen Dokumente verzeichnet. Da das Web viele Milliarden Dokumente enthält und es die Aufgabe von Websuchmaschinen ist, eine möglichst vollständige Abdeckung dieser Inhalte zu erreichen, ist die Erstellung und Pflege dieser Datenbanken sehr aufwendig und teuer. Daraus entsteht das Problem, dass solche Investitionen nur von wenigen Unternehmen geleistet werden können und andere Suchmaschinen, die sich dem Nutzer als vermeintlich eigenständige Suchlösungen präsentieren, keine eigenen Datenbestände (mehr) unterhalten, sondern auf den Index einer anderen Suchmaschine zurückgreifen.

Es ist zu unterscheiden zwischen dem Aufbau und der Pflege eines solchen Indexes. Beim Aufbau ist der grundlegende Datenbestand zu erstellen; hier liegt das Problem vor allem in der vollständigen Erfassung des Status quo. Bei der Pflege des Indexes spielt vor allem Aktualität eine Rolle: Neue Dokumente müssen gefunden, nicht mehr vorhandene Dokumente gelöscht und veränderte Dokumente neu indexiert werden. Dabei wird deutlich, dass ein Webindex zu jedem Zeitpunkt bereits wieder veraltet ist. Am treffendsten lässt sich dies mit der Formulierung von der „Lokalen Kopie des Web“ [1] beschreiben. Allerdings ist eine vollständige und aktuelle Kopie einer sich stetig verändernden Dokumentenmenge von unbekannter Größe schlicht nicht möglich. Die Aufgabe der Suchmaschinen ist daher eher, eine möglichst gute Annäherung an dieses Ideal zu erreichen.

In diesem Kapitel wird unter einem Suchmaschinenindex ein umfassender Datenbestand des Webs verstanden, der keine thematischen oder geografischen Grenzen hat. Im Gegensatz dazu stehen die Indizes der Spezialsuchmaschinen, die selbstgewählt solche Beschränkungen aufweisen [2]. Ebenso geht es nicht um die zahlreichen „begleitenden" Indices wie Bilder, Video, Maps, News usw. Eine heutige Websuchmaschine zeichnet sich aber eben dadurch aus, dass sie viele verschiedene Indices hat, deren Ergebnisse wiederum in der Ergebnispräsentation zusammengeführt werden (siehe [3]). Dabei soll allerdings nicht verschwiegen werden, dass manche zusätzlichen Indices, wie etwa ein Bestand der im Web vorhandenen Bilder, kaum ohne den vorherigen Aufbau eines Indexes der Webseiten (in die die Bilder eingebettet sind) zu erstellen sind.

Es gibt weitere Webindices, die nicht oder zumindest nicht primär zum Zweck der Suche eingesetzt werden. Beispielweise werden solche Indices für Analysezwecke erstellt, um etwa Trends im Netz zu beobachten ([4], [5], siehe Abschnitt 5.4). Während der Schwerpunkt in diesem Kapitel auf den Suchmaschinenindices liegt, werden die weiteren Webindices zumindest am Rande mitbehandelt. Letztlich sind der Aufbau und der Prozess der Erstellung vergleichbar; bei Suchmaschinen kommen allerdings jene Komponenten hinzu, die eine direkte Abfrage (in Echtzeit mit nur geringen Antwortzeiten) ermöglichen.

Der Rest dieses Kapitels ist folgendermaßen gegliedert: Zuerst wird beschrieben, wie Suchmaschinen ihre Daten erfassen, die in den Index geschrieben werden. Es werden unterschiedliche Methoden erläutert und vor allem das Crawling wird beschrieben. Anschließend wird auf die Evaluierung von Suchmaschinenindices eingegangen: Wie lässt sich die Qualität der Datenbestände eigentlich messen? Im nächsten Abschnitt wird der Suchmaschinenmarkt dargestellt, der unterschieden wird in einen Markt für vollständige Suchmaschinen (die von den Endnutzern abgefragt werden können) und einen Markt der Suchmaschinenindices, dessen Anbieter sich an Suchmaschinen- bzw. Suchportalbetreiber wenden. Anschließend wird ein Überblick über die bedeutenden Suchmaschinenindices gegeben, danach wird nach dem Zugang zu solchen Indices gefragt: Unter welchen Bedingungen ist es einem Anbieter möglich, auf einen anderen Suchmaschinenindex zuzugreifen und welche Daten können in welcher Form verwendet werden? Im Fazit werden die Ergebnisse des Kapitels zusammengefasst und diskutiert: Welche Konsequenzen ergeben sich für Forschung, Praxis und Politik aus der (strukturellen) Situation bei den Suchmaschinenindices? Das Kernergebnis hier ist die Forderung nach einem zu fairen Bedingungen für alle Anbieter zugänglichen Suchmaschinenindex, auf dem neue Suchmaschinen aufgebaut werden können, was die Vielfalt beim Zugang zu Informationen über Suchmaschinen stärken kann.

2. Aufbau von Suchmaschinenindices

Die Aufnahme von Inhalten in den Datenbestand einer Suchmaschine wird als *content acquisition* bezeichnet. Diese geschieht vor allem über das sogenannte *Crawling*, also das Entdecken neuer Dokumente über das Verfolgen von Links in bereits bekannten Dokumenten ([6], [7], S. 48 ff.). Allerdings können Inhalte auch auf andere Weise in den Datenbestand einer Suchmaschine gelangen, etwa durch das Auswerten von sogenannten *Feeds* und den Zugriff auf strukturierte Daten aus Datenbanken. Diese drei Methoden werden in den folgenden Abschnitten erläutert.

2.1. Crawling

Crawling bezeichnet das Auffinden von Dokumenten im Web durch das Verfolgen von Links auf bereits bekannten Seiten. Idealtypisch wird eine einzige Ausgangsseite indexiert und alle auf dieser Seite vorhandenen Links werden verfolgt; auf jeder dadurch gefundenen Seite werden wiederum alle Links verfolgt usw. Dieser Prozess wird so lange fortgesetzt, bis alle vorhandenen Dokumente gefunden wurden. Der Prozess beginnt dann von vorne, um neue und aktualisierte Dokumente zu finden.

Dieses Modell wäre in der Praxis jedoch nur gültig, wenn sich jedes Dokument von dem als Startpunkt verwendeten Dokument aus durch Verlinkungen erreichen ließe. Schon früh konnte jedoch gezeigt werden, dass auch bei der Verwendung von Mengen von Startseiten (*seed set*) bei Weitem nicht alle Dokumente des Webs erreicht werden können [8]. Wenn auch aufgrund dieses Umstands im Index Lücken auftauchen können, so bildet dieses Verfahren doch die effizienteste Methode, von Inhalten des Webs überhaupt Kenntnis zu erlangen.

Crawling beruht darauf, dass Suchmaschinen Dokumente selbstständig entdecken. Damit steht dieses Verfahren anderen Methoden der Content Acquisition gegenüber, die entweder Daten strukturiert auf Basis vorher getroffener Entscheidungen erfassen (Bsp. manuelle Erfassung aller Artikel aus bestimmten Tageszeitungen für eine Pressedatenbank) oder sich auf die Zulieferung durch Externe verlassen Feeds (s. u.).

Die Aufgabe des Crawlings besteht nicht nur darin, neue Dokumente aufzufinden, sondern auch bereits bekannte Dokumente zu überprüfen. Veränderte Dokumente müssen neu erfasst und gelöschte Dokumente aus dem Index entfernt werden.

Allerdings ist zu beachten, dass über das Crawling nicht einmal alle Dokumente, die potenziell über dieses Verfahren zugänglich sind, auch erfasst werden können. So gibt es Dokumente, die früher einmal verlinkt waren, zu denen inzwischen allerdings keine Verlinkungen mehr bestehen. Dies ist beispielsweise oft der Fall, wenn ganze Websites neu erstellt werden (*Relaunch*), die alten Dokumente aber auf dem Server belassen werden. In diesen Fällen kann nur eine Suchmaschine, die die Seiten bereits in der Vergangenheit entdeckt hat, eine vollständige Indexierung anbieten.

Ebenso wichtig sind historische Daten für die Bestimmung des Erstellungsdatums eines Dokuments. Da das Erstellungsdatum über die Dokumentinformationen bzw. seine Meta-Daten nicht zuverlässig festgestellt werden kann [9], ist es hier sinnvoll, das Datum des ersten Auffindens mit einzubeziehen. Allerdings kann auch dieses Verfahren zu gravierenden Problemen führen, die beispielhaft von Schuler [10] anhand von Google News beschrieben werden.

2.1.1. Steuerung des Crawlings durch Websitebetreiber

Das Crawling kann von Websitebetreibern zumindest zum Teil gesteuert werden. Dies kann auf der einen Seite durch einfache Anweisungen an die Suchmaschinen-Crawler in der robots.txt-Datei oder in den Meta-Daten einzelner Dokumente erfolgen. Diese beiden Verfahren werden unter der Bezeichnung *Robots Exclusion Standard* zusammengefasst. Ihnen ist gemeinsam, dass es sich um freiwillige Übereinkünfte handelt, denen die großen Suchmaschinen zwar folgen, bei denen aber keine Garantie besteht, dass etwa vom Websitebetreiber ausgeschlossene Dokumente auch tatsächlich nicht indexiert werden.

Auf der anderen Seite kann das Crawling über XML-Sitemaps gesteuert werden (siehe http://www.sitemaps.org). Diese erlauben genauere Anweisungen an die Crawler

der bekannten Suchmaschinen; auch hier gibt es allerdings keine Garantie dafür, dass die Anweisungen von den (bzw. von allen) Suchmaschinen befolgt werden.

2.1.2. Probleme des Crawlings

Trotz des relativ einfachen Modells der Linkverfolgung ergeben sich beim Crawling Probleme, von denen hier die drei wichtigsten beschrieben werden sollen.

Bei *Duplicate Content* handelt es sich um Inhalte, die bereits auf anderen Seiten vorhanden sind oder die diesen Inhalten sehr ähnlich sind. Es gilt, diese Inhalte bereits im Prozess des Crawlings zu erkennen, damit nicht erst Websites mit unter Umständen Tausenden von Dokumenten indexiert werden müssen, um dann schließlich zu erkennen, dass diese Inhalte bereits im Index vorhanden sind. Ein bekanntes Beispiel sind die zahlreichen Wikipedia-Clones, die schlicht die Inhalte aus Wikipedia kopieren und in anderem Layout (und meist mit Werbung ergänzt) präsentieren. Es zeigt sich, dass selbst die großen Suchmaschinen nicht in der Lage sind, diese Inhalte vollständig aus ihren Datenbeständen herauszuhalten (Abb. 1).

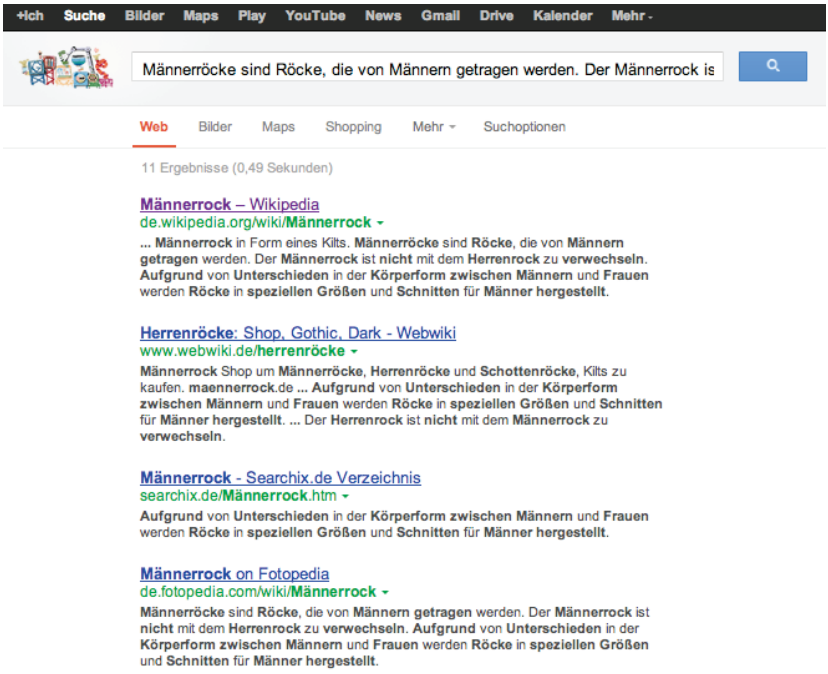

Abb. 1. Wikipedia-Clones in den Suchergebnissen von Google

Ein großes Problem für Suchmaschinen ist sogenannter *Spam*, also Inhalte, die von den Suchmaschinen nicht erwünscht sind, aber gerade zu dem Zweck erstellt werden, in den Suchmaschinen zu erscheinen. Es handelt sich dabei meist um automatisch generierte Seiten, die keine eigenständige Inhalte haben und auf den Verkauf von (oft fragwürdigen) Produkten und Dienstleistungen zielen. Die Ersteller dieser Seiten verdienen dann entweder direkt durch den Verkauf, durch Werbung oder durch das Weiterleiten der Nutzer auf eine Seite, wo ein Produkt oder eine Dienstleistung gekauft werden kann (sog. *Affiliate-Modell*).

Spider Traps sind Fallen, die dem Crawler gestellt werden, sei es absichtlich oder unabsichtlich. Das Problem besteht meist darin, dass große Dokumentenmengen indexiert werden, die für die Generierung guter Suchergebnisse nicht relevant sind. Ein einfaches Beispiel einer Spider Trap ist ein automatisch generierter Kalender, der einen Monat anzeigt und jeweils einen Link auf den nächsten Monat enthält. Wird dieser Link ausgewählt, wird eine neue Seite mit einer Darstellung des nächsten Monats generiert. Im Prinzip lassen sich so unendlich viele Seiten generieren, die ein Crawler, der darauf eingestellt ist, möglichst alle Dokumente auf einer Website zu erfassen, indexieren würde.

Schon die drei beschriebenen Problembereiche zeigen, dass sich ein Crawler zwar relativ leicht aufsetzen lässt, es jedoch schwierig ist, ein sowohl gezieltes als auch möglichst vollständiges Crawling zu betreiben. Ausführliche Beschreibungen des Crawlings finden sich in den Büchern von Baeza-Yates und Ribeiro-Neto ([6], S. 515-544) sowie Croft, Metzler und Strohman ([11], S. 31-73). Zu ethischen Fragen des Crawlings siehe Thelwall und Stuart [12].

2.2. Feeds

Feeds bieten eine Möglichkeit, dem Index einer Suchmaschine Daten in strukturierter Form hinzuzufügen. Im Gegensatz zum Crawling werden Dokumente nicht über das Verfolgen von Links aufgefunden, sondern neue Dokumente werden der Suchmaschine dadurch bekannt, dass sie in regelmäßigen Abständen bei den Betreibern von Webangeboten, die Feeds bereitstellen, anfragt, ob neue Dokumente vorhanden sind oder diese neuen Dokumente von den Anbietern direkt gemeldet werden. Man spricht im ersten Fall von einem „Ping", d. h., eine kurze Anfrage nach neuen Dokumenten wird gestellt und nur, wenn solche tatsächlich vorhanden sind, werden Daten in größerem Umfang übermittelt. Die Übermittlung erfolgt dann in strukturierter Form, d. h., es ist aufseiten der Suchmaschine nicht mehr nötig, aus unstrukturierten bzw. teilstrukturierten Dokumenten erst Strukturinformationen zu extrahieren und die Dokumente dann in den Index zu schreiben.

Feeds finden vor allem in Spezialsuchen wie der Blogsuche und der Produktsuche Verwendung. In Letztere werden ganze Produktkataloge in strukturierter Form eingespielt, was die Weiterverarbeitung und Zusammenführung der Daten unterschiedlicher Anbieter erheblich erleichtert. Suchmaschinen wie Google geben ein Standardformat vor, in dem solche Produktkataloge eingespielt werden können.

2.3. Zugriff auf strukturierte Daten aus Datenbanken

Vor allem bei Spezialsuchmaschinen [2] wird die direkte Erfassung von Dokumenten aus Datenbanken eingesetzt, prinzipiell ist sie aber auch bei Websuchmaschinen möglich. Dabei werden die Daten aus einer Datenbank in den Index der Suchmaschine integriert. Dies kann entweder dadurch geschehen, dass zwischen Suchmaschine und Datenbank eine Schnittstelle geschaffen wird, mit deren Hilfe die Daten aus der Datenbank vollständig an die Suchmaschine übermittelt werden, oder aber durch eine Erweiterung des Crawlings. Dabei sendet die Suchmaschine, sobald sie auf ein Formular auf einer Website stößt, möglichst passende Anfragen an die dahinterstehende Datenbank und indexiert die daraufhin ausgegebenen Dokumente [13]. Dieses Verfahren kann vor allem eingesetzt werden, wenn sich in der Datenbank direkt Volltextdokumente befinden; es eignet sich weniger für die Erfassung beispielsweise von bibliografischen Daten (oder allgemeiner: von Daten, die nur im Kontext der Abfrage einer Datenbank sinnvoll sind).

2.4. Vollständigkeit der Indices

Suchmaschinen streben an, einen möglichst vollständigen Index des Webs zu erstellen. Dazu benötigen sie einen Datenbestand, der idealerweise ein vollständiges und aktuelles Abbild des Webs enthält. Erst, wenn dieser Datenbestand aufbereitet ist, spricht man von dem Index einer Suchmaschine.

Suchmaschinen sind jedoch weit von diesem Ideal entfernt. Auf der einen Seite sind sie durch technische und finanzielle Gegebenheiten beschränkt, auf der anderen Seite schließen sie bestimmte Inhalte bewusst aus. In den folgenden Abschnitten werden diese Beschränkungen benannt und kurz beschrieben.

2.4.1. Technische und finanzielle Beschränkungen

Um das Web mit dem Anspruch der vollständigen Abdeckung zu crawlen, bedarf es enormer Bandbreite und großen Speicherplatzes [14]. Wie weiter unten beschrieben wird, sind die benötigten Ressourcen tatsächlich so groß, dass nur wenige Unternehmen überhaupt in der Lage sind, die dafür benötigte technische Infrastruktur aufzubauen. Hinzu kommt, dass dieser Aufbau sehr teuer und die Kosten für die Pflege des Webindexes sehr hoch sind, was sich wiederum nur wenige Unternehmen leisten können. Die Erstellung eines Webindexes ist also auf der einen Seite als technisches, auf der anderen Seite aber auch als ein finanzielles Problem anzusehen.

2.4.2. Filterung aufgrund gesetzlicher Bestimmungen

Suchmaschinen bieten in ihren Datenbeständen nicht die technisch eigentlich mögliche Vollständigkeit. Vielmehr werden Inhalte aufgrund gesetzlicher Bestimmungen gesperrt und tauchen nicht in den Suchergebnissen auf. Zum einen handelt es sich dabei um verbotene Inhalte (in Deutschland bspw. Websites, die den Holocaust leugnen), zum anderen kann es sich aber auch um als jugendgefährdend eingestufte Inhalte handeln (wie sie etwa von der Bundesprüfstelle für jugendgefährdende Medien klassifiziert werden). Diese werden in den Suchergebnissen nicht angezeigt; teils finden sich Hinweise auf ausgelassene Suchergebnisse, ohne diese jedoch konkret zu benennen. Abb. 2 zeigt eine solche Meldung der Suchmaschine Google, die am Ende der Trefferliste erscheint. Um diese Meldung zu sehen, muss erst die letzte Trefferseite aufgerufen werden (Google zeigt maximal 1.000 Ergebnisse pro Suchanfrage an).

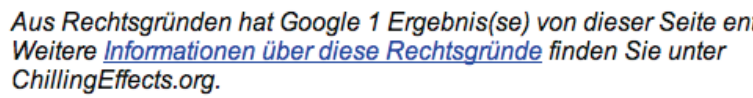

Abb. 2. Hinweis auf ein ausgeschlossenes Ergebnis am Ende der Trefferliste von Google

Inhalte, die aufgrund gesetzlicher Bestimmungen nicht angezeigt werden, lassen sich oft erreichen, indem schlicht eine andere Länderversion der gleichen Suchmaschine aufgerufen und dort gesucht wird. Die Dokumente sind also im Index enthalten, werden jedoch bei Suchanfragen aus bestimmten Ländern unterdrückt.

2.4.3. Spam

Unter Spam können verschiedene Arten von Dokumenten verstanden werden. Zumindest eine der folgenden Bedingungen muss erfüllt sein, damit man von Spam sprechen kann:

- Nutzer sollen auf ein unseriöses Angebot gelockt werden;
- Nutzern sollen falsche Tatsachen vorgespielt werden, um sie auf bestimmte Angebote zu locken;
- Nutzern soll Schadsoftware untergejubelt werden.

Spamdokumente werden meist automatisch in großer Anzahl erstellt. Das Ziel dieser Dokumente ist es, in den Suchmaschinen gerankt zu werden, damit Nutzer auf sie stoßen und dann eines der oben genannten Ziele erreicht werden kann. Es ist offensichtlich, dass es für Suchmaschinen wichtig ist, solche Dokumente möglichst schon im Crawling auszuschließen. Dabei werden sowohl automatisierte Verfahren eingesetzt, die teilweise schon im Prozess des Crawlings verwendet werden, als auch manuelle Verfahren[1]. Beiden Fällen gemeinsam ist, dass von außen nicht nachvollzogen werden kann, nach welchen Kriterien genau die Suchmaschinen Spamseiten ausfiltern. Die Grenzen zwischen Seiten, die in den Index aufgenommen werden sollen, und Spamseiten sind auch fließend: So finden sich auf der einen Seite in allen Websuchmaschinen Dokumente, die als Spam anzusehen sind, auf der anderen Seite kann keineswegs ausgeschlossen werden, dass Seiten irrtümlich als Spam klassifiziert wurden und daher von der Suchmaschine nicht angezeigt werden.

2.4.4. Filterung aufgrund von Unerwünschtheit

Suchmaschinen filtern Inhalte auch aufgrund eigener Kriterien, die sich unter dem Begriff Unerwünschtheit zusammenfassen lassen. Dabei handelt es sich um Inhalte, bei denen der Suchmaschinenbetreiber entscheidet, dass sie nicht in den Index aufgenommen (bzw. später im Ranking abgewertet) werden sollen. Auch wenn solche Entscheidungen durchaus wünschenswert sein können (siehe z. B. die Diskussion um „child model websites“ in [15]), so sind sie doch nicht transparent und basieren allein auf nicht veröffentlichten Kriterien der Suchmaschinenbetreiber.

Zwischen Unerwünschtheit und Spam besteht eine fließende Grenze. So gibt es durchaus Seiten, die für sich reklamieren könnten, originäre Inhalte bereitzustellen, jedoch von den Suchmaschinen trotzdem nicht indexiert werden.

Für die Evaluierung von Suchmaschinenindices (siehe Abschnitt 3) bedeutet der Ausschluss bestimmter Seiten aus dem Index, dass die Frage nach der Vollständigkeit der Datenbestände eigentlich nicht beantwortet werden kann: Solange man keine klaren und von allen Suchmaschinen akzeptierten Kriterien für bzw. gegen die Aufnahme in den Index hat, kann auch nicht gemessen werden, ob ein individueller Index vollständig ist.

[1] Siehe http://www.google.com/insidesearch/howsearchworks/fighting-spam.html.

2.4.5. Das Invisible Web

Sherman und Price [16] definieren das Invisible Web als denjenigen Teil des Webs, der von den allgemeinen Suchmaschinen nicht erfasst wird. Damit sind sowohl Dokumente, die von den Suchmaschinen aufgrund technischer Gegebenheiten nicht indexiert werden *können*, als auch Dokumente, die aus welchen Gründen auch immer von den Suchmaschinen nicht indexiert *werden*, dem Invisible Web zuzurechnen. Ebenso zählen alle Datenbanken, die über das Web abfragbar sind, zum Invisible Web. Bergman [17] hingegen spricht vom Deep Web, welches nur aus ebendiesen Datenbanken besteht.

In der Praxis werden die beiden Begriffe Invisible Web und Deep Web allerdings synonym gebraucht. Da die Suchmaschinen gegenüber dem von Sherman und Price beschriebenen Stand des Jahres 2001 große Fortschritte gemacht haben, kann man inzwischen die über das Web zugänglichen Datenbanken als den wesentlichen Teil des Invisible Web bezeichnen. Allerdings ist nicht bekannt, wie groß dieser für die allgemeinen Suchmaschinen nicht zugängliche Teil der Webinhalte ist. Bergmans Untersuchung von 2001 [17] hierzu wird hier immer noch häufig zitiert, auch wenn gezeigt werden konnte, dass die darin präsentierten Zahlen sicherlich zu hoch angesetzt wurden [18]. Nichtsdestotrotz handelt es sich bei den Datenbankinhalten, die *via* Web zugänglich sind (im Gegensatz zu den Inhalten, die *im* Web zugänglich sind; vgl. [19]), um einen wichtigen Bereich, der bis heute in den allgemeinen Suchmaschinen nicht verfügbar ist.

3. Evaluierung von Suchmaschinenindices

Nachdem die Suchmaschinen-indices mit ihrem Aufbau und den damit verbundenen Problemen beschrieben wurden, soll nun gezeigt werden, anhand welcher Kriterien sich solche Indices evaluieren lassen und welche empirischen Ergebnisse über die Qualität der Indices der bekannten Suchmaschinen vorliegen.

Nach dem Modell von Lewandowski & Höchstötter [20] gibt es vier Säulen der Suchmaschinenqualität:

- Qualität des Indexes,
- Qualität der Suchresultate,
- Qualität spezieller Suchfunktionen und
- Nutzbarkeit von Suchmaschinen (Usability).

In dem hier behandelten Bereich Qualität des Indexes können unter anderem die Größe des Indexes, seine Abdeckung des World Wide Web, seine Aktualität und eventuell vorhandene Verzerrungen (*bias*) gemessen werden.

3.1. Größe

Die Messung der Größe von Suchmaschinenindices ist problematisch, da erstens keine Übereinkunft darüber herrscht, welche Dokumente des Webs überhaupt indexiert werden sollen, und zweitens selbst bei Bestehen einer solchen Übereinkunft nicht klar wäre, wie groß das Web insgesamt überhaupt ist, d. h. von welcher Ausgangsmenge man bei solchen Messungen überhaupt ausgehen müsste.

Während sich die Suchmaschinenbetreiber früher gerne mit der Größe ihres jeweiligen Datenbestands brüsteten [21], ist dieser Vergleich heute nicht mehr üblich. Das

hat unter anderem damit zu tun, dass das Web mittlerweile nicht mehr nur aus statischen HTML-Dokumenten besteht, sondern Dokumente zu einem großen Teil dynamisch erzeugt werden, d. h., dass sie erst im Moment einer Anfrage dynamisch aus verschiedenen Elementen zusammengestellt werden. Da sich eine solche Zusammenstellung beliebig machen lässt, können potenziell unendliche Mengen von Dokumenten erzeugt werden, ohne dass inhaltlich irgendetwas Neues entstehen würde.

Zur Messung der Größe von Suchmaschinenindices lassen sich verschiedene Methoden einsetzen: So kann man einfach die teilweise von den Betreibern veröffentlichten Zahlen verwenden. Auch wenn die großen Suchmaschinen seit einigen Jahren keine Angaben mehr über die Größe ihrer Datenbestände veröffentlichen und diese Zahlen, gerade weil sie ja von den Betreibern selbst kommen, mit Vorsicht zu genießen sind, so sind sie doch eine einfache – und oft auch die einzige – Möglichkeit, Werte für einen Vergleich zu erhalten.

Für alle anderen Methoden werden Samplingverfahren benötigt; entweder auf der Basis von Suchanfragen oder auf der Basis von Serveradressen. Letztlich ist es noch möglich, die Größe von Suchmaschinenindices auf der Basis von Überschneidungen zu bestimmen ([22], [23], 24]).

Auch wenn die zitierten Studien bereits älter sind, so ist doch davon auszugehen, dass ihre Kernergebnisse weiterhin gültig sind: Die Größe des Webs und seine Abdeckung durch Suchmaschinen ist nicht genau zu bestimmen und ein Vergleich der Datenbestände verschiedener Suchmaschinen zeigt, dass keine Suchmaschine einen vollständigen Index des Webs besitzt.

3.2. Aktualität

Eine kontinuierliche Aktualisierung und Ergänzung der Indices ist nötig, da sich das Web beständig und relativ schnell verändert [25]. Neben der Größe ist also auch die Aktualität ein wesentliches Qualitätsmerkmal eines Suchmaschinenindexes. Allerdings konnte gezeigt werden, dass es den bekannten Suchmaschinen nicht gelingt, ihre Indices aktuell zu halten, und dass beispielsweise Seiten, die täglich aktualisiert werden, von den Suchmaschinen nur zum Teil zeitnah erfasst werden ([26], [27].

3.3. Verzerrungen in der Abdeckung

Selbst wenn eine Suchmaschine einen sowohl großen (wenn auch nicht vollständigen) als auch aktuellen Index erstellt, so bedeutet dies noch nicht, dass dieser Index das Web auch gleichmäßig abbildet [28]. Vielmehr ist nach Verzerrungen zu fragen, beispielsweise nach Sprachen oder Ländern.

In mehreren Studien ([29], [30]) konnte beispielsweise gezeigt werden, dass Websites aus unterschiedlichen Ländern von den bekannten Suchmaschinen in unterschiedlichem Maße abgedeckt wurden: US-Seiten wurden sowohl mit einer höheren Wahrscheinlichkeit *überhaupt indexiert* (d. h. mindestens ein Dokument der Website war durch die Suchmaschinen auffindbar) als auch *tiefer indexiert* (d. h. es wurden mehr Dokumente pro Website indexiert).

4. Der Suchmaschinenmarkt und der Markt für Suchmaschinenindices

Der Markt für Websuchmaschinen ist stark konzentriert und zeichnet sich vor allem in Europa durch die Dominanz eines einzigen Anbieters (Google) aus [31]. Die Situation in den USA zeigt sich mit drei wesentlichen Anbietern (Google, Bing und Yahoo) etwas anders [32], jedoch ist hier zu bedenken, dass Yahoo zwar ein eigenständiges Suchangebot betreibt, seine Suchergebnisse jedoch von Bing bezieht [33].

Die im Folgenden dargestellte Marktaufteilung nach Suchmaschinen, die auf einem eigenen Index basieren, und solchen, die auf der Zulieferung von Suchergebnissen einer anderen Suchmaschine beruhen, verdeutlicht, dass das wesentliche Problem des Marktzugangs (im Sinne eines vielfältigen Suchmaschinenmarkts mit dem Vorteil der verschiedenartigen, wenngleich relevanten, Suchergebnisse mehrerer Anbieter) in der Erstellung und Pflege des eigenen Indexes zu sehen ist. Dieser Aufwand rechnet sich für die meisten Anbieter nicht; daraus ist letztlich zu folgern, dass der Schlüssel zur Förderung von Innovation im Bereich Suche in der Bereitstellung eines unter fairen Bedingungen für alle zugänglichen Webindexes zu sehen ist.

Es gibt Anbieter, die eine eigene Suchmaschine mit einem eigenen Index betreiben. Die prominentesten Beispiele sind Google und Bing. Beide Anbieter geben im Rahmen von Geschäftsvereinbarungen allerdings ihre Indices (bzw. genauer: Suchergebnisse aus diesen Indices) an Partner weiter. Dabei handelt es sich um Suchmaschinen, die ohne eigenen Index arbeiten. Zu nennen sind hier alle bekannten Portale, die eine Websuche anbieten, wie etwa T-Online, AOL und Web.de.

Der Erfolg dieser Partnerschaften rührt daher, dass es eben sehr teuer und technisch anspruchsvoll ist, einen Webindex zu betreiben, während das Aufsetzen eines Suchinterfaces, welches auf die Daten eines anderen Anbieters zugreift, weit einfacher ist. Dazu kommt, dass das beschriebene Modell für beide Seiten ein sehr gutes Geschäft ist: Die Einnahmen aus der Werbung, die zu den Suchergebnissen ausgeliefert wird, wird zwischen den beiden Geschäftspartnern nach einem vorher festgelegten Schlüssel geteilt. Für die Suchmaschine, welche die Ergebnisse liefert, bedeutet das Ausliefern zusätzlicher Ergebnisse weder einen großen zusätzlichen Aufwand noch große Kosten: Bei immensen Kosten für die Erstellung und die Pflege des Indexes und die Bereitstellung der Infrastruktur fallen Kosten für die Auslieferung der Ergebnisseiten kaum mehr ins Gewicht.

Für die Suchmaschine, die die Ergebnisse von einem anderen Anbieter bezieht, bedeutet eine solche Vereinbarung, dass sie selbst die enormen Kosten für die Erstellung und die Pflege des Indexes nicht mehr tragen muss. Vielmehr kann sie – genügende Reichweite bei den Nutzern vorausgesetzt – mit recht geringem technischen Aufwand Geld verdienen. Es ist daher kein Wunder, dass durch dieses sehr attraktive Geschäftsmodell der Suchmaschinenmarkt in allen wichtigen Ländern deutlich bereinigt wurde bis hin zu einer Aufteilung des Marktes zwischen Google und Bing in den USA bzw. einem Quasi-Monopol von Google in den europäischen Ländern. [2]

Das prominenteste Beispiel einer solchen beschriebenen Partnerschaft ist die zwischen Bing und Yahoo. Yahoo zeigt seit 2009 Ergebnisse von Bing an, wodurch diese Suchmaschine ihren Anteil auf dem US-Markt mit einem Schlag verdoppeln konnte [34]. Auch wenn die Vereinbarung bislang noch nicht die ursprünglich festgesetzten Ziele erreicht (und Zweifel bestehen, ob die Vorgaben jemals erreicht werden können,

[2] Zur historischen Entwicklung dieses Modells in Deutschland siehe auch den Beitrag von Thomas Dominikowski in diesem Band.

siehe [35]), so dürfte sie sich letztlich doch für beide Seiten auszahlen. Die Schwäche ist auch weniger bei den Suchergebnissen zu suchen als bei der Monetarisierung durch die Werbetreffer.

An dem Fall Yahoo ist besonders interessant, dass dieses Portal erst 2004 eine eigene Suchmaschine gestartet hatte. Davor schon bezog es seine Suchergebnisse von Fremdanbietern, wobei mehrmals der Lieferant gewechselt wurde. Dann wurde die Entwicklung einer eigenen Suchtechnologie als strategisch wichtig betrachtet und von 2004 bis 2009 wurde die eigene Suchmaschine betrieben. In diesem Jahr schließlich wurde der Vertrag mit Microsoft geschlossen. Dieses Unternehmen hat eine ähnliche Geschichte hinter sich: Lange Jahre wurde eine eigene Suchmaschine nicht als strategisch wichtig angesehen, bis schließlich im Jahr 2006 mit „Live Search" der erste (erfolglose) Versuch unternommen und dann im Jahr 2009 Bing gestartet wurde.

Aktuell wird diskutiert, unter welchen Umständen Yahoo seinen Vertrag über die Verwendung von Bing-Ergebnissen auflösen könnte [36]. Abgesehen von allen vertraglichen Regelungen stellt sich dabei allerdings die Frage, von welchem Anbieter Yahoo dann seine Ergebnisse beziehen könnte: Die eigene Suchtechnologie wurde ja bereits vor Jahren aufgegeben, und als ernstzunehmende Alternative zu Bing bleibt eigentlich nur Google. Eine Vereinbarung zwischen Yahoo und Google wurde allerdings bereits im Jahr 2008 diskutiert, dann allerdings vonseiten Googles aus Bedenken gegenüber einem möglichen Kartellverfahren nicht weiter verfolgt [37].

5. Übersicht über Suchmaschinenindices

In diesem Abschnitt sollen die wichtigsten Suchmaschinenindices kurz beschrieben werden. Dabei wird von den Suchmaschinen ausgegangen, die auf Basis ihres eigenen Indexes arbeiten. Zusätzlich werden einige Indices genannt, die nicht (oder nicht primär) erstellt wurden, um eine Suchmaschine auf ihrer Basis zu betreiben.

5.1. Suchmaschinen mit eigenem Index

Wie bereits erwähnt, ist bei Weitem nicht selbstverständlich, dass Suchmaschinen auf einem eigenen Webindex basieren. Dies ist den Nutzern allerdings oft nicht bekannt und kann dazu führen, dass, wenn denn bei einer Recherche noch eine alternative Suchmaschine abgefragt wird, diese die gleichen (oder fast identischen) Ergebnisse anzeigt, da sie eben auf dem gleichen Index beruht.

Bei den Suchmaschinen mit eigenem Index sind zuerst die beiden bedeutendsten Suchmaschinen weltweit, Google und Bing, zu nennen. Beide verfügen über sehr große Datenbestände des Webs, auch wenn über die genaue Größe und Vollständigkeit kaum Angaben gemacht werden können (siehe Abschnitt 3). Auch die Betreiber selbst machen keine Angaben mehr zur Größe ihrer Datenbestände, was aufgrund der schwierigen Zählung auch einhellig begrüßt wurde [21]. Sowohl Google als auch Bing zeichnen sich durch eine internationale Abdeckung und damit verbunden durch den Anspruch, das komplette Web abzudecken, aus.

Weitere Anbieter sind durch ihre Interfaces und ihre Anpassungen an die jeweilige Sprache und Region erst einmal auf nationale Märkte fokussiert; sie sind in ihrem Index jedoch nicht auf diese Märkte beschränkt, auch wenn sich die tatsächliche Abdeckung bestimmter Länder oder Regionen nur schwer messen lässt. Unter diesen Suchmaschinen sind vor allem Yandex aus Russland und Baidu aus China zu nennen. Die-

sen ist gemein, dass sie in ihren Heimatländern dominierende Marktpositionen erreicht haben, sich zumindest bislang jedoch noch nicht international etabliert haben. Weitere Suchmaschinen mit nennenswerten eigenen Indices sind Blekko und Exalead; in der Literatur findet sich auch die koreanische Suchmaschine Naver.

5.2. Suchmaschinen ohne eigenen Index

Suchmaschinen ohne eigenen Index beziehen ihre Suchergebnisse von einer anderen Suchmaschine. Zu unterscheiden ist zwischen den eigentlichen Suchergebnissen und den Werbetreffern (Adwords). Teilweise beziehen Anbieter nur Letztere von einem Suchmaschinenbetreiber, während sie ihre Suchergebnisse selbst generieren oder ein Angebot betreiben, in dem die (Web-) Suche keine Rolle spielt.

Abb. 3 zeigt anhand des deutschen Markts beispielhaft, welche Angebote Suchergebnisse bzw. Werbetreffer von den beiden großen Suchmaschinen Google und Bing erhalten. Unter den „Suchergebnisempfängern" finden sich alle namhaften Suchportale, die auch in den üblichen Aufstellungen zu den Marktanteilen auf dem Suchmaschinenmarkt auftauchen (so etwa in der Auflistung von Webhits, [38]). Deren Marktanteile sind bei einer Zusammenstellung der Suchergebnisquellen dann allerdings den jeweils liefernden Suchmaschinen zuzurechnen.

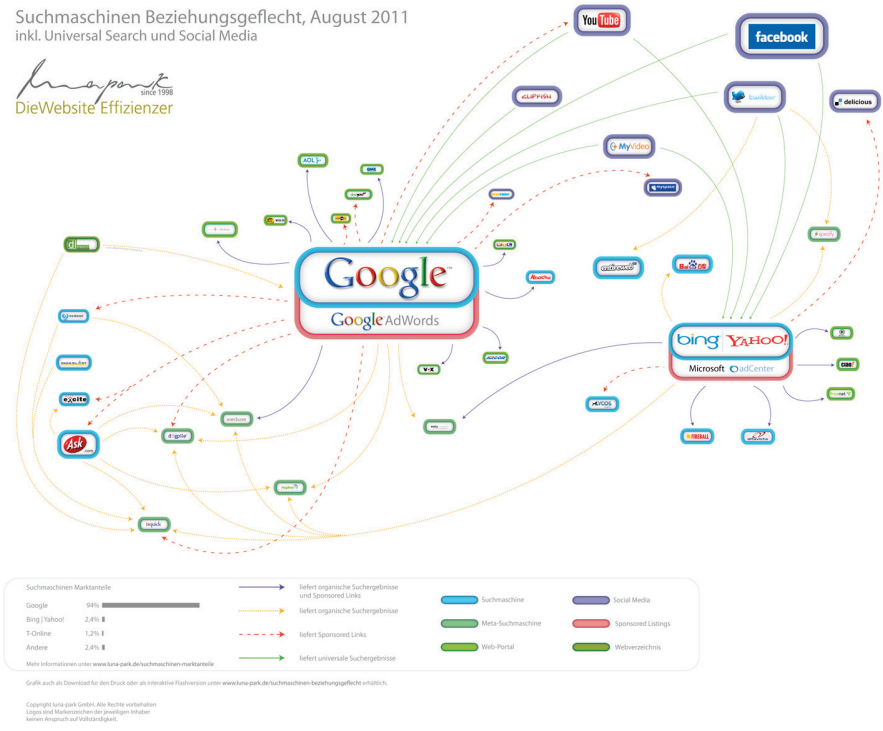

Abb. 3. Beziehungsgeflecht der Suchmaschinen (2011) [39]

Wenn eine Suchmaschine ihre primären Suchergebnisse von einem anderen Anbieter bezieht, so muss das nicht bedeuten, dass diese Suchmaschine keinen Mehrwert für ihre Nutzer bietet. Auf der einen Seite kann eine solche Suchmaschine ein anderes Suchinterface anbieten, das besser auf die Bedürfnisse ihrer Nutzerschaft abgestimmt ist; auf der anderen Seite ist es möglich, Suchergebnisse aus unterschiedlichen – auch eigenen und zugespielten – Quellen zu kombinieren. So zeigt beispielsweise die Suche von T-Online (suche.t-online.de) zwar primäre Webergebnisse von Google, reichert diese jedoch mit eigenen Universal-Search-Ergebnissen an. Genauso geht man bei Yahoo vor, wo primäre Ergebnisse von Bing kommen, allerdings um Ergebnisse aus eigenen Diensten angereichert werden. Dazu kommt, dass Yahoo davon ausgeht, dass der Kampf um die Nutzer weniger über die individuellen Suchergebnisse als über das bestmögliche Interface entschieden wird [40]. Ob Yahoo damit allerdings auf Dauer als eigenständige Suchmaschine wahrgenommen werden wird, bleibt abzuwarten.

Zu den Suchmaschinen ohne eigenen Index zählen auch die Meta-Suchmaschinen. Diese greifen auf die Indices mehrerer Suchmaschinen zurück und fragen bei diesen jeweils die Topergebnisse zu einer Suchanfrage ab. Die zurückgegebenen Ergebnisse werden dann in einem neuen Ranking zusammengefasst. Der Vorteil der Meta-Suchmaschinen soll in der größeren Abdeckung und in einem besseren Ranking liegen; diese Vorteile sind jedoch empirisch nicht nachgewiesen und werden auch auf theoretischer Ebene angezweifelt [41].

5.3. Öffentliche Webindices

Den im letzten Abschnitt diskutierten Suchmaschinen ist gemeinsam, dass sie keinen Zugriff auf den vollständigen Index der gebenden Suchmaschine(n) erhalten, sondern jeweils nur eine bestimmte Anzahl von Topergebnissen abfragen können. Dem entgegen stehen Indices, die vollständig zugänglich gemacht werden. An erster Stelle ist hier Common Crawl[3] zu nennen, ein Non-Profit-Projekt, das einen Webindex für jeden Interessierten verfügbar machen will. Die Suchmaschine Blekko trägt zu diesem Index bei, indem sie ihren Crawl an die Stiftung weitergibt.[4]

Daneben gibt es einige Crawls, die für Forschungszwecke bereitgestellt werden, beispielsweise im Lemur-Projekt[5]. Dabei handelt es sich jedoch um statische Crawls, die nicht laufend aktualisiert werden.

Für beide Arten von öffentlich verfügbaren Crawls gilt, dass sie sich nicht für den praktischen Einsatz bei einer Suchmaschine eignen. Speziell der Ansatz von Common Crawl ist aber als wichtig anzusehen, da dieses Projekt der Vorreiter für die Zugänglichmachung von Webindices sein könnte. Seine Bedeutung ist daher weniger danach zu bemessen, ob es tatsächlich gelingt, einen möglichst umfangreichen Index für die Zwecke Suche und Analyse zur Verfügung zu stellen (so das offizielle Ziel), sondern vielmehr danach, ob es gelingt, das Thema öffentlich zugängliche Webindices in die Öffentlichkeit zu bringen.

[3] http://commoncrawl.org/our-work/

[4] http://www.heise.de/newsticker/meldung/Blekko-schenkt-Crawler-Daten-der-Common-Crawl-Stiftung-1771423.html

[5] http://lemurproject.org/clueweb12/specs.php

5.4. Webindices für Analysezwecke

Indices des Webs lassen sich auch zu anderen Zwecken nutzen als zur Suche. Zu nennen ist vor allem das Webmonitoring [4], das in vielen Fällen großer Datenmengen bedarf. Es ist davon auszugehen, dass die Anforderungen an das Webmonitoring in Zukunft noch steigen werden und viele Analysen erst sinnvoll mit einem möglichst vollständigen Index des Webs durchführbar sind. Als Beispiel eines Anbieters, der sich zunächst auf die Websuche fokussierte, dann aber das Potenzial der Webanalyse auf Basis eines eigenen umfangreichen Webindex erkannt hat, ist Exalead zu nennen. Hier ist die Websuche[6] eher als Demo für die Software bzw. Dienstleistungen des Anbieters[7] zu sehen.

Auch unter Webmonitoring bzw. Webanalyse fallen viele (sozialwissenschaftliche) Forschungsprojekte, die sich der Daten aus Webindices bedienen ([42], [43], 44].

Im Bereich der Suchmaschinenoptimierung (SEO) ist es insbesondere für größere Projekte von Bedeutung, komplexe Verlinkungsstrukturen im Web zu analysieren. Da Links im Web unidirektional sind, lassen sich verlinkende Seiten nur ermitteln, wenn man zuvor einen Link-Graphen mittels Crawling aufgebaut hat. Spezialisierte Anbieter erkennen zunehmend die Potenziale eines eigenen, explizit für Analysezwecke erstellten Webindexes. So bietet etwa SEOmoz umfangreiche Tools für Suchmaschinenoptimierer an, deren Datenbasis ein eigener Webindex ist. Dieser mag nicht die Größe der Indices der großen Universalsuchmaschinen erreichen, erlaubt aber doch komplexe Auswertungen, die mittels der üblichen, frei verfügbaren Tools zur Linkanalyse nicht möglich wären.

Exklusiv für die Analyse erstellte Indices können nicht nur für diesen Zweck entsprechend aufgebaut werden, sondern sie haben auch den Vorteil, dass kein Aufwand die Durchsuchbarkeit in Echtzeit gesteckt werden muss. Während in der Suche Antworten auf Suchanfragen innerhalb von Sekunden (bzw. von weniger als einer Sekunde) gegeben werden müssen, ist die Zeit, um eine Analyse durchzuführen, als weniger kritisch anzusehen.

6. Zugang zu Suchmaschinenindices

Es konnte gezeigt werden, dass nur relativ wenige Webindices existieren und der Aufbau eines neuen Index mit erheblichen Kosten verbunden ist. Für Unternehmen (und auch Non-Profits) ist es daher wünschenswert, Zugriff auf fremde Indices zu erhalten und auf den aus diesen Indices gewonnenen Daten eigene Dienste aufzubauen.

In der Praxis gestaltet sich der Zugang zu fremden Indices allerdings schwierig. Auch in den Fällen, in denen man Zugriff auf einen Index bekommen kann, ist dieser beschränkt; dadurch sind auch die Dienste, die sich auf Basis dieser Indices erstellen lassen, von vornherein beschränkt. Es gibt bislang keinen Webindex, der fremden Anbietern (auch gegen Bezahlung) einen umfassenden Zugriff auf seine Daten erlauben würde.

Die verschiedenen Möglichkeiten, auf Suchmaschinenindices für eigene Dienste zuzugreifen, werden in den folgenden Abschnitten beschrieben.

[6] http://www.exalead.com/search/
[7] http://www.3ds.com/products/exalead/

6.1. Programmierschnittstellen (APIs)

Programmierschnittstellen (sog. APIs; *Application Programming Interfaces*) erlauben den automatischen Zugriff auf Suchmaschinenindices und die Weiterverarbeitung der so abgefragten Daten unter den Bedingungen des Anbieters.

Suchanfragen können über die API automatisch übermittelt werden; dabei kann spezifiziert werden, wie viele Treffer zurückgegeben werden, teilweise auch, in welcher Form die Trefferbeschreibungen erfolgen sollen.

Wie bei der manuellen Abfrage der Suchmaschinen ist die maximale Trefferzahl allerdings beschränkt (in der Regel auf die ersten 1.000 Treffer). Damit lassen sich die APIs nur schwer nutzen, um Analysedienste aufzusetzen, bei denen unter Umständen weit mehr Treffer benötigt werden.

Der Zugriff über die APIs erlaubt nur die Abfrage einer bestimmten Menge der Toptreffer für die gestellten Suchanfragen, d. h., die ursprüngliche Reihung der Treffer durch die gebende Suchmaschine kann zwar verändert werden, nicht jedoch auf der Basis einer vollständigen Treffermenge (in dem Sinn, dass alle der gebenden Suchmaschine bekannten, potenziell relevanten Dokumente zurückgegeben werden), sondern nur ein unter Umständen kleiner Teil davon. Gerade für Suchmaschinen, die alternative Ergebnissets bieten wollen, ist dies jedoch nicht ausreichend.

Erschwerend kommt hinzu, dass die APIs der Suchmaschinen (vor allem die von Google, [45]) nicht die gleichen Treffer liefern, die bei einer manuellen Suchanfrage angezeigt würden. Es ist nicht klar, ob der vollständige Index abgefragt wird oder nur ein Teil.

Die Suchmaschinen bieten APIs in der Regel auch nicht für alle Kollektionen an, sodass beispielsweise zwar auf den Webindex zugegriffen werden kann, jedoch nicht auf die Nachrichten- oder Videobestände.

Suchmaschinen, die einen Zugriff auf ihren Webindex mittels API erlauben, sind beispielsweise Google, Bing und Yandex. Einzig Bing bietet einen allerdings kostenpflichtigen Zugang, was auch Anfragen in großem Umfang möglich macht. Bei den anderen Anbietern besteht eine Beschränkung hinsichtlich der maximalen Anzahl der Anfragen pro Tag, was eine umfangreichere kommerzielle Nutzung unmöglich macht.

6.2. Scraping

Unter Scraping versteht man das automatische Erfassen von HTML-Seiten und die (Rück-) Gewinnung von Informationen aus diesen. Im Bereich der Suchmaschinen werden automatisiert Suchanfragen an eine bestimmte Suchmaschine geschickt und eine bestimmte Anzahl von Suchergebnissen angefordert (Details zum Scraping von Suchergebnisseiten finden sich in [46] und [47]). Die von der Suchmaschine zurückgegebenen HTML-Seiten werden analysiert und beispielsweise die URLs aller ausgegebenen Treffer in eine Datenbank geschrieben. Dieses Verfahren simuliert im Gegensatz zum Zugriff über APIs echte Nutzeranfragen, d. h., die Suchmaschine gibt eine HTML-Seite zurück, wie sie auch auf die Anfrage eines Nutzers ausgegeben werden würde. Man erhält also die „echten" Suchergebnisse, die dann weiterverarbeitet werden können.

Ein Nachteil des Scrapings ist, dass bei jeder Veränderung des HTML-Codes der Suchergebnisseiten der Suchmaschinen der Scraper entsprechend angepasst werden muss. Da die Websuchmaschinen ihre Suchergebnisseiten kontinuierlich verändern, ist eine regelmäßige Anpassung in recht kurzen Abständen nötig. Ein weiterer gravierender Nachteil ist, dass kommerzielle Suchmaschinen in der Regel keine automatisch ge-

stellten Suchanfragen (oder nur bis zu einer bestimmten Menge) erlauben. In der Praxis existieren allerdings zahlreiche (auch kommerzielle) Dienste, die besonders auf Daten, die auf Suchergebnisseiten von Google gescraped wurden, aufbauen. Die Verletzung der Nutzungsbestimmungen durch diese Anbieter ist allerdings auch als Ausdruck dafür zu sehen, dass kein vollständiger und zuverlässiger (auch kostenpflichtiger) Zugriff auf die Suchmaschinenindices möglich ist.

Wie bei den Programmierschnittstellen ist auch über das Scraping kein kompletter Zugriff auf den Index der jeweiligen Suchmaschine möglich, sondern nur die Abfrage der Topergebnisse bis zu einer jeweils definierten Position.

6.3. Kooperationen

In Abschnitt 5.2 wurde bereits auf die Partnerschaften von Suchmaschinen, die letztlich nur aus einem Suchinterface bestehen, mit Zulieferern von Suchergebnissen (in der Regel eine der großen Suchmaschinen) eingegangen.

Die vertragliche Gestaltung der Vereinbarungen steht den jeweiligen Partnern natürlich prinzipiell frei. Die bestehenden Verträge enthalten jedoch eine wichtige Beschränkung: Zwar darf der Bezieher der Suchergebnisse den Ergebnisblock der gebenden Suchmaschine in seine Suchergebnisseiten einbinden, innerhalb des Ergebnisblocks dürfen jedoch keine Veränderungen vorgenommen werden. Das heißt, dass die gebende Suchmaschine in der Regel Blöcke von jeweils zehn Suchergebnissen liefert, diese Blöcke von der nehmenden Suchmaschine jedoch nicht aufgespalten, verändert oder mit anderen Elementen gemischt werden dürfen. Damit ist klar, dass die nehmenden Suchmaschinen

1. keinen Zugriff auf den Index der gebenden Suchmaschine erhalten, sondern nur jeweils auf eine geringe Anzahl von Topergebnissen, und
2. keine eigenen Dienste auf Basis der zugelieferten Suchergebnisse aufbauen können.

Nicht einmal im Rahmen einer vertraglichen Partnerschaft ist es also möglich, direkt auf den Index einer großen Suchmaschine zuzugreifen, um einen eigenen Suchdienst aufzusetzen. Eine Ausnahme ist die Suchmaschine Duck Duck Go, die Daten sowohl von Yahoo (über Yahoo BOSS) als auch von Yandex erhält und diese mit weiteren Quellen und Daten aus einem eigenen Crawl mischt.[8] Die Details der zugrunde liegenden Vereinbarung (insbesondere mit Yandex) sind jedoch nicht bekannt.

7. Fazit

Die Grundlage einer Suchmaschine ist der Index. In diesem Kapitel wurde gezeigt, dass ein Suchmaschinenindex nur mit enormem finanziellen und technischen Aufwand erstellt werden kann. Allerdings wird sowohl für den Betrieb einer Websuchmaschine als auch für komplexe Analysedienste ein möglichst vollständiger und aktueller Index des Webs benötigt.

Neue Suchmaschinen scheitern an der Erstellung des Indexes. Man kann sehen, dass sich neue Suchdienste in der Regel mit bestimmten Teilbereichen des Webs (oder anderen Datenbeständen) befassen, nicht mit dem kompletten Web. Auf der einen Seite

[8] http://help.duckduckgo.com/custo-mer/portal/articles/216399

mag dies daran liegen, dass mit Google eine zu große Konkurrenz besteht, gegen die anzutreten aussichtslos erscheint. Auf der anderen Seite ist ein Grund aber auch in dem enormen technischen wie finanziellen Aufwand zu sehen, der benötigt wird, um einen Webindex überhaupt aufzubauen.

Daraus ergibt sich, dass es weniger eine Notwendigkeit für den Aufbau bzw. die Förderung alternativer Suchmaschinen gibt (diese Diskussion wurde in der Vergangenheit bereits geführt) als vielmehr die Notwendigkeit, einen alternativen Index des Webs aufzubauen, auf den unter fairen Bedingungen zugegriffen werden kann.

7.1. Implikationen für die Forschung

Für die Forschung ergeben sich neben der Weiterentwicklung bzw. Verfeinerungen der Techniken zur Indexierung des Webs zahlreiche Fragestellungen in der Evaluierung der bestehenden Suchmaschinenindices. Auch wenn dieses Thema in diesem Kapitel nur angerissen werden konnte, so wurde doch deutlich, dass sowohl auf theoretischer als auch auf empirischer Ebene weitere Forschung nötig ist, um Möglichkeiten und Grenzen der Indices zu bestimmen sowie um Verfahren zur Messung der Indexqualität zu entwickeln.

7.2. Implikationen für die Praxis

Für die Praxis ergeben sich bereits aus der heutigen Zugänglichkeit von Ergebnissen aus den Indices bekannter Suchmaschinen (siehe Abschnitt 6.1) zahlreiche Möglichkeiten für den Aufbau neuer und innovativer Dienste. Chancen sind insbesondere in der Kombination von Daten der Suchmaschinen mit Daten aus eigenen (Such-) Diensten zu sehen.

7.3. Politischer Handlungsbedarf

Um einen fairen Zugang zu den Informationen des World Wide Web zu gewährleisten, bedarf es der Suchmaschinen und damit der ihnen zugrunde liegenden Indices. Die aktuelle Situation auf dem Suchmaschinenmarkt, die im Wesentlichen durch die Monopolstellung eines einzigen Anbietern gekennzeichnet ist, hat sich vor allem aufgrund des enormen Aufwands, Webindices zu erstellen, ergeben. Dazu kommt, dass durch das Vermarktungsmodell der Suchmaschinenindices („Partnerindex“) ein Modell geschaffen wurde, das Anbieter, die bislang eine Suchmaschine mit eigenem Index betrieben hatten, aus wirtschaftlichen Erwägungen dazu veranlasste, dieses aufzugeben. Dies führte dazu, dass in Deutschland kein Anbieter mehr existiert, der einen eigenen Webindex von konkurrenzfähiger Größe besitzt; in Europa kommt einzig Exalead in Betracht.

Innovationen im Bereich Websuche oder in der Analyse von Webdaten benötigen aber solche Indices. Da die Einstiegskosten jedoch so hoch sind, dass sich diese wohl kein Anbieter leisten kann (international werden eigentlich nur zwei bedeutende Indices betrieben; einer von Google, einer von Microsoft), können innovative Dienste in diesen Bereichen nicht entstehen. Ein Ausweg ist in der Förderung eines zu fairen Bedingungen für alle zugänglichen Webindex zu sehen, auf den beliebige Dienste aufgesetzt werden können. Aufgrund des enormen Finanzbedarfs und des weit länderübergreifenden Nutzens ist eine solche Förderung auf EU-Ebene zu empfehlen.

Literaturverzeichnis

[1] Risvik, K.M., Michelsen, R.: Search engines and web dynamics. Computer Networks. 39, 289-302 (2002).

[2] Lewandowski, D.: Spezialsuchmaschinen. In: Lewandowski, D. (Hrsg.) Handbuch Internet-Suchmaschinen. S. 53-69. Akademische Verlagsgesellschaft Aka, Heidelberg (2009).

[3] Quirmbach, S.: Universal Search - Kontextuelle Einbindung von unterschiedlicher Quellen und Auswirkungen auf das User Interface. In: Lewandowski, D. (Hrsg.) Handbuch Internet-Suchmaschinen. S. 220-248. Akademische Verlagsgesellschaft Aka, Heidelberg (2009).

[4] Höchstötter, N., Lüderwald, K.: Web Monitoring. In: Lewandowski, D. (Hrsg.) Handbuch Internet-Suchmaschinen 2: Neue Entwicklungen in der Web-Suche. S. 289-322. Akademische Verlagsanstalt AKA, Heidelberg (2011).

[5] Pietruszkiewicz, W.: The Computational Analysis of Web Search Statistics in the Intelligent Framework Supporting Decision Making. In: Lewandowski, D. (Hrsg.) Web Search Engine Research (Library and Information Science, Volume 4). S. 79-102. Emerald, Bingley (2012).

[6] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval: The Concepts And Technology Behind Search. Addison Wesley, Harlow (2011).

[7] Lewandowski, D.: Web Information Retrieval: Technologien zur Informationssuche im Internet. Deutsche Gesellschaft f. Informationswissenschaft u. Informationspraxis, Frankfurt am Main (2005).

[8] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Computer networks. 33, 309-320 (2000).

[9] Lewandowski, D.: Date-restricted queries in web search engines. Online Information Review. 28, 420-427 (2004).

[10] Schuler, T.: Google News: Automatischer Absturz, http://www.sueddeutsche.de/digital/google-news-automatischer-absturz-1.705184 [7.5.2013].

[11] Croft, W.B., Metzler, D., Strohman, T.: Search Engines: Information retrieval in practice. Pearson, Boston, MA (2010).

[12] Thelwall, M., Stuart, D.: Web crawling ethics revisited: Cost, privacy, and denial of service. Journal of the American Society for Information Science and Technology. 57, 1-14 (2006).

[13] Cafarella, M.J., Madhavan, J., Halevy, A.: Web-Scale Extraction of Structured Data. SIGMOD Record. 34, 55–61 (2008).

[14] Patterson, A.: Why writing your own search engine is hard. Queue. 2, 49-53 (2004).

[15] Watters, P.A., Lueg, C., Spiranovic, C., Prichard, J.: Patterns of ownership of child model sites: Profiling the profiteers and consumers of child exploitation material. First Monday. 18, (2013).

[16] Sherman, C., Price, G.: The Invisible Web: Finding Hidden Internet Resources Search Engines Can't See. Cyberage Books (2001).

[17] Bergman, M.K.: The deep Web: Surfacing hidden value. Journal of Electronic Publishing. 7, 1-17 (2001).

[18] Lewandowski, D., Mayr, P.: Exploring the academic invisible web. Library Hi Tech. 24, 529-539 (2006).

[19] Stock, W.: Weltregionen des Internet: Digitale Informationen im WWW und via WWW. Password. 18, 26-28 (2003).

[20] Lewandowski, D., Höchstötter, N.: Qualitätsmessung bei Suchmaschinen – System- und nutzer- bezogene Evaluationsmaße. Informatik-Spektrum. 30, 159-169 (2007).

[21] Sullivan, D.: End Of Size Wars? Google Says Most Comprehensive But Drops Home Page Count. http://searchenginewatch.com/article/2066323/End-Of-Size-Wars-Google-Says-Most-Comprehensive-But-Drops-Home-Page-Count [7.5.2013]

[22] Lawrence, S., Giles, C.L.: Searching the world wide web. Science. 280, 98–100 (1998).

[23] Bharat, K., Broder, A.: A technique for measuring the relative size and overlap of public Web search engines. Computer Networks and ISDN Systems. 30, 379–388 (1998).

[24] Gulli, A., Signorini, A.: The indexable web is more than 11.5 billion pages. 14th international conference on World Wide Web. S. 902–903. ACM, Chiba, Japan (2005).

[25] Ntoulas, A., Cho, J., Olston, C.: What's new on the web?: the evolution of the web from a search engine perspective. Proceedings of the 13th international conference on World Wide Web. S. 1-12. ACM, UCLA Computer Science Carnegie Mellon University (2004).

[26] Lewandowski, D., Wahlig, H., Meyer-Bautor, G.: The freshness of web search engine databases. Journal of Information Science. 32, 131-148 (2006).

[27] Lewandowski, D.: A three-year study on the freshness of Web search engine databases. Journal of Information Science. 34, 817-831 (2008).

[28] Mowshowitz, A.: Measuring search engine bias. Information Processing & Management. 41, 1193-1205 (2005).

[29] Vaughan, L., Thelwall, M.: Search engine coverage bias: Evidence and possible causes. Information Processing & Management. 40, 693-707 (2004).

[30] Vaughan, L., Zhang, Y.: Equal representation by search engines? A comparison of websites across countries and domains. Journal of Computer-Mediated Communication. 12, 888-909 (2007).
[31] Maaß, C., Skusa, A., Heß, A., Pietsch, G.: Der Markt für Internet-Suchmaschinen. In: Lewandowski, D. (Hrsg.) Handbuch Internet-Suchmaschinen. S. 3-17. Akademische Verlagsgesellschaft Aka, Heidelberg (2009).
[32] ComScore: comScore Releases January 2013 U.S. Search Engine Rankings, http://www.comscore.com/Insights/Press_Releases/2013/2/comScore_Releases_January_2013_U.S._Search_Engine_Rankings [7.5.2013].
[33] Clay, B.: Search Engine Relationship Chart, http://www.bruceclay.com/searchenginechart.pdf [7.5.2013].
[34] Press Release: comScore Reports Global Search Market Growth of 46 Percent in 2009, http://comscore.com/Press_Events/Press_Releases/2010/1/Global_Search_Market_Grows_46_Percent_in_2009 [7.5.2013].
[35] Sullivan, D.: Why Yahoo Will Never Reach The "Revenue Per Search" That Microsoft Promised. http://marketingland.com/yahoo-microsoft-rps-guarantee-42680 [7.5.2013].
[36] Sullivan, D.: Even If Yahoo Wants To Leave Microsoft, Here's Why It Can't. http://searchengineland.com/even-if-yahoo-wants-to-leave-microsoft-heres-why-it-cant-158684 [7.5.2013].
[37] Google Cancelled Yahoo Search Deal To Avoid "Monopoly" Designation, http://searchengineland.com/google-cancelled-yahoo-search-deal-to-avoid-monopoly-designation-15735 [7.5.2013].
[38] Webhits: Webhits Web-Barometer, http://www.webhits.de/deutsch/index.shtml?webstats.html [7.5.2013].
[39] Beziehungsgeflecht der Suchmaschinen, http://www.luna-park.de/blog/2287-suchmaschinen-beziehungsgeflecht/ [7.5.2013].
[40] Gesenhues, A.: Yahoo Announces New Search Tools & Upgrades In Their Continued Battle To Win More Of The Search Market. http://searchengineland.com/yahoo-announces-new-search-tools-upgrades-in-their-continued-battle-to-win-more-of-the-search-market-158799 [7.5.2013].
[41] Thomas, P.: To What Problem Is Distributed Information Retrieval the Solution? Journal of the American Society for Information Science & Technology. 63, 1471-1476 (2012).
[42] Thelwall, M.: Quantitative comparisons of search engine results. Journal of the American Society for Information Science and Technology. 59, 1702-1710 (2008).
[43] Thelwall, M., Ruschenburg, T.: Grundlagen und Forschungsfelder der Webometrie. Information. 57, 401-406 (2006).
[44] Thelwall, M.: Extracting accurate and complete results from seach engines: case study Windows Live. Journal of the American Society for Information Science and Technology. 59, 38–50 (2008).
[45] Tosques, F., Mayr, P.: Programmierschnittstellen der kommerziellen Suchmaschinen. Handbuch Internet-Suchmaschinen. 116-147 (2009).
[46] Lewandowski, D., Sünkler, S.: Relevance Assessment Tool Ein Werkzeug zum Design von Retrievaltests sowie zur weitgehend automatisierten Erfassung, Aufbereitung und Auswertung von Daten. In: Ockenfeld, M.; Weller, K.; Peters, I.: Social Media und Web Science: Das Web als Lebensraum. Proceedings der 2. DGI-Konferenz. S. 237-249 (2012).
[47] Sünkler, S.: Prototypische Entwicklung einer Software für die Erfassung und Analyse explorativer Suchen in Verbindung mit Tests zur Retrievaleffektivität. Master Thesis, Hochschule für Angewandte Wissenschaften Hamburg. http://opus.haw-hamburg.de/volltexte/2012/1845/pdf/Suenkler_Sebastian_20120229.pdf [7.5.2013] (2012).