INSTITUT FÜR INFORMATIK
DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Diplomarbeit

Evaluierung von Suchmaschinen für den Einsatz im BMW-Intranet

Bearbeiter:	Stephan Hager
Aufgabensteller:	Prof. Dr. Heinz-Gerd Hegering
Betreuer:	Dr. Kirsten Heiler Regina Grafe BMW AG

INSTITUT FÜR INFORMATIK
DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Diplomarbeit

Evaluierung von Suchmaschinen für den Einsatz im BMW-Intranet

Bearbeiter:	Stephan Hager
Aufgabensteller:	Prof. Dr. Heinz-Gerd Hegering
Betreuer:	Dr. Kirsten Heiler Regina Grafe BMW AG

Abgabetermin: 15. August 1997

Hiermit versichere ich, daß ich die vorliegende Diplomarbeit selbständig verfaßt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 15. August 1997

. . . .
(Stephan Hager)

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Einführung

Der Wandel von der Produktions- zur Kommunikationsgesellschaft bewirkt einen enormen Bedeutungsanstieg der Information. Daher wird der Besitz von Informationen immer relevanter. Es wird immer wichtiger, auf möglichst viele Informationen rasch zugreifen zu können. So wird z.B. ein Informationsvorsprung einem Unternehmen auch einen Wettbewerbsvorteil bringen.

Definition. *Information nennt man den abstrakten Gehalt einer Aussage, Beschreibung, Anweisung, Nachricht oder Mitteilung. Die äußere Form der Darstellung nennt man Repräsentation. [Broy 92]*

Somit macht erst die Zuordnung einer Bedeutung eine Repräsentation zu einer Information. D.h. auch Buchstaben und Wörter sind nur Repräsentation und sind ohne Deutungsfestlegung bedeutungsleer. Ein großes Problem bei der Suche nach Information ist daher, daß man nicht nach der Repräsentation, sondern nach deren Bedeutung sucht. In natürlichen Sprachen kommt es vor, daß zwei Wörter dieselbe Schreibweise, aber verschiedene Bedeutungen haben. Dieses Phänomen wird Homonymie genannt und führt zu gewissen Schwierigkeiten bei der Suche nach spezieller Information. [Bauer 91]

Andererseits weckt auch ein und dasselbe Wort in verschiedenen Zusammenhängen unterschiedliche Assoziationen. Da die meisten Veröffentlichungen aber aus Repräsentationen, Buchstaben und Wörtern, bestehen, kann bei einer Suche in einem Dokument nur deren Vorkommen oder Nichtvorkommen überprüft werden. Notwendig wäre aber eine semantische Analyse und nicht eine syntaktische. All dies erschwert die Suche nach Informationen. Noch schwieriger wird die Suche nach Musik, Bildern oder Videos. Hier ist die Information nicht durch Buchstaben, sondern durch Töne und Bildpunkte repräsentiert. Hinzu kommt die Tatsache, daß die Menge der Informationen stetig steigt.

Diese Problematik ist nicht neu und wurde nicht erst durch den Einsatz von Computer-Netzen hervorgerufen. Sucht man z.B. in einem Buch eine bestimmte

Information, so wird man weder das Buch von vorne durchlesen, bis man zu der gesuchten Stelle gekommen ist, noch wahllos irgend welche Seiten aufschlagen, um zu schauen, ob man zufällig die richtige Seite getroffen hat, sondern im Register den entsprechenden Begriff suchen. Daher ist es nur konsequent, wenn man sich bei der Informations-Suche in den Computer-Netzen ähnlicher Hilfsmittel bedient. Da es erheblichen Aufwand bereitet, ein Netz, bestehend aus vielen Rechnern, zu durchsuchen, liegt es nahe, dafür Programme zu verwenden. Diese werden Suchmaschinen genannt. Dabei liegt das Hauptaugenmerk nicht auf dem Suchen, sondern vielmehr auf dem Finden von Information. Im Jargon werden diese Programme oft auch als Robots, Search-Engines, Wanderer, Crawler, Worms, Spider, ... bezeichnet. Da es derzeit noch keine festgeschriebene Nomenklatur gibt, werden diese Bezeichnungen im folgenden nicht mehr unterschieden.

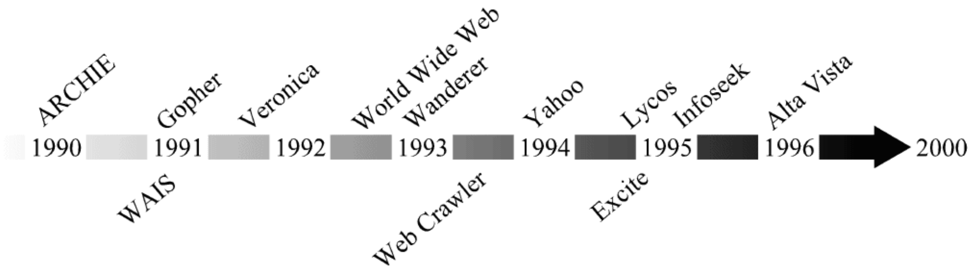

Abbildung 1.1: Die Entwicklung von Informationsbeschaffungswerkzeugen in Datennetzen [Kyas 97]

Mit dem rasanten Wachstum der Datennetze an Umfang und an Bedeutung nimmt auch die Anzahl der Programme zur Informationssuche zu.

1.2 Rechnernetze

Im Jahre 1969 finanzierte die Defence Advanced Research Projects Agency (DARPA) ein Forschungs- und Entwicklungsprojekt mit dem Ziel ein experimentelles Paketvermittlungsnetz zu entwickeln. Dieses Netz wurde ARPANET genannt, und aus diesem ist das heutige Internet entstanden. [Hunt 95] Die Grundlage für das Internet ist der TCP/IP-Protokollstapel. Das Internet ist ein Verbund vieler lokaler Netzwerke, die das IP-Protokoll nutzen und so ein weltweites Netzwerk bilden. Das IP-Protokoll bietet eine Reihen von Diensten an: z.B. Email, FTP, Telnet, Usenet. Ferner wurde 1989-1991 eine graphische Hypertext-Anwendung (World Wide Web) am Europäischen Institut für Teilchenphysik (CERN) entwickelt. „In Verbindung mit dem vom NCSA entwickelten Mosaic-Viewer wurde es im WWW möglich, eine aus mehreren Seiten bestehende Web-Site mit Informationen aus Text, Bildern, Klang und sogar Video sowie Links zu anderen Web-Pages einzurichten. Im ersten Jahr nach Freigabe von Mosaic stieg die Zahl der WWW-Server von 100 auf 7000.“ [Tanenbaum 97] Derzeit gibt es über eine Million WWW-Server weltweit. Zur Gestaltung der

WWW-Seiten steht die Sprache HTML zur Verfügung. Mit dieser kann jeder Internet-Benutzer seine eigenen WWW-Seiten gestalten und weltweit veröffentlichen. Andere Anwender können diese Seiten mittels eines WWW-Browsers abrufen. Das WWW erhöhte die Popularität des Internets durch seine graphische Benutzeroberfläche. Der im WWW produzierte Datenverkehr übertrifft längst den aller anderen Internet-Dienste, und das WWW wächst weiterhin schneller als das Internet selber. [Maurer 96] Immer mehr Firmen richten WWW-Server ein und stellen so Informationen weltweit zur Verfügung.

„Informationen, die neu ins Web kommen, werden bei keiner zentralen Stelle angemeldet. Entsprechend gibt es auch niemanden, der Überblick über die wuchernden Datenmengen hat.“ [PC-Welt 4/97] Dies erschwert die gezielte Suche nach Informationen. Daher gibt es inzwischen eine fast unüberschaubare Anzahl von Suchmaschinen im Internet, die alle versprechen, die Suche zu erleichtern.

Da viele Unternehmen für ihre Unternehmens-Netze die gleichen Transportprotokolle und die gleiche Technologie wie im Internet nutzen, entstand der Begriff des Intranets. “Intranets sind unternehmensinterne, auf dem Transportprotokoll TCP/IP basierende Netze, die auf der Basis von offenen Internetstandards eine leistungsfähige Infrastruktur für Informationsaustausch, Kommunikation und Applikationen bilden.“ [Kyas 97]

Viele Unternehmen betreiben daher inzwischen interne WWW-Server für ihr Intranet. Viele Anwendungen, die früher textuell ausgeführt wurden, werden in das WWW verlagert. So kann z.B. heute eine Datenbank-Anfrage über das WWW erfolgen und nicht mehr über ein Text-Terminal.

1.3 Umfeld der Diplomarbeit

Die Diplomarbeit betrachtet nur die Anforderungen, die sich aus der speziellen Situation der BMW AG ergeben. Daher wird nur deren Intranet betrachtet.

1.3.1 Intranet der BMW AG

Bei der BMW AG wird derzeit ein Intranet für unternehmensweite, interne Informationsdienste aufgebaut, das u.a. auf WWW-Technologien basiert. In einem heterogenen und geographisch weit verteilten Unternehmensnetz wie dem, das die diversen Standorte der BMW AG weltweit miteinander verbindet, gewinnt das schnelle, gezielte Auffinden von Information innerhalb des im Intranet ständig wachsenden Angebots immer mehr an Bedeutung; in naher Zukunft sind allein auf den diversen UNIX-Servern, die an das Intranet angeschlossen sind, mehrere tausend Web-Seiten zu erwarten.

1.3.2 Aufgabenstellung

In der vorliegenden Diplomarbeit werden die konkreten Anforderungen, die sich aus den unternehmensspezifischen Gegebenheiten bei der BMW AG ergeben, ermittelt und geeignet aufbereitet. Auf diesen Anforderungen basierend, werden Bewertungskriterien aufgestellt, anhand derer derzeit verfügbare Suchmaschinen auf ihre Anwendbarkeit im BMW-Umfeld hin analysiert und einander gegenübergestellt werden. Besonderes Augenmerk wird dabei u.a. auf eine flexible Suchlogik (z.B. kategorienbasiert bzgl. bestimmter Gruppen) und eine benutzerfreundliche Schnittstelle gerichtet. Eine weitere wichtige Anforderung ist die Berücksichtigung von Security-Gesichtspunkten. So sollte eine Konfiguration möglich sein, die sicherstellt, daß geschützte Seiten, die nur für einen bestimmten Mitarbeiterkreis zugänglich sind, auch nicht durchsucht werden. Ferner werden weitergehende Anforderungen, wie z.B. die Möglichkeit des Dokumenten-Managements von WORD-Dokumenten, mit in Betracht gezogen.

Die aufgrund der Anforderungs- und Werkzeuganalyse in Frage kommenden Suchmaschinen wurden vor Ort installiert und mit Hilfe von Pilottestern bei der BMW AG weiter auf ihre Einsetzbarkeit hin untersucht.

1.4 Grundsätzliche Probleme

Die Verwendung von Suchmaschinen löst leider noch nicht alle Probleme zufriedenstellend. So erhält der Anwender meist eine viel zu große Anzahl an gefundenen Textstellen. Ferner sind oft viele Treffer unbrauchbar, da der Benutzer den Begriff in einem anderen Kontext suchte wie in dem angegebenen Dokument.

Ferner lassen sich die Grenzen eines zusammenhängenden Dokuments nicht immer eindeutig bestimmen. [Koch 96] Immer öfter werden Filialdokumente verwendet. Gibt der Anwender einen Suchbegriff ein, so erhält er meist jedes Unterkapitel, in dem der Suchbegriff vorkommt, als eigenen Treffer angezeigt. Besser wäre es, nur das Inhaltsverzeichnis angezeigt zu bekommen.

Ein prinzipielles Problem ist die Aktualität und Korrektheit der Dokumente. Dies ist kein spezielles Problem der Suchmaschinen, sondern des WWW, besonders im Internet, da dort jeder die Dokumente veröffentlichen kann, die er möchte.

1.5 Aufbau und Ergebnisse der Arbeit

Zuerst werden die Grundlagen der Funktionsweisen der verschiedenen Typen von Suchmaschinen erklärt. Im folgenden werden dann die Bewertungskriterien dargelegt und erläutert. Im nächsten Kapitel findet eine Bewertung und Gegenüberstellung der in Frage kommenden Suchmaschinen statt. Anschließend folgt ein kurzer Test der installierten Suchmaschinen. Zum Schluß wird ein

Ausblick auf die zukünftige Entwicklung bei Suchmaschinen gegeben.

Bei dem derzeitigen Stand der Entwicklung wird im Rahmen dieser Arbeit das Produkt Alta Vista von digital für den Einsatz im BMW-Intranet vorgeschlagen.

Kapitel 2

Funktionsweise von Suchmaschinen

Prinzipiell können alle Programme, die Informationen in einem Rechnernetz suchen, als Suchmaschinen bezeichnet werden. Diese unterscheiden sich hauptsächlich durch ihre unterschiedliche Funktionsweise.

Während bei roboterbasierten Suchmaschinen die Daten nach und nach von einem Roboter gesammelt werden, müssen bei einer katalogbasierten die Informationen für den Katalog manuell zusammengetragen werden. Die agentenbasierten Suchmaschinen ersparen dem Benutzer Routine-Anfragen, indem sie das Netz selbständig nach neuen Informationen durchsuchen. Bedingt durch die immer größer werdenden Datenmengen spezialisieren sich inzwischen etliche Suchmaschinen auf bestimmte Gebiete. Dies kann thematischer oder regionaler Art sein. So durchsucht z.B. die Suchmaschine Flipper nur deutsche WWW-Seiten. Und gewisse Spezialkataloge bieten nur Dokumente zu einem bestimmten Thema an.

Bei Intranet-Lösungen werden oft auch Kombinationen von verschiedenen Suchmaschinen-Typen eingesetzt. Agentenbasierte Suchmaschinen sind in aller Regel Erweiterungen zu roboterbasierten Suchmaschinen.

2.1 Unterschied Internet - Intranet

Obwohl im Intranet und im Internet die gleichen Transportprotokolle verwendet werden, ergeben sich große Unterschiede bezüglich der Anforderungen an die Suchmaschinen.

Im Internet ist die Benutzung der Suchmaschinen in der Regel kostenlos, ihre Betreiber finanzieren sich aus der Werbung auf den WWW-Seiten. Eine Suchmaschine für ein Intranet muß von den betreffenden Unternehmen bezahlt werden, da in einem firmeninternen Intranet keine kommerzielle Werbung stattfindet.

Wegen der Größe des Internets ist es fast nicht möglich, wirklich alle Informationen aktuell zu erfassen. Der Hauptunterschied der verschiedenen Suchdienste im Internet liegt daher in der Anzahl der indizierten Seiten und in der Zeit, die benötigt wird, das Internet einmal abzusuchen.

In einem Intranet hingegen erwartet man, daß alle Informationen für den Benutzer verfügbar sind. Obwohl auch in einem Intranet immer neue Daten hinzukommen und es einige Zeit dauert, bis diese erfaßt sind, kann man dennoch davon ausgehen, daß bei einer Suche über das gesamte Netz letztlich alle relevanten Daten gefunden werden.

Im Internet entstanden vor einiger Zeit sogenannte Metasuchmaschinen. Diese senden eine Anfrage gleichzeitig an mehrere Suchmaschinen. So wird versucht, den Nachteil, daß keine Suchmaschine alles indiziert hat, zu beseitigen. Für ein Intranet ist dies nicht praktikabel, da dann mehrere Suchmaschinen im Betrieb sein müßten. Dies würde nur die Netzlast erhöhen und keine Verbesserung bringen, da ja jede Suchmaschine für sich schon das Netz umfassend durchsuchen kann. Ferner steigen die Kosten für die benötigten Ressourcen und der Administrationsaufwand, wenn mehrere Indizes verwendet werden.

Durch die Dominanz des WWW im Internet werden dort nur HTML-Dokumente und teilweise Newsgroups durchsucht. In einem Intranet hingegen sollen auch andere Dokumente aus Standard-Softwareprodukten (wie z.B. WORD, EXCEL) und weitere Informationsquellen wie z.B. Filesysteme oder CD-ROMs durchsucht werden.

2.2 Roboterbasierte Suchmaschinen

Es ist unpraktikabel, bei jeder Such-Anfrage das gesamte Intranet nach dem Suchbegriff abzusuchen. Deshalb wird von der Suchmaschine ein Index aufgebaut, der alle wichtigen Informationen, die im gesamten Intranet verfügbar sind, enthält. Auf entsprechende Anfrage der Benutzer hin wird dann nur noch der Index und nicht das Netz durchsucht.

Die Funktionsweise dieses Typs von Suchmaschine gliedert sich in drei Hauptphasen.

- Suchen der Informationsseiten (Roboter)
- Ablegen der Information in einer Datenbank (Indexer)
- Abfrageschnittstelle für den Benutzer (User-Interface)

Zuerst versucht ein Roboter von einer Startseite aus, die der Administrator festlegt, mit geeigneten Algorithmen, alle WWW-Seiten im Intranet zu finden. (vergl. [Koster 97]) Dann werden die gefundenen Seiten durchsucht, und die relevanten Daten werden in einer Datenbank abgelegt. Dem Benutzer wird eine Abfrageschnittstelle an die Datenbank geboten. So kann der Benutzer im Intra-

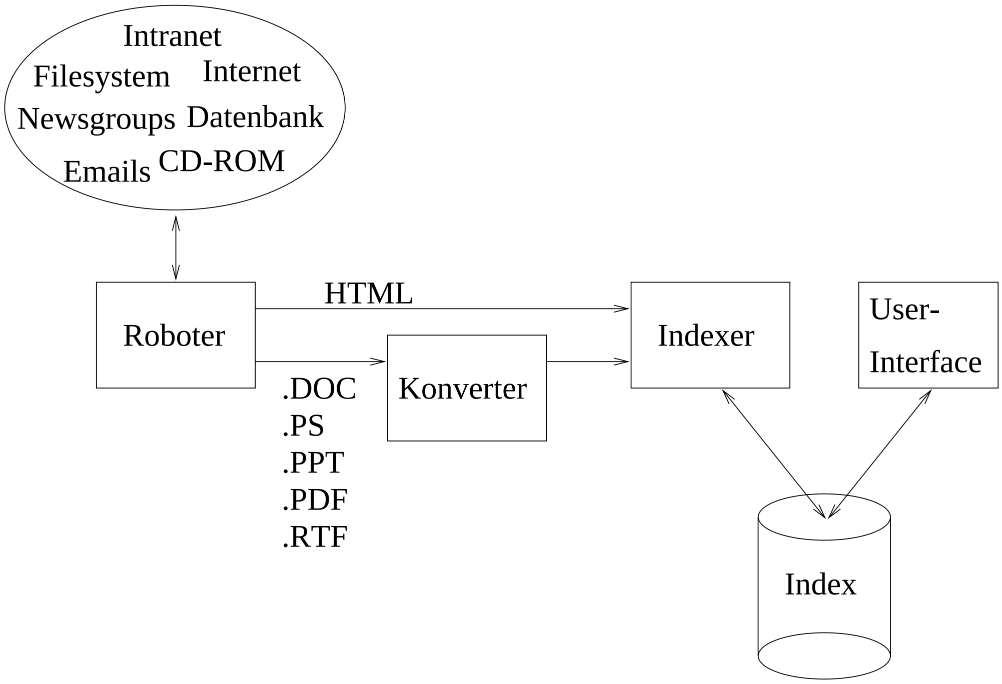

Abbildung 2.1: Aufbau einer roboterbasierten Suchmaschine nach [digital 97]

net suchen lassen, ohne daß zu diesem Zeitpunkt das gesamte Intranet durchsucht wird. Der Benutzer sieht nur das User-Interface, die Indizierung läuft automatisch und ist für ihn verborgen. Ein typischer Vertreter dieser Klasse ist das Produkt Alta Vista der Firma digital.

2.2.1 Datensuche

Der Roboter durchsucht in regelmäßigen Abständen die angegebenen WWW-Server, deren Dokumente indiziert werden sollen. Normalerweise gibt der Administrator eine Start-Seite an, bei der der Roboter mit seiner Suche beginnen soll. Von hier aus verfolgt er alle Links zu anderen Seiten und von dort rekursiv weiter, bis er das komplette Intranet durchsucht hat. Hierbei werden auch die Links verfolgt, die nicht auf denselben WWW-Server zeigen. Um die Suche auf bestimmte WWW-Server im Intranet einzuschränken, kann der Administrator festlegen, welche WWW-Server durchsucht werden sollen. Dies geschieht entweder, indem explizit eine Liste der zu durchsuchenden Server angegeben wird oder der Administrator gewisse Regeln festlegt, die die Server bestimmen.

Oft werden alle WWW-Server im Intranet und zusätzlich ein paar ausgewählte Internet-Server durchsucht. Es können aber auch andere Rechner innerhalb des Intranets durchsucht werden, z.B. News-Server oder File-Server, wenn diese allgemein verfügbar sind. Dieser Vorgang kann dann ausgeführt werden, wenn die Netzlast am geringsten ist. Oder es wird explizit eine Zeitspanne angeben, zu der der Roboter nicht aktiv sein soll. Damit kann zu Spitzenzeiten die Netzlast vermindert werden. Es muß nur sichergestellt sein, daß alle Server, die durchsucht werden sollen, immer aktiv sind. Da ein WWW-Server aber immer erreichbar sein sollte, ist dieser Nachteil nur bei Workstations gewichtig, die sonst abgeschaltet werden.

Das WWW kann als gerichteter Graph abstrahiert werden. Die Dokumente sind die Knoten und die Links von einer auf eine andere Seite die Kanten. Eine Ausnahme bildet nur der Hyper Wave Server, der bidirektionale Links verwendet. Wegen dieser Abstraktion kann man die Suchstrategien aus der Graphentheorie anwenden. Bei folgenden zwei Verfahren ist sichergestellt, daß alle referenzierten Seiten gefunden und Zyklen erkannt werden. Oft haben die Roboter auch noch eine „Notbremse“ eingebaut, die dafür sorgt, daß die Suche nach einer gewissen Anzahl von durchsuchten Seiten abbricht.

- Tiefensuche: „Das heißt, daß man von einem Knoten, der gerade besucht wird, erst zu einem noch nicht besuchten Nachbarknoten geht und dort den Algorithmus rekursiv fortsetzt.“ [Duden 93] Zuerst entfernt man sich bei der Suche immer weiter vom Ausgangspunkt. Erst wenn man in einer Sackgasse gelandet ist, macht man in der Nähe des Ausgangspunktes weiter. (Abb. 2.2)

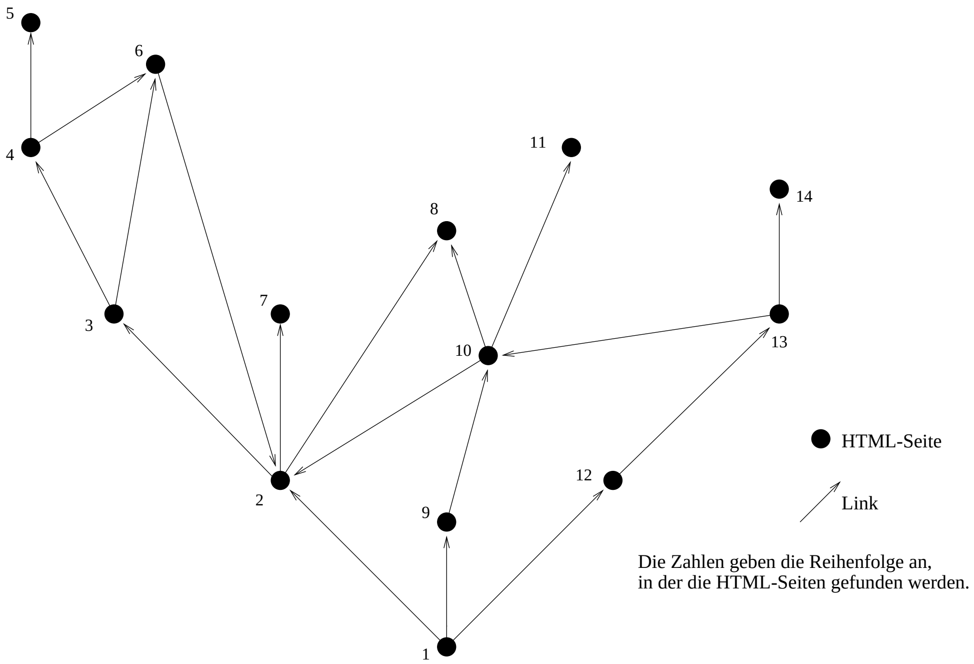

Abbildung 2.2: Tiefensuche

- Breitensuche: „Man geht von einem Knoten, der gerade besucht wird, zuerst zu allen Söhnen, bevor deren Söhne besucht werde.“ [Duden 93] Hier werden zuerst alle Seiten in der Nähe des Ausgangspunkt gesucht und dann erst die entfernteren. (Abb. 2.3)

Die Suchstrategie entscheidet, welche Seiten zuerst gefunden werden. In einem

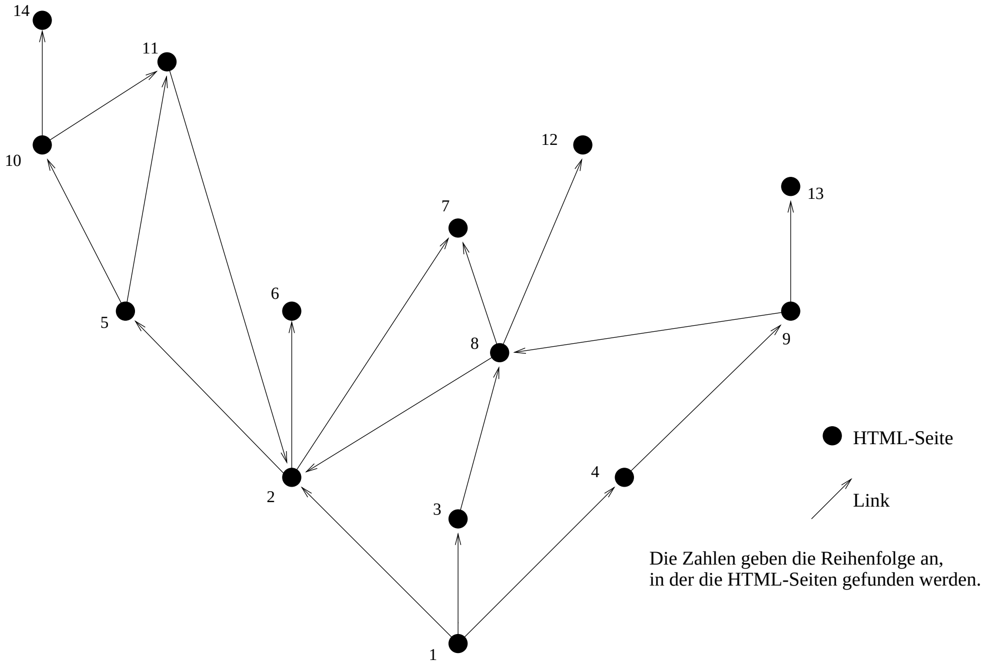

Abbildung 2.3: Breitensuche

Intranet wird aber erwartet, daß die ausgewählten Rechner vollständig durchsucht und alle Informationen gefunden werden. Daher ist die Suchstrategie nur für die Reihenfolge verantwortlich, in der die Informationen gefunden werden.

Leider kann man in der Regel nicht in Erfahrung bringen, wie lange es dauert, bis eine neu erstellte Seite von dem Roboter gefunden wird. Da dies aber nie sofort sein kann, ist der Datenbestand immer etwas veraltet.

Abhängig von der benötigten Aktualität muß festgelegt werden, wie häufig das Intranet durchsucht wird.

2.2.2 Indizierung

Wurde eine Seite gefunden, so muß sie nach den wichtigen Begriffen durchsucht werden, und diese werden dann indiziert in einer Datenbank abgelegt. Sollen auch Dokumente durchsucht werden, die keine HTML-Dokumente sind, müssen auch diese indiziert werden. Dies ist nicht so einfach, da vor der Indizierung die Dateien erst konvertiert werden müssen. So muß z.B. ein PostScript-Dokument oder eine WORD-Datei in eine Text-Datei umgewandelt werden. Es ist daher wichtig, daß der Hersteller gute Konverter mit seinem Produkt mitliefert.
Es ist nicht einfach, alle wichtigen Wörter zu erfassen und gleichzeitig den Datenbestand möglichst gering zu halten. Am wichtigsten sind die Daten aus Titel und den Überschriften. Um aber eine gute Volltextrecherche machen zu können, müssen sämtliche Informationen verarbeitet werden. Daher ist es sehr wichtig, daß die Suchmaschine nicht nach ein paar Zeilen Text abbricht und den Rest des Textes für unwichtig erachtet.

Teilweise führen Suchmaschinen eine lexikalische Analyse durch. Dies hat den Nachteil, daß Wörter, die nicht in diesem Lexikon enthalten sind, nicht indiziert werden. Manche Suchmaschinen entfernen aus Speicherplatzgründen zuerst alle Füllwörter wie „der“, „die“, „das“, „es“, ... Hier tritt aber die Schwierigkeit auf, daß verschiedene Sprachen auch verschiedene Füllwörter enthalten. So wird eine englische Suchmaschine Wörter wie „the“ und „a“ entfernen, eine deutsche Suchmaschine aber „das“ und „ein“.

Manche Suchmaschinen umgehen das Problem, indem sie keine Wörter entfernen und alle in den Index aufnehmen. Bei dem Weglassen von Wörtern ergibt sich als weitere Problematik, daß dann eine Suche nach diesen Wörtern, aber auch nach Phrasen, in denen diese Wörter vorkommen, nicht mehr möglich ist. Textstücke wie „to be or not to be“ würden nicht gefunden werden. Am besten wäre daher eine semantische Analyse des Textes. Damit könnten Begriffe aus dem Index entfernt werden, die zwar im Text vorhanden sind, inhaltlich aber nichts mit dem Artikel zu tun haben. Dies ist aber bei den heutigen Suchmaschinen noch nicht möglich. Kommt ein Wort auf sehr vielen Seiten vor, hat es fast keinen Informationsgehalt mehr.

Natürlich können auch mehrere Indizes aufgebaut werden. Hierbei muß der Administrator eine geeignete Zuordnung zu den einzelnen Datenbanken treffen. So können z.B. alle Dokumente eines WWW-Servers in einem eigenen Index abgelegt werden. Die Aufteilung kann auch thematisch stattfinden. Selbstverständlich muß sich die Suche dann über alle Indizes erstrecken können.

2.2.3 User-Interface

Der Benutzer setzt mit Hilfe seines WWW-Browsers eine Suchanfrage an die Suchmaschine ab. Die gesuchten Begriffe werden mit dem Inhalt der Datenbank verglichen und das Ergebnis im WWW-Browser angezeigt.

In der Regel kann nicht nur ein Suchbegriff eingegeben werden, sondern es können komplexe Suchanfragen mit Hilfe von Booleschen Operatoren formuliert werden. So hat der Benutzer die Möglichkeit, eine ziemlich präzise Anfrage zu stellen, bei der er auch die Information erhält, die er erwartet.

Der Roboter und der Indexer sind für den Anwender verborgen. Er sieht nur das User-Interface. Deshalb muß ein besonderes Augenmerk auf seine Gestaltung gelegt werden.

Manche Suchmaschinen haben eine Kontextsuche integriert. Hierbei wird nicht nur nach dem Suchbegriff selber, sondern auch nach Synonymen gesucht. Teilweise werden Synonym-Wörterbücher verwendet, oder es wird bei der Indizierung ein Kontext erstellt. Wörter, die immer in einem gemeinsamen Umfeld auftreten, gehören zu einem Kontext. So wird z.B. bei einer Kontextsuche über „BMW“ nicht nur nach dem Wort „BMW“, sondern auch nach „Automobil“ und ähnlichen Begriffen gesucht. Da ein Kontext aber nicht statisch sein muß - eine Firma kann mit einer anderen fusionieren, eine Firma kann ihr Geschäftsfeld verlagern - muß es möglich sein, den Kontext jederzeit den neuen Gegebenheiten anzupassen.

2.3 Katalogbasierte Suchmaschinen

Die Kataloge, oft auch Verzeichnisse genannt, charakterisieren sich vor allem durch ihren hierarchischen Aufbau. Der Anwender wählt aus einer Menge von Themen das aus, das er für seine Suche am geeignetsten hält. Er erhält darauf eine neue Liste von Unterthemen, aus der er dann wieder eine Auswahl treffen kann. Dies setzt er solange fort, bis er auf die gewünschte Information getroffen ist.

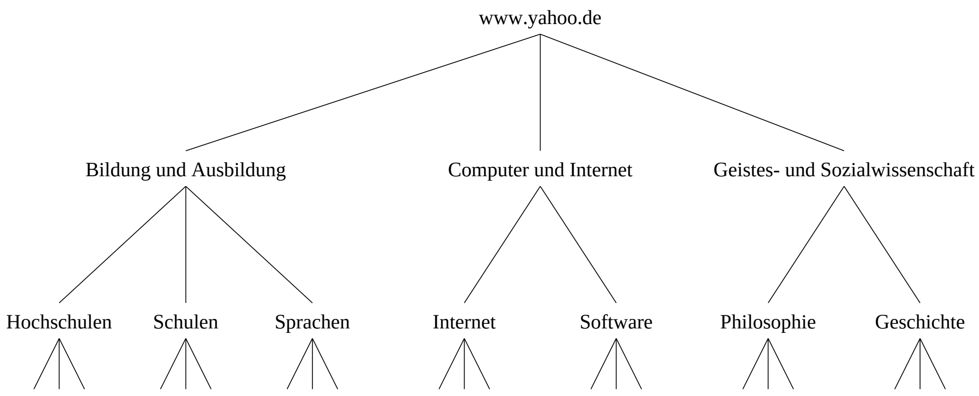

Abbildung 2.4: Prinzipieller Aufbau eines Katalogs

Die Informationen müssen in aller Regel von Hand klassifiziert werden. Um so einen thematisch sortierten Katalog erstellen zu können, ist es nötig, das entsprechende Fachwissen zu haben, um die Dokumente richtig einzuordnen. „Die Verzeichnisse kommen nicht durch die Tätigkeit von Maschinen, sondern durch die Arbeitsleistung von Menschen zustande. Deshalb sind sie natürlich nie so umfangreich wie Indizes, führen aber aufgrund ihres Ordnungsschemas oft schneller ans Ziel.“ [PC-Welt 4/97]

Oft werden Kataloge mit einer roboterbasierten Suchmaschine kombiniert. So kann der Benutzer entweder eine Anfrage über den gesamten Katalog starten, ohne die Oberbegriffe zu kennen, oder eine Such-Anfrage auf ein bestimmtes Thema eingrenzen. Ein typisches Beispiel für einen Internet-Katalog ist das Produkt Yahoo. (siehe Abb. 2.4)

Prinzipiell kann auch für ein Intranet ein Katalog eingerichtet werden. Dazu ist es aber notwendig, daß alle Dokumente zentral registriert und klassifiziert werden. Jeder, der ein neues Dokument veröffentlicht, muß dieses dann durch eine geeignete Mitteilung dem Katalog bekannt geben. Der Administrator muß dafür die Themen festlegen, nach denen die Dokumente klassifiziert werden. Will ein Benutzer sein Dokument in den Katalog eintragen lassen, muß er ein Dokument einem bestimmten Thema zuordnen. Dies geschieht meist mittels eines HTML-Formulars.

Bei einem schon bestehendem Intranet mit vielen Informationen ist der Aufwand sehr hoch, alle schon bestehenden Dokumente und die täglich hinzukommenden thematisch zu sortieren.

2.4 Agentenbasierte Suchmaschinen

Agentenbasierte Suchmaschinen sind dann von großer Bedeutung, wenn jemand zu einem Thema über längere Zeit hinweg auf dem Laufenden gehalten werden will. Dabei nimmt der Benutzer nicht interaktiv wie bei den roboter- und katalogbasierten Suchmaschinen die Suche vor, sondern er delegiert seine Anfrage an einen Agenten. Regelmäßig notwendige Suchanfragen entfallen. Der Benutzer definiert sich einmal einen Agenten, der ihn über neue Entwicklungen zu einem bestimmten Thema informiert. Um dem Agenten die Themengebiete mitzuteilen, kann ein bestimmtes Suchprofil erstellt werden, in das Suchbegriffe eingetragen werden können. Ein Agent ist ein Programm, das autonom im Auftrag eines Nutzers handelt. In der Regel erstellt der Benutzer mit einem WWW-Browser sein persönliches Benutzerprofil, das alle für ihn wichtigen Suchbegriffe enthält.

Wird dann im überwachten Netz ein Dokument mit einem dieser Suchbegriffe veröffentlicht oder ein bestehendes Dokument verändert, wird der Benutzer benachrichtigt. Der Benutzer kann festlegen, wie er über interessante Entwicklungen informiert wird. Der Agent liefert den Benutzern z.B. über Fax, Email oder eine dynamisch generierte WWW-Page die relevanten Informationen. Jeder Benutzer kann sich seine privaten Suchprofile erstellen, die von anderen weder gelesen noch verändert werden können.

Erstellt der Benutzer z.B. einen Agenten, der alle Neuerungen auf dem Gebiet des Motorenbaus melden soll, so erhält der Benutzer eine Nachricht, falls zu diesem Thema neue Informationen im Netz verfügbar sind. Die agentenbasierten Suchmaschinen sind meist eine Ergänzung zu roboterbasierten Suchmaschinen. Ein bekanntes Beispiel ist Search '97 von Verity.

Sowohl bei den roboterbasierten wie auch bei den agentenbasierten Suchmaschinen sind Roboter im Netz unterwegs. Der Hauptunterschied ist nur, daß der Roboter einer agentenbasierten Suchmaschine nur bestimmte Informationen für eine bestimmte Person sucht, während der andere Roboter alle Informationen sucht, um daraus den Index zu generieren.

Kapitel 3

Bewertungskriterien für Suchmaschinen

Vor dem Einsatz eines Datenverarbeitungssystems müssen dessen Vor- und Nachteile abgewogen werden.

Um überhaupt eine Evaluierung bestehender Produkte durchführen zu können, ist es nötig, geeignete Bewertungskriterien zur Verfügung zu haben. Da es derzeit noch keine allgemeingültigen Kriterien für die Bewertung von Suchmaschinen gibt, mußte zuerst ein Kriterienkatalog mit Gewichtung erstellt werden. Eine einfache Kosten-Nutzen-Rechnung kann hier nicht aufgestellt werden.

Die aufgeführten Bewertungskriterien sind als Hilfe bei der Auswahl einer Suchmaschine gedacht. In erster Linie sind diese Kriterien für das Intranet der BMW AG aufgestellt worden, sie sind aber auch auf andere Unternehmen übertragbar. Die meisten Kriterien sind global zu verwenden, andere, wie z.B. das der Umlaute, nur lokal für deutsche Anforderungen.

Entscheidend für eine Gewichtung ist natürlich, inwiefern das Unternehmen von der Aktualität und Vollständigkeit der durch die Suchmaschine verfügbaren Informationen abhängig ist. Eine Zeitungsredaktion hat z.B. ganz andere Maßstäbe wie ein Unternehmen, das Autos produziert.

Dabei richten sich die Kriterien nicht nur nach den Funktionen der derzeit existierenden Suchmaschinen, sondern es werden allgemeine Forderungen aufgestellt, die nicht immer alle von existierenden Produkten erfüllt werden. Der Bewertungskatalog muß noch aktuell sein, auch wenn schon wieder neue Produkte oder Weiterentwicklungen auf dem Markt sind. Daher sind auch Forderungen zulässig, deren Lösung heute noch gar nicht in Sicht sind.

Ausgangspunkt für das Finden der Kriterien war die Analyse existierender Produkte. Weitere Ideen konnten aus dem Studium entsprechender Fachliteratur gewonnen werden. Zusätzlich wurde eine Bedarfsanalyse durchgeführt. In Zusammenarbeit mit BMW-Angestellten wurden die Haupkriterien aufgestellt.

Allerdings beschreibt die Fachliteratur fast ausschließlich Suchmaschinen für das Internet. Die spezielle Problematik für Intranets (vgl. 2.1) wird kaum erfaßt. Ferner befassen sich die meisten Artikel fast ausschließlich mit Statistiken, welche Suchmaschine bei einem bestimmten Suchbegriff die meisten Treffer vorweisen kann oder sie vergleichen die Funktionalität der einzelnen Internet-Suchmaschinen. Die Kriterien für Suchmaschinen im Internet können dabei nicht einfach übernommen werden. Teilweise ergeben sich Schwerpunkt-Verschiebungen bei der Übertragung der Kriterien auf Suchmaschinen im Intranet. Andere Internet-Kriterien sind hingegen für ein Intranet völlig irrelevant (z.B. Werbung).

Es wurden die gängigsten kommerziellen und nichtkommerziellen Suchmaschinen auf ihre Funktionsweise und ihre Funktionalität untersucht. Gerade auch fehlende und fehlerhafte Funktionen bei bestehenden Suchmaschinen trugen zur Aufstellung mancher Forderung bei.

Die Erwartungen der Anwender an eine Suchmaschine sind sehr unterschiedlich, so daß diese erst sortiert und bewertet werden müssen. Forderungen wie z.B. Schnelligkeit und gute Handhabung sind nicht so leicht in meßbare Kriterien umzusetzen. Die verschiedenen Anwender-Gruppen haben auch unterschiedliche Anforderungen an eine Suchmaschine. So hat z.B. eine Sekretärin andere Wünsche als ein Entwickler oder ein Manager.

Dabei kristallisierte sich heraus, welche Funktionen unbedingt benötigt werden und welche nur eine Ergänzung darstellen.

Die Kriterien sind nicht isoliert zu betrachten, da sie sich oft gegenseitig beeinflussen oder gar widersprechen. So erfordert ein größerer Funktionsumfang oft auch einen größeren Aufwand für die Administration und die Schulung der Anwender.

Um einen besseren Überblick zu vermitteln, wurden die einzelnen Kriterien thematisch klassifiziert.

3.1 Tabellarische Übersicht

Die folgenden Tabellen geben einen ersten Überblick über die Bewertungskriterien. Anschließend werden die aufgeführten Kriterien näher erläutert. Die einzelnen Kriterien wurden den Anforderungen für das Intranet der BMW AG entsprechend gewichtet. Für andere Unternehmen kann sich durchaus auch eine andere Gewichtung ergeben, die dann evtl. auch zu einer anderen Entscheidung führen kann.

Die Gewichtung wurde nach folgendem Schema durchgeführt:

sehr wichtig	$\oplus\oplus$
wichtig	$\oplus$
irrelevant	-

Unterstützte Informationsquellen	**Gewicht**
Intranet	$\oplus\oplus$
Internet	$\oplus$
Filesysteme	$\oplus\oplus$
Newsgroups	$\oplus$
Emails	-
Datenbanken	$\oplus$
CD-ROMs	$\oplus$

Indizierung	**Gewicht**
Vollständigkeit	$\oplus\oplus$
Behandlung von HTML-Seiten	$\oplus\oplus$
Verhinderung der Indizierung bestimmter Seiten	$\oplus$
Volltextrecherche	$\oplus\oplus$
Zeitpunkt der Indizierung	$\oplus\oplus$
Reindizierung	$\oplus\oplus$
Behandlung von nicht statischen Seiten	$\oplus$
Netzlast während der Indizierung	$\oplus$
Verteilte Datenbanken	$\oplus$

Sicherheit	**Gewicht**
Geschützte Dokumente	$\oplus\oplus$
Sicherheitsmechanismen	$\oplus\oplus$
Einschränkung der Suche auf bestimmte Server	$\oplus\oplus$

User-Interface	**Gewicht**
Benutzerfreundliche Eingabe	$\oplus\oplus$
Einschränkung der Suche auf bestimmte Bereiche	$\oplus\oplus$
Hilfesystem	$\oplus\oplus$
Verknüpfung von Suchbegriffen	$\oplus\oplus$
Suchmaske	$\oplus\oplus$

Suche	**Gewicht**
Unscharfe Suche	⊕⊕
Groß- und Kleinschreibung	⊕⊕
Umlaute	⊕⊕
Sonderzeichen	⊕
Unterstützte Sprachen	⊕⊕
Gewichtete Suchbegriffe	⊕
Wildcards	⊕⊕
Phrasensuche	⊕⊕
Verschiedene Suchmodi	⊕⊕
Zeitspanne	⊕
Kennwörter	⊕⊕
Boolesche Operationen	⊕⊕
Suche nach Schlüsselwörtern	⊕⊕
Kommentare	⊕
Kontextsuche	⊕

Suchergebnis	**Gewicht**
Geschwindigkeit	⊕⊕
Anzahl der Treffer	⊕⊕
Art der angezeigten Informationen	⊕⊕
Ranking	⊕
Markierung des Suchbegriffes	⊕

Dokumentenmanagement	**Gewicht**
Unterstützte Dokumentarten	⊕⊕
Anzeige der Dokumente	⊕

Agent	**Gewicht**
Agent als Erweiterung	⊕
Netzlast	⊕⊕
Benutzereigene Agenten	⊕
Übermittlung des Ergebnisses	⊕

Schnittstellen	**Gewicht**
Datenbanken	⊕
Programme	⊕

Administration	**Gewicht**
Installation	⊕⊕
Installationsmedium	-
Konfiguration	⊕⊕
Log-Files	⊕
Integriertes Management	⊕
Überwachung der Indizierung	⊕⊕

Verhalten in kritischen Situationen	**Gewicht**
Abfrage des Zustandes	$\oplus\oplus$
Benutzermeldungen in kritischen Situationen	$\oplus\oplus$
Verhalten bei Systemausfall	$\oplus\oplus$
Fehlermeldungen	$\oplus\oplus$

Systemanforderungen	**Gewicht**
Plattform	$\oplus$
Benötigte Rechenleistung	$\oplus$
Benötigter Speicherbedarf	$\oplus$
Unterstützte Browser	-
Unterstützte WWW-Server	-

Dokumentation	**Gewicht**
Handbuch	$\oplus\oplus$
Online-Handbuch	$\oplus$
Sprachen	$\oplus$
Fortbildungskurse	-

Support	**Gewicht**
Hotline	$\oplus\oplus$
Update	$\oplus\oplus$
Patches	$\oplus$

Preis	**Gewicht**
Preis	-

3.2 Beschreibung der Kriterien

3.2.1 Unterstützte Informationsquellen

In einem unternehmensweiten Intranet ist es notwendig, daß nicht nur die Informationen indiziert werden, die in WWW-Seiten abgelegt sind, sondern daß auch andere Informationsquellen durchsucht werden können. Nicht alle Suchmaschinen unterstützen sämtliche Informationsquellen, die für die BMW AG nötig sind.

- Intranet
  Alle internen WWW-Server des Unternehmens sollen in die Suche integriert sein. Dadurch ist sichergestellt, daß die Benutzer dieselben Informationen suchen und finden können, die sie auch sonst in ihrem Intranet vorfinden. Der Administrator kann so eine Liste der zu durchsuchenden WWW-Server erstellen. Um nicht lange Listen mit WWW-Servern erstellen zu müssen, sollte die Möglichkeit vorhanden sein, über bestimmte

Auswahlregeln die Server festzulegen. So werden z.B. mit *.muc alle Server indiziert, deren Domain-Name „muc“ lautet.

- Internet
Es sollte möglich sein, bestimmte, ausgewählte Internet-Server in die Suche mit aufnehmen zu können. Da es nicht sinnvoll ist, das gesamte Internet zu indizieren, ist es nötig, daß die Suche auf wenige ausgewählte Server beschränkt bleibt. Die Indizierung des kompletten Internets ist auch gar nicht nötig, da es dafür schon eine Vielzahl von Suchmaschinen gibt.

- Filesysteme
Alle Daten, die allgemein in Netz verfügbar sind, sollen auch in den Index mit aufgenommen werden. Hierbei könnte man auch an die Einrichtung eines oder mehrerer File-Server denken, auf dem die Dateien gespeichert sind, die für das gesamte Unternehmen zugänglich sein sollen. Hierfür bietet sich eine themenbezogene bzw. abteilungsbezogene Aufteilung an.

- Newsgroups
Newsgroups sind Diskussionsforen über ein bestimmtes Thema. Interessierte Leute können in einer Newsgroup Mitglied werden und so die veröffentlichten Artikel lesen und selbst welche schreiben. Neben den schätzungsweise 10.000 öffentlichen Newsgroups, können auch firmeninterne in einem Intranet angeboten werden. [Tanenbaum 97] Gerade das Durchsuchen dieser Newsgroups ist besonders wichtig.

- Emails
Werden Emails in die Suche integriert, so kann der Benutzer auch in diesen nach Informationen suchen. Diese an sich nicht notwendige Eigenschaft ist aber gerade bei seiner steigenden Anzahl von Emails oft nützlich. Bei der Indizierung von Emails muß sichergestellt sein, daß jeder Benutzer nur seine eigenen Emails durchsuchen kann und nicht die von anderen Benutzern.

- Datenbanken
Für eine umfassende Informationssuche ist es nötig, auch die Daten, die in Datenbanken abgelegt werden, zu durchsuchen. Etliche Suchmaschinen bieten daher ein Datenbank-Gateway an, mit dem auch Informationen aus den gängigsten Datenbanken durchsucht werden können. (vgl. Schnittstellen / Datenbanken)

- CD-ROMs
Die CD-ROM gewinnt gerade in letzter Zeit immer mehr an Bedeutung. Daher werden Informationen, die sich nicht oft oder nie ändern, auf CD-ROM geschrieben. Oft sind auch Bibliotheken auf CD-ROM verfügbar. Damit die Anwender auf die CD-ROMs zugreifen können, müssen diese ständig verfügbar sein, d.h. jede CD-ROM muß sich in einem eigenen Laufwerk befinden und darf nicht entfernt werden. Die Informationen können nur dann mit einer Suchmaschine gefunden werden, wenn die CD-ROMs auch indiziert werden.

3.2.2 Indizierung

- Vollständigkeit
  Die Vollständigkeit ist eine der zentralen Anforderungen an eine Intranet-Suchmaschine. Alle WWW-Server im Intranet müssen komplett durchsucht werden. Es muß sichergestellt sein, daß auch wirklich alle Seiten gefunden und indiziert werden. Im Gegensatz zu einer Suchmaschine im Internet kann man im Intranet wegen der im Vergleich zum Internet relativ geringen Menge an Daten erwarten, daß alle Informationsquellen vollständig indiziert werden. Eine Überprüfung auf Vollständigkeit ist kaum machbar, da immer nur das Fehlen von Informationen, nie aber das Vorhandensein auffällt.
  Darüber hinaus ist es gut, wenn die Benutzer neu erstellte Seiten der Suchmaschine mitteilen können. Ein Benutzer kann damit wichtige Seiten gezielt durchsuchen und in den Index aufnehmen lassen. Diese Dokumente sind dann schneller verfügbar, als wenn der Roboter „zufällig“ auf sie stoßen würden. Wählt der Benutzer Seiten aus, die normal nicht indiziert werden, weil sie z.B. außerhalb des Intranets liegen, so darf die Eingabe nicht berücksichtigt werden.

- Behandlung von HTML-Seiten
  Grundvoraussetzung für die Indizierung ist das richtige Interpretieren der HTML-Seiten. Es ist absolut erforderlich, daß der derzeitig gültige HTML-Standard unterstützt wird. Leider halten sich die Hersteller von WWW-Browsern nicht immer an die existierenden Standards, sondern erweitern diese durch eigene Funktionen. Daher ist es wünschenswert, wenn diese Zusatz-Funktionen ebenfalls unterstützt würden.
  Es darf dem Roboter keine Probleme bereiten, Seiten zu finden, die mit Frames gestaltet sind. Ferner müssen auch HTML-Seiten, die von verschiedenen Editoren erstellt wurden, fehlerfrei erkannt werden. Gerade heute werden immer mehr HTML-Dokumente automatisch generiert.

- Verhinderung der Indizierung bestimmter Seiten
  Nicht alle Benutzer wollen, daß ihre HTML-Seiten von Robotern durchsucht werden. Die Personen, die HTML-Seiten erstellen, sollten daher in der Lage sein, zu verhindern, daß bestimmte Seiten in den Index aufgenommen werden. Hierzu gibt es zwei Möglichkeiten:
  „Das Robots Exclusion Protocol ist ein Internet-Standard, in dem festgelegt wird, wie ein Web-Server-Administrator Einfluß darauf nehmen kann, welche Datenbereiche seines Servers von einem Roboter erfaßt werden dürfen und in welchem Umfang.“ [Kirchgesser 97] Dazu wird eine Datei „robots.txt“ im Root-Verzeichnis des entsprechenden Servers angelegt. Da diese eine Datei das gesamte Verhalten des Servers gegenüber dem Roboter festlegt, hat nur der Administrator das Recht, diese Datei zu ändern, und nicht der einzelne Benutzer. Daher gibt es noch eine weitere Möglichkeit, den Roboter von eigenen Seiten fernzuhalten. Mit Hilfe von Robots META tag kann jeder Benutzer beim Erstellen seiner HTML-Seiten festlegen, ob seine Seiten indiziert werden sollen oder nicht.

Die Suchmaschine sollte beide Möglichkeiten unterstützen. Mit der ersten kann z.B. verhindert werden, daß CGI-Skripts durchsucht werden und mit der zweiten Möglichkeit kann jeder Benutzer für sich entscheiden, ob seine HTML-Seite in den Index aufgenommen werden soll oder nicht.

- Volltextrecherche
  Um eine Volltextrecherche durchführen zu können, muß auch die gesamte Information, die in einem Dokument vorhanden ist, verarbeitet werden. Nur so kann der Benutzer davon ausgehen, daß er auch alle Dokumente findet, die seinen Suchbegriff enthalten. Zum Teil wird keine Volltextindizierung durchgeführt, sondern es werden nur der Titel, die Links, die URL und die ersten 20 Zeilen zusammenhängender Text in die Datenbank aufgenommen. (Beispiel Lycos)
  Da nicht alle HTML-Seiten überhaupt einen Titel haben, und manche sehr ungünstig gewählt sind, kann der Informationsgehalt des Titels sehr gering sein. Daher kann in einem Intranet nur eine Lösung akzeptiert werden, bei der keine wesentliche Information unberücksichtigt bleibt.

- Zeitpunkt der Indizierung
  Der Administrator muß die Möglichkeit haben, die Indizierung automatisch in regelmäßigen Abständen laufen zu lassen. Hier ist eine umfassende Konfiguration nötig. Es sollte möglich sein, daß die Indizierung z.B. jede Stunde, jeden Tag oder einmal in der Woche zu einer bestimmten Stunde stattfindet.
  Als Alternative zu einer regelmäßigen Indizierung gibt es noch die kontinuierliche Indizierung. Hier ist der Roboter immer auf der Suche nach neuen Dokumenten. In diesem Fall muß es aber auch die Möglichkeit geben, daß der Administrator den Index löscht und neu aufbaut.
  Es sollte auch möglich sein, daß der Roboter zu bestimmten Hauptlast-Zeiten nicht aktiv ist, um das Netz nicht zu überlasten. Je flexibler die Möglichkeiten der Konfiguration sind, desto besser kann der Administrator die Suchmaschine an die Unternehmens-Bedürfnisse anpassen. Da im allgemeinen der Benutzer keine Mitteilung bekommt, wenn seine Seite in den Index aufgenommen wurde, ist es wünschenswert, daß es systemweit eine maximale Zeitspanne gibt, nach der eine neu erstellte Seite im Index aufgenommen ist.

- Reindizierung
  Bei großen Datenbeständen ist es nicht möglich, den Index jedesmal neu aufbauen zu lassen, da dies viel Zeit benötigt und die Anwender in dieser Zeit keinen Zugriff auf den kompletten Index haben. Daher werden nur die Änderungen berücksichtigt. Wichtig ist, daß die Suchmaschine nicht nur neue und veränderte Seiten berücksichtigt, sondern auch gelöschte Seiten aus dem Index entfernt. Auf Seiten, die nicht mehr existieren, darf auch kein Link mehr zeigen. Seiten, die nur zeitweilig nicht erreichbar sind, dürfen aber nicht aus dem Index entfernt werden.
  Vorteilhaft ist auch eine Möglichkeit für den Administrator, gelöschte Seiten von Hand aus dem Index zu entfernen. Die Suchmaschine sollte er-

kennen, ob sich eine Seite seit der letzten Indizierung verändert hat, oder nicht. Sollte es dennoch nötig sein, einen Index neu aufzubauen, ist es erforderlich, daß für den Zeitraum, bis der neue Index wieder alle Informationen enthält, eine Kopie des alten zur Verfügung steht.

- Behandlung von nicht statischen Seiten
  Die Suchmaschine muß in der Lage sein, nicht statische Seiten sinnvoll zu behandeln. Auch dynamisch erzeugte Seiten soll sie durchsuchen können. Ebenso sollten auch Links verfolgt werden, die der Benutzer mit Hilfe einer sensitiven Grafik auswählen kann. Da manche HTML-Seiten nur so erreichbar sind, würden diese sonst nicht indiziert. Auf der anderen Seite sollen Formulare und Anwendungen, die jedesmal mit anderen Daten aufgerufen werden, nicht durchsucht werden. So liefert z.B. eine Datenbank-Anfrage mit einem entsprechendem CGI-Script verschiedene Ergebnisse abhängig von der Eingabe.
  Enthält eine WWW-Seite ein Java-Applet, über das der Anwender neue Seiten auswählen kann, so ist es notwendig, daß auch diese Seiten indiziert werden.
  Alle Informationen, die der Benutzer von Hand aufsuchen kann, müssen auch mit der Suchmaschine zur Verfügung stehen.

- Netzlast während der Indizierung
  Die Netzlast sollte so gering wie möglich sein. Gerade bei einer kontinuierlichen Indizierung muß die Last so gering sein, daß der übrige Betrieb nicht beeinträchtigt wird. Aber auch beim regelmäßigen Aufbau eines Indexes dürfen andere Anwendungen nicht beeinträchtigt werden. So kann z.B. dann das Netz indiziert werden, wenn die Netzlast nicht so hoch ist wie zu anderen Zeiten.
  Daher ist es gut, wenn der Administrator die Last, die durch die Indizierung erzeugt wird, beeinflussen kann. Eine geringere Last hat meist zur Folge, daß es länger dauert, bis das Netz indiziert ist. Der Administrator muß hier durch eine geeignete Wahl der Parameter einen leistungsfähigen Betrieb sicherstellen.

- Verteilte Datenbanken
  Wenn es möglich ist, verschiedene Teilbereiche des Netzes gesondert zu durchsuchen, ist es erforderlich, die Indizes in verschiedenen Datenbanken zu verwalten. Dazu sollte keine zweite Suchmaschine installiert werden müssen. So können z.B. verschiedene Abteilungen ihren eigenen Index erstellen und pflegen, der dann auch auf einem Abteilungs-Server abgelegt ist. Dennoch können alle Benutzer des Intranets eine Suche über alle Indizes des Unternehmens starten. Bei einem solchen Konzept ist aber ein verteiltes Management erforderlich.
  Denkbar sind auch Konzepte, bei denen bei jedem WWW-Server ein eigener Index erstellt wird und der Anwender dann wählt, über welche Server sich die Suche erstrecken soll. Dazu ist es auch erforderlich, daß mit einer Oberfläche Indizes verschiedener Hersteller abgefragt werden können.

3.2.3 Sicherheit

Die Sicherheit spielt in den Netzwerken eine entscheidende Rolle. Durch die Suchmaschine darf keine Sicherheitslücke entstehen. Die Benutzer sollen genau die Daten durchsuchen können, zu denen sie auch sonst Zugang haben. Daher muß die Suchmaschine gewisse Mindestanforderungen erfüllen.

- Geschützte Dokumente
  Geschützte Seiten sollen nur von autorisierten Personen durchsucht werden können. Seiten, die für den Benutzer verboten sind, dürfen nicht in der Liste der gefundenen Dokumente auftauchen. Bei manch einer Suchmaschine erhält sonst der nicht autorisierte Benutzer einen kleinen Einblick in das Dokument, da die meisten Suchmaschinen ein kleines Textstück um den Suchbegriff herum anzeigen. Selbst wenn der Benutzer nicht den gesamten Text sehen kann, so weiß er aber dennoch, daß es dieses geheime Dokument gibt.
  Wenn die Trefferliste vom Administrator so konfigurierbar ist, daß keine Textstelle angezeigt wird, kann der Benutzer zumindest keine Textfragmente sehen. Oft werden aber auch Dokumente, die nicht für alle lesbar sein dürfen, gar nicht indiziert. So treten keine Sicherheitsprobleme auf.
  Für ein umfassendes Sicherheitskonzept ist es erforderlich, bei der Indizierung die Informationen über die Rechte der Dateien mit in die Datenbank aufzunehmen. Sinnvoll ist die Einrichtung von Benutzergruppen. So könnten auch nicht öffentlich zugängliche Daten indiziert werden, wenn sichergestellt ist, daß diese nur von den autorisierten Personen durchsucht werden können. Aus Sicherheitsgründen ist es gut, diese sicherheitsrelevanten Daten in einem eigenen Index zu speichern. Dann kann garantiert werden, daß nur eine bestimmte Benutzergruppe Zugang zu diesen Daten hat.

- Sicherheitsmechanismen
  Die Suchmaschine sollte sich leicht in das Sicherheitskonzept eines Unternehmens eingliedern lassen. Dazu ist erforderlich, daß die Sicherheitsmechanismen des Betriebssystems verwendet werden. Wenn Dateien, die nicht für alle lesbar sind, nicht indiziert werden, sind auch keine zusätzlichen Sicherheitsmechanismen nötig.

- Einschränkung der Suche auf bestimmte Server
  Es muß möglich sein, die Suche auf bestimmte Server einzuschränken. Es sollte die Möglichkeit geben, Regeln anzugeben, die festlegen, wie die URL der zu durchsuchenden Seite aufgebaut sein muß.
  Es muß verhindert werden, daß ein Link von dem firmeneigenen Intranet in das weltweite Internet verfolgt wird, damit nicht versehentlich das gesamte Internet abgesucht wird. Da der Rechner, auf dem die Suchmaschine installiert ist, nicht für solche Datenmengen geeignet ist, würde der Versuch einer Indizierung des Internet große Probleme verursachen.

3.2.4 User-Interface

Das User-Interface ist besonders für die Akzeptanz der Suchmaschine bei den Anwendern wichtig. Ein übersichtlicher Bildschirmaufbau ist selbstverständlich. Da die Benutzer normalerweise ihre Anfragen mit einem WWW-Browser stellen, ist das User-Interface eine HTML-Seite, bzw. wird über ein CGI-Skript automatisch generiert. Diese kann der Administrator flexibel den Bedürfnissen der Benutzer anpassen. Wenn die mitgelieferten Abfrage-Seiten jedoch schon sorgfältig und gut gestaltet sind, ist der Arbeitsaufwand für den Administrator geringer, als wenn er die Seiten neu gestalten muß.
Bei all den Gestaltungsmöglichkeiten darf die eigentliche Aufgabe des Programms, das Suchen und Finden von Information, nicht vernachlässigt werden.

- Benutzerfreundliche Eingabe
  Das User-Interface muß so gestaltet sein, daß auch Benutzer, die die Suchmaschine nur selten benötigen, sich damit zurecht finden. Um dies zu erreichen, muß die Suchmaschine überwiegend selbsterklärend sein. Das User-Interface sollte so gestaltet sein, daß das Eingabefeld für den Suchbegriff das zentrale Element auf der Seite darstellt. Das Eingabefeld für den Suchbegriff muß so groß sein, daß auch längere Begriffe eingegeben werden können, ohne daß der Wortanfang verschwindet.

- Einschränkung der Suche auf bestimmte Bereiche
  Es sollte möglich sein, nicht nur das gesamte Intranet und alle Informationsquellen durchzusuchen, sondern auch nur Teilbereiche. Dazu ist es nötig, daß die Benutzer die Teilbereiche auswählen können, auf die sie ihre Suche begrenzen möchten. So ist es möglich, daß bestimmte Benutzergruppen nur in ihrem speziellen Themenkreis suchen. Es sollte aber auch möglich sein, nur Seiten, die mit einer bestimmten URL beginnen, zu durchsuchen.

- Hilfesystem
  Damit sich auch ein ungeübter Benutzer auf Anhieb mit der Suchmaschine zurechtfindet, ist ein komfortables und ausführliches Hilfesystem nötig. Besonders wichtig ist dabei, dem Benutzer zu vermitteln, wie er am besten seine Suchanfragen formuliert, um besonders gute Ergebnisse zu erhalten. Daher dürfen aussagekräftige Beispiele auf keinen Fall fehlen.
  Ferner soll der Anwender durch eine problembezogene Hilfe unterstützt werden. Ist z.B. die Anzahl der Treffer sehr hoch, so wird dem Anwender eine konkrete Hifestellung gegeben, wie er seine Anfrage besser formulieren kann.
  Da nicht jeder Benutzer genügend gut Englisch beherrscht, sollte die Hilfe auch in Deutsch vorhanden sein. In einem weltweiten Unternehmen muß sichergestellt sein, daß auch mehrere Sprachen für ein Hilfesystem verfügbar sind. Die Benutzer sollten die Sprache für die Hilfe selber wählen können.

- Verknüpfung von Suchbegriffen
  Um komplexere Suchanfragen starten zu können, müssen auch mehrere Suchbegriffe miteinander verknüpft werden können. Es muß möglich sein, daß nicht nur eine Art von Verknüpfungen verwendet werden kann, sondern alle erlaubten Verknüpfungen beliebig miteinander kombiniert werden können. Es reicht also nicht, nur Radiobuttons für AND und OR auf der Seite anzubringen.

- Suchmaske
  Der Administrator soll die Suchmaske möglichst frei festlegen können. Damit eine Suchmaschine in ein firmeninternes Intranet eingebunden werden kann, muß es möglich sein, die Abfrage-Seite frei zu gestalten. Der Administrator kann die Abfrage-Seite so gestalten, daß sich die Benutzer möglichst gut zurechtfinden. So können z.B. firmeninterne Beispiele besser geeignet sein als die normalen Standard-Beispiele. Oft sind HTML-Seiten unter gewissen firmeninternen Layout-Vorgaben erstellt. Die Suchmaschine könnte so in dieses Konzept eingegliedert werden.

3.2.5 Suche

Jede Suchmaschine hat ihre eigene Syntax und ihre eigene Regeln. Es ist daher unabdingbar, daß die Benutzer genau informiert werden, wie sie ihre Suchanfrage stellen müssen, um die Dokumente zu finden, die die gewünschten Informationen enthalten. Die vielen Möglichkeiten, die der Benutzer bei der Formulierung hat, sollen dazu dienen, eine möglichst präzise Anfrage zu stellen. Dabei sollte der Anwender keine kryptischen Befehle auswendig lernen müssen, sondern sich einer möglichst natürlichsprachlichen Anfrage-Sprache bedienen können. Das Programm muß so fehlertolerant sein, daß kein Anwender-Fehler den Ausfall der Suchmaschine verursachen kann. Ein einmal gemachter Fehler darf sich nicht auf weitere Such-Anfragen auswirken.

- Unscharfe Suche
  Bei einer unscharfen Suche werden auch Wörter, die sich um ein oder mehrere Zeichen von dem Suchbegriff unterscheiden, gefunden. „Im WWW etwas zu suchen, von dem nicht weiß, ob es im Singular oder Plural, als Kompositum oder in der Form mehrere Wörter, in der alten oder neuen Rechtschreibung oder womöglich in fehlerhafter Orthographie gespeichert ist, das ist mit konventionellen Suchverfahren nahezu aussichtslos. Doch mit Systemen zur unscharfen Suche findet man die gewünschte Information in den meisten Fällen trotzdem.“ [c't 4/97]
  So können auch Dokumente gefunden werden, bei denen sich der Autor bei einem Wort verschrieben hat. Man kann nicht davon ausgehen, daß die Informationen alle fehlerfrei veröffentlicht werden. Ein häufiger Fehler ist das Vertauschen von zwei Buchstaben, so daß eigentlich zwei Buchstaben falsch sind. Wird hier nur eine Toleranz von einem Fehler zugelassen, so wird der Begriff in Dokument nicht gefunden.

Je mehr Unterschiede aber zugelassen werden, um so größer wird die Anzahl der Treffer, da dann immer mehr Begriffe gefunden werden, die mit dem Suchbegriff keine semantische Übereinstimmung haben. Der Benutzer muß die Anzahl der Fehler pro Wort selbst einstellen können.

- Groß- und Kleinschreibung
  Um einen besseren Erfolg zu haben, erscheint es sinnvoll, bei der Suche nach Groß- und Kleinschreibung zu unterscheiden. So hat der Benutzer die Möglichkeit, seine Suche etwas zu präzisieren. Die verschiedenen Suchmaschinen haben unterschiedliche Syntaxkonventionen für die Groß- und Kleinschreibung. Daher muß der Anwender wissen, ob die Groß- und Kleinschreibung berücksichtigt wird oder nicht. Am besten ist es, wenn der Anwender die Wahl hat, ob sein Suchbegriff unabhängig oder abhängig von der Groß- und Kleinschreibung interpretiert wird.

- Umlaute
  Da im deutschen Alphabet Umlaute enthalten sind, und viele Dokumente diese auch enthalten, ist es notwendig, daß auch die Suche nach Wörtern mit Umlauten möglich ist. Hier ergibt sich die Schwierigkeit, daß aber nicht alle Benutzer eine Tastatur mit Umlauten zur Verfügung haben. Daher muß bei der Eingabe von „ae“ auch ein Wort gefunden werden, das „ä“ enthält. Nur so kann sichergestellt werden, daß alle Benutzer auch Suchbegriffe mit Umlauten verwenden können. Ferner ist es in HTML üblich z.B. den Umlaut „ä“ auch „ä“ zu schreiben. Auch diese Form muß von der Suchmaschine richtig interpretiert werden.

- Sonderzeichen
  Die Suchmaschine sollte auch die Suche nach Sonderzeichen unterstützen. Es sollte möglich sein, alle an einer gängigen Tastatur verfügbaren Zeichen in eine Suchanfrage einzubeziehen.

- Unterstützte Sprachen
  Die Suchmaschine sollte keine Einschränkung auf irgendeine bestimmte Sprache haben. Diese Forderung ist so aber sehr schwierig zu erfüllen besonders, wenn Sprachen mit anderen Zeichen verwendet werden, wie z.B. in der japanischen Sprache. Daher soll es genügen, wenn folgende Sprachen unterstützt werden: Englisch, Deutsch, Französisch. Es sollte deshalb keine Rolle spielen, in welcher dieser Sprachen die Information verfaßt ist. Wenn bei der Indexierung die Füllwörter entfernt werden, ist zu beachten, daß jede Sprache eigene Füllwörter hat.

- Gewichtete Suchbegriffe
  Unterstützt die Suchmaschine eine Gewichtung der Suchbegriffe, so können den einzelnen Suchbegriffen unterschiedliche Prioritäten zugeordnet werden. So kann der Anwender festlegen, daß ihm z.B. ein Suchbegriff dreimal so wichtig ist wie ein anderer. Der Benutzer kann damit erreichen, daß der eine Begriff auf jeden Fall in dem Dokument enthalten sein muß, ein anderer aber nicht zwingend in dem Dokument vorkommt.

- Wildcards
  Oft ist es zweckmäßig nicht den exakten Suchbegriff anzugeben, sondern nur den Anfang des Wortes, um dann auch die Dokumente zu finden, in denen der Suchbegriff z.B. im Plural oder in einem anderen Fall verwendet wurde. Entweder ergänzt die Suchmaschine den Suchbegriff automatisch oder der Benutzer muß ein Sonderzeichen dafür eingeben (z.B. „*“). Da jedoch das uneingeschränkte Benutzen von Wildcards leicht zu einer unüberschaubaren Menge von Treffern führt, wird bei manchen Suchmaschinen der Einsatz eingeschränkt. So darf ein „*“ erst nach ein paar anderen Zeichen im Suchbegriff auftauchen oder ein „*“ ersetzt nur eine bestimmte Anzahl von Buchstaben. Neben „*“ kann z.B. auch „?“ verwendet werden. Der Unterschied liegt darin, daß dann nur ein Zeichen in einem Wort beliebig sein darf.

- Phrasensuche
  Unter Phrasensuche versteht man die Suche nach einem kurzen zusammenhängenden Text. Diese ist nur möglich, wenn die Suchmaschine eine Volltextindizierung durchführt und keine Füllwörter ignoriert. Es werden nur die Dokumente gefunden, in denen die Suchbegriffe in genau derselben Reihenfolge auftreten, wie in der Suchphrase. Eine einfache Suche würde die Wörter nicht als zusammengehörig erkennen und nur nach den einzelnen Wörtern suchen. Damit die Suchmaschine erkennt, daß der Benutzer nach einer Phrase sucht, werden die Suchbegriffe meist mit „“ eingeschlossen.

- Verschiedene Suchmodi
  In einem großen Unternehmen ist das Wissen der verschiedenen Benutzer sehr unterschiedlich. Ein Teil arbeitet sehr selten mit dem Computer und ein anderer fast ausschließlich. Daher ist es sinnvoll, für verschiedene Benutzergruppen verschiedene Suchmodi zu haben. So können die Benutzer, die nur einfache Anfragen stellen, eine einfache Suchmaske verwenden und werden nicht durch zu viele Details verwirrt. Die Experten sollen die Möglichkeiten haben, komplexere Suchanfragen zu stellen. Die verschiedenen Suchmodi werden vom Administrator eingerichtet und sollen frei gestaltet sein dürfen.

- Zeitspanne
  Bei einem großen Datenbestand ist es nötig, den Erstellungszeitraum der Dokumente einzugrenzen. So werden nur Dokumente gefunden, die in dem angegebenen Zeitraum veröffentlicht wurden. Weiß der Benutzer das ungefähre Datum, an dem das Dokument veröffentlicht wurde, so kann die Anzahl der durch die Suchmaschine angezeigten Dokumente oft erheblich eingeschränkt werden. So hat der Benutzer z.B. die Möglichkeit, sich Informationen zu einem bestimmten Thema suchen zu lassen, die nicht älter als ein Jahr sein dürfen.

- Kennwörter
  Um zu einem besseren Suchergebnis zu kommen, kann es sinnvoll sein,

die Suche eines Begriffs auf Kennwörter zu begrenzen. Weiß der Benutzer, daß der Suchbegriff im Titel des Dokumentes vorkommt, so kann er z.B. die Suche auf den Titel begrenzen. Dokumente, in denen der Begriff an einer anderen Stelle auftritt, werden nicht angezeigt. Sucht der Benutzer nach einem bestimmten Bild, dessen Namen er weiß, so kann er die Suche auf Bilder einschränken. Genauso können alle Seiten gesucht werden, die Links zu einer bestimmten HTML-Seite enthalten. Es gibt verschiedene Kennwörter:
z.B. url:..., author:..., title:..., link:..., image:..., applet:...[digital 97]

- Boolesche Operationen
  Zur Erstellung komplexer Suchanfragen ist es nötig, mehrere Suchbegriffe mit Booleschen Operatoren zu verknüpfen. Auf jeden Fall müssen die Operatoren „OR", „AND" und „NOT" vorhanden sein. Weitere Operatoren wie „NEAR" sind zwar nicht unbedingt notwendig, aber oft ganz nützlich.
  Um komplexe Anfragen formulieren zu können ist die Verwendung von Klammern unverzichtbar. Nur so kann eine größere Anzahl von unterschiedlichen Operatoren verknüpft werden. Da oft nicht eindeutig ersichtlich ist, welcher Operator stärker bindet als ein anderer, ist im Zweifelsfall immer zu der Verwendung von Klammern zu raten. Die Anzahl der möglichen Klammerebenen sollte nicht zu gering sein.
  Bei manchen Suchmaschinen wird „+" als logisches UND verwendet. Damit ist z.B. eine Suche nach C++ nicht möglich, ohne daß alle Informationen über C auch angezeigt werden.

- Suche nach Schlüsselwörtern
  Schlüsselwörter sind Wörter, die eine bestimmte Bedeutung bei den Anfragen an die Suchmaschine haben. Ein Beispiel sind die Booleschen Operatoren. Um auch eine Suche nach diesen Schlüsselwörtern zuzulassen, muß der Begriff dann anders eingegeben werden, als wenn „OR" als Boolescher Operator verwendet wird. Die Suchmaschine muß daher unterscheiden können, welches „OR" gemeint ist.

- Kommentare
  In HTML gibt es die Möglichkeit, Kommentare einzugeben. Diese werden vom WWW-Browser ignoriert und nicht angezeigt. Daher muß auch die Suchmaschine den Kommentar ignorieren, und nicht in ihren Index mit aufnehmen. Sonst kann es leicht zu einer Verwirrung der Benutzer kommen, wenn ein Dokument in der Trefferliste erscheint, der WWW-Browser den gesuchten Begriff aber nicht am Bildschirm darstellt.

- Kontextsuche
  Mit Hilfe der Kontextsuche gelingt es, nicht nur den exakten Suchbegriff zu finden, sondern auch Wörter mit derselben oder ähnlichen Bedeutung. Dies kann durch ein Synonymwörterbuch, das evtl. noch von Hand ergänzt werden kann, realisiert sein. Wenn nach einem Begriff gesucht wird, wird auch ein Wort, das aus dem gleichen Umfeld ist, gefunden. Bei der Suche

nach z.B. Raumfahrt werden auch alle Artikel, die das Wort Ariane enthalten, gefunden.
Da sich der Zusammenhang bestimmter Wörter zueinander immer wieder ändert, sollten diese Abhängigkeiten von Hand verändert und mit einer Gewichtung versehen werden können. Da jeder Benutzer andere Assoziationen hat, müssen diese Einstellungen benutzerspezifisch sein. Dies bedeutet aber einen sehr großen Aufwand, der sich nur dann rechtfertigt, wenn das Suchen von Informationen eine große Rolle bei der täglichen Arbeit spielt.

3.2.6 Suchergebnis

Der Wunsch des Benutzers ist eine schnelle und umfassende Antwort auf seine Such-Anfrage. Eine Suchmaschine, die nach ein paar Minuten Tausende von Dokumenten unsortiert anzeigt, wird keine besondere Resonanz finden. Daher ist es wichtig, daß gewisse Kriterien erfüllt werden. Erhält der Anwender einmal zu viele Ergebnisse, so muß er die Anfrage präzisieren können.

- Geschwindigkeit
  Um ein vernünftiges Arbeiten mit der Suchmaschine sicherzustellen, darf es nicht zu lange dauern, bis der Benutzer eine Antwort auf seine Suchanfrage bekommt. Auch bei großen Datenbeständen muß die Antwortzeit so kurz sein, daß der Benutzer nicht entnervt die Suche abbricht und von Hand anfängt, sich die gesuchte Information zu beschaffen. Die Geschwindigkeit ist natürlich auch von der Rechenleistung und der Größe des Hauptspeichers abhängig.
  Die Bestimmung der Antwortzeiten ist relativ schwierig, da die Übertragung der HTML-Seite über das Netz nicht mitgerechnet werden darf. Andererseits interessiert den Nutzer am meisten die gesamte Wartezeit. Um die Antwortzeit so gering wie möglich zu halten ist es nötig die Suchmaschine mit einer schnellen Netzanbindung zu versehen.

- Anzahl der Treffer
  Um übersichtliche Trefferlisten zu erhalten, ist es nicht sinnvoll, alle Treffer auf einer Seite untereinander anzuordnen. Daher sollte der Benutzer sowohl die maximale Anzahl der Treffer insgesamt als auch die maximale Anzahl der Treffer pro Seite einstellen können. Gerade wenn die Treffer in einer gewissen Rangfolge ausgegeben werden, sind die letzten Trefferstellen nahezu unbrauchbar.

- Art der angezeigten Informationen
  Gerade in Abhängigkeit von der Anzahl der Treffer pro Seite ist darauf zu achten, daß die Seiten nicht zu groß werden. Daher ist es entscheidend, welche Informationen bei dem Suchergebnis angezeigt werden. Auf jeden Fall muß der Titel des Dokumentes und die URL, unter der das Dokument liegt, erscheinen. Aufgrund der URL kann der Benutzer verifizieren, auf welchem Server sein gesuchtes Dokument liegt. Wird die URL hingegen

in einer internen, kryptischen Form angezeigt, hat sie für den Benutzer keinen Nutzen mehr.
Der Benutzer sollte Art und Umfang der angezeigten Informationen einstellen können. Werden nur die gefunden Dokumente angezeigt, wird wesentlich weniger Platz benötigt, als wenn noch bei jedem Dokument ein kurzer Textausschnitt angezeigt wird. Dabei kann der Textausschnitt ein Teil am Anfang des Dokumentes sein, oder ein Stück um den gesuchten Begriff herum. Vorteilhaft ist es, wenn der Autor des Dokumentes den Textausschnitt bestimmen kann.
Die zusätzlichen Informationen können jedoch dem Benutzer helfen, zu entscheiden, welches Dokument für ihn am wichtigsten sein kann. Wird die Größe der Datei mit angezeigt, so kann der Benutzer feststellen, ob die Datei evtl. zum Laden zu groß ist. Auf Grund des Indizierungsdatums kann leicht festgestellt werden, wie aktuell ein Dokument ist. Eine Anzeige, die nur die Anzahl der Treffer angibt, erscheint unwichtig, da sie höchstens für statistische Zwecke entscheidend sein kann, nicht aber zum Finden von Informationen.

- Ranking
  Damit der Benutzer nicht die wichtigstem Dokumente zuletzt erhält, ist eine Sortierung der Treffer unverzichtbar. Der Text, der am besten zu der Anfrage des Benutzers paßt, wird in der Trefferliste an erster Stelle ausgegeben. Die Gewichtung sollte nicht starr vorgegeben sein, sondern vom Benutzer verändert werden können.
  Leider verwenden manche Suchmaschinen statistische Analysen, mit deren Hilfe die Priorität der Wörter festgelegt wird. Sucht der Benutzer nach seltenen Begriffen, erhält er die Fundstellen nicht immer in der von ihm erwarteten Reihenfolge. Besser ist es daher, dem Benutzer die Bewertung selbst zu überlassen. Hat die Suchmaschine ein Ranking-Feld, so werden Dokumente früher ausgegeben, die diesen Begriff enthalten, als andere Dokumente. Werden die Treffer sortiert angezeigt, so ist es sinnvoll, dem Benutzer auch diese Prioritäts-Zahl anzuzeigen. Damit erfährt der Benutzer, daß ein bestimmtes Dokument z.B. zu 70% seiner Suchanfrage entspricht. So kann er entscheiden, ab welcher Priorität die Dokumente für ihn unnütz sind.

- Markierung des Suchbegriffes
  Es ist sehr wünschenswert, wenn der gefundene Suchbegriff im Dokument markiert wird. So findet der Benutzer leichter die Teile des Dokumentes, die den Begriff enthalten. Wenn der Suchbegriff nicht markiert wird, muß der Benutzer das Dokument evtl. noch nach dem Begriff durchsuchen. Gerade bei größeren Dokumenten ist diese Vorgehensweise nicht ideal. Ob der Begriff markiert wird oder nicht, hängt aber auch von der Art des Dokumentes ab, da nicht alle Dateiformate und deren Editoren Markierungen unterstützen.

3.2.7 Dokumentenmanagement

Eine Suchmaschine kann und soll kein Dokumentenmanagement-System ersetzten. Erforderlich hingegen ist es, daß sich die Suche über alle Dokumente erstrecken kann.

- Unterstützte Dokumentarten
  Es sollen nicht nur HTML-Seiten, sondern auch Dokumente von gängigen WINDOWS-Anwendungen (z.B. WORD, EXCEL, PDF) durchsucht werden. Vorteilhaft ist es, wenn auch in PostScript-Dateien gesucht werden kann. Die möglichen Dokumentarten sollten bei der Konfiguration einstellbar sein. Der Administrator legt damit fest, welche Dokumentarten in den Index aufgenommen werden, und welche nicht.
  Um überhaupt Dokumente, die nicht im HTML-Format sind, indizieren zu können, müssen diese Dateien konvertiert werden. Wichtig ist, daß diese Konverter fehlerfrei arbeiten, und das entsprechende Datei-Format vollständig verstehen. So darf es z.B. keine Schwierigkeiten bereiten, wenn in WORD die Dateien mit der „Schnellspeicherfunktion“ gespeichert werden. Da sich die Datei-Formate ständig weiterentwickeln und neue hinzukommen, muß der Hersteller sicherstellen, auch in Zukunft diese Formate zu unterstützen.

- Anzeige der Dokumente
  Die Dokumente sollten so originalgetreu wie möglich angezeigt werden. Gerade HTML-Dokumente werden nicht konvertiert, da sonst die Gefahr besteht, daß die Benutzer ihre eigenen Seiten nicht wiedererkennen.
  Dokumente, die keine HTML-Seiten sind, können nicht ohne weiteres in einem WWW-Browser angezeigt werden. Es gibt verschiedene Möglichkeiten, dem Benutzer den Text verfügbar zu machen. Es kann das Dokument in ein Format konvertiert werden, das ausgegeben werden kann, oder es werden die vorhanden Plug-Ins verwendet. Bei den Konvertern ist darauf zu achten, daß durch die Konvertierung nicht zu viel Layout-Information verloren geht. Wird das Originaldokument ohne Verwendung von Konvertern angezeigt, ist eine Markierung des Suchbegriffes nicht realisierbar.
  Sobald sich die entsprechenden Dateiformate ändern, müssen aber auch die dazugehörigen Konverter angepaßt werden. Es ist auch möglich, daß das entsprechende Anwendungsprogramm die Datei lädt. Somit ist es z.B. denkbar, daß ein Benutzer an einem PC mit WINDOWS 95 ein WORD-Dokument gefunden hat und jetzt betrachten will. Beim Klicken mit der Maus auf die gewünschte Datei wird WORD mit dem gewünschten Text gestartet.

3.2.8 Agent

- Agent als Erweiterung (vgl. auch 2.4)
  Teilweise ist bei Suchmaschinen ein Agent als Erweiterung vorgesehen.

Das Vorhandensein eines Agenten ist nicht zwingend notwendig, aber gerade für Benutzer, die regelmäßig ähnliche Suchanfragen starten, ist ein Agent eine erhebliche Erleichterung. So entfällt das regelmäßige Suchen nach immer denselben Begriffen. Hat der Benutzer einmal ein Suchprofil erstellt, bekommt er die gewünschten Informationen automatisch mitgeteilt.

- Netzlast
  Der Agent darf keine besonders hohe Netzlast erzeugen. Da die Agenten während des normalen Betriebes im Einsatz sind, ist darauf zu achten, daß die Netzlast dadurch nicht zu groß wird. Die übrigen Anwendungen sollen durch den Einsatz von Agenten nicht gestört oder behindert werden. Dazu ist es z.B. nötig, die Anzahl der Agenten pro Benutzer zu begrenzen.

- Benutzereigene Agenten
  Jeder Benutzer sollte in der Lage sein, eigene Agenten zu generieren. Hierbei muß es möglich sein, daß jeder Benutzer seine Agenten so schützen kann, daß sie nicht von anderen Benutzern gelesen werden können. Dazu sind geeignete Systemfunktionen für einen sicheren Paßwortschutz nötig.

- Übermittlung des Ergebnisses
  Nicht alle Benutzer verbringen den ganzen Tag an einem festen Arbeitsplatz. Daher ist es nötig, die Mitteilung von gefundener Information auf verschiedenste Weise dem Benutzer mitzuteilen. Der Anwender entscheidet, wie ihm das Ergebnis mitgeteilt wird. Durch die steigende Verbreitung von mobilen Systemen wird es immer wichtiger, die Leute unterwegs zu erreichen. Die Vernetzung von Notebooks und der Einsatz von Handys ermöglichen es, Emails und Faxe von unterwegs zu empfangen, oder die dynamisch generierte WWW-Page abzurufen. So ist der Benutzer jederzeit über aktuelle Informationen auf dem laufenden.

3.2.9 Schnittstellen

- Datenbanken
  Um auch nach Daten in Datenbanken suchen zu können, benötigt die Suchmaschine ein Gateway zu einer Datenbank. Das Datenbank-Gateway sollte alle gängigen Datenbanken unterstützen, um eine möglichst große Flexibilität zu erreichen. Für gewisse Anwendungen kann eine solche Anbindung ganz sinnvoll sei, aber wenn jede Datenbank in einem Unternehmen indiziert wird, wird der Index der Suchmaschine sehr groß und die Performance leidet darunter.

- Programme
  Zur Integration der Suchmaschine ist es nötig, die Suche aus einem anderen Anwender-Programm heraus starten zu können. Die gefundenen Dokumente sollten dann mit dem Programm weiterverarbeitet werden können.

3.2.10 Administration

Die Suchmaschine muß einen geringen Administrations-Aufwand haben. Im „normalen“ Betrieb sollte der Administrator kaum Arbeit mit der Suchmaschine haben. Zu den häufigsten Administrations-Aufgaben zählt das Überwachen und Steuern des Indizierers und eine Abfrage des Status der Suchmaschine. Dazu sind komfortable Tools erforderlich. Bei Fehlermeldungen von Benutzern muß der Administrator schnell den entsprechenden Fehler finden und beheben können. Sollte die Suchmaschine einmal nicht verfügbar sein, so müssen die Anwender informiert werden. Bei kritischen Aufgaben, wie z.B. dem Löschen des Indexes, sind Sicherheits-Abfragen erforderlich.

- Installation
  In aller Regel wird eine Suchmaschine nur einmal installiert, so daß die Installationsdauer nicht die entscheidende Rolle spielt. Viel wichtiger ist, daß die Installation unkompliziert und selbsterklärend ist. Das Installationsprogramm installiert die Suchmaschine auf einem gewünschten Rechner in Interaktion mit dem Administrator. Dieser legt z.B. fest, unter welchem Pfad die Suchmaschine installiert wird. Dieses Programm unterscheidet sich in der Regel für die verschiedenen Betriebssysteme. Aus Wartungs- und Managementgründen muß es auch möglich sein, die Suchmaschine über Netzwerk zu installieren.

- Installationsmedium
  Bei den heutigen Programmgrößen ist es am bequemsten, wenn eine CD zur Installation mitgeliefert wird.

- Konfiguration
  Im Gegensatz zum Installieren wird der Administrator die Suchmaschine öfters neu konfigurieren. Daher ist hier eine graphische Oberfläche wünschenswert. Mehr Flexibilität erreicht man, wenn dazu kein eigener Client nötig ist, sondern die Konfiguration mit einem WWW-Browser durchgeführt werden kann. Wird die Konfiguration verändert, so darf es nicht nötig sein, den Index neu aufbauen zu müssen.
  Durch geeignete Sicherheitmechanismen muß verhindert werden, daß Benutzer die Konfiguration verändern können.

- Log-Files
  Alle Zustandsänderungen und Fehlermeldungen müssen in Log-Files geschrieben werden, die dann später ausgewertet werden können. Daher ist es erforderlich, daß das „Common Log Format“ unterstützt wird, das auch von den meisten WWW-Servern verwendet wird. Da Log-Files meist sehr groß werden, ist es besser, wenn die verschiedenen Ereignisse in verschiedene Log-Files geschrieben werden.

- Integriertes Management
  Es ist erforderlich, daß sich die Suchmaschine in ein integriertes Management einbinden läßt. Von einer Plattform aus müssen mehrere Such-

maschinen gemanagt werden können. Zum anderen soll die Suchmaschine von einer Management-Plattform überwacht und gesteuert werden. „Die Forderung nach integriertem, unternehmensübergreifendem Management resultiert also auch aus der dringenden wirtschaftlichen Notwendigkeit, Corporate Networks zuverlässig und sicher betreiben zu können.“ [Hegering 93] Die Suchmaschine ist in einem Intranet ein Dienst wie viele andere auch, die sicher und zuverlässig betrieben werden müssen.

- Überwachung der Indizierung
  Der Administrator sollte in der Lage sein, zu überprüfen, wann ein bestimmter Server das letzte Mal indiziert wurde. Nur so kann er den Benutzern mitteilen, wann deren neu erstellte Seite in den Index aufgenommen wurde. Wird ein Server aus welchen Gründen auch immer längere Zeit nicht indiziert, so kann der Administrator die nötigen Schritte unternehmen, um eine möglichst aktuelle Suchmaschine betreiben zu können. Ferner soll der Administrator wissen, welcher Server gerade indiziert wird.

3.2.11 Verhalten in kritischen Situationen

Um eine möglichst hohe Verfügbarkeit zu haben, sollte die Suchmaschine stabil laufen und nicht ausfallen. Es gelten die üblichen Erwartungen, die man bei jeder Software voraussetzt.

- Abfrage des Zustandes
  Der Administrator muß den Zustand der Suchmaschine jederzeit abfragen können. Nur so kann er im Falle von Problemen die Situation richtig einschätzen. So sollte z.B. erkennbar sein, wann die letzte Indizierung stattgefunden hat.

- Benutzermeldungen in kritischen Situationen
  Die Anzahl der maximal gleichzeitigen Anfragen sollte nicht durch die Suchmaschine begrenzt sein. Sollte dennoch eine Anfrage abgewiesen werden, muß der Benutzer eine Meldung erhalten, daß seine Anfrage nicht bearbeitet werden kann. Sollte der Index (aus welchen Gründen auch immer) nicht verfügbar sein, so ist ebenfalls der Benutzer zu informieren, daß keine Daten vorhanden sind.

- Verhalten bei Systemausfall
  Die Suchmaschine sollte eine Möglichkeit haben, die Datenbank zu reorganisieren, für den Fall, daß während der Indizierung der Server ausfällt oder ein andere Fehler auftritt. Gerade wenn die Indizierung sehr lange dauert, ist es sehr ungünstig, wenn alle bis dahin gewonnen Daten verloren wären und der Index vollkommen neu aufgebaut werden müßte.

- Fehlermeldungen
  Damit die Benutzer über den Zustand der Suchmaschine informiert sind,

müssen entsprechende Fehlermeldungen ausgegeben werden. Es ist wichtig, daß die Meldungen eindeutig und präzise sind, so daß sie leicht verstanden werden können und bei einer Fehlersuche hilfreich sind. Die Fehlermeldungen sollen vom System automatisch angezeigt werden, im Bedarfsfall vom Administrator aber verändert oder ergänzt werden können.

3.2.12 Systemanforderungen

- Plattform
  Die Suchmaschine sollte auf vielen Plattformen verfügbar sein. Am wichtigsten erscheint die Unterstützung für diverse UNIX-Systeme und WINDOWS NT. Voraussetzung für eine leistungsstarke Suchmaschine ist eine entsprechende Hardware. Die Suchmaschine darf die Rechnerarchitektur nicht einschränken.
  Besonders wichtig ist in diesem Zusammenhang auch die Skalierbarkeit. Oft ist es notwendig, die Rechenleistung später zu erhöhen. Sinnvoll ist es auch die Suchmaschine als verteilte Anwendung auf mehrere Rechner zu verteilen. Es ist sehr schwierig, die Entwicklung im Bereich WWW auch nur für ein paar Jahre vorauszusehen.

- Benötigte Rechenleistung
  Um ein System sorgfältig planen zu können, ist es erforderlich, daß man bei einer Angabe des Datenvolumen ungefähr weiß, welche Rechenleistung benötigt wird.

- Benötigter Speicherbedarf (Hauptspeicher, Festplatte)
  Der Hersteller sollte angeben, wie groß der Hauptspeicher sein soll, um ein reibungsloses Arbeiten sicherzustellen. Ferner soll angegeben sein, wie groß der Speicherbedarf des Programms ist. Um den benötigten Speicherbedarf für den Index abschätzen zu können, ist es sehr hilfreich, wenn der Hersteller ungefähr angibt, wie groß der Index im Vergleich zu der Anzahl an Dokumenten werden kann. Der benötigte Speicherbedarf hängt natürlich auch von der Anzahl der Benutzer und deren Anfragen ab.

- Unterstützte Browser
  Die meisten Suchmaschinen verwenden einen WWW-Browser sowohl für die Administration, als auch für die Benutzerschnittstelle. Es sollten daher alle gängigen WWW-Browser verwenden werden können.

- Unterstützte WWW-Server
  Üblicherweise benötigen die Suchmaschinen einen WWW-Server, der auf demselben Rechner wie die Suchmaschine installiert sein muß. Dieser stellt dann die Anfrage-Seiten unternehmensweit zur Verfügung. Um mit dem Kauf einer Suchmaschine nicht gleichzeitig auf einen bestimmten WWW-Server festgelegt zu sein, sollte die Suchmaschine mit möglichst allen bekannten WWW-Servern zusammenarbeiten. Es ist auch möglich, daß ein WWW-Server in die Suchmaschine integriert ist. In diesem Fall wird kein weiterer WWW-Server benötigt.

3.2.13 Dokumentation

- Handbuch
  Gerade für einen Administrator ist ein umfassendes Handbuch von großem Nutzen. Das Handbuch muß gut gegliedert sein und sollte so verfaßt sein, daß es ohne Probleme gelesen und verstanden werden kann. Das Handbuch sollte auch mit der aktuell ausgelieferten Version so übereinstimmen, daß nicht im Handbuch andere Bildschirmmasken zu sehen sind als am Bildschirm. Das erscheint zwar selbstverständlich, ist aber leider häufig nicht verwirklicht. Oft ist eine kurze Einführung gut geeignet, um einen kleinen Überblick über die Suchmaschine zu vermitteln.

- Online-Handbuch
  Ein Online-Handbuch kann und soll auch nie ein Handbuch aus Papier ersetzen, kann aber durchaus eine gute Ergänzung sein. Bei längeren Textabschnitten ist das Lesen am Bildschirm sehr mühsam. Oft wird das Handbuch parallel zur Suchmaschine am Bildschirm benutzt. Da ist es sehr aufwendig, wenn sowohl das Handbuch wie auch die Suchmaschine am Bildschirm angeordnet werden müssen.

- Sprache
  Es ist wünschenswert, wenn die Handbücher nicht nur in Englisch, sondern auch in Deutsch und anderen Sprachen wie z.B. Französisch verfügbar sind.

- Fortbildungskurse
  Für die Einführungsphase der Suchmaschine ist es hilfreich, wenn der Hersteller der Suchmaschine einen Fortbildungskurs halten würde, um den Benutzern die Arbeit mit dem neuen Programm zu erleichtern.

3.2.14 Support

Die meisten Hersteller lassen sich den Support gesondert bezahlen. Daher kann man dann auch gewisse Leistungen beanspruchen.

- Hotline
  Da es immer wieder Probleme geben kann, ist es absolut nötig, daß über eine Hotline alle Fragen an den Hersteller gerichtet werden können. Die Hotline muß das nötige Know-how haben, um auch bei komplexen Problemstellungen weiterhelfen zu können.

- Update
  Um nicht nach ein paar Jahren ein veraltetes Produkt einsetzen zu müssen, ist es nötig, daß der Hersteller in regelmäßigen Abständen ein Update herausbringt. In den neuen Versionen sollten die bisher bekannten Fehler alle behoben sein, so daß eine neue Version immer eine Verbesserung darstellt.

- Patches
  Um akute Fehler am Programm zu beheben, ist es nötig, daß der Hersteller Patches zu Verfügung stellt. Dies kann evtl. via FTP erfolgen.

3.2.15 Preis

Der Preis von Suchmaschinen ist nicht einfach zu vergleichen. Gerade, wenn die Hersteller das Produkt in verschiedene Module unterteilt haben, ist eine genaue Gegenüberstellung fast unmöglich, da die einzelnen Module einen verschiedenen Funktionsumfang haben. Oft ist der Preis von der Menge der Informationsseiten und der Anzahl der Benutzer abhängig. Entscheidend ist auch, ob der Support im Kaufpreis schon enthalten ist, oder nicht.

3.3 Besondere Anforderungen der BMW AG

- Übergreifende Suche
  Es soll nicht nur der WWW-Server durchsucht werden, sondern auch andere Server, deren Daten allgemein im Netz verfügbar sind. (File-Server)

- Dokumentenmanagement
  Es müssen auch Dokumente aus gängigen WINDOWS-Anwendungen (z.B. WORD, EXCEL) durchsucht werden können.

- Suche
  Umlaute müssen unterstützt werden, und es darf keine Beschränkung der Sprache geben.

- Administration
  Der Administrationsaufwand muß möglichst gering sein. Die Suchmaschine soll nach erfolgreicher Installation und Konfiguration ohne nennenswerten Aufwand betrieben werden.

- Systemanforderungen
  Da bei der BMW AG der Browser von Netscape als Standard verwendet wird, ist die Unterstützung eines anderen WWW-Browsers nicht erforderlich.
  Ebenso wird derzeit im Intranet exklusiv der „Netscape Enterprise Server“ eingesetzt. Andere WWW-Server müssen daher nicht unterstützt werden.

3.4 Qualität der veröffentlichten Dokumente

Die Qualität einer Suchmaschine hängt aber auch zu einem entscheidenden Teil von der Qualität der vorhandenen Daten, die indiziert werden, ab. Alleine die Existenz von Dokumentiteln kann die Trefferquote schon erheblich steigern. Beim Verfassen von HTML-Seiten kann der Benutzer zusätzliche Kennwörter eingeben, die die Indizierung beeinflussen. Dazu wird die META tag Funktion verwendet. [Weibel 96] So kann z.B. eine kurze Beschreibung des Dokumentes angegeben werden. Da es aber keine zwingenden Vorschriften gibt, wie ein Dokument verfaßt sein muß, ist es sinnvoll, unternehmensweite Regeln für die Veröffentlichung von Dokumenten festzulegen.

Kapitel 4

Evaluierung von Suchmaschinen an Hand der aufgestellten Kriterien

Im Rahmen dieser Arbeit wurden eine Reihe von Suchmaschinen untersucht. Eine Grundvoraussetzung für den Einsatz im Intranet der BMW AG ist, daß es sich um ein kommerzielles Produkt handelt. Nur so ist eine Weiterentwicklung des Produktes und ein ausreichender Support sichergestellt. Alle frei verfügbaren Suchmaschinen schieden deshalb von Anfang an aus.

Im Internet gibt es eine fast unüberschaubare Anzahl von Suchmaschinen. Dies liegt auch daran, daß viele Anbieter dieselbe oder ähnliche Technologien verwenden wie ihre Konkurenz. Die Anzahl der Produkte, die für ein Intranet angeboten werden, ist wesentlich geringer.

Die Evaluierung kann sich nur auf den Ist-Zustand der untersuchten Produkte beziehen. Funktionen, die heute fehlen oder fehlerhaft sind, können schon in der nächsten Version implementiert sein. Eine Produktentscheidung nur aufgrund einer Evaluierung gestaltet sich schwierig, da die bestehenden Produkte ständig erweitert werden.

4.1 Suchmaschine von Netscape

Der „Catalog Server" von Netscape ist als Produkt bei dem Netscape Enterprise Server enthalten. Die Technologie der Suchmaschine stammt von Verity, beinhaltet aber einen stark eingeschränkten Funktionsumfang. Da jedoch nur HTML-Dokumente durchsucht werden können und keine WORD- oder ähnliche Dateiformate unterstützt werden, kommt diese Suchmaschine für einen Einsatz bei BMW nicht in Betracht.

4.2 Suchmaschine von PLS

Die Firma PLS stellte in einer Präsentation ihr Produkt bei der BMW AG vor. Es stellte sich heraus, daß die gefundenen Dokumente schlecht dargestellt werden. Darüber hinaus waren die Mitarbeiter von PLS, die die Präsentation durchführten, nicht in der Lage, die offenen Fragen befriedigend zu beantworten. Deshalb wurde die Entscheidung getroffen, das Produkt nicht weiter zu untersuchen.

4.3 Gegenüberstellung von Search '97 und Alta Vista

Zuerst werden die beiden Suchmaschinen, die für einen Einsatz bei der BMW AG in die engere Auswahl kommen, gegenübergestellt. Anschließend wird dann diese Bewertung genauer dargelegt und es werden die Produkte detaillierter beschrieben. Die folgenden Tabellen sollen einen groben Überblick über die Funktionsfähigkeiten beider Suchmaschinen geben.

Erfüllt ein Programm die aufgestellten Kriterien, erhält es ein $\oplus$, werden die Kriterien jedoch nicht erfüllt, wurde das Produkt mit $\ominus$ bewertet. Es konnten allerdings nicht alle Funktionen im einzelnen verifiziert werden. Konnte nicht eindeutig entschieden werden, ob das Kriterium erfüllt ist oder nicht, wird in der entsprechenden Spalte ein „?“ eingetragen.

Für die Erläuterungen der Funktionalität sei auch auf die entsprechenden WWW-Server von digital [1] und Verity [2] verwiesen.

Unterstützte Informationsquellen	**Search '97**	**Alta Vista**
Intranet	$\oplus$	$\oplus$
Internet	$\oplus$	$\oplus$
Filesysteme	$\oplus$	$\oplus$
Newsgroups	$\oplus$	$\oplus$
Emails	$\oplus$	$\ominus$
Datenbanken	$\oplus$	$\ominus$
CD-ROMs	$\oplus$	$\oplus$

[1] http://www.altavista.digital.com
[2] http://www.verity.com

Indizierung	**Search '97**	**Alta Vista**
Vollständigkeit	⊕	⊕
Behandlung von HTML-Seiten	⊖	⊕
Verhinderung der Indizierung bestimmter Seiten	⊖	⊕
Volltextrecherche	⊕	⊕
Zeitpunkt der Indizierung	⊕	⊕
Reindizierung	⊕	⊕
Behandlung von nicht statischen Seiten	⊖	⊖
Netzlast während der Indizierung	?	?
Verteilte Datenbanken	⊖	⊖

Sicherheit	**Search '97**	**Alta Vista**
Geschützte Dokumente	⊕	⊕
Sicherheitsmechanismen	⊕	⊕
Einschränkung der Suche auf bestimmte Server	⊕	⊕

User-Interface	**Search '97**	**Alta Vista**
Benutzerfreundliche Eingabe	⊕	⊕
Einschränkung der Suche auf bestimmte Bereiche	⊕	⊕
Hilfesystem	⊕	⊕
Verknüpfung von Suchbegriffen	⊕	⊕
Suchmaske	⊕⊕	⊕

Suche	**Search '97**	**Alta Vista**
Unscharfe Suche	⊖	⊖
Groß- und Kleinschreibung	⊕	⊕
Umlaute	⊕	⊕
Sonderzeichen	⊕	⊕
Unterstützte Sprachen	⊕	⊕
Gewichtete Suchbegriffe	⊕	⊕
Wildcards	⊕	⊕
Phrasensuche	⊕	⊕
Verschiedene Suchmodi	⊕	⊕
Zeitspanne	⊕	⊕
Kennwörter	⊕	⊕
Boolesche Operationen	⊕	⊕
Suche nach Schlüsselwörtern	⊕	⊕
Kommentare	⊖	⊕
Kontextsuche	⊕	⊖

Suchergebnis	**Search '97**	**Alta Vista**
Geschwindigkeit	?	?
Anzahl der Treffer	⊕	⊕
Art der angezeigten Informationen	⊕	⊕
Ranking	⊕	⊕
Markierung des Suchbegriffes	⊕	⊖

Dokumentenmanagement	**Search '97**	**Alta Vista**
Unterstützte Dokumentarten	⊕	⊕
Anzeige der Dokumente	⊖	⊕

Agent	**Search '97**	**Alta Vista**
Agent als Erweiterung	⊕	⊖
Netzlast	?	
Benutzereigene Agenten	⊕	
Übermittlung des Ergebnisses	⊕	

Schnittstellen	**Search '97**	**Alta Vista**
Datenbanken	⊕	⊖
Programme	⊕	⊖

Administration	**Search '97**	**Alta Vista**
Installation	⊖	⊕
Installationsmedium	⊕	⊕
Konfiguration	⊕	⊕
Log-Files	⊕	⊕
Integriertes Management	⊖	⊖
Überwachung der Indizierung	⊕	⊖

Verhalten in kritischen Situationen	**Search '97**	**Alta Vista**
Abfrage des Zustandes	⊕	⊕
Benutzermeldungen in kritischen Situationen	⊖	⊕
Verhalten bei Systemausfall	⊕	⊖
Fehlermeldungen	⊕	⊕

Systemanforderungen	**Search '97**	**Alta Vista**
Plattform	⊕	⊖
Benötigte Rechenleistung	⊖	⊕
Benötigter Speicherbedarf	⊕	⊕
Unterstützte Browser	⊕	⊕
Unterstützte WWW-Server	⊕	⊕

Dokumentation	Search '97	Alta Vista
Handbuch	$\oplus$	?
Online-Handbuch	$\oplus$	$\oplus$
Sprachen	$\ominus$	$\ominus$
Fortbildungskurse	$\oplus$	?

Support	Search '97	Alta Vista
Hotline	$\oplus$	$\oplus$
Update	$\oplus$	$\oplus$
Patches	$\oplus$	?

Preis	Search '97	Alta Vista
Preis	?	?

4.4 Search '97 von Verity

Verity bietet im Bereich von Suchwerkzeugen für Intranets die Produktfamilie Search '97 an. Diese beinhaltet unter anderem folgende Produkte, die für einen Einsatz im BMW-Intranet geeignet sein könnten:

- Information Server
  Der Information Server ist das Kernstück der Suchmaschine, und dient als ihre Plattform.

- Remote Spider
  Um die WWW-Server im Intranet indizieren zu können, wird der Remote Spider benötigt.

- Agent Server
  Mit dem Agent Server ist es möglich, benutzerspezifische Agenten zu generieren.

- Information Gateways
  Mit den Information Gateways lassen sich auch noch Informationen aus Datenbanken oder anderen Applikationen indizieren und die gefundenen Dokumente anzeigen.

4.4.1 Unterstützte Informationsquellen

- Intranet
  Alle WWW-Server im Intranet können indiziert werden. Der Administrator legt dazu eine Liste der entsprechenden Rechner an.

- Internet
  Es können bestimmte Server ausgewählt werden, die auch indiziert werden

sollen, die sich nicht im Intranet, sondern im Internet befinden. Es werden dann nur diese Server indiziert. Links zu anderen Servern im Internet werden nicht verfolgt.

- Filesysteme
  Alle Daten, die von dem Rechner, auf dem die Suchmaschine installiert ist, erreicht werden können, können auch indiziert werden.

- Newsgroups
  Alle gewünschten Newsgroups können in den Index aufgenommen werden.

- Emails
  Der Anwender kann sich seine Emails indizieren lassen. Da jeder nur Zugang zu den eigenen Mails hat, kann sich die Suche natürlich auch nur über die eigenen erstrecken.

- Datenbanken
  Mit dem Datenbank Gateway können alle gängigen Datenbanken durchsucht werden.

- CD-ROMs
  Es ist eine Indizierung möglich, wenn sich die CD-ROMs ständig in einem CD-ROM Laufwerk befinden und der Rechner, auf dem die Suchmaschine installiert ist, auf diese zugreifen kann.

4.4.2 Indizierung

- Vollständigkeit
  Der Hersteller sichert zu, daß alle angegebenen Informationsquellen vollständig durchsucht werden. Bei einem großen Kommunikations-Netz wie bei der BMW AG kann aber nicht überprüft werden, ob wirklich alle Daten in den Index aufgenommen wurden.

- Behandlung von HTML-Seiten
  Alle gefundenen Dateien werden in ein internes Format konvertiert. Dabei werden unnötigerweise auch HTML-Seiten von diesem Konverter verarbeitet. Dies führt zu Problemen, wenn der Konverter HTML-Code anders interpretiert als der WWW-Browser.

- Verhinderung der Indizierung bestimmter Seiten
  Der Benutzer hat keine Möglichkeit, bei der Erstellung einer HTML-Seite anzugeben, daß diese nicht von dem Roboter durchsucht wird.

- Volltextrecherche
  Es wird eine Volltextindizierung durchgeführt. Alle in den Dokumenten vorkommenden Wörter werden in den Index aufgenommen.

- Zeitpunkt der Indizierung
  Die Abstände, in denen die Indizierung durchgeführt wird, ist frei wählbar.

So kann der Administrator die Indizierung von Hand aktivieren, oder regelmäßig zu einem bestimmten Zeitpunkt ausführen lassen.

- Reindizierung
  Der Indizierer unterstützt inkrementelles Indizieren. Es ist aber auch möglich, den vorhandenen Index zu löschen, und den Index komplett neu aufzubauen. Durch diese Methode wird sichergestellt, daß keine veralteten Links mehr in der Datenbank existieren.

- Behandlung von nicht statischen Seiten
  Dokumente, die nur über sensitive Grafiken erreicht werden können, werden nicht indiziert. Daraus folgt, daß auch alle Seiten, die von diesen referenziert werden, nicht erfaßt sind. Dokumente, die nur über Java-Applets erreicht werden können, werden nicht indiziert.

- Netzlast während der Indizierung
  Die Netzlast während der Indizierung ist so hoch, daß ein Arbeiten in diesem Zeitraum mit diesem Rechner nicht möglich ist.

- Verteilte Datenbanken
  Es ist möglich, das Netz nicht nur als Ganzes zu durchsuchen, sondern auch abschnittsweise. Die dann entstehenden Indizes werden auch Kollektionen genannt. Die verschieden Kollektionen können sich auf verschiedenen Datenträgern befinden. Die Indizes müssen aber dennoch von einem Rechner und von einem Administrator verwaltet werden. Ein abteilungsbezogenes, verteiltes Management es daher nicht möglich.

4.4.3 Sicherheit

- Geschützte Dokumente
  Da der Suchvorgang mit der Kennung als WWW-User durchgeführt wird, werden in dem Index nur die Dokumente erfaßt, die von allen gelesen werden dürfen. So kann es nicht passieren, daß ein nicht autorisierter Benutzer das Dokument lesen kann. Es gibt auch keinen Hinweise auf deren Existenz. Die Einrichtung bestimmter Benutzergruppen ist nicht vorgesehen.

- Sicherheitsmechanismen
  Da nur Daten verarbeitet werden, die frei zugänglich sind, sind keine Sicherheitsmechanismen vorhanden.

- Einschränkung der Suche auf bestimmte Server
  Die Suche beschränkt sich immer auf die Kollektionen. Bei deren Erstellung kann der gewünschte Rechnername angegeben werden. Ferner kann ein Teil der URL festgelegt werden, der bei den durchsuchten Seiten auftauchen muß (z.B. http://www.bmw.muc/). So ist sichergestellt, daß ein Link ins Internet nicht verfolgt wird.

4.4.4 User-Interface

Die Abfrage-Seite für den Benutzer kann sehr frei konfiguriert werden. So obliegt es dem Administrator, die Seiten übersichtlich und benutzerfreundlich zu gestalten. Es werden auch ein paar Seiten mitgeliefert, die nur mehr den eigenen Bedürfnissen angepaßt werden müssen. Für die Suchmaschine von Verity ist auch ein eigener Client erhältlich. Hier wurde aber nur das User-Interface, das mit einem WWW-Browser zusammenarbeitet, betrachtet.

- Benutzerfreundliche Eingabe
  Die vorhanden Beispiele sind übersichtlich gestaltet. Das Eingabe-Feld ist so lang, daß auch mehrere Verknüpfungen von Suchbegriffen möglich sind.
- Einschränkung der Suche auf bestimmte Bereiche
  Der Anwender kann auswählen, in welcher Kollektion er suchen möchte. Sind die Kollektion nach thematischen Gesichtspunkten erstellt worden, kann der Anwender die Suche auf ausgewählte Themenbereiche eingrenzen.

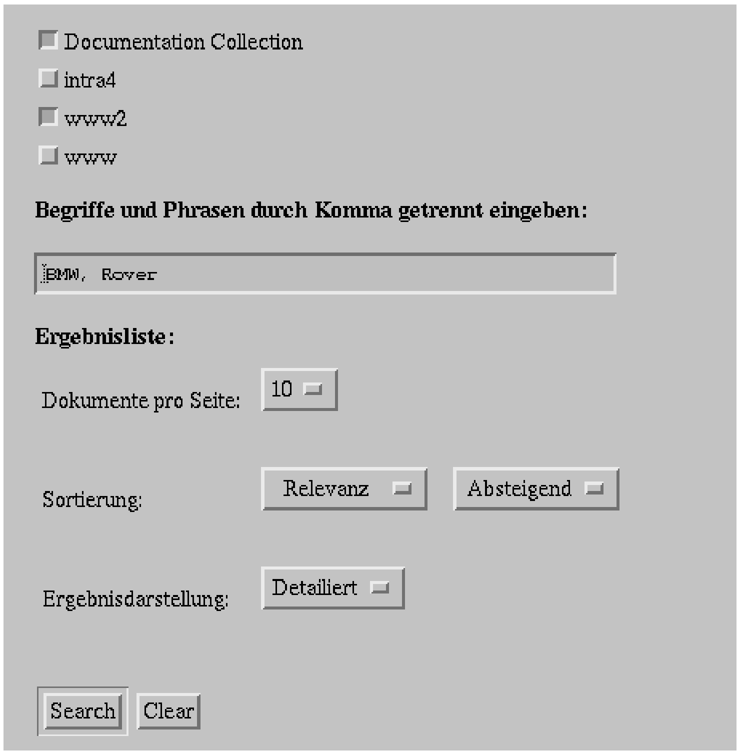

Abbildung 4.1: Beispiel einer Suchmaske von Verity

- Hilfesystem
  Ein umfangreiches Hilfesystem ist vorhanden. Gegebenenfalls kann der Administrator die Anfrage-Seite um eigene Beispiele erweitern.
- Verknüpfung von Suchbegriffen
  Standardmäßig ist vorgesehen, daß der Benutzer im Eingabe-Feld die

Suchbegriffe und die Operatoren eingibt. Prinzipiell könnten aber das User-Interface mit „AND“ und „OR“ Buttons ergänzt werden. Mit den Standard-Seiten ohne Buttons erhält der Anwender eine größere Flexibilität.

- Suchmaske
  Da die Anfrage-Seite in HTML verfaßt ist, kann der Administrator die Seite so abändern, daß sie den firmeninternen Layout-Vorgaben entspricht.

4.4.5 Suche

- Unscharfe Suche
  Der Anwender hat keine Möglichkeit anzugeben, ob und wieviele Fehler in seinem Suchbegriff enthalten sein dürfen. Somit werden Wörter, die fehlerhaft veröffentlicht wurden, nicht gefunden.

- Groß- und Kleinschreibung
  Im Standard-Modus werden alle Wörter, die mit dem Suchbegriff übereinstimmen, unabhängig von ihrer Groß- und Kleinschreibung gefunden. Es gibt aber die Möglichkeit mit einem „CASE“-Modifier die Suche auf die exakte Schreibweise einzuschränken.

- Umlaute
  Die Suchmaschine von Verity hat teilweise Probleme mit der richtigen Behandlung von Umlauten. So werden z.B. bei der Trefferliste die Umlaute nicht angezeigt, wenn es sich bei dem gefundenen Dokument um ein WORD-Dokument handelt. Um alle interessierenden Dokumente zu finden, muß bei der Suche die Schreibung eingegeben werden, die der jeweilige Autor verwendet hat. Für den Benutzer einer Tastatur mit Umlauten bedeutet das doppelte Eingabe, für einen Benutzer einer Tastatur ohne Umlaute führt es dazu, daß in der Regel nicht alle Dokumente gefunden werden.

- Sonderzeichen
  Alle auf der Tastatur befindliche Zeichen können in die Suche integriert werden.

- Unterstützte Sprachen
  Es gibt keinerlei Einschränkungen auf bestimmte Sprachen. Der Benutzer muß nur in der Lage sein, die Zeichen, nach denen er sucht, auch auf seiner Tastatur einzugeben.

- Gewichtete Suchbegriffe
  Eine direkte Gewichtung der Suchbegriffe ist nicht vorgesehen. Der „NOT“ Operator legt nur fest, daß der folgende Suchbegriff nicht in den Dokument enthalten sein darf. Es gibt aber keine Möglichkeit den einzelnen Begriffen Prioritäten zuzuweisen.

- Wildcards
  Wildcards sind nicht erforderlich, da die Suchbegriffe automatisch ergänzt werden. Wird ein Wort in Anführungszeichen geschrieben, so wird nur nach der exakten Schreibweise gesucht.

- Phrasensuche
  Die Suche nach Phrasen wird unterstützt. Dazu werden die gesuchten Wörter mit Anführungszeichen geklammert.

- Verschiedene Suchmodi
  Der Administrator kann verschiedene Suchmodi für bestimmte Benutzergruppen einrichten. Gegebenenfalls muß er dazu dann die Seiten für die einzelnen Suchmodi selbst gestalten oder modifizieren.

- Zeitspanne
  Prinzipiell kann die Suche auf ein bestimmtes Datum oder einen Entstehungszeitraum eingeschränkt werden. Dazu muß evtl. die Suchmaske vom Administrator angepaßt werden.

- Kennwörter
  Mit der Zonen-Suche kann der Anwender in allen Bereichen des HTML-Dokumentes suchen. Will er z.B. seine Suche auf den Titel einschränken gibt er nach dem Suchbegriff „<IN> title“ ein. Genauso kann in Bereichen wie „H1“, „H2“ gesucht werden.

- Boolesche Operationen
  Folgende Operatoren werden unterstützt: AND, OR, NOT und NEAR. Ferner gibt es noch den „,“-Operator. Dieser sucht nach Dokumenten, die mindestens einen der Suchbegriffe enthalten. Die Ausgabe erfolgt dann in der Reihenfolge, daß das Dokument mit den meisten Übereinstimmungen auch als erstes angezeigt wird. Die Anzahl der Klammerebenen ist nicht beschränkt.

- Suche nach Schlüsselwörtern
  Soll z.B. nach dem Wort „OR“ gesucht werden, muß es in Anführungszeichen gesetzt werden, um es nicht mit dem Operator zu verwechseln.

- Kommentare
  Es werden auch Dokumente in der Trefferliste angezeigt, bei denen der Suchbegriff nur im Kommentar des HTML-Dokumentes auftrat. Dies kann die Anwender leicht verunsichern.

- Kontextsuche
  Die Suchmaschine verfügt über eine Kontextsuche. Hierbei kann ein Synonymwörterbuch verwendet werden, oder der Benutzer definiert sich entsprechende Assoziationen von bestimmten Wörtern selbst. Die Funktionalität ist sehr hoch und erhöht bestimmt die Anzahl der gefundenen Dokumente. Auf der anderen Seite ist der Administrations-Aufwand dafür sehr hoch, da nicht alle Personen dieselben Assoziationen haben und sich auch der Zusammenhang der Wörter untereinander ständig ändert.

4.4.6 Suchergebnis

- Geschwindigkeit
  Eine direkte Messung der Geschwindigkeit ist nicht möglich, da zu der Zeit, die benötigt wird, um die Trefferliste zu erzeugen noch die Zeit zur Übertragung von der Suchmaschine zu dem Benutzer-Rechner hinzukommt.

- Anzahl der Treffer
  Die Anzahl der Treffer kann in der Suchmaske eingestellt werden. Der Administrator legt dabei fest, in welchen Schritten der Anwender die Anzahl der Dokumente pro Seite auswählen kann.

- Art der angezeigten Informationen
  Der Administrator kann in der Suchmaske vorsehen, daß die Anwender die Art der angezeigten Informationen auswählen können.

- Ranking
  Die Treffer werden sortiert ausgegeben. Damit der Benutzer ungefähr erkennt, welche Dokumente noch brauchbar sind wird eine Bewertung der Treffer angezeigt. Hier kann zwischen einer Anzeige mit „*“ und einer in Prozent ausgewählt werden.

- Markierung des Suchbegriffes
  Wenn die Art des Dokumentes es zuläßt, werden die Suchbegriffe im Dokument markiert. So werden z.B. in einem PDF-Dokument alle vorkommenden Suchbegriffe grau hinterlegt.

4.4.7 Dokumentenmanagement

- Unterstützte Dokumentarten
  Es werden über 200 verschiedene Dateiformate unterstützt. Darunter auch HTML, PDF, ASCII, DOC. Teilweise treten Probleme auf, wenn WORD-Dokumente mit der „Schnellspeicherfunktion“ abgespeichert werden. Der Administrator muß darauf achten, daß, wenn neue Versionen von z.B. Textverarbeitungsprogrammen und somit neue Dateiformate eingesetzt werden, auch neue Konverter für die Suchmaschine benötigt werden.

- Anzeige der Dokumente
  Die Anzeige der Dokumente ist teilweise nicht korrekt. Da alle Dokumente in ein internes Format konvertiert werden, sehen die Dokumente manchmal anders aus, als der Anwender sie verfaßt hat. Der Konverter kann Informationen vernachlässigen und so das Layout verändern. WORD-Dokumente werden z.B. in PDF-Dokumente umgewandelt und als solche dargestellt.

4.4.8 Agent

- Agent als Erweiterung
  Verity bietet den Agent Server in seiner Produktpalette an. Für das System-Management hat der Administrator eine Reihe von Funktionen wie z.B. das Suchen, Löschen und Editieren von Agenten zur Verfügung. Relevanz-Schwellwerte ermöglichen die individuelle Gewichtung von Begriffen innerhalb einer Profilabfrage.

- Netzlast
  Die Netzlast kann ohne aufwendige Meßapparaturen (wie z.B. Netzanalysator) nicht gemessen werden.
  Der Administrator kann die maximale Anzahl der Agenten pro Benutzer festlegen und so die Netzbelastung einschränken.

- Benutzereigene Agenten
  Jeder Benutzer kann sich seine eigenen Agenten erstellen, von denen er dann auf neue Informationen aufmerksam gemacht wird.

- Übermittlung des Ergebnisses
  Das Ergebnis kann über folgende Medien übermittelt werden: HTML, Email, Pager, Fax. Der Administrator legt fest, welche dieser Möglichkeiten in dem jeweiligen System aktiv sind.

4.4.9 Schnittstellen

- Datenbanken
  Verity unterstützt ein direktes Datenbank-Gateway für Oracle auf der Basis von OCI, sowie ein ODBC-Gateway mit Treibern für den Zugriff auf andere relationale Systeme wie Adabas, Informix, Ingres und Sybase. [Verity 96]

- Programme
  Mit dem Applikations-Gateway kann die Suchmaschine leicht in andere Programme integriert werden.

4.4.10 Administration

- Installation
  Unter UNIX findet die Installation zeilenorientiert statt. Der Administrator kann fehlerhafte Eingaben nur durch eine Neuinstallation rückgängig machen. So werden z.B. ohne Abfrage in den Konfigurationsdateien des WWW-Servers Eintragungen hinzugefügt. Es wird kein Deinstallationsprogramm mitgeliefert.

- Installationsmedium
  Die Suchmaschine wird auf einer CD-ROM ausgeliefert.

- Konfiguration
  Die Konfiguration kann mit einem WWW-Browser durchgeführt werden. Die verschiedene Aufgaben sind auf mehrere HTML-Seiten aufgeteilt.

- Log-Files
  Die Suchmaschine schreibt ihre Einträge in das Log-File des WWW-Servers. Es wird das Common Log Format verwendet.

- Integriertes Management
  Eine Einbettung in eine Management-Plattform ist nicht möglich.

- Überwachung der Indizierung
  Eine Überwachung der Indizierung ist nur eingeschränkt möglich. So kann der Administrator z.B. erkennen, welcher WWW-Server gerade indiziert wird.

4.4.11 Verhalten in kritischen Situationen

- Abfrage des Zustandes
  Während der Indizierung wird der Administrator über alle wichtigen Ereignisse informiert. Der Status der Suchmaschine kann jederzeit abgefragt werden.

- Benutzermeldungen in kritischen Situationen
  Kann die Suchmaschine wegen Überlastung keine neuen Anfragen mehr annehmen, so erhält der Benutzer eine entsprechende Nachricht.

- Verhalten bei Systemausfall
  Falls die Datenbank z.B. durch einen Systemausfall beschädigt wurde, kann diese mit einer Reorganisation wieder funktionsfähig gemacht werden.

- Fehlermeldungen
  Wird der Index gerade neu aufgebaut, so erhält der Benutzer eine Meldung, daß er warten muß, bis der Index aufgebaut ist.

4.4.12 Systemanforderungen

- Plattform
  Die Suchmaschine ist für folgende Plattformen erhältlich: WINDOWS NT, SPARC SunOS, SPARC Solaris, RS6000 IBM/AIX, HP/UX und DEC Alpha OSF1.

- Benötigte Rechenleistung
  Der Hersteller liefert keine Angaben über die benötigte Rechenleistung der Computers, auf dem die Suchmaschine installiert wird.

- Benötigter Speicherbedarf (Hauptspeicher, Festplatte)
  Dar Programm benötigt ca. 65 MB Speicher auf der Festplatte. Für die Größe des Index wird folgende Faustregel angegeben: Indexgröße = 1/3 der zu indizierenden Daten. Der Hauptspeicher sollte so konzipiert werden, daß ein Teil des Indexes immer im Speicher gehalten werden kann.

- Unterstützte Browser
  Das Produkt benötigt den WWW-Browser Netscape Navigator ab V3.x oder Microsoft Internet Explorer ab V3.x.

- Unterstützte WWW-Server
  Der Netscape Enterprise Server und die WWW-Server von Apache und NCSA werden unterstützt.

4.4.13 Dokumentation

- Handbuch
  Es wird ein Satz Handbücher in englischer Sprache mitgeliefert. Dieser besteht unter anderem aus Installations-Anleitung, User's Guide und Administrations-Handbuch.

- Online-Handbuch
  Ein Online-Handbuch ist nur in englischer Sprache vorhanden.

- Sprache
  Die Dokumentationen sind nur in Englisch verfügbar.

- Fortbildungskurse
  Wenn Fortbildungskurse gewünscht werden, kann darüber verhandelt werden.

4.4.14 Support

Wenn der Kunde den Support vereinbart, stehen ihm folgende Dienste zur Verfügung.

- Hotline
  Für Fragen und Hilfe bei Problemen steht eine Telephon-Hotline zur Verfügung.

- Update
  Bei der Weiterentwicklung des Produktes können Updates erworben werden.

- Patches
  Zur Behebung von Programmfehlern werden Patches bereitgestellt. Diese kann dann der Administrator via FTP herunterladen und installieren.

4.4.15 Preis

Die Preise für die Search '97-Produktfamilie sind in folgender Tabelle zusammengefaßt und beziehen sich auf einen Rechner mit einem Prozessor: (Stand: März 1997)

Information Server:	14.900 DM
Remote Spider für 10 WWW-Server:	14.900 DM
Information Gateways:	10.900 DM
Agent Server:	20.000 DM
Toolkit:	80.000 DM

Support (Update, Upgrade, Hotline): 18% der Kaufsumme

4.5 Alta Vista von digital

Die Firma digital vertreibt als Suchmaschine für Intranets das Produkt „Alta Vista Search Private eXtensions“. In der Testphase war aber erst eine Beta-Version des Produkts erhältlich. Es bleibt daher zu hoffen, daß gewisse Fehler in der endgültigen Version beseitigt sind.

4.5.1 Unterstützte Informationsquellen

- Intranet
  Alle WWW-Server innerhalb eines Intranet können in die Suche integriert werden. Der Administrator legt an Hand von Regeln fest, welche Server indiziert werden sollen und welche nicht. Ferner werden ein oder mehrere Start-Server benannt, von denen die Suche nach den Dokumenten beginnt. Alternativ kann auch eine Liste mit allen Servern angelegt werden, die indiziert werden dürfen.

- Internet
  Server, die sich nicht im Intranet befinden, können ebenfalls durchsucht werden, wenn die Regeln für die Suche nach den Dokumenten dies zulassen.

- Filesysteme
  Das gesamte Filesystem des Rechners, auf dem die Suchmaschine installiert ist, kann durchsucht werden.

- Newsgroups
  Die Indizierung von weltweiten und firmeninternen Newsgroups ist möglich.

- Emails
  Emails können in eine Suche nicht integriert werden.

- Datenbanken
  Derzeit gibt es noch keine Gateways zu Datenbanken.

- CD-ROMs
  Alle CD-ROMs, die sich permanent in einem allgemein zugänglichen CD-ROM Laufwerk befinden, können indiziert werden.

4.5.2 Indizierung

- Vollständigkeit
  Es ist davon auszugehen, daß die Indizierung vollständig ist, obwohl es teilweise sehr lange dauert, bis die Dokumente in den Index aufgenommen werden.

- Behandlung von HTML-Seiten
  HTML-Seiten werden korrekt behandelt. Es wird dafür kein Konverter verwendet. Nur die Indizierung von Seiten, die nur über Frames erreicht werden können, bereiten in der derzeitigen Version gewisse Probleme. Dieser Fehler soll aber nach Firmenangaben baldmöglichst behoben werden.

- Verhinderung der Indizierung bestimmter Seiten
  Da die Software den „Standard for Roboter Exclusion“ befolgt, kann der Administrator die Indizierung bestimmter Seiten verhindern. Dazu ist im Root-Verzeichnis eine Datei „robots.txt“ anzulegen. Hier können für den Roboter Regeln angegeben werden, welche Dokumente er nicht durchsuchen darf.

- Volltextrecherche
  Das Produkt ermöglicht eine Volltextrecherche, da es sämtliche Informationen der Dokumente indiziert. Es werden alle Wörter indiziert, unabhängig davon, ob das Wort in einem Wörterbuch existiert oder nicht.

- Zeitpunkt der Indizierung
  Der Administrator kann die Zeit, in der der Roboter Daten sammelt sehr flexibel beeinflussen. Entweder er läßt den Roboter kontinuierlich Daten suchen, oder er legt mit Hilfe eines Schedulers genau fest, zu welchen Zeiten das Netz nach neuen Informationen durchsucht wird. Der Roboter kann jederzeit von Hand gestoppt werden, um z.B. ein Problem zu beheben.

- Reindizierung
  Der Roboter fügt nur neue Informationen in den Index hinzu. Werden Dokumente gelöscht, so werden die entsprechenden Einträge nicht aus dem Index gelöscht. Um diese „toten“ Links zu beseitigen, muß die Datenbank neu aufgebaut werden. Beim Neuaufbau eines Index wird aber der alte gelöscht. Als Konsequenz können die Anwender in dem Zeitraum, bis der neue Indexer alle Server durchsucht hat, deutlich weniger Treffer erhalten. Ferner kann der Administrator bestimmte Dokumente aus dem Index von Hand löschen.

- Behandlung von nicht statischen Seiten
  Dokumente, die nur über dynamische Seiten erreicht werden können, werden nicht indiziert.

- Netzlast während der Indizierung
  Eine signifikante Erhöhung der Netzlast konnte nicht festgestellt werden. Der Administrator hat außerdem die Möglichkeit über die Policie des Roboters die Zeit einzustellen, die der Roboter nach dem Holen eines Dokumentes verstreichen läßt. Bei einer niedrigen Such-Frequenz kann es aber lange dauern, bis das ganze Intranet indiziert wird. Darunter leidet natürlich dann die Aktualität.

- Verteilte Datenbanken
  Alle Daten werden in einem zentralem Index abgelegt. Ein verteiltes Management ist daher nicht möglich.

4.5.3 Sicherheit

- Geschützte Dokumente
  Es werden nur Dokumente indiziert, die von allen Benutzer gelesen werden können. Somit kann nur nach Dokumenten gesucht werden, die allgemein zugänglich sind. Es ist nicht vorgesehen, verschiedene Benutzergruppen einzurichten.

- Sicherheitsmechanismen
  Da nur frei zugängliche Informationen verarbeitet werden, sind keinerlei Sicherheitsmechanismen notwendig.

- Einschränkung der Suche auf bestimmte Server
  Der Administrator kann die Suche durch bestimmte Regeln einschränken. So kann durch „include“ oder „exclude“ Befehle genau festgelegt werden, welche Server und welche Verzeichnisse indiziert werden und welche nicht. Ein Link zu einem Server, der nicht den Regeln entspricht, wird nicht verfolgt.

4.5.4 User-Interface

Das User-Interface kann von dem Administrator frei und flexibel gestaltet werden. Die Software enthält zwei Beispiel-Abfrage-Seiten, die den vollen Funktionsumfang darstellen.

- Benutzerfreundliche Eingabe
  Die mitgelieferten Abfrage-Seiten sind gut gegliedert und übersichtlich gestaltet. Das Eingabe-Feld ermöglicht die Eingabe von langen Suchbegriffen mit mehreren Verknüpfungen.

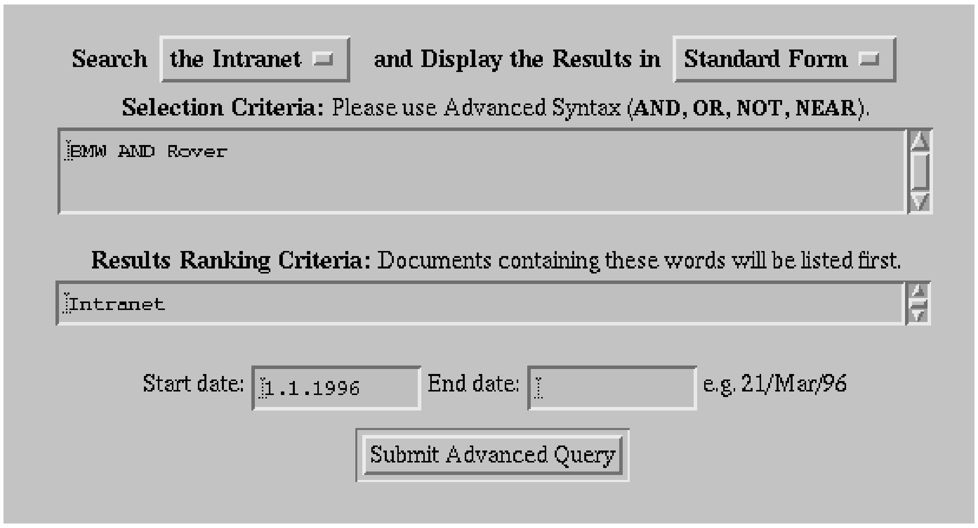

Abbildung 4.2: Erweiterte Suche mit Alta Vista

- Einschränkung der Suche auf bestimmte Bereiche
  Obwohl es nur eine Datenbank gibt, kann dennoch die Suche auf bestimmte Bereiche eingeschränkt werden. Evtl. kann durch die Verwendung von Java oder Java-Skript die Abfrage-Seite so gestaltet werden, daß der Benutzer nur noch auf bestimmten Servern oder in bestimmten Verzeichnissen sucht, die er über Buttons auswählen kann.

- Hilfesystem
  Dem Anwender steht ein komfortables Hilfesystem zur Verfügung. Es hilft vor allem bei der Formulierung effizienter Anfragen an die Suchmaschine. So erfährt der Benutzer, welche Möglichkeiten er insgesamt hat, und welche im jeweiligen Fall die besten sind. Die Hilfe ist in mehreren Sprachen (z.B. Deutsch, Englisch) erhältlich, wobei der Benutzer die Sprache auswählen kann. Zur Unterstützung der Anwender kann der Administrator noch eigene firmenspezifische Beispiele hinzufügen. Ebenso erhält der Administrator durch ein eigenes Hilfesystem Unterstützung bei seinen Administrations-Aufgaben. Diese Hilfe ist nur in Englisch verfügbar.

- Verknüpfung von Suchbegriffen
  Die Suchbegriffe werden in der Regel in dem Eingabe-Feld durch die Verwendung von geeigneten Operatoren miteinander verknüpft. Die Anfrage-Seiten können aber auch so geändert werden, daß die Verknüpfung durch Buttons realisiert ist.

- Suchmaske
  Der Administrator kann alle Abfrage-Seiten verändern und nach persönlichen Bedürfnissen anpassen. Somit ist es ein Leichtes, das User-Interface in die Firmen-Umgebung zu integrieren.

4.5.5 Suche

- Unscharfe Suche
  Eine fehlertolerante Suche ist nicht möglich.
- Groß- und Kleinschreibung
  Werden bei einer Anfrage ausschließlich Kleinbuchstaben eingegeben, so wird bei der Suche die Groß- und Kleinschreibung nicht unterschieden. Wenn der Suchbegriff Großbuchstaben enthält, wird die Suche unter Berücksichtigung der Groß- und Kleinschreibung durchgeführt.
- Umlaute
  Die Suche nach Wörtern mit Umlauten ist problemlos möglich. Hat der Anwender eine Tastatur mit Umlauten, so kann er den exakten Suchbegriff eingeben. Aber auch bei der Eingabe von z.B. „ae“ werden Wörter mit „ä“ gefunden. Wurde bei der Erstellung des HTML-Dokumentes z.B. „aumlt;“ geschrieben, so wird dieses Wort ebenfalls gefunden.
- Sonderzeichen
  Die Suchmaschine unterstützt alle Zeichen im Zeichensatz ISO Latin-1. Damit können alle Zeichen dieses Zeichensatzes für Such-Anfragen verwendet werden.
- Unterstützte Sprachen
  Es gibt keine Einschränkungen für bestimmte Sprachen, solange diese den Zeichensatz ISO Latin-1 verwenden. Das Suchen z.B. in japanischen Texten ist daher nicht möglich.
- Gewichtete Suchbegriffe
  Der Anwender hat keine Möglichkeit, den Suchbegriffen Prioritäten zu geben.
- Wildcards
  Zur Suche von Wörtern mit demselben Muster können Wildcards verwendet werden. Bei Alta Vista wird hierfür das Zeichen „*“ verwendet. Damit die Anzahl der Treffer nicht zu groß wird, müssen sich vor dem Platzhalter aber mindestens drei Zeichen befinden.
- Phrasensuche
  Um nach einer Phrase suchen zu lassen, muß diese in Anführungszeichen gesetzt werden.
- Verschiedene Suchmodi
  Die Suchmaschine verfügt über zwei verschiedene Suchmodi, die sich zwar in ihrer Leistungsfähigkeit nicht unterscheiden, aber in der Mächtigkeit der Eingabe der Such-Anfrage.
- Zeitspanne
  Der Anwender kann die Suche auf eine bestimmte Periode einschränken. Dafür stehen ihm die zwei Felder Anfangsdatum und Enddatum zur Verfügung.

- Kennwörter
  Die Verwendung von Kennwörtern wird unterstützt. So kann die Suche leicht auf z.B. einen bestimmten Server eingeschränkt werden. Folgende Kennwörter werden von der Suchmaschine unterstützt: anchor:..., applet:..., domain:..., host:..., image:..., link:..., text:..., title:..., url:...

- Boolesche Operationen
  Der Anwender kann bei diesem Produkt zwischen zwei verschiedenen Abfragen wählen. Diese unterscheiden sich vor allen in ihrem Funktionsumfang.
  Einfache Suche: Werden Begriffe durch Leerzeichen voneinander getrennt eingegeben, so werden Dokumente gesucht, die einen oder alle Begriffe enthalten. Die Dokumente, die die meisten Suchbegriffe enthalten, werden an erster Stelle in der Trefferliste angezeigt. Ferner gibt es noch den „+“ und den „-“ Operator. Dabei legt „+“ fest, daß dieser Suchbegriff in Dokument enthalten sein muß, während „-“ festlegt, daß Dokumente, die diesen Suchbegriff enthalten, ausgeschlossen sind.
  Erweiterte Suche: Hier kann sich der Anwender der Booleschen-Operatoren wie „AND“, „OR“ und „NOT“ bedienen. Der „NEAR“ Operator sucht Dokumente, in denen sowohl die angegebenen Wörter enthalten sind, als auch zwischen den zwei Wörtern nicht mehr als zehn andere Wörter stehen dürfen. Bei der Interpretation einer Anfrage mit mehreren Operatoren folgt die Software einer Standard-Reihenfolge, aber es ist dennoch sinnvoll, bei komplexen Anfragen Klammern zu verwenden, um die gewünschte Reihenfolge anzugeben.

- Suche nach Schlüsselwörtern
  Um nach Schlüsselwörtern zu suchen, müssen diese nur in Anführungszeichen gesetzt werden. Enthält der Suchbegriff z.B. ein „+“, so ist es am einfachsten, die erweiterte Suche zu verwenden, da dort das „+“ kein Operator ist. (z.B. Suche nach C++)

- Kommentare
  Die Suchmaschine indiziert den Kommentar der HTML-Seiten nicht.

- Kontextsuche
  Die Suchmaschine von digital verfügt über keine Kontextsuche. Es gibt weder ein Synonymwörterbuch, noch kann sich der Anwender eigene Assoziationen definieren.

4.5.6 Suchergebnis

- Geschwindigkeit
  Auf Grund der hohen Komplexität einer Geschwindigkeitsmessung können hier keine Angaben gemacht werden.

- Anzahl der Treffer
  Die Anzahl der Treffer, die auf einer Seite angezeigt werden, kann der Administrator verändern.

- Art der angezeigten Informationen
  Der Anwender kann bei seiner Suche angeben, wie detailliert die Trefferliste sein soll.
  Standardformat: Es werden die ersten Zeilen des Dokuments, die Größe der Datei und das Erstellungsdatum angezeigt. Der Titel und die URL sind ferner mit einem Link zu dem entsprechenden Dokument versehen.
  Kompaktes Format: Für diese Darstellung wird immer nur eine Zeile für einen Treffer verwendet. Die ersten Wörter des Dokuments und das Erstellungsdatum werden angezeigt. Der Titel stellt einen Link zu dem Dokument her. Zählung: Hier wird nur die Gesamtzahl der Treffer angezeigt, ohne weitere Information. Der Administrator kann diese Voreinstellungen anpassen und das Format der Trefferlisten ändern. Dazu sind eine Reihe von internen Variablen vorhanden.

- Ranking
  Bei der einfachen Suche verwendet die Suchmaschine feste Ranking-Kriterien. Wird ein Wort am Anfang eines Dokuments oder im Titel gefunden, hat es eine hohe Priorität. Ein Wort, das selten im Index vorhanden ist, hat eine größere Bedeutung als Wörter, die häufig vertreten sind. Werden mehrere Suchbegriffe eingegeben, so hat das Dokument die größte Priorität, in dem die meisten der Begriffe enthalten sind.

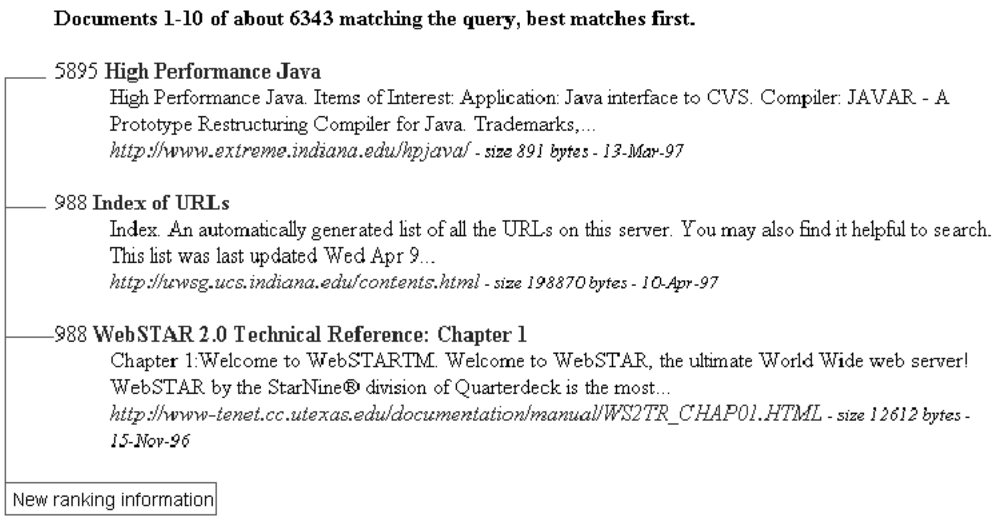

Abbildung 4.3: Ranking bei Alta Vista [digital 97]

  Bei der erweiterten Suche werden die Treffer unsortiert ausgegeben, wenn der Anwender keine Ranking-Regeln angibt. Um Ergebnisse nach einer Rangfolge zu erhalten, kann man ein oder mehrere Wörter in das Ranking-Feld eintragen. Dokumente, die diese Wörter enthalten, werden dann vor den übrigen angezeigt.

- Markierung des Suchbegriffes
  Eine Markierung des Suchbegriffes in den gefundenen Dokumenten ist nicht vorgesehen.

4.5.7 Dokumentenmanagement

- Unterstützte Dokumentarten
  Alta Vista unterstützt eine Vielzahl von verschiedenen Dokumentarten. Der Administrator legt bei der Erstellung des Indexes fest, welche der möglichen Dokumentarten auch indiziert werden sollen. Die wichtigsten sind: TXT, HTML, DOC, RTF, PS, EPS, PDF. Im Test war das Indizieren von PDF-Dokumenten noch nicht möglich. Dies wird aber laut Herstellerangaben bei der nächsten Version beseitigt.

- Anzeige der Dokumente
  Die Dokumente werden entsprechend den Einstellungen im WWW-Browser angezeigt. Gegebenenfalls ist dazu ein entsprechendes Plug-In notwendig.

4.5.8 Agent

- Agent als Erweiterung
  Digital hat derzeit keinen Agenten in seiner Produktpalette, plant aber für die Zukunft eine solche Erweiterung.

4.5.9 Schnittstellen

- Datenbanken
  Es gibt keine Möglichkeit, eine Datenbank in die Informations-Suche zu integrieren.

- Programme
  Die Suchmaschine kann nicht aus anderen Programmen heraus aufgerufen werden. Die erhaltene Treffer-Liste kann nur als HTML-Dokument abgespeichert und weiterverarbeitet werden. Eine umfassende Integration in eine entsprechende Programm-Umgebung ist daher nicht möglich.

4.5.10 Administration

- Installation
  Das Programm läßt sich einfach installieren. Dazu muß nur eine Setup-Prozedur aufgerufen werden. Es werden folgende Informationen erfragt: Der Pfad, unter dem die Software installiert werden soll, und die Kennung und das Paßwort für den Administrator. Am Ende der Installation startet auf Wunsch ein Programm mit der Konfiguration. Um die Software notfalls wieder zu entfernen, wird ein Uninstall-Programm mitgeliefert.

- Installationsmedium
  Das Programm befindet sich auf einer CD-ROM.

- Konfiguration
  Die Konfiguration findet mit einem WWW-Browser statt. Damit kann der Administrator den Zustand der Suchmaschine überwachen und den Index neu aufbauen oder verändern.

- Log-Files
  Das Programm verwendet insgesamt sechs verschiedene Log-Files. In die entsprechenden Dateien werden sämtliche Fehlermeldungen der Suchmaschine, alle angeforderten Informationen, Probleme bei der Indizierung und der Konvertierung geschrieben. Unter WINDOWS NT steht ein Programm Event Log zur Verfügung, mit dessen Hilfe der Administrator die Ereignisse beobachten kann.

- Integriertes Management
  Eine Einbettung in eine Management-Plattform ist nicht möglich.

- Überwachung der Indizierung
  Der Administrator kann eine Statistik erhalten, in der die Gesamtzahl der indizierten Dokumente ausgegeben wird. Ferner kann er die Anzahl der WWW-Server abfragen, die durchsucht wurden sind. Durch die Abfrage der Konverter-Statistik kann der Administrator in Erfahrung bringen, wieviele Dokumente von welcher Dateiart indiziert wurden.

4.5.11 Verhalten in kritischen Situationen

- Abfrage des Zustandes
  Bei Problemen kann der Administrator den Zustand abfragen. Er bekommt dann die Meldung, ob der Roboter noch aktiv ist, oder nicht.

- Benutzermeldungen in kritischen Situationen
  Können Anfragen wegen Überlastung der Suchmaschine nicht erledigt werden, so erhalten die Anwender eine Mitteilung.

- Verhalten bei Systemausfall
  Wird der Index durch einen Systemausfall zerstört, so muß der Index neu aufgebaut werden. Es findet keine Reorganisation statt.

- Fehlermeldungen
  Die Anwender, die eine Suche starten wollen, erhalten eine kurze Information, wenn der Index derzeit nicht verfügbar ist.

4.5.12 Systemanforderungen

- Plattform
  Derzeit ist die Software nur für WINDOWS NT verfügbar. Digital plant jedoch die Portierung auf Solaris.

- Benötigte Rechenleistung
  Als Minderst-Anforderung gibt der Hersteller an: Intel Pentium mit 133 MHz Prozessor oder ein Alpha System 21064 mit 200 MHz.

- Benötigter Speicherbedarf (Hauptspeicher, Festplatte)
  In Abhängigkeit der Größe des Intranets gibt der Hersteller folgende Empfehlung: [digital 97]

Intranet Size	Number of Users	Number of Pages	RAM	Disk Memory
Small	1000	10.000	64 MB	1 GB
Medium	25.000	300.000	128 MB	2 GB
Large	50.000	500.000	256 MB	5 GB

- Unterstützte Browser
  Unterstützt werden die WWW-Browser Microsoft Internet Explorer ab Version 2.0 und Netscape Navigator ab Version 2.0.

- Unterstützte WWW-Server
  Da bei diesem Produkt ein eigener WWW-Server enthalten ist, wird kein anderer WWW-Server benötigt.

4.5.13 Dokumentation

- Handbuch
  Ein Handbuch in gedruckter Form liegt derzeit noch nicht vor, ist aber vorgesehen.

- Online-Handbuch
  Für den Administrator steht ein umfassendes Online-Handbuch in HTML zur Verfügung. Es gliedert sich in mehrere Dokumente: Allgemeine Einführung, Erste Hilfe für die Installation, Hinweise zur Erstellung eines Index und der Konfiguration, Problembewältigung, Änderungen am User-Interface.

- Sprache
  Im Gegensatz zur Online-Hilfe ist das Handbuch nur in englischer Sprache verfügbar.

- Fortbildungskurse
  Diese müßten bei Bedarf zusätzlich vereinbart werden.

4.5.14 Support

- Hotline
  Digital bietet eine Hotline zur Behebung von Problemen an. Über deren Qualität läßt sich bisher keine Aussage treffen, da sich keine Notwendigkeit für einen Anruf ergab.

- Update
  Wenn neue Versionen des Produktes erscheinen, kann ein Update erworben werden.

- Patches
  Es ist anzunehmen, daß fehlerhafte Versionen durch die Bereitstellung von Patches berichtigt werden.

4.5.15 Preis

Preis für Alta Vista Search Private eXtensions (Stand: Juni 1997)
DM 25.000 DM

4.6 Fazit

Beide Suchmaschinen erfüllen grundsätzlich die Anforderungen an eine Suchmaschine für das Intranet der BMW AG.

Das Produkt von Verity ist mit seiner Kontextsuche das mächtigere der beiden, wobei diese Funktionen einen erheblichen Administrations- und Schulungsaufwand erwarten lassen.

Besonders positiv fiel bei der Suchmaschine von Verity auf, daß die Suche auf bestimmte Kollektionen beschränkt werden kann. Bei Alta Vista kann dies jedoch durch ein geeignetes Java-Skript nachgebildet werden. Ein Beispiel wurde der BMW AG von einem Mitarbeiter von digital zur Verfügung gestellt.

Ausschlaggebend für die Empfehlung für die Suchmaschine Alta Vista sind vor allem auch die deutlich bessere Darstellung der Dokumente im Vergleich zu Verity.
Verity hat teilweise Probleme mit der Darstellung von Umlauten und HTML-Seiten, die von HTML-Editoren erstellt wurden. Die Dokumente werden nicht immer so angezeigt, wie sie erstellt wurden. Bei der Suchmaschine Alta Vista werden die Dokumente nicht konvertiert und erscheinen daher im Original.

Abzuwarten bleibt freilich noch, ob alle Probleme, die erkannt wurden, auch behoben werden.

Kapitel 5

Installation und Test

5.1 Search '97 von Verity

5.1.1 Installation

Die erste Testinstallation wurde auf einer Sparc II mit Solaris als Betriebssystem durchgeführt. Als Installationsmedium wurde eine CD geliefert. Unter UNIX erfolgt die Installation rein textuell, so daß diese auch von einem anderen Rechner aus durchgeführt werden kann. Während der Installation werden Daten, wie Betriebssystem, Servername, verwendeter WWW-Server, ... erfragt. Es kann bei der Schnittstelle zwischen CGI und NSAPI gewählt werden. Derzeit wird NSAPI allerdings nur für den Administrator unterstützt und für den Benutzer wird immer CGI verwendet. Bei BMW ist derzeit der Netscape Enterprise Server im Einsatz. Weiter werden die Server von NCSA und Apache unterstützt. Leider kann ein Eingabefehler nicht mehr korrigiert werden, wenn er einmal mit „Return“ bestätigt wurde. Es ist auch nirgends beschrieben, wie man die entsprechenden Konfigurationsdateien von Hand ändern kann. Daher hilft hier nur eine Neuinstallation. Dabei können Probleme auftauchen. Es wird nämlich keine völlständige Neuinstallation durchgeführt. Manche Informationen werden aus den Konfigurationsdateien gelesen, die bei der Erstinstallation angelegt wurden. Sind darunter Informationen, die bei der Erstinstallation falsch eingegeben wurden, so muß zusätzlich das Installationsverzeichnis gelöscht werden. Wurde die Installation erfolgreich durchgeführt, wird der WWW-Server neu gestartet. Speicherplatz der Suchmaschine ca. 65 MB. Installationsdauer: ca. 20 Minuten (falls keine Komplikationen auftreten).

5.1.2 Konfiguration

Die Konfiguration wird mit Hilfe eines WWW-Browsers durchgeführt. Im Gegensatz zur Installation steht hier eine ausführliche Online-Hilfe zur Verfügung.

5.2 Alta Vista von digital

5.2.1 Installation

Die Suchmaschine wurde auf einem Alpha-Rechner von digital mit WINDOWS NT als Betriebssystem installiert. Für die Testinstallation wurde der Rechner von digital zur Verfügung gestellt. Die Installation von Alta Vista Search Intranet eXtention 97 Beta for WINDOWS NT wurde von Mitarbeitern der Firma digital durchgeführt. Diese verlief wie unter WINDOWS NT üblich graphisch unterstützt. Der Administrator wird zur Eingabe eines Administrator-Paßworts aufgefordert. So wird ein unberechtigter Zugriff auf die Konfiguration verhindert.

5.2.2 Konfiguration

Die Konfiguration findet mit einem WWW-Browser statt. Hierbei kann auch eine Online-Hilfe verwendet werden. Ferner stehen unter WINDOWS NT noch zwei Programme zur Verfügung, mit denen die Ereignisse bzw. die Fehler der Suchmaschine beobachtet werden können.

Abbildung 5.1: Server Status bei Alta Vista

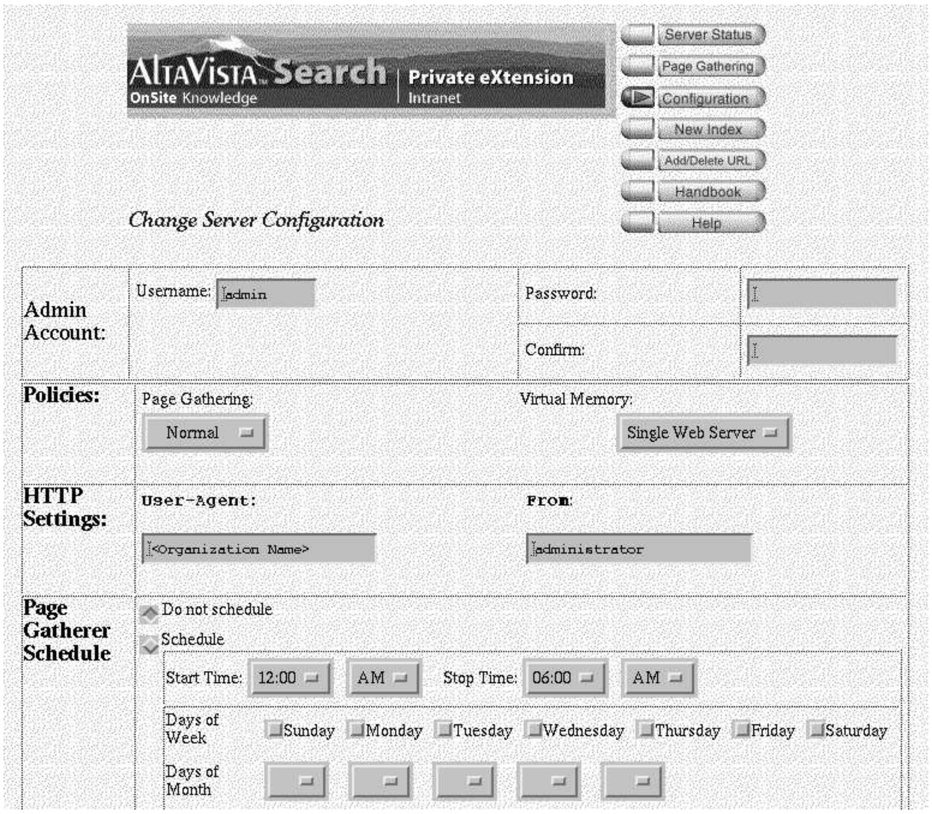

Abbildung 5.2: Konfiguration bei Alta Vista

5.3 Ergebnis der Tests

Im Test wurden die Standard-Anfrage-Seiten verwendet. Ein Administrator kann diese Seiten jedoch leicht ändern und den eigenen Wünschen anpassen. Dies wurde in den Tests aber nicht gemacht.

Alle WWW-Server am Standort München (ohne Werke) wurden indiziert. (ca. 30)

Die Anwender begrüßten eine Einschränkung auf einen bestimmten Such-Bereich, wie dies bei Search '97 von Verity realisiert ist. Es ist deshalb anzuraten, dies auf jeden Fall bei Alta Vista nachzubilden.

Die Indizierung bei Verity verlief relativ kurz (ca. 30 Minuten), aber dafür kann der Rechner in dieser Zeit nicht mehr anderweitig verwendet werden. Bei Alta Vista findet eine ständige Indizierung statt. Hier war kein Anstieg der Netzlast erkennbar.

Beide Suchmaschinen waren im Test sehr schnell. Die Anwender hatten fast keine Wartezeit.

Kapitel 6

Ausblick

Kommerzielle Suchmaschinen sind erst seit kurzem verfügbar und stehen deshalb am Anfang ihrer Entwicklung. Dabei sind unterschiedliche Weiterentwicklungen denkbar.

Denn die Firmen, die Suchmaschinen in ihren Intranets einsetzen wollen, können sehr unterschiedliche Bedürfnisse haben. Es gibt - in der Regel kleinere und mittlere - Firmen, die sehr stark branchenspezifisch spezialisiert sind. Sie haben völlig andere Schwerpunkte bei der Informationsbeschaffung wie beispielsweise große Konzerne, die eine breite Produktpalette herstellen.

6.1 Branchenspezifische Suchmaschinen für kleine Intranets

Um eine hohe Effizienz zu erreichen, könnte man sich hier gut eine katalogbasierte Suchmaschine vorstellen. Mit der Suchmaschine wird bereits ein branchenspezifischer Katalog mitgeliefert und das Unternehmen, das die Suchmaschine einsetzt, verpflichtet seine Mitarbeiter, jedes in Frage kommende Dokument nach diesem Katalog zu klassifizieren und entsprechend einzubinden. Da die Thematik im allgemeinen nicht zu umfangreich sein wird, werden die Mitarbeiter über das nötige Wissen verfügen, um die Klassifikation durchzuführen.

Dabei tritt natürlich das Problem auf, daß gerade in sehr innovationsfreudigen Branchen der Katalog immer wieder den neuen Entwicklungen angepaßt werden muß. Ist das einzelne Unternehmen selbst, dazu nicht in der Lage, so muß die Suchmaschine vom Hersteller mit einem aktualisierten Katalog als Update angeboten werden. Mit ihm muß dann eine Reorganisation des bereits existierenden Katalogs durchführbar sein.

In der Zukunft sollte es auch möglich sein, Programme zu entwickeln, die selbstständig eine Klassifikation von Dokumenten durchführen. In diesem Fall wird dann der Katalog automatisch generiert.

6.2 Thematische Suchmaschinen mit Datenbankanbindug

Für Juristen und Mediziner gibt es z.B. bereits große kommerzielle Datenbanken, die ständig aktualisiert werden. Es gibt inzwischen auch in Deutschland große Anwaltskanzleien, die geographisch getrennte Niederlassungen haben, deren Mitarbeiter die Informationen ihrer vielen Kollegen, aber auch das in den Datenbanken gespeicherte Wissen, möglichst direkt abrufbereit haben sollten.

Das spezifische „Kanzleiwissen“ wird problemlos im hauseigenen Intranet gespeichert werden können. Dazu muß dann aber eine schnelle und effektive Einbettung der entsprechenden Datenbank möglich sein. Für den Nutzer bräuchte gar nicht ersichtlich sein, ob die Suchmaschine im eigenen Intranet oder in einer Datenbank sucht. Es ist nur eine Such-Plattform nötig, obwohl in verschiedenen Informationsquellen gesucht wird. Analoges gilt z.B. für die Mitarbeiter an Kliniken oder medizinischen Forschungseinrichtungen.

6.3 Globale Suchmaschinen für große Intranets

Bei ständig steigenden Datenmengen wird es immer schwieriger, die gesuchte Information auch zu finden. Daher müssen neue Konzepte gefunden werden, um nach Informationen zu suchen.

Neue Such-Werkzeuge müssen die semantische Bedeutung eines Textes erkennen können. Nur so bekommt der Anwender die Information, die er wirklich braucht. Dazu müssen auch Verfahren aus der Künstlichen Intelligenz angewendet werden. Erst wenn die Infrastruktur der Netze selbst ein gewisses Maß an Intelligenz hat und den Sinn einer Anfrage zumindest in Ansätzen verstehen kann, werden die Informationen gefunden, die der Anwender sucht. [Kyas 97]

Gerade für Bilder und Musik sind solche Verfahren Grundvoraussetzung für ein sinnvolles Suchen.

Der Anwender soll keine kryptischen Befehle auswendig lernen müssen, sondern mit eigenen Worten die Anfrage möglichst natürlich-sprachlich formulieren. In einem Dialog mit dem Programm ist die Suche so einzugrenzen und zu präzisieren, daß die richtige Information gefunden werden kann. Ist die Suchanfrage zu ungenau formuliert, erhält der Anwender Verbesserungsvorschläge.

Die Zukunft liegt nicht in einer gigantischen Suchmaschine mit einem großen Index, sondern in lokal-verteilten nach Themen sortierten Indizes. Für diesen Zweck wurde das Programm „Harvest“ [Hardy 96] entwickelt. Aber auch bei dezentralen Lösungen werden nur Buchstaben verglichen und nicht Informationen.

Es gibt schon erste Ansätze diese Probleme zu lösen. Mit einem phonetisch motivierten Suchverfahren werden Wörter gefunden, die ähnlich klingen. Dies hilft vor allem, wenn die Schreibweise eines Wortes nicht genau bekannt ist, bzw. im veröffentlichten Dokument falsch ist.

Eine Suchmaschine sollte „wissen“, wonach der Anwender sucht, und sie sollte die Bedeutung der veröffentlichten Dokumente „kennen“. Auf eine derartige „intelligente“ Suchmaschine wird man wohl noch eine Weile warten müssen.

Literaturverzeichnis

[Bauer 91] Friedrich L. Bauer und Gerhard Goos, *Informatik 1*, Springer-Verlag, 1991.

[Broy 92] Manfred Broy, *Informatik Eine grunglegende Einführung Teil 1*, Springer-Verlag, 1992.

[c't 4/97] Reinhard Rapp, „Text-Detektor“, *c't*, 4, April 1997.

[digital 97] digital, *Alta Vista Search 97 Intranet Handbook: Product Overview*, 1997.

[Duden 93] Lektorat des BI-Wissenschaftsverlag (Hrsg.), *Duden Informatik*, Dudenverlag, 1993.

[Hardy 96] Darren R. Hardy, Michael F. Schwartz und Duane Wessels, „Harvest User's Manual“, Technischer Bericht, Department of Computer Science University of Colorado, 1996, http://harvest.transarc.com/afs/transarc.com/public/trg/Harvest/doc.html

[Hegering 93] Heinz-Gerd Hegering und Sebastian Abeck, *Integriertes Netz- und Systemmanagement*, Addison-Wesley, 1993.

[Hunt 95] Craig Hunt, *TCP/IP*, O'Reilly/International Thomson Verlag, 1995.

[Kirchgesser 97] Ulrike Kirchgesser, „Suchen im Internet“, Technischer Bericht, Leibniz-Rechenzentrum München, 1997, http://www.lrz.de/suchen/

[Koch 96] Traugott Koch, „Suchmaschinen im Internet“, Technischer Bericht, 1996, http://www.ub2.lu.se/tk/demos/DO9603-manus.html

[Koster 97] Martijn Koster, „The Web Robots Page“, Technischer Bericht, 1997, http://info.webcrawler.com/mak/projects/robots/

[Kyas 97] Othmar Kyas, *Corporate Intranets*, International Thomson Publishing, 1997.

[Maurer 96] Rainer Maurer, *HTML- und CGI-Programmierung*, dpunkt, 1996.

[PC-Welt 4/97] Wolfgang Sommergut, „Suchen im Internet“, *PC-Welt*, 4, April 1997.

[Tanenbaum 97] Andrew S. Tanenbaum, *Computernetzwerke*, Prentice Hall, 1997.

[Verity 96] Verity, *Verity Search'97 Informationsbroschüre*, 1996.

[Weibel 96] Stuart Weibel, *A Proposed Convention for Embedding Metadate in HTML*, 1996, http://www.w3.org/pub/WWW/Search/9605-Indexing-Workshop/ReportOutcomes/S6Group2.html

Abbildungsverzeichnis