Diese Dokumentation wurde verfasst von:

Bernhard Kreinz

Systemspezialist / Webmaster

(Version 1.0 - August 1999)

Themenübersicht:

1.0 Vorwort

Dieser kleine Workshop soll dir Gelegenheit geben, dein Verständnis für die Suchdienste im Internet zu vertiefen. Nach welchen Kriterien werten Suchmaschinen Ergebnisse aus ... oder warum erscheint mein Dokument nicht zuoberst ? Wie funktionieren sogenannte Searchengines oder wie müssen Dokumente gekennzeichnet werden damit sie sinnvoll gefunden werden ? - Wir werden uns in diesem Workshop mit den verschiedensten Themen auseinandersetzen, nicht alle werden direkt in Zusammenhang mit Soft- oder Hardware stehen. So stellt sich beim Thema "Suchen" im Internet/Intranet ein grundlegendes philosophisches Problem. Wer sucht hat bereits eine Vorstellung von dem, was er zu finden hofft. Diese Vorstellung deckt sich aber nicht immer mit dem, was man schriftlich in einem oder mehreren Begriff zu fassen vermag. Suchbegriffe und Suchergebnisse stehen nämlich nicht immer im gewünschten Verhältnis. So muss man wissen wem man seine Abfragen in den Schlund wirft. Wenn ich mit einem Japaner deutsch rede, dann ist es nicht selbstverständlich, dass er mich versteht. Jeder Searchengine - Betreiber bietet in der Regel einen "Help"-Link an, welcher zu den notwendigen Informationen führt, die du brauchst, um den jeweiligen Bestand erfolgreich abzufragen. "Suchen" ist eine Kunst für sich.

Als Literatur empfehle ich dir die "Suchfibel" im Internet. Dort steht wirklich alles wissenswerte rund um das Thema. Die gesamte Ausgabe ist online im Netz abrufbar ! (-> unter http://www.suchfibel.de)

Spätestens nach diesem Workshop wirst du die Funktionsweise oder Arbeitsmethode von Searchengines begriffen haben. Mein Ziel ist es dir beizubringen, dass Sucheingabe und Suchergebnis in einem direkten Zusammenhang stehen. Da der Datenbestand im Internet stetig steigt, aber keine Suchmaschine alles aktuell halten kann, geschweige denn vollständig, kann auch nicht nur eine Suchmaschine konsultiert werden, ausser ...

Ich unterbreche hier, weil das Dokument sonst ausarten würde. Wie schon erwähnt lege ich euch die "Suchfibel" sehr nahe.

Dieses Skript strukturiert sich folgendermassen:

Alle für das SIZ Zertifikat relevanten Angaben in Bezug auf das Suchen im Internet sind in diesem Skript zusammengefasst !

In einem ersten Teil werden wir über Suchstrategien reden müssen und uns ein wenig mit der Geschichte des Suchens auseinandersetzen. Was heisst den Suchen in Datenbeständen. Ich suche grundsätzlich nach einem "String" in einem bestimmten Kontext. Was meine ich damit ?

Bin ich gerade mal schnell an einer Information interessiert oder brauche ich Daten für eine Recherche. Wie gehe ich jeweils vor und wo beginne ich mit meiner Suche ?
In einem zweiten Teil werden wir versuchen Ordnung in das Begriffschaos zu bringen. Fulltext, Katalog, Meta Engine, Spezialisiert Suchmaschinen ... u.a.
Im Anhang findest du dann noch verschiedene andere verwandte Kapitel.

Vielleicht noch ein paar Worte zu meiner Person ... Nein, die spar ich mir. Das kann man unter http://www.barnes.ch nachlesen. Zusammenfassend lässt sich vorwegnehmen, daß der Stoff, den du hier vermittelt bekommst, zu deinen Kernkompetenzen zählt.

Du wirst sehen, daß der Workshop sich in zwei grundsätzliche Kategorien aufteilen lässt:

Produktespezifischer Workshop
Produkteübergreifendes Verständnis und Know-how

Dieses Dokument wird daher sehr allgemein gehalten und soll als Leitfaden dienen für deine weitere Tätigkeit . Wir wollen die Aspekte in diesem Dokument vorhanden wissen, welche für alle Produkte (in diesem Themenbereich) gelten. Wie soll ich denn wissen mit welcher Search Engine du arbeiten willst oder musst? Handelt es sich um einen Katalog oder oder arbeitest du mit einer Fulltext-Search-Engine ? So werden wir uns an den Standards orientieren, welche von verschiedenen "Institutionen" wie Altavista, Yahoo usw. zu den jeweiligen Themengebieten vorgegeben werden, dies meint deren Verwendungszweck.
Näheres dazu aber im nächsten Kapitel.

Jetzt beginnen wir aber ....

2.0 Die Suchstrategie oder die Kunst des Suchens

Wir unterscheiden hier grundsätzlich zwischen folgenden Arten des Suchens:

- ➔ kurzfristige Suche (nach dem Motto "auf zur schnellen Info")
- ➔ schnelle Recherche (wenig Zeit zum Suchen aber viel Information als Resultat einer Abfrage)
- ➔ intensive Recherche

Suchen ist ja nicht gleich Suchen. Wie vieles im Leben suche in nach etwas Bestimmtem. Diese Etwas kann vielleicht sogar benannt werden und zwar in einem Begriff (Namen). Wenn ich also nach etwas bestimmten suche, dann setzte ich voraus, dass dieses in irgendeiner Form auch vorhanden ist, d.h. auch gefunden werden kann. Abgesehen von den Sprachunterscheidungen (englisch, deutsch, spanisch, japanisch usw.) werden wir auch mit anderen Schwierigkeiten zu kämpfen haben, doch dazu später mehr. Jetzt wollen wir uns mit einem grundsätzlichen Ansatz beschäftigen, nämlich mit der Motivation. Wann suchen wir was und zu welchem Zweck.

2.1 Die schnelle Info oder Die kurzfristige Suche

Annahme:
Wenn ich mal schnell bei Ziff-Davis vorbeischaue, um den Newsticker zu konsultieren, dann werde ich, sofern ich kein Bookmark gesetzt habe, entweder über die Homepage http://www.zdnet.de oder über die URL http://www.zdnet.de/news direkt einsteigen. Ich weiss also was ich wo zu erwarten habe.

Suche nach News (auch Wirtschaft oder Finanz-Börsen) verläuft in der Regel nach diesem Schema. Ich weiss wo ich welche Informationen finde, d.h. der Anbieter ist mir bekannt. Hier geht es in der Regel um tagesaktuelle Informationen, welche in Form eines Listings oder Tickers angeboten werden.

Wenn ich nur mal schnell nach Hintergrundinformationen zu einem bestimmten Thema suchen will, dann treten erstmal andere Kriterien in den Vordergrund. Menschlich, allzumenschlich ist, dass jedermann Vorlieben hat. So werden Suchmaschinen besucht und angefragt, aber nicht anhand von Kriterien wir "Thema", Umfang, Geschwindigkeit, sondern mittels dem Faktor Gewohnheit.

Dann wird auch nicht zwischen Katalog und Volltextsuchmaschine unterschieden.

Da "Yahoo" zu den Pionieren gehört, wird auch gerne mal dort nachgeschlagen, auch wenn die Aktualität und der Umfang geradezu bescheiden sind. Man kennt

Yahoo als Label. Mal schnell innerhalb des Kataloges in die gewünschte Kategorie einsteigen und einen Suchbegriff eintippen führt aber trotzdem zu brauchbaren Ergebnissen.
Bei Volltext-Suchmaschinen wie "Altavista" muss für ein sinnvolles Ergebnis eine gezieltere Suchsyntax verwendet werden.

Dieses und mehr zu Abfrageoptionen findest du später in diesem Skript.

2.2 Schnelle Recherche - Wenig Zeit und viele Infos

Ähnlich wie in 2.1 verhält es sich mit diesem Ansatz.

Annahme:
Ich habe wenig Zeit, will aber ein Maximum an relevanten Informationen zum Thema das mich interessiert.

Natürlich muss auch hier zwischen Suchmaschinen und anderen Informationsanbietern unterschieden werden. Informationsanbieter wie Reuters oder aber auch Yahoo, bieten zu den unterschiedlichen News auch weiterführende Links zu Hintergrundinformationen an. Dies gilt ebenso für Tageszeitungen oder allgemeiner für alle Informationsanbieter, welche wir bereits aus anderen Medien kennen und welche im Internet Präsents zeigen. Für viele Unternehmen ist das Internet lediglich ein weiterer Vertriebskanal. So werden Informationen auch verkauft. Wie zum Beispiel Echtzeit-Online-Kurse von Aktien.

So gilt auch hier die Unterscheidung zwischen Suche nach Tagesaktualität mit vertieftem Hintergrund, welche in der Regel kostenpflichtig sind, oder der Suche in einem unbekannten Fundus an Informationen mittels einer Abfrage, d.h bei einer Suche in unstrukturierten Datenbeständen.

Die Empfohlene Vorgehensweise lautet hier:
Als erstes in einem Katalog nachschlagen (wie z.B. Yahoo oder einem ähnlichen Dienst - web.de. Wenn dort nichts (oder wenig brauchbares) gefunden wird, dann empfiehlt sich eine Suche in einer Volltext-SearchEngine.

2.3 Intensive Recherche

Hier muss wohl aus allem geschöpft werden, was das Internet hergibt. Angefangen von Standardlösungen wie Suchmaschinen können auch Bibliotheke oder spezialisierte Suchmaschinen angezapft werden. Es gilt halt immer noch - viele Dokumente sind auch heute noch nicht in digitalisierter Form abrufbar,

sodass in grossen Bibliotheken wenigstens eine Referenz auf ein Dokument in herkömmlicher Form abrufbar ist. Meistens kann dann eine Onlineresreservation getätigt werden oder sogar eine Heimlieferung (zeitlich beschränkt) verlangt werden. Man gibt sich zwar allerorts Mühe konventionelle Datenbestände in digitaler Form anzubieten, allerdings ist dieses Vorhaben sehr kostspielig und zeitintensiv, zumal die Texte dann auch entsprechend gekennzeichnet werden müssen um wieder gefunden zu werden.

3.0 Kleine Geschichte zum Suchen und Finden

Wir werden an dieser Stelle nicht die Historie von den verschiedenen Suchmaschinen angehen, sondern den Ansatz verfolgen der dieser Methode (meint das Suchen) vorangeht. Wie bereits im Vorwort beschrieben ist das Suchen eines der wesentlichen Merkmale menschlichen Daseins und mit eine Faktor für dessen rasante Weiterentwicklung (in kognitiver Hinsicht).

Das Suchen selber muss wie auch immer im Kontext des gewünschten Zieles gesehen werden. Wenn ich nach einer Nadel in einem Heuhaufen suche, dann gibt es verschiedene Ansätze:
Ich kann einen getrockneten Grashalm nach dem anderen auf die Seite legen. Irgendwann finde ich die gesuchte Nadel sicher. Ich kann aber auch mit einem grossen Magneten dahinter gehen und so die Suchzeit verkürzen. In beiden Fällen bin ich systematisch vorgegangen. Beide Varianten führen zum Ziel. Die zweite, schnellere Alternative aber unterschiedet sich qualitativ von der ersten insofern, als dass hier mit einer Eigenschaft des gesuchten Objektes gearbeitet wurde.

Informationstechnisch gesehen unterscheiden wir folgende Kategorien:

Suchen in Texten
Suchen in einem Dateisystem
Suchen in einer Datenbank
Suchen in einer Metadatenbank
Über strukturierte und unstrukturierte Daten

3.1 Suche in Texten

In der Informationstechnologie wurde von Beginn an mit Suchmöglichkeiten gearbeitet. Jeder trivialste Texteditor bietet eine Möglichkeit zur Suche innerhalb des Textes nach einem beliebigen String. Ebenfalls jeder Browser hat diese Option welche mittels CTRL+F aufgerufen werden kann. (Achtung bei Frames, dort muss vor der Tastenkombination zusätzlich das gewünschte Frameset mit dem Mauszeiger aktiviert werden.)

3.2 Suche in einem Dateisystem

Alle Betriebsysteme bieten Suchmöglichkeiten an welche von der Suche nach Dateinamen bis zur Suche innerhalb von Dateien gehen. Letzteres ist allerdings sehr rechenintensiv und kann praktisch den ganzen Computer lahmlegen, da für eine Suche innerhalb von Dateien nicht nur alle Dateiattribute, sondern jede einzelne Datei selber geöffnet und zeichenweise mit dem Suchstring verglichen werden muss, um anschliessend vielleicht als Treffer ausgewiesen werden zu können. Der Vorteil aber liegt auf der Hand. Das Suchergebnis bezieht sich immer auf den aktuellsten Stand. Wenn kein Treffer ausgegeben wird, darf man annehmen, dass das Gesuchte auch nicht vorhanden ist. Hier können aber Konflikte auftauchen, wenn verschiedene Benutzer gleichzeitig auf dieselbe Datei zugreifen wollen oder müssen. Je mehr Dateien vorhanden sind, desto länger dauert die Geschichte und es wird auf der Festplatte gerattert was das Zeug hält.

3.3 Die Suche innerhalb einer Indexdatei

Um die Probleme mit 3.2 in der Griff zu kriegen,versuchte man sich mit folgendem Ansatz. Man trennte den Suchprozess in zwei Teile auf. Der eine Teil erstellte eine Indexdatei mit allen relevanten Angaben zu den indizierten Dateien. Der andere Teil sollte den eigentlichen Suchvorgang durchführen, dies meint die Suche innerhalb der Indexdatei.
Das Suchergebnis verweist dann schliessend wieder auf das Original (die Quelldatei).

Vor- und Nachteile:

Bei diesem Verfahren muss nur noch eine Datei durchsucht werden, was den ganzen Suchprozess massiv beschleunigt. Was allerdings in den jeweiligen Index aufgenommen wird ist von Suchmaschine zu Suchmaschine verschieden. So indiziert "Indexserver" lediglich die ersten 320 Zeichen innerhalb des <body> Tag. Was in den Metatags <meta name=content/description/keywords value="Beschreibung"> steht findet auch seine Beachtung. Würde alles indiziert (ein komplettes Abbild der Informationen) so müsste der Index gleich gross wie die Summe der Quelldateien sein. Mit ein paar ausgeklügelten Algorithmen kann der Index allerdings kompriemiert werden, doch auch dann noch muss dafür ein Computer mit genügend Plattenplatz zur Verfügung stehen. Übersteigt die Indexgrösse die Grösse des physisch vorhandenen Memory (RAM), dann muss auf die Festplatte ausgelagert werden. Dies kann aber wieder zu starken Performace-Einbussen führen. Da aber immer nur die Indexdatei durchsicht wird und das Ursprungsdokument zwischenzeitlich geändert werden kann, ist der Index immer "veraltet", d.h. die Indexsoftware muss immer wieder von neuem Beginnen. Irgendwann ist der Aufwand so gross, dass nur noch der bestehende Index gepflegt wird und keine neue Seite mehr aufgenommen werden kann, oder

aber altes wird nicht mehr upgedated. Um derlei Probleme wissen die Suchmaschinenbetreiber aber genügend.

Dennoch arbeiten die meisten Volltext-Suchmaschinen nach diesem Prinzip. Wie zum Beispiel "Altavista" oder Lycos. Grundsätzlich gilt oben genanntes Prinzip, dennoch bieten die Dienste auch Abfragen auf Felder an. Dies meint das Vermögen auf spezifische Eigenheiten von HTML Dokumenten einzugehen, wie z.B. den Tag <title> oder eben die Metatags. Auch der DNS Name innerhalb der URL (meint den Hostname) dient als wertvolles Merkmal.

3.4 Die Suche innerhalb einer Datenbank

Die Zuordnung von Inhalten in bestimmte Kategorien (oder Oberbegriffe) scheint für Ordnungsfanatiker selbstverständlich. Wie aber können wir der Menge an unstrukturierten Daten begegnen, welche im Internet anzutreffen sind ? Dass bestimmte Inhalte nicht irgendwo in einem HTML Dukument stehen, sondern in bestimmten Felder eingetragen werden, um anschliessend sinnvoll ausgewertet zu werden klingt gut. Doch wer passt schon immer die metatags description, keywords oder content an ? Der <title> wird ja meistens schon ungenügend gesetzt.
Die Zuordnung von Informationen zu Feldern ermöglicht natürlich ein genaueres Suchen. Wenn ich in einem Namensfeld nach "Müller" suche, dann erhalte ich als Resultat eine Liste mit Personen und dem Namen Müller und nicht die Berufsbezeichnung - eine Volltextsuchmaschine würde letzteres ebenfalls als Treffer ausgeben. Das Arbeiten mit Tabellen Feldnamen und Feldwerten bietet viele andere, zusätzlichen Möglichkeiten. Man kann Felder in Beziehung zueinander setzten und Datensätze zusammenfassen. Mehrere Datensätze ergeben eine Datenbank.

Beispiele für DB orientierte Systeme im Internet sind Bibliotheken und häufig auch Websites von Printmedien. Z.B. Tageszeitungen usw. Websites auf der Basis von Lotus Domino sind meistens von Natur aus dynamisch generiert.

Bemerkenswert ist, dass auf allen relevanten Betriebsystemen Webserver mit Datenbankservern gekoppelt werden können.

3.5 Struktur von Daten - Exzerpt aus der Suchfibel©

Wenn nicht explizit mit Felder (DB) gearbeitet wird, sondern ein beliebiges Dokument (bzw. sein Inhalt) indiziert werden soll, dann muss man sich gewisse Methoden überlegen wie dies zu realisieren ist, ohne dass man die eigenen Resourcen überfordert.

Triviales Beispiel wie ein Index aufgebaut werden kann oder wie man etwas Struktur in unstrukturierte Daten gibt soll untenstehende Tabelle verdeutlichen.

Bei ein paar Millionen Dokumenten und damit auch Stichworten ist diese (virtuelle) Tabelle natürlich ein wenig größer... Mit Hilfe dieser Indextechnik läßt sich die Größe der indexierten Dokumente auf ca. 4% reduzieren.

4.0 Über Kataloge und die Volltextsuche

Die Anbieter von Suchmaschinen unterteilen sich in folgende Kategorien:

Volltextsuche
Katalog
Kooperation
Spezialisierte Suchmaschinen
Das HTML Dokument - Was wichtig ist für eine maschinelle Indizierung

- ➔ Katalog - Grundlage für den Aufbau liefert der menschliche Geist
- ➔ Volltext Searchengine - ausgeklügelte Software, leistungsfähige Rechner und viel Bandbreite sind Voraussetzung für diesen Ansatz.

Grundsätzlich unterscheiden sich Katalog und "Fulltext" Searchengine folgendermassen. Der Katalog versucht Struktur in seinen Datenbestand zu bringen. Die Fulltextengines greifen sich einen Punkt (Startseite) und verfolgen weitere Links.

Je länger je mehr beginnen verschiedene Anbieter beide Konzepte zu verschmelzen oder aber es werden Kooperationen gebildet. (Yahoo mit Inktomi , web.de mit Lycos)

4.1 Katalog

Wenn ein bestimmtes Dokument oder gar eine Website in einen Katalog wie z.B. Yahoo aufgenommen werden soll, dann wird dieses Anliegen zuerst von Hand in der gewünschten Kategorie platziert und anschliessend redaktionell weiterbearbeitet ... und zwar von einem Yahoo-Mitarbeiter. Das Ausnützen von bestimmten Eigenschaften des HTML Dokumentes ist hier nicht von Bedeutung, weder der <title> noch <Meta> Tags werden hier beachtet, Sondern lediglich die Eingaben, welche bei der Registrierung angegeben wurden. Das heisst, dass hier ebenfalls mit Datenbanken und Felder gearbeitet wird. Die Felder sind aber durch den Kataloganbieter definiert.

Vorteile bei Aufnahme in einen Katalog:

- ➔ der menschliche Geist entscheidet die Zuordung einer Seite zu einem bestimmten Inhalt bzw. Kategorie. In der Folge sollte der Inhalt besser wiedergefunden werden, da der Aufbau in der Regel hierarchisch geordnet ist und man so das Suchgebiet immer mehr einschränken kann.

➔ Zu jedem Eintrag wird redaktionell eine Beschreibung angefügt. In anderen Fällen werden zusätzlich Stichworte verlangt.

Nachteile von Katalogen

➔ da redaktionelle Arbeit notwendig ist, verzögert sich die Aufnahme in den Katalog um mehrere Wochen.
➔ Der Katalog kann fast nicht mehr aktuell gehalten werden. (Es gibt zwar Programme die checken können, ob der Link noch aktuell ist - ob sich der Inhalt wesentlich geändert hat wird kaum überprüft.
➔ Der Umfang des Kataloges kann mit dem einer Fulltext-Searchengine nicht mithalten. Das Wachstum ist schlicht zu gross. (Meint: Die Anzahl von Seiten wächst stetig)
➔ Probleme bei Webseiten mit dynamischem Inhalt. (Dynamische Webseiten haben ja gerade den Vorteil immer aktuell zu sein ....)
➔ Es gibt Kataloge die keine Verknüpfungen in der Abfrage zulassen.
➔ Das Definieren von Kategorien ist eine wissenschaftliche Disziplin für sich.

... aber das schauen wir uns am besten einmal "online" an ....

4.2 Volltext Suchmaschinen - Fulltext Search Engine

Die Fulltext Searchengine arbeitet heute nach folgendem Prinzip. Es werden drei Komponenten gebraucht. Eine Komponente durchsucht das Web (Crawler, Spider etc.) und übergibt die Resultate einer Indexierungssoftware. Die dritte Komponente ist die Suchsoftware welche der Internetanwender gebraucht und welche den Index abfragt.

Über "add URL" kann in verschiedenen Volltext-Suchmaschinen eine Startseite angegeben werden, welche in den Index aufgenommen werden soll.

Vorteil von Fulltext Search Engines:

➔ Da hier automatisiert vorgegangen wird, ist dieses Verfahren sehr effizient.
➔ Wer die Suchsyntax beherrscht, findet hier die meisten Resultate, da dieser Typus von Searchengine in der Regel den grössten Fundus in sich birgt.
➔ Immerhin wird eine einmal eingetragene Url auch später wieder besucht. Das garantiert wenigstens ein Minimum an Aktualität.
➔ Hier werden in der Regel die verschiedenen (später aufgelisteten) Metatags berücksichtigt.
➔ Volltext ist mit Vorsicht zu geniessen ... in der Regel handelt es sich lediglich um die ersten 320 Zeichen im Body des Dokuments.

➔ Es gibt verschiedene Anbieter (z.B. Altavista), welche sogar verlinkte Dokumente anderen Typs indizieren. Z.B. pdf,doc,xls,nsf u.s.w.

Nachteil von Fulltext Search Engines:

➔ will man eine solche Suchmaschine selber betreiben, muss auf eine entsprechende Infrastruktur geachtet werden. Gegebenenfalls müssen die unterschiedlichen Komponenten auf unterschiedlichen Systemen betrieben werden.
➔ Auch sie werden mit der Informationsflut nicht fertig und können nur einen bestimmten Teil abdecken.
➔ Fulltext beschränkt sich in der Regel auf die ersten 320 Zeichen im Body Tag
➔ Das Ranking wird zum Teil manipuliert
➔ Die Kenntnis der Abfragesyntax ist Voraussetzung für eine erfolgreiche Abfrage
➔ Die Terminologie des Autors muss berücksichtigt werden, d.h. man muss seine Abfrage eventuell ändern.
➔ Da der Datenbestand immer grösser wird und vielbesuchte Searchengines alle Anfragen bewältigen wollen, werden pro Abfrage nur eine bestimmte Zeitspanne zugelassen. Je mehr aktuelle Anfragen anstehen desto kürzer ist das Zeitfenster. Es kann also sein, dass die Anzahl Treffer (bei geleicher Abfrage) unterschiedlich sind.

4.3 Metasuchmaschinen

Metasuchmaschinen machen eigentlich nichts anderes, als Suchanfragen an andere Suchmaschinen weiterzuleiten und die Resultate gesammelt wiederzugeben. Der eigentliche Challenge bei diesem Unterfangen liegt darin, bei komplexeren Abfragen die Syntax des Zielsystems zu beachten und gegebenenfalls umzustrukturieren.

4.4 Kooperation

Wie bereits erwähnt gibt es in der Weiterentwicklung von Suchmaschinen verschiedene Ansätze. Angefangen von Metasuchmaschinen bis hin zu Kooperationen. Eine Kooperation meint im einfachsten Fall: Falls es zu einer Abfrage kein Ergebnis gibt, wird die Abfrage an eine "befreundete" Dsuchmaschine weitergeleitet - in der Hoffnung dort auf ein Ergebnis zu stossen. Damit soll der Frust beim Anwender etwas gemässigt werden.

5.0 Wie eine Suchmaschine gefüttert werden will

Mathematik und Logik - schon mal gehört ?
Boole'sche[1] Verknüpfungen
Trunkierung
Suchen in Feldern

Grundsätzliche Kenntnisse der einfachen Mengenlehre (Mathematik) und der Logik sind unabdingbar für den gezielten Umgang mit Suchmaschinen. So werden hier ansetzen und ein bisschen üben.

➔ Verknüpfungen:

Unter einer Verknüpfung (Operator) versteht man eine logische Beziehung zwischen zwei Objekten (Hier Begriffen).

5.1.1 Die UND Verknüpfung:

(A UND B UND C) ist dann wahr, wenn A wahr ist und B wahr ist und C wahr ist.

5.1.2 Die ODER Verknüpfung:

(A ODER B) UND C ist dann wahr, wenn A wahr ist oder B wahr ist (oder beide) und auf jedenfall C wahr ist.

5.1.3 NICHT - Die Negation:

A UND (B ODER C) NICHT D

A muss sein . B oder C ist egal (mindestens eines von beiden). Aber bitte ohne D

Fazit:

[1] **Der Mathematiker George Boole, 1815 in England geboren, wurde mit seinem Werk The Mathematical Analysis of Logic bekannt. Diese Publikation demonstrierte erstmals, daß sich die Aristotelische Logik als algebraische Strukturen darstellen läßt. Boole: "Wir sollten nicht länger Logik und Metaphysik, sondern Logik und Mathematik miteinander verbinden."**

Mit UND müssen alle Ausdrücke vorkommen (wahr sein) und bei ODER muss mindestens einer der Begriffe vorkommen - es dürfen aber auch alle zutreffen, welche mit ODER verbunden werden. Mit NICHT werden Begriffe ausgeschlossen.

Achtung:

Andere Länder andere Sprachen. In der Regel werden die Operatoren (so nennt man die Verknüpfungen) in englischer Sprache verwendet und zwar in Grossbuchstaben, um sie von anderen Worten unterscheiden zu können.

Deutsch	Englisch
UND	AND
ODER	OR
NICHT	NOT

Soviel zur einfachen Megenlehre. Jetzt kommen wir zu etwas, das sich "Modal Logik" nennt und eine eigentliche Weiterentwicklung der einfachen Logik darstellt. Dies ist eine eigentliche Untertreibung, aber die Thematik "Modal Logik" würde den Rahmen der Vorlesung sprengen.

5.1.4 Vom Plus und Minus

Wir lehnen uns jetzt an oben genannten Beispielen an.

+A +B +C -> meint sowohl A als auch B und C müssen wahr sein.

+(A B) +C -D -> A oder B müssen wahr sein (oder beide) und auf jeden Fall C, wobei D nicht vorkommen darf.

Man kann hier Analogien zu den anderen Verknüpfungen ziehen, sodass ich mir ein detaillierte Erklärung spare.

5.1.5 Der Nachbar

Es gibt da noch einen Operator, den man einsetzen kann. Er heisst NEAR und meint einen relativen Abstand im Text zu anderen Suchbegriffen.

A NEAR B

A muss also in der Nähe von B vorkommen. Unter Nähe werden von Suchmaschine zu Suchmaschine unterschiedliche Massstäbe angesetzt. Die Zahlen hier schwanken zwischen 20 und 200 Worten.

5.1.6 Die Abkürzung

Was hier unter Abkürzung verstanden wird nennen die Profis Trunkierung. Die sogenannte Arbeit mit Wildcards (Platzhaltern) ist sehr beliebt kann aber auch zu nicht geahnten Ergebnissen führen.

In der Regel ist der Platzhalter, welcher gesetzt werden kann ein Asteriks *. Es kann aber auch mal das Prozentzeichen vorkommen.

Bei der Eingabe von Auto* wird beispielsweise nach Automobil, Automation, Automatismus, Autoreisezug und nach wer weiß noch gesucht. Wildcards lassen sich auch elegant einsetzen, um nach verschiedenen Schreibweisen eines Fachbegriffes gleichzeitig zu suchen. Gra*ik such nach Graphik und Grafik zugleich.

5.1.7 Phrasen

Untern Phrase versteht man eine besondernen Ausdruck. Beispiel: "Bernhard Kreinz" . Wie mehrfach erwähnt vergeleichen Searchengine Strings, welche man übergibt zeichenweise. Der Ausdruck innerhalb der Anführungszeichen wird deshalb wie ein String behandelt. SO kann schlussendlich auch nach "Max und Moritz" gesucht werden.

5.1.8 Grosse und kleine Buchstaben

Wird ein Buchstabe gross geschrieben, wird der Begriff als "case-sensitiv" markiert ! Gross- oder Kleinschreibung ist demnach relevant ! Im Zweifelsfalle lieber nur kleine Buchstaben eingeben.

5.1.9 Stopwords

Es gibt verschiedene Wort nach denen grundsätzlich nicht gesucht werden kann. Sei es weil sie zu oft vorkommen (z.B der, die das) ode weil sie bereits als Operatoren verwendet werden (Z.B. UND ODER NICHT)

Es gibt auch Suchmaschinen welche nur Worte indizieren welche mehr als 3 Zeichen beinhalten.

Sonderzeichen werden auch oft ignoriert. Wie z.B $ oder & . Meistens werden diese Zeichen im Programmablauf verwendet.

5.1.10 Die Felder

Es gibt ja im HTML Dokument verschiedene Tags, welche von verschiedenen Suchmaschinen als Feldnamen verwendet werden und deren Inhalt so im Index abgelegt werden. Es gibt aber auch andere Informationen welche besonders erkennbar sind wie z.B. Dateiendungen wie gif oder Hostname, Link u.v.m.
Nach diesen kann dann auch gesucht werden:

- ➔ "link:www.barnes.ch" findet zum Beispiel alle Dokumente welche einen Link zu www.barnes.ch aufweisen.
- ➔ "title:Suchfibel" findet entsprechende Dokumente
- ➔ u.s.w.

Näheres dazu findet sich in der Hilfe zu der jeweiligen Suchengine.

6.0 Die Antwort einer Suchmaschine

Da Suchmaschinen auch nur von Menschen programmiert werden, ist auch das Ranking oder die Auswertung nur im jeweiligen Kontext (meint die betreffende Suchmaschine) zu interpretieren.

Faktoren welche hier eine Rolle spielen:

- ➔ Geld - wer zahlt ist weiter oben
- ➔ URL - Title - Keywords, Description
- ➔ Die ersten 200 bis 300 Zeichen im Body
- ➔ Stopwords (negativ !)