Soekia – ein Blick hinter die Kulissen von Suchmaschinen

Einsatz einer didaktischen Suchmaschine im Unterricht

Informationen im Internet zu finden ist ein Kinderspiel! Wer sich für Informationen zur Entwicklung des Ozonlochs interessiert, gibt auf `www.google.de` die Suchanfrage „Entwicklung Ozonloch“ ein, drückt den Button „Google Suche“ und erhält in Sekundenbruchteilen rund 5 800 Treffer (Stand Oktober 2003). Aber Achtung: Ein anderer Benutzer gibt vielleicht den Suchbegriff „Entwicklung des Ozonlochs“ ein. Auf diese Anfrage liefert Google nur noch rund 780 Treffer. Nur noch 150 Treffer werden bei der Anfrage „Entwicklung des Ozonloches“ gefunden. Während für uns Menschen die verschiedenen Deklinationsformen des Wortes „Ozonloch“ semantisch gleich sind, betrachtet Google „Ozonloch“, „Ozonlochs“ und „Ozonloches“ als völlig verschiedene Begriffe. Gerade bei Recherchen zu wissenschaftlichen Themen ist die Gefahr deshalb gross, relevante Informationen zu übersehen. Informationsbeschaffung im Internet beschränkt sich keineswegs auf die Handhabung eines Web-Browsers. Effiziente und effektive Informationsbeschaffung ist eine anspruchsvolle Aufgabe, die eine fundierte Ausbildung voraussetzt. Dazu gehört auch ein Verständnis für die Funktionsweise von Suchmaschinen. Wie eine Suchmaschine arbeitet, bleibt aber den Benutzern weitgehend verborgen: Das Erfassen von Webseiten durch Webroboter (Crawling / Spidering), das Erstellen einer effizienten Datenstruktur für die Suche (Indexierung, Index), das Finden zu einer Benutzeranfrage passender Dokumente (Matching) und die Präsentation der gefunden Dokumente in einer guten Reihenfolge (Rangierung). Die didaktische Suchmaschine Soekia ermöglicht einen Blick hinter die Kulissen von Suchmaschinen. Schülerinnen und Schüler können beispielsweise selber den Index der Suchmaschine inspizieren. Oder Sie können die erfassten Dokumente variieren und die Auswirkungen auf die Rangliste beobachten. Das vertiefte Verständnis für die Funktionsweise einer Suchmaschine erlaubt es, Suchanfragen gezielter zu stellen und sowohl die Ausbeute als auch die Präzision bei der Internet-Recherche zu verbessern. Soekia steht zum kostenlosen Download auf dem Bildungsserver SwissEduc (`www.swisseduc.ch/informatik/soekia/`) zur Verfügung.

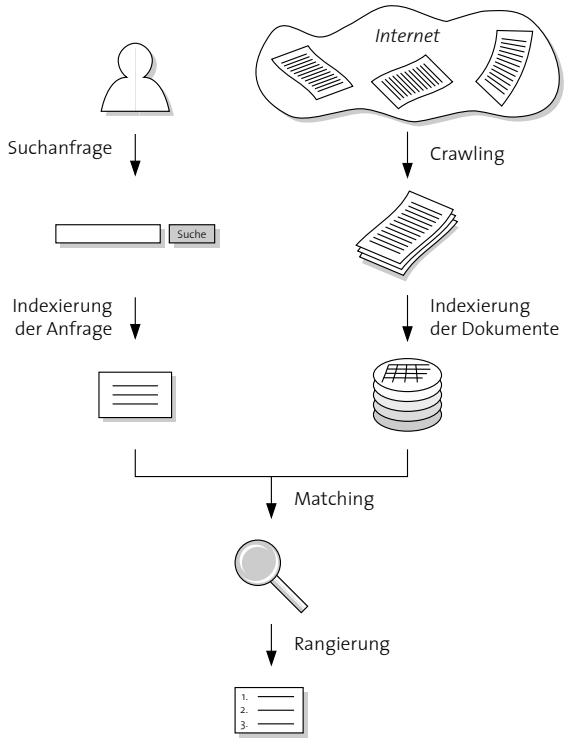


Abbildung 1: Komponenten einer Suchmaschine

1 Der Mensch als entscheidender Faktor bei der Suche

Bereits 1971 hielt Saracevic in [Sar71] fest: „The human factor, i. e. variations introduced by human decision-making, seems to be the major factor affecting performance of every and all components of an information retrieval system.“ Aus vielen Untersuchungen zum Benutzerverhalten bei Suchmaschinen (z. B. [WI01] oder [SJWS02]) ist bekannt, dass die Benutzer die angeboten Möglichkeiten von Suchdiensten nur selten richtig nutzen:

- User verwenden zwar oft Suchmaschinen, planen die Suche aber kaum. Insbesondere legen sich nur wenige User vorgängig Rechenschaft ab, ob es sich bei der Fragestellung um eine offene Frage oder eine geschlossene Frage handelt. Offene Fragen (z. B. Was sind die Ursachen für Hooliganismus?) bedingen eine ausbeuteorientierte Suche und eignen sich beispielsweise nicht für die einfache Suche bei Google. Geschlossene Fragen hingegen (z. B. Wie hoch ist der Eiffelturm?) sind präzisionsorientiert und können ohne Weiteres mit der einfachen Google Suche und einigen spezifischen Suchbegriffen (z. B. Eiffelturm Höhe Meter) effizient beantwortet werden.
- Die User verwenden einfache Suchanfragen mit in der Regel nur 1–2 Suchbegriffen, die zudem oft sehr unspezifisch sind.

- User inspizieren nur die ersten Treffer in der Rangliste und nutzen die Interaktion mit dem Suchdienst (z. B. Relevanzfeedback, manuelle Anfrageerweiterung) kaum.
- User schätzen die Glaubwürdigkeit der Informationen auf dem Web als hoch ein und sind sich nicht bewusst, dass sie oft einen Grossteil der an und für sich relevanten Dokumente übersehen.

Ein grosser Teil dieses unzulänglichen Verhaltens der Benutzer beruht wohl auf der Tatsache, dass nur die wenigsten eine Vorstellung von der Funktionsweise einer Suchmaschine haben. Die meisten Leute sind der falschen Ansicht, dass Suchmaschinen bei einer Anfrage das Internet in Echtzeit durchsuchen. Eine klare Vorstellung vom Aufbau und der Funktionsweise eines Indexes könnte hier vielen Missverständnissen vorbeugen. Mit der didaktischen Suchmaschine Soekia kann die Funktionsweise von Suchmaschinen in ihren Grundzügen transparent gemacht werden. Im Folgenden werden die wichtigsten Komponenten einer Suchmaschine und einer Suche anhand von Soekia aufgezeigt.

2 Erfassen und Speichern aller Dokumente einer Kollektion

Es gibt Millionen von Web-Servern, die Hunderte Millionen von Webseiten anbieten. Irgendwann entsteht auf irgendeinem dieser Server eine neue Seite, oder es wird eine der bestehenden Seiten geändert. Eine Suchmaschine muss diese Dokumente erfassen. Hier kommt der Web-Roboter (auch *Spider* oder *Crawler* genannt) ins Spiel. Der Web-Roboter ist ein Programm, das zur Aufgabe hat, Webseiten zu finden. Dazu nützt der Roboter die Eigenschaft des World Wide Web aus, dass die Dokumente über Hyperlinks miteinander verbunden sind. In einer Tabelle legt eine für den Suchdienst verantwortliche Person die Startpunkte für die Suche nach Webseiten fest. Der Web-Roboter geht durch diese Liste mit URLs und bezieht die zugehörigen Seiten aus dem Internet. Dann wird jede Seite nach weiterführenden Verweisen (Hyperlinks) untersucht. Die gefundenen Hyperlinks landen ebenfalls in der Tabelle mit den URLs, damit der Web-Roboter über die schon besuchten Seiten Bescheid weiss. Später werden auch die neu eingetragenen Seiten nach weiteren Verweisen untersucht. Auf diese Weise arbeitet sich der Web-Roboter immer weiter in die Tiefen des WWW vor. Früher oder später findet der Web-Roboter somit alle Seiten, die auf irgend einem Weg von den Startseiten aus erreicht werden können.

Diesen Prozess des Erfassens von Webseiten kann man mit Soekia nicht zeigen. Erfahrungsgemäss ist dieser Sachverhalt für Schülerinnen und Schüler aber auch leicht nachvollziebar. Wir alle kennen die Methoden, um ein Labyrinth abzuwandern. Genau so funktioniert ein Webroboter.

Damit die Benutzer nicht selbst HTML-Dokumente zusammenzustellen müssen, stehen auf der Soekia Web-Seite für verschiedene Fragestellungen Beispiel-Kollektionen zur Verfügung. Somit erübrigt sich das Zusammentragen von Dokumenten aus dem Web.

Abbildung 2: Soekia Programm-Fenster

3 Indexierung: Erstellen einer effizienten Datenstruktur für die Suche

Die von einer Suchmaschine erfassten Dokumente bilden die *Dokumenten-Kollektion*, in welcher die Suchmaschine sucht. Eine Ablage der einzelnen Webseiten und Dokumente in unveränderter Form ist nicht zweckmässig. Die Suchmaschine erstellt deshalb einen Index der Dokumenten-Kollektion. Der Index entspricht ziemlich genau dem Stichwortverzeichnis am Ende eines Buches und umfasst die in den Dokumenten vorkommenden Begriffe samt einem Verweis auf die entsprechenden Dokumente.

Soekia stellt den Index in einer lesbaren Form dar. Abbildung 3 zeigt einen Ausschnitt eines Indexes. Neben jedem Begriff steht, wie oft der Begriff in der Kollektion vorkommt und in wie vielen Dokumenten er auftritt. Angezeigt wird auch die gesamte Anzahl Begriffe in der Kollektion. Mit dieser Index-Darstellung lassen sich zahlreiche Fragen untersuchen. Wie verändert sich der Index beim Hinzufügen von gleichartigen Dokumenten zu einer Dokumenten-Kollektion, wie bei artfremden Dokumenten? Wie wirkt sich Wort-Normalisierung auf den Index aus? Wie unterscheiden sich die Index-Einträge von allgemeinen und spezifischen Begriffen?

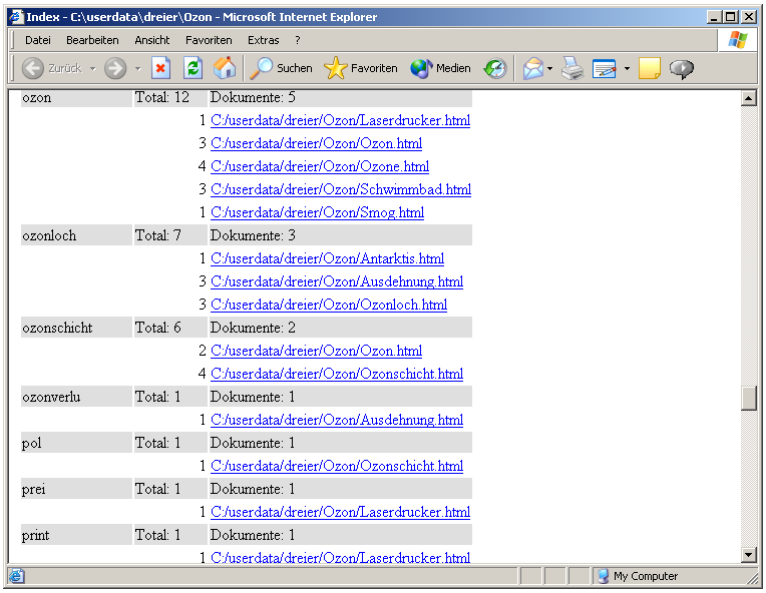


Abbildung 3: Index der Beispiel-Kollektion „Ozon“ (Ausschnitt)

4 Matching: Finden der zu einer Benutzeranfrage passenden Dokumente

Sinnvolle quantitative Vergleiche setzen immer einen dem Vergleich vorausgehenden Normalisierungsprozess voraus. Wir kennen das aus dem Alltag bestens: Um den Preis zweier Produkte zu vergleichen, müssen wir zuerst den Preis für gleich grosse Mengen des Produktes bestimmen. Genau so verhält es sich bei Suchdiensten im Internet. Viele der grösseren Suchdienste (z. B. Google) vergleichen Wörter strikt Buchstabe für Buchstabe. „Ozonloch“ ist somit ein anderer Begriff als „Ozonloches“ oder „Ozonlöcher“. Viele an und für sich relevante Dokumente werden bei der Anfrage „Ozonloch“ übersehen.

Ein besseres Verfahren reduziert Wörter auf den Wortstamm. Aus „Ozonloches“ und „Ozonlöcher“ wird bei diesem auch *Stemming* genannten Prozess der Wortstamm „Ozonloch“. Wichtig ist, dass die Suchmaschine dieses Stemming-Verfahren nicht nur beim Indexieren der Dokumente anwendet, sondern auch die Begriffe einer Suchanfrage demselben Prozess unterzieht. Nur so findet eine Anfrage mit „Ozonloches“ ein Dokument mit „Ozonlöcher“.

Die von einer Suchmaschine vorgenommene Normalisierung kann verschieden aufwändig sein. Eine umfangreiche Normalisierung beinhaltet unter anderem eine Wortzerlegung („Ozonloch“ wird als „Ozon“ und „Loch“ im Index erfasst) und eine Reduktion auf den Wortstamm („gingen“ wird unter „gehen“ im Index erfasst). Die grossen Suchmaschinen unterstützen aus Gründen der Rechnerleistung meistens nur eine sehr einfache Normalisierung: Grossbuchstaben werden in Kleinbuchstaben umgewandelt oder Umlaute wie „ä“ auf „a“ abgebildet.

Soekia macht den Normalisierungsprozess transparent. Standardmässig werden Gross-

buchstaben und Umlaute normalisiert. Obendrein kann die Schülerin wählen, ob auch ein beschränktes, pseudo-linguistisches Wort-Stemming durchgeführt wird. Beim Stemming trennt Soekia 40 häufige Endungen ab, darunter Substantivendungen wie -heit, -keit und -ung, Adjektivendungen wie -bar, -er, -sten und -lich sowie Verbalendungen -end, -et, -st. Zusätzlich werden die Vorsilben ge-, ver- und un- abgetrennt. Dabei kann sowohl *over-stemming* wie auch *under-stemming* auftreten. Beim *over-stemming* werden Wörter auf einen gemeinsamen Stamm reduziert, obwohl sie semantisch nicht miteinander verwandt sind. Beispiel: Das Substantiv „Versicherung“ und das Reflexivpronomen „sich“ werden beide auf „sich“ abgebildet. Das Gegenteil heisst *under-stemming* und ist im Deutschen ohne Wörterbuch nicht zu verhindern. Stark gebeugte Verben wie „gehen, gingen, gegangen“ lassen sich nicht durch Abtrennen von Endungen auf denselben Stamm reduzieren. Für die englische Sprache verwendet Soekia den berühmten Porter-Algorithmus [Por80].

Ausserdem besteht in Soekia die Möglichkeit, häufige Wörter (sog. Stoppwörter) zu eliminieren. Stoppwörter wie „auf, der, die, das, ein, in“ haben wenig Informationsgehalt. Deren Erfassung würde aber den Index stark anwachsen lassen.

Dass eine gute Normalisierung für eine erfolgreiche Suche entscheidend ist, kann mit kleinen Experimenten in Soekia sehr gut gezeigt werden. In unserer Beispiel-Kollektion „Ozon“ findet man mit der Suchanfrage „Ozonloch“ ohne Stemming nur ein Dokument, mit Stemming hingegen drei (vgl. Abb. 4).

5 Rangierung: Präsentation der gefundenen Dokumente in der Reihenfolge ihrer Relevanz

Zum Suchbegriff „Ozonloch“ wurden 22 000 Dokumente gefunden; für einen Benutzer ein Ding der Unmöglichkeit, alle diese Dokumente zu inspizieren. Für die Qualität einer Suchmaschine ausschlaggebend ist deshalb auch die Reihenfolge, in welcher die gefundenen Dokumente dem Benutzer angezeigt werden. Die relevantesten Treffer sollen in der Rangliste weit oben erscheinen. Untersuchungen des Benutzerverhaltens haben gezeigt, dass weniger als ein Drittel der Benutzer drei oder mehr Dokumente in der Rangliste einer Suchmaschine besuchen. Die Tendenz der letzten Jahre geht sogar dahin, dass nur noch der erste und allenfalls zweite Treffer in der Rangliste inspiziert wird [SJWS02].

Für jede Suchmaschine ist es also ein Muss, die relevantesten Dokumente ganz oben in der Rangliste aufzuführen. Nun wird keine Suchmaschine je genau wissen können, was für den Anfragesteller relevant und was irrelevant ist. Trotzdem verfügen heutige Suchsysteme über ausgeklügelte Verfahren zum Erstellen der Ranglisten. Das Zauberwort heisst *relevance ranking*. Damit meint man das Anordnen von Dokumenten gemäss absteigender Relevanz bezüglich einer Anfrage.

Relevance Ranking läuft in zwei Schritten ab: Nachdem die Benutzerin eine Suchanfrage gestellt hat, werden alle verfügbaren Dokumente mit der Anfrage verglichen. Bei diesem Vergleich entsteht für jedes Dokument ein Relevanzwert. Je höher der Relevanzwert ausfällt, desto wahrscheinlicher stuft das Suchsystem das Dokument bezüglich der Anfrage als relevant ein. Nun sortiert das Suchsystem die Dokumente aufgrund des Relevanzwertes in absteigender Reihenfolge. Die so entstehende geordnete Liste wird

Abbildung 4: Suchanfrage ohne Wort-Stemming (oben) und mit Wort-Stemming (unten)

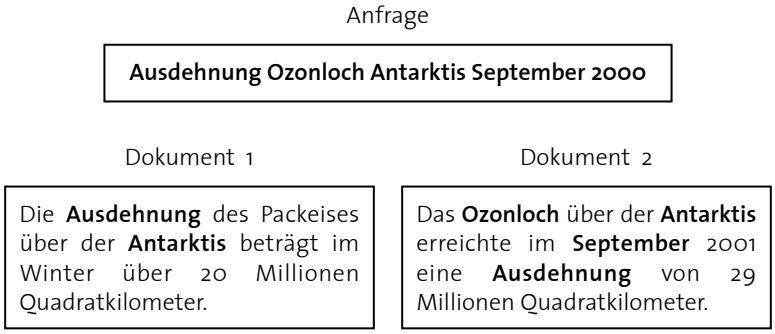
Rangliste genannt. Die Suchmaschine präsentiert die Rangliste der Benutzerin, die je nach ihrem Bedürfnis wenige oder viele Dokumente daraus auswählt und genauer betrachtet.

Man kann sich zwei vollkommen unterschiedliche Bedürfnisse bei einer Recherche vorstellen: Die Physikerin auf der Suche nach dem Zahlenwert von Pi auf 40 Stellen genau gibt sich mit *einem einzigen* relevanten Dokument zufrieden. Sie ist an einer hohen *Präzision* interessiert. Ein Patentanwalt hingegen muss abklären, ob für eine neue Erfindung bereits ein Patent existiert. Deshalb möchte er natürlich *möglichst alle* relevanten Dokumente auffinden, die ähnliche Erfindungen beschreiben. Er ist an einer möglichst hohen *Ausbeute* interessiert. Er wird also einen grösseren Teil der Rangliste in Betracht ziehen als die Physikerin. Durch das Sortieren der Dokumente in der Rangliste gemäss ihrer Relevanzwerte wird diesen zwei völlig entgegengesetzten Bedürfnissen gleichzeitig Rechnung getragen.

Wie geht nun ein Suchsystem konkret vor, um die Relevanz eines Dokuments bezüglich

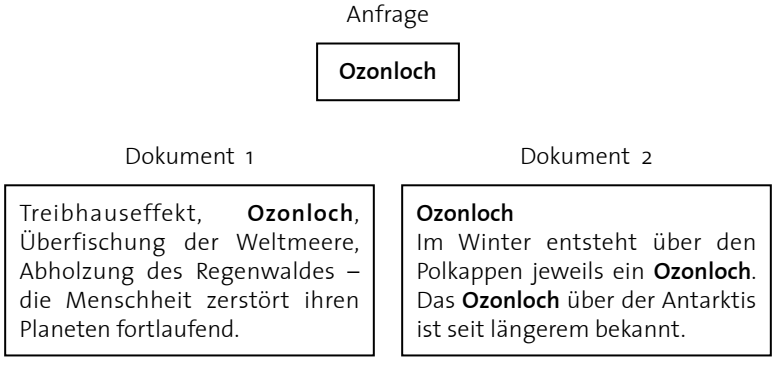
einer Anfrage zu berechnen? Das Vorgehen basiert auf einer wichtigen Annahme: Die Vorkommen von Suchbegriffen in einem Dokument geben Hinweise auf die Relevanz dieses Dokuments. Diese Annahme bildet die theoretische Grundlage für wissenschaftliche Modelle zur Berechnung der Relevanz. Die teilweise komplexen mathematischen Hintergründe sollen uns hier nicht interessieren. Wir illustrieren die drei wichtigsten Rangierungsregeln (gemäss [HNS00]) anhand von Beispielen in Soekia.

Rangierungsprinzip 1: Je mehr Suchbegriffe in einem Dokument vorkommen, desto relevanter ist das Dokument.



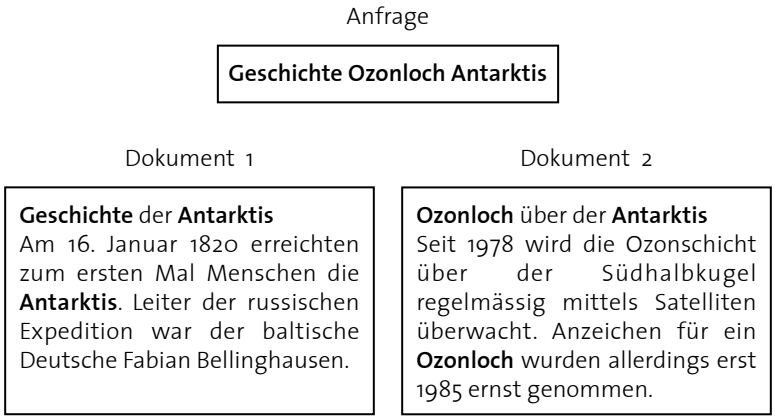
Das zweite Dokument enthält vier der fünf Suchbegriffe und wird deshalb als relevanter eingestuft als das erste Dokument, welches nur zwei Suchbegriffe enthält. Für die Anfragestellerin hat dieses Rangierungsprinzip eine unmittelbare Konsequenz: Je mehr gut gewählte Suchbegriffe die Anfrage enthält, desto besser kann die Suchmaschine die Rangliste erstellen. Es lohnt sich daher, viele Suchbegriffe zu verwenden.

Rangierungsprinzip 2: Je häufiger ein Suchbegriff in einem Dokument vorkommt, desto relevanter ist das Dokument.



Das zweite Dokument enthält den Suchbegriff drei Mal und wird deshalb als relevanter eingestuft als das erste Dokument, welches den Begriff nur ein Mal enthält. Die Konsequenz für die Anfragestellerin lautet: Überlegen Sie sich, welche Begriffe im gewünschten Dokument häufig vorkommen.

Rangierungsprinzip 3: Dokumente, die seltene Suchbegriffe enthalten, sind relevanter als Dokumente, die häufige Suchbegriffe enthalten.



Beide Dokumente enthalten gleich viele Suchbegriffe und die Begriffe tauchen gleich häufig auf. Das zweite Dokument wird als relevanter eingestuft, weil der Begriff „Ozonloch“ spezifischer ist als der Begriff „Geschichte“. „Spezifisch“ bedeutet hier, dass der Suchbegriff in der gesamten Dokumenten-Kollektion selten vorkommt. Google findet zu „Geschichte“ 3,4 Millionen Seiten, zu „Ozonloch“ aber nur 20 000. Als Anfragesteller muss man deshalb möglichst nach seltenen, für das gewünschte Dokument treffenden Begriffen suchen.

Die oben beschriebenen drei Rangierungsprinzipien werden von den meisten Suchmaschinen berücksichtigt. Daneben gibt es noch unzählige andere Kriterien, die von Suchmaschine zu Suchmaschine variieren. Mit Soekia nicht illustrieren lassen sich natürlich die dokumentenunabhängigen Rangierungsprinzipen, etwa das von Google verwendete Page Ranking von Webseiten aufgrund ihrer Popularität.

6 Qualität der Suche: Ausbeute versus Präzision

Hat man eine Suche erfolgreich abgeschlossen, sollte man sich auch über die Qualität der Antworten Rechenschaft ablegen. Bei offenen Fragestellungen, also im Zusammenhang mit umfangreichen Recherchen, spielt die Ausbeute eine grosse Rolle. Wie viele der in der Dokumenten-Kollektion zur gestellten Anfrage relevanten Dokumente wurden auch gefunden? Bei den grossen Suchmaschinen kann keine Aussage zur Ausbeute einer Suche gemacht werden. Man weiss nie, ob nicht noch weitere relevante Dokumente vorhanden wären. Bei Soekia ist aber die Dokumenten-Kollektion vorgegeben und kontrolliert. Man kann also zu einer Fragestellung in einem Experiment beispielsweise 20 relevante Dokumente in der Dokumenten-Kollektion „verstecken“ und dann beobachten, wie viele Dokumente die Schülerinnen und Schüler finden. Damit kann relativ einfach das Bewusstsein für den Trade-Off zwischen Ausbeute und Präzision geschärft werden.

Ein verstärktes Bewusstsein für die Problematik geringer Ausbeute scheint uns gerade heute sehr wichtig. Google, die zur Zeit am meisten verwendete Suchmaschine, arbeitet in der einfachen Suche strikt präzisionsorientiert. Alle eingegebenen Suchbegriffe werden automatisch mit AND verknüpft, d. h. es werden nur Dokumente angezeigt, die *alle* Suchbegriffe enthalten. Als Konsequenz führen mehr Suchbegriffe zu einer kleineren Anzahl von Treffern und zu einer gefährlichen Einschränkung der Ausbeute. Nur die wenigsten Nutzer von Google sind sich dieser Gefahr bewusst.

7 Methodik der Informationssuche als Bestandteil der Allgemeinbildung

Der Beschaffung von Informationen im Internet kommt eine immer grössere Bedeutung zu. Mit der zunehmenden Miniaturisierung bzw. Ubiquitious Computing werden wir bald in jeder Situation Zugriff auf die Informationen im Web haben. Das Handy wird bald über einen Google-Button verfügen, die Windschutzscheibe im Auto nicht nur die aktuelle Position, sondern auch weitere Informationen wie Hotels in der näheren Umgebung, Öffnungszeiten von Geschäften etc. anzeigen. Mit dieser technischen Entwicklung haben die Benutzer nicht Schritt gehalten.

Eine Schlüsselkompetenz in der Informationsgesellschaft wird es sein, die richtigen Fragen zu stellen, effizient und effektiv zu suchen und die gefundenen Informationen kritisch zu hinterfragen. Dazu ist ein Verständnis für die den Suchmaschinen zugrunde liegenden Konzepte notwendig. Wer eine Vorstellung zur Bildung und Form des Indexes einer Suchmaschine hat, wer die Kriterien bei der Erstellung der Trefferranglisten versteht, kann gezieltere Suchanfragen stellen. Soekia ist ein einfaches Werkzeug, das es im Unterricht erlaubt, aus der *black box* einer Suchmaschine zumindest teilweise eine *white box* zu machen.

Literatur

[HNS00] Hartmann, Werner, Michael Näf und Peter Schäuble: *Informationsbeschaffung im Internet, Grundlegende Konzepte verstehen und umsetzen.* Orell Füssli, Zürich, 2000.

[Por80] Porter, Martin F.: *An algorithm for suffix stripping.* Program, 14(3):130–137, 1980.

[Sar71] Saracevic, Tefko: *Selected results from an inquiry into testing of information retrieval systems.* Journal of the American Society for Information Science, 22(2):126–139, 1971.

[SJWS02] Spink, Amanda, Bernard J. Jansen, Dietmar Wolfram und Tefko Saracevic: *From E-Sex to E-Commerce: Web Search Changes.* IEEE Computer, 35(3):107–109, 2002.

[WI01] White, Marilyn Domas und Mirja Iivonen: *Questions as a factor in web search strategy.* Information Processing and Management, 37(5):721–740, 2001.