1.4 Verarbeitung von Suchanfragen

Damit Suchmaschinen gut funktionieren, müssen Nutzer die Worte eingeben, die auf der Webseite vorkommen sollen. Nicht das Web, der Index der Suchmaschine wird nach Übereinstimmungen mit der Suchanfrage durchsucht. Grundsätzlich gilt: Suchbegriffe werden von Google und Co. auf die gleiche Weise verarbeitet wie die Wörter, die auf der Webseite geschrieben stehen. In beiden Fällen muss ein von Menschenhand geschriebener Text in ein einheitliches Format gebracht werden. Nur so kann die Suchmaschine den Suchbegriff mit dem Index abgleichen.

Bei Suchanfragen, die aus sehr vielen Begriffen bestehen (Long-Tail-Suchanfragen), entfernt die Suchmaschine sehr häufig verwendete Begriffe wie beispielsweise wo oder wie und auch bestimmte einzelne Zahlen und Buchstaben, weil sie nach Ansicht von Google die Qualität der Treffer nicht erhöhen, gleichzeitig aber die Suchgeschwindigkeit verlangsamen.

Nachdem die überflüssigen Begriffe in Long-Tail-Suchanfragen entfernt wurden, verbindet die Suchmaschine die restlichen Wörter mit dem Suchoperator AND. Die großen Suchmaschinen setzen den AND-Operator automatisch. Er bewirkt, dass die Maschine nach Webseiten sucht, auf denen alle Suchbegriffe geschrieben stehen. Sollte die Suchanfrage noch andere, spezielle Operatoren enthalten, etwa die Einschränkung bei Google, nur nach bestimmten Dateiformaten zu suchen, dann werden diese Suchoperatoren anhand einer Liste von reservierten Zeichen erkannt und in den Suchprozess entsprechend einbezogen.

LENZ, Suchmaschinenoptimiert schreiben. ISBN 978-3-86764-284-2

Suchtipps:

- **Stoppwörter in die Suche einbeziehen**
  Bei der Suchanfrage »Star Wars Episode I« wird »I« eliminiert. Mit dem »+«-Zeichen vor dem Stoppwort kann man die Eliminierung verhindern (Star Wars Episode +I).

- **Exakte Wortgruppensuche**
  Ähnlich wie beim »+«-Zeichen: Setzt man die »Suchbegriffe in Anführungszeichen«, sucht Google nach Webseiten, auf denen die Begriffe »in exakt der gleichen Reihenfolge« geschrieben stehen.

- **Ausschluss von Begriffen**
  Falls ein Suchbegriff mehrere Bedeutungen hat, kann man die Suchanfrage mit dem »-«-Zeichen eingrenzen (Paris -Hilton).

Matching

Matching nennt man den Vorgang, wenn die Maschine nach der Eingabe eines Suchbegriffs nach Übereinstimmungen im Index sucht. Wie Sebastian Erlhofer schreibt, werden die Wörter im Index nicht tabellenartig, sondern zusammen mit der Hitlist in verkürzter Form und nicht in ihrer orthografischen Erscheinungsform gespeichert (Erlhofer: 113). Jede Zeichenabfolge zwischen zwei Wortseparatoren wird codiert. Auch hier spielt es für das grundsätzliche Verständnis keine Rolle, welche Codes Google verwendet. Das Schema ist stets gleich. Wurde der Begriff spätrömische mit der Nummer 33 indexiert, dekadenz mit der Nummer 56, dann übersetzt die Suchmaschine diese Wörter als Suchwörter in dieselben Nummern. Das Matching für die Suchanfrage »spätrömische dekadenz« würde dann vereinfacht dargestellt folgendermaßen aussehen:

LENZ, Suchmaschinenoptimiert schreiben. ISBN 978-3-86764-284-2

Keyword	Index
spätrömische	56615, 5863, 1562, 5656, 9531
dekadenz	5863, 16851, 9531, 6815
Matching	9531, 5863

Matching für »spätrömische dekadenz«.

Die Webseiten mit den Nummern 56615, 5863, 1562, 5656 und 9531 enthalten den Begriff spätrömische. Das Wort dekadenz hat die Suchmaschine auf den Seiten mit den Nummern 5863, 16851, 9531 und 6815 entdeckt. Bestünde der Google-Index nur aus diesen acht Webseiten, würde die Suchmaschine genau zwei Treffer auflisten: 5863 und 9531.

Bevor die Treffer in der Ergebnisliste angezeigt werden, müssen sie noch gerankt werden. Wichtig ist hierbei vor allem die Reputation der Website, aber auch die Bedeutung, die der Suchbegriff für die inhaltliche Beschreibung der Webseite einnimmt.

Die Bedeutung des Suchbegriffs wurde in der Hitlist im direkten Index vermerkt. Besteht eine Suchanfrage aus nur einem Wort (dekadenz), der Index wie im Beispiel oben aus acht Webseiten, ist das Ranking relativ einfach: Die angesehenste Seite, die den Begriff dekadenz mehrfach und an besonders wichtigen Stellen enthält, ist in Bezug auf den Suchbegriff relevanter als eine unbedeutende Seite, auf der dekadenz nur einmal am Ende des Fließtextes geschrieben steht. Komplizierter wird es in der Praxis, also wenn eine Suchanfrage aus mehreren Wörtern besteht, der Index mehr als drei Milliarden Web-Dokumente enthält. Im dritten Teil dieses Buchs werden wir das Thema Ranking ausführlich besprechen.

Polyseme und Homonyme

Was wissen wir bis jetzt? Der Crawler durchsucht die Web-Server nach neuen Hyperlinks. Mittels DNS löst er die neu gefundenen URLs in IP-Adressen auf. Anschließend fordert er den Server per HTTP-Request mit der GET-Methode zum Übertragen der Nutzinformation auf. Der Server

LENZ, Suchmaschinenoptimiert schreiben. ISBN 978-3-86764-284-2

schickt die Nutzinformation per HTTP-Response zurück an den Crawler, der sie mit einer eindeutigen Nummer versieht und an den Parser weiterleitet. Der Parser entfernt Auszeichnungssprache und Programmiercode. Mithilfe der Wortseparatoren überführt er den Zeichenstrom in einen Wortstrom. Die extrahierten Wörter werden mit Zusatzinformationen versehen und anschließend in komprimierter Form im Index gespeichert. Suchbegriffe werden von der Suchmaschine in ein mit dem Index vergleichbares Format umgewandelt. Nur so kann die Suchmaschine nach Übereinstimmungen suchen. Und genau hier tritt ein Problem auf. Beispiel Paris.

Paris kann eine Stadt sein, eine Figur aus der griechischen Mythologie, der Vorname einer prominenten Hotelerbin und noch vieles mehr. Sprachwissenschaftler sprechen von Polysemen oder Homonymen und meinen Wörter, die trotz gleicher Schreibweise unterschiedliche Bedeutungen haben. Die Liste ist lang. Mit Bank kann eine Sitzbank im Park oder ein Geldinstitut gemeint sein. Der Begriff Ton hat mit Musik zu tun und mit Töpferei.

Mit Sprachmarken beziehungsweise separaten Indexen für jede natürliche Sprache versuchen Google und Co. mehrdeutige Wörter in den Griff zu bekommen. Natürliche Sprachen nennt man in der Linguistik Sprachen, die aus einer historischen Entwicklung heraus entstanden sind. Auszeichnungs- und Programmiersprachen sind künstliche Sprachen.

Identifizierung der natürlichen Sprache

Die natürlichen Sprachen im Web sind so vielseitig wie die Kulturen und Sprachen auf der Welt. Laut einer Studie des Unternehmens Global Reach waren 2004 die englischen Muttersprachler am häufigsten vertreten, dicht gefolgt von den Chinesen, Spaniern, Japanern und den Deutschen an fünfter Stelle (Erlhofer: 94). Die meisten Nutzer bevorzugen jedoch Webseiten in ihrer Muttersprache. Anhand der IP-Adresse erkennt Google den geografischen Standort des Suchenden. Deutsche Nutzer werden so automatisch auf die deutsche Google-Seite verwiesen. Damit die Suchmaschine beim Matching die Seiten in deutscher Sprache bevorzugen kann, muss sie zuerst die auf den Webseiten verwendete natürliche Sprache identifizieren. Sebastian Erlhofer zufolge kombinieren Suchmaschinen zu diesem Zweck statistische Methoden mit einer Wörterbucherkennung. Die statistischen Methoden beruhen auf sogenannten Hidden-Markov-Modellen. In einer vorangegangenen Trainingsphase analysiert die Suchmaschine beispielhafte Texte einer natürlichen Sprache und arbeitet typische Muster heraus. Ein

LENZ, Suchmaschinenoptimiert schreiben. ISBN 978-3-86764-284-2

überdurchschnittliches Auftreten von deutschen Umlauten beispielsweise lässt auf einen Text in deutscher Sprache schließen.

Besonders verlässlich sind die Hidden-Markov-Modelle innerhalb von Fließtexten. Bei stichwortartigen Textinhalten oder Texten, die verstärkt aus Eigennamen, Lehnwörtern oder Fachtermini bestehen, treten die sprachtypischen Muster nicht deutlich genug hervor. Hier, so Erlhofer, kann der Abgleich der im Text auftretenden Begriffe mit einem gut gepflegten Wörterbuch Abhilfe schaffen (Erlhofer: 96). Zusätzlich können Websitebetreiber im Quelltext das Sprachen-Tag setzen. Es wird von Suchmaschinen jedoch nur in Ausnahmefällen berücksichtigt, etwa wenn verschiedene natürliche Sprachen auf derselben Seite verwendet wurden.

```
<meta name=„content-language“ content=„de“/>
```

Ohne Sprachmarken beziehungsweise Länderindexe würde die Zahl der Polyseme und Homonyme eskalieren. Welcher deutsche Nutzer, der nach Informationen über die Baumart Buche sucht, möchte sich schon spanische Seiten über die Aussackung der Speiseröhre bei Vögeln ansehen, also Webseiten über den Vogel-Kropf, der auf Spanisch ebenfalls Buche heißt? Oder Public Viewing. Wenn deutsche Nutzer danach suchen, meinen sie in erster Linie öffentliche Orte zum Fußballgucken. Amerikaner wundern sich darüber. In Amerika bedeutet Public Viewing Leichenschau.

Wer sich dennoch kosmopolitisch informieren möchte, kann die Voreinstellungen deaktivieren und die Suchsprache manuell eingeben. Der entsprechende Link (Sucheinstellungen) befindet sich auf der Google-Startseite oben rechts. Bei der Suchmaschine Bing findet man ihn an der gleichen Stelle. Er heißt dort passenderweise Deutschland.

LENZ, Suchmaschinenoptimiert schreiben. ISBN 978-3-86764-284-2

Suchsprache bei Google manuell auswählen.

Word Stemming und Mehrwortgruppenidentifikation

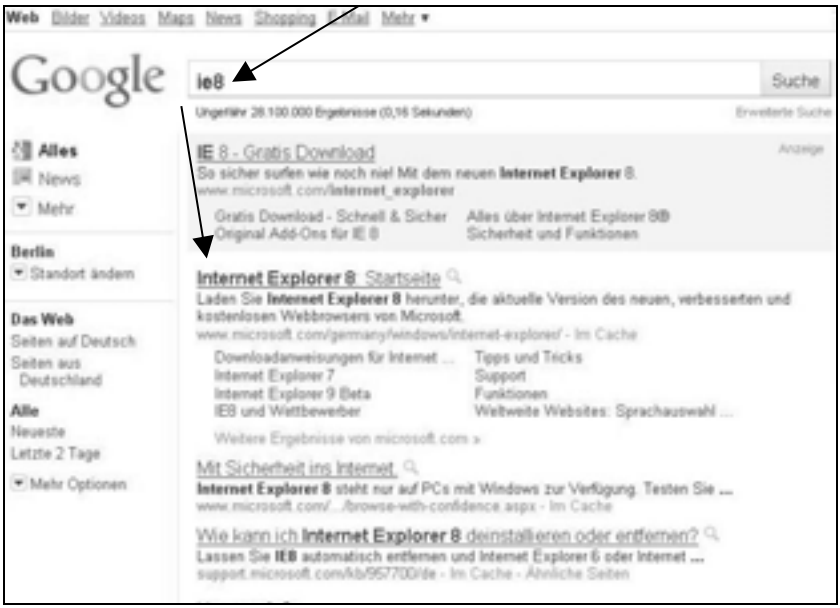
Im praktischen Einsatz sind die Abläufe bei der Index-Erstellung und der Verarbeitung von Suchanfragen sehr viel komplexer. Das System ist nicht perfekt, die Suchmaschinenbetreiber testen immer wieder neue Methoden. So werden heute in einigen Fällen synonyme Begriffe in die Suche mit einbezogen. In der Trefferliste werden dann auch Webseiten angezeigt, die den Suchbegriff nicht enthalten (sofern ausreichend andere Beweise dafür vorliegen, dass die Seite relevant ist).

Hintergrund ist ein algorithmisches Verfahren mit dem lexikalisch verwandte Begriffe zusammengeführt werden (Word Stemming). Der von der Suchmaschine errechnete Wortstamm ist ein künstlicher und nicht identisch mit dem grammatikalisch korrekten Wortstamm. An der englischen Sprache wurden Stemming-Algorithmen erstmals ausprobiert. Dem korrekten Wortstamm beauty beispielsweise entspricht der künstlich errechnete Wortstamm beauti.

Es gibt viele verschiedene Stemming-Algorithmen. Google weiß, dass mit der Abkürzung IE8 der Browser Internet Explorer 8 gemeint ist, und findet mit der Sucheingabe IE8 Webseiten, auf denen ausschließlich Internet Explorer 8 geschrieben steht. Allerdings werden Stemming-Algorithmen sehr unregelmäßig eingesetzt, weil sie die Trefferqualität verringern, sodass Such-

LENZ, Suchmaschinenoptimiert schreiben. ISBN 978-3-86764-284-2

maschinenoptimierer immer noch davon ausgehen, dass die exakte Abbildung des Suchbegriffs auf der Webseite die optimale Voraussetzung für das Ranking ist. Ob und wie Google einen Begriff stemmt, erkennt man daran, welche Begriffe in der Trefferliste durch Fettung hervorgehoben sind.



Google stemmt den Begriff Internet Explorer 8.

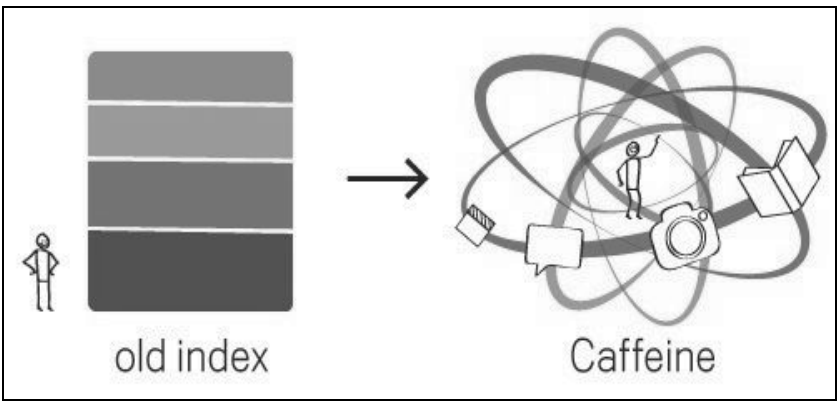
Eine weitere Möglichkeit, die Qualität der Treffer zu erhöhen, ist die algorithmische Mehrwortgruppenidentifikation. Um Mehrwortgruppen in ihre Komponenten zu zerlegen (Marktforschungsinstitut), nutzen Suchmaschinen ein spezielles Mehrwortgruppen-Wörterbuch, das nach einem robusten Verfahren automatisch erstellt werden kann. Parameter wie der Abstand zwischen Komponenten, die Reihenfolge und die Satzstruktur werden laut Erlhofer berücksichtigt.

Hundertprozentig zuverlässige Ergebnisse liefert das Verfahren nicht, weshalb jeweils Wahrscheinlichkeitswerte errechnet werden, die dann als Gewichtung im Retrieval-Verfahren berücksichtigt werden (Erlhofer: 100f).

LENZ, Suchmaschinenoptimiert schreiben. ISBN 978-3-86764-284-2

Google Caffeine

Ende des Jahres 2009 gab es die Google-Suchmaschine zweimal im Web. Ein beschleunigtes Indexierungsverfahren sollte getestet werden. Das hat der Testversion sehr schnell den Spitznamen Caffeine eingebracht. Abgesehen vom URL war Google Caffeine nicht zu unterscheiden von der normalen Google-Seite. Tests hatten gezeigt, dass die Ergebnisliste tatsächlich schneller vorlag – wenn auch im Millisekundenbereich, sodass es vom Nutzer nicht bemerkt wurde. Außerdem fand Google Caffeine mehr Seiten. Die erste Trefferliste war fast identisch mit der normalen Trefferliste. Unterschiede hatte man auf der zweiten und dritten Ergebnisseite festgestellt. Ein paar Monate später hat Google sein Indexierungssystem dann tatsächlich umgestellt.



Schema des beschleunigten Indexierungssystems.

Die Grafik hat Google zusammen mit der Systemumstellung veröffentlicht:

- http://googleblog.blogspot.com/2010/06/our-new-search-index-caffeine.html

Der linke Teil steht für das alte Indexierungssystem. Vor Google Caffeine wurden die Inhalte je nach Indexzugehörigkeit unterschiedlich häufig aktualisiert. Nachrichten-Websites wurden (und werden) von den Crawlern quasi permanent nach neuen Inhalten durchsucht, der Hauptindex dagegen nur

LENZ, Suchmaschinenoptimiert schreiben. ISBN 978-3-86764-284-2

»every couple of weeks«. Das hatte zur Folge, dass neuer Content in einigen Fällen auch erst einige Wochen nach der Veröffentlichung gefunden werden konnte.

Seit der Systemumstellung im August 2010 werden auch Videos, Fotos und andere Web-Inhalte schneller indexiert. Das verdeutlicht der rechte Teil der Grafik. Sobald die Crawler neuen Content finden, wird er den bestehenden Indexen hinzugefügt.

Wie Google das neue Indexierungssystem mit allen technischen Finessen im Detail realisiert, bleibt Googles Betriebsgeheimnis und hat für unsere Zwecke wenig Bedeutung. Es geht nicht um Entwicklerfragen. Es geht um die Grundlagen der Netztechnik; was man als Autor auf der Textebene beachten sollte, und wie sich die Netztechnik auf das Journalismussystem auswirkt. Daran hat sich auch nach der Systemumstellung nichts geändert. Im Gegenteil. Google findet jetzt noch mehr Webseiten und indexiert seit 2010 auch Echtzeit-Inhalte, also öffentlich einsehbare Twitter-Kurzmitteilungen oder Facebook-Statusupdates.

Christian Jakubetz schreibt in »Crossmedia«, dass Journalisten die neue Nutzungssituation im Web beachten müssen. Journalisten müssten den Nutzer dort abholen, wo er sich befindet, und ihm das geben, was er will, wann er will, wo er will (Jakubetz: 13).

Was die Nutzer wollen, wo sie sich aufhalten ist schon lange kein Geheimnis mehr. Laut ARD/ZDF-Onlinestudie 2009 gehört die Suchmaschinennutzung seit Jahren zu den häufigsten Internetanwendungen. Und wenn die Nutzer für die Informationsbeschaffung Suchmaschinen nutzen, sollten diejenigen, die Informationen anbieten, den Nutzern auch dorthin folgen.

Je nachdem, wo (Medientyp, Zielgruppe, Medienmarke) über was (Hard News, Soft News) und mit welcher Absicht (informieren, unterhalten, beraten) berichtet wird, wählen Journalisten eine Textsorte aus dem Repertoire der klassischen Textsorten aus. Jede dieser Textsorten folgt einem bestimmten formalen Aufbau und Stil. Grundlage dieser Textsorten ist die Technik der Massenmedien. Die Netztechnik erfordert eine andere Sprache und konzeptionelle Ausrichtung.

LENZ, Suchmaschinenoptimiert schreiben. ISBN 978-3-86764-284-2