BIRKHÄUSER

Biowissenschaftlich recherchieren

Über den Einsatz von Datenbanken und anderen Ressourcen der Bioinformatik

Nicola Gaedeke

Birkhäuser
Basel · Boston · Berlin

Autorin:

Nicola Gaedeke
- BioTools.info -
Neuwerker Weg 4
D-14167 Berlin

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

ISBN 978-3-7643-8525-5 Birkhäuser Verlag AG, Basel – Boston – Berlin

Ein Unternehmen der Fachverlagsgruppe Springer Science+Business Media
Gedruckt auf säurefreiem Papier, hergestellt aus chlorfrei gebleichtem Zellstoff. TCF ∞
Umschlaggestaltung: Alexander Faust, Basel, Schweiz
Printed in Germany

ISBN: 978-3-7643-8525-5 e-ISBN: 978-3-7643-8526-2

9 8 7 6 5 4 3 2 1 www.birkhauser.ch

Inhaltsverzeichnis

Vorwort

Dieses Buch ist ein Leitfaden für die Informationssuche im Bereich der Lebenswissenschaften, mit einem Schwerpunkt auf molekularbiologischen Daten. Es basiert auf einer erprobten Fortbildung zur „Fachkraft für Bioinformatik", die vom Gläsernen Labor in Berlin-Buch mehrmals im Jahr angeboten wird (http://www.glaesernes-labor.de/).

Der Fokus in diesem Buch liegt auf den Datenbanken und Ressourcen des National Center for Biotechnology Information (NCBI). Das hat zwei Gründe. Zum einen sind die Webseiten des NCBI stark frequentiert. So hatte die Homepage im Jahre 2002 allein über 28 Mill. Anfragen von über 240.000 Besuchern täglich. Mit BLAST, dem Basic Local Alignment Search Tool, wurden täglich über 100.000 Sequenzähnlichkeitssuchen durchgeführt. Der zweite Grund ergibt sich aus meinen persönlichen Erfahrungen. In den Jahren 2000 – 2002 habe ich in enger Zusammenarbeit mit dem NCBI einen Kurs für „Bioinformatics Information Specialists" entwickelt, der seitdem einmal im Jahr angeboten wird (http://www.ncbi.nlm.nih.gov/Class/NAWBIS/). Die meisten der Kursteilnehmer arbeiten in einer medizinischen Bibliothek einer US-Amerikanischen Universität und geben dort einen „User Support Service" für bioinformatische Fragestellungen. Meine größten Erfahrungen liegen daher bei den Tools des NCBI. In derselben Zeit habe ich einen Bioinformatics Support Service an der Universität von Utah in „Salt Lake City" angeboten, der sehr gut angenommen wurde. Und obwohl die Seiten des NCBI eine einfache Bedienung suggerieren, zeigten mir die Fragen der Anwender, dass viele Einstellungen, Möglichkeiten und Bedeutungen der Anwendungen nicht bekannt sind und relevante Informationen oft gar nicht gefunden werden. Ich habe in diesem Buch versucht, die Suchmöglichkeiten am NCBI zu erläutern, sowie die Einstiegsseiten für die weniger bekannten Ressourcen aufzuzeigen. Auch die Seiten des NCBI ändern sich. Oft kommen neue Ressourcen hinzu. Der Leser ist nach der Lektüre dieses Buches jedoch mit den Prinzipien der Suchoberflächen am NCBI vertraut und kann neue Ressourcen hoffentlich leichter einordnen. Im Allgemeinen gilt, dass sich aus jeder neuen Methode im Bereich der Lebenswissenschaften, die eine große Menge an Datensätzen produziert, immer neue Datenbanken ergeben werden, die recherchiert werden müssen. Die Datensätze werden komplexer werden, wie wir es z.B. schon aus der Genexpressionsanalyse kennen. Die Suchoberflächen sollen jedoch – so will es der Anwender – so einfach wie möglich sein. Wie könnten also die Recherchemöglichkeiten nach komplexen Daten aussehen und wie kann das

Ergebnis so zuverlässig ausfallen wie bei einer Sequenzrecherche? Und wie könnte eine Informationssuche nach komplexen Zusammenhängen, wie es die Systembiologie erfordert, durchgeführt werden? Vielleicht sind diese Fragen nur eine technische Herausforderung (an die programmierenden Bioinformatiker), vielleicht bedarf es aber auch in Zukunft aufmerksamer Anwender, die das Ziel der Informationssuche nicht aus dem Auge verlieren und geeignete Suchstrategien entwickeln können. Dieses Buch richtet sich an alle, die zu aufmerksamen Anwendern werden oder ihre Kenntnisse über Suchstrategien und Ressourcen in der Bioinformatik auffrischen und erweitern wollen. Ein sicherer Umgang mit dem Internet wird für die Übungen in diesem Buch vorausgesetzt.

An dieser Stelle möchte ich mich bei allen bisherigen Kursteilnehmern bedanken, besonders aber bei Monika Jung und Sunita Singh, die maßgeblich zum Gelingen dieses Buches beigetragen haben. Ein ebenfalls großer Dank gilt Herrn Dr. Ulrich Scheller, dem Leiter des Gläsernen Labors in Berlin-Buch, der meine Idee für eine Fortbildung für Laborpersonal zur „Fachkraft für Bioinformatik“ aufgegriffen hat und die Voraussetzung für eine zertifizierte Weiterbildung (TÜV-Akademie) geschaffen hat.

Das Glossar am Ende des Buches ist sehr ausführlich, da mir persönlich viele Glossare zu klein sind und gerade im Bereich der Bioinformatik viele Abkürzungen für IT-Begriffe, für molekulare Daten, aber auch für Institute wie Allgemeinwissen behandelt werden.

Ich bemühe mich, die Weblinks zu Datenbanken und Ressourcen der Bioinformatik auf meiner Webseite (http://www.biotools.info) aktuell zu halten. Über Vorschläge des Lesers zur Vervollständigung und Aktualisierung dieser Seiten würde ich mich sehr freuen.

Berlin, im Juli 2007

Nicola Gaedeke
- BioTools.info -

1

Die Informationssuche im World Wide Web (WWW)

Die Informationsflut ist heute größer als jemals zuvor. Die klassischen Informationsquellen werden durch das World Wide Web (WWW) abgelöst. Das Medium „WWW“ unterscheidet sich von den herkömmlichen Informationsquellen dadurch, dass sich der Anwender selbst aktiv auf die Informationssuche begibt. Die Herausforderung, der wir uns bei der Benutzung des WWW stellen müssen, ist daher, die Informationen zu filtern, um das für uns Wichtige und Richtige zu finden. Dieses Buch soll Ihnen helfen, diese Filter zu definieren. So kann die Informationssuche im WWW schon mit ein paar zusätzlichen Gedanken zur Suchstrategie effizienter werden. Oftmals ist eine Suche im WWW gar nicht der richtige Ansatz, da viele Informationen in Datenbanken hinterlegt sind und die Suche daher direkt in der Datenbank erfolgen sollte. Was früher nur in Bibliotheken oder bei Datenbankanbietern über eine Telnet-Verbindung möglich war, wie z. B. die Suche in Medline über Silverplatter, ist heute über eine Datenbankrecherche im Internet möglich. Da die Erstellung einer Suchstrategie für die Suche in einer Datenbank ebenso gilt wie für eine effiziente Suche im WWW, werden in diesem Buch zuerst die Grundlagen einer Suchanfrage vorgestellt.

Funktion des Internets

Das Internet ist ein Zusammenschluss von individuellen regionalen Netzwerken. Über diese Netzwerke, die von Universitäten, Firmen oder Online-Diensten betrieben werden, können verschiedene Dienste aufgerufen werden, die sich durch unterschiedliche Funktionen am Bedarf des Anwenders orientieren. Die vier bekanntesten Dienste des Internets sind hier erwähnt:

(1) Das **World Wide Web** (**WWW**) ist der am häufigsten genutzte Dienst im Internet. Die Anwendung des Internets ist hier durch eine grafische Aufbereitung für den Nutzer erleichtert. Die Dokumente/Webseiten stehen im HTML-Format zur Verfügung.

   a. **HTML** (**Hypertext Markup Language**) ist eine kodierte Sprache zur Darstellung von Webseiten über einen Internet-Browser. Weitere Kodierungssprachen oder zusätzliche Software können auf HTML-Seiten eingebunden werden, wie z.B. CGI (Common Gateway Interface für Animationen) und Java/ Java-Script (für PopUps).
   b. **HTTP** (**Hypertext Transportation Protocol**) ist das Protokoll zur Abfrage für HTML-Dokumente auf der Basis von ASCII Sequenzen.

(2) **FTP** (**File Transfer Protocol**) dient dem Datenaustausch zwischen verschiedenen Rechnern.
(3) **E-Mail/Mailinglisten**
(4) **Newsgroups**

Um eine Informationsübertragung im Internet zu gewährleisten, benötigen Sender und Empfänger eine einheitliche Sprache, ein so genanntes Protokoll. Die grundlegenden Protokolle des Internets sind **TCP** und **IP**. Sie werden meistens gemeinsam genannt, da sie sich ergänzen und somit eine Einheit bilden. 30% der verschickten Daten sind reine Protokolldaten. Ihre Funktionen sind folgende:

- Das **TCP** (**Transmission Control Protocol**) teilt die Daten in ungefähr gleich große Blöcke auf und übergibt sie dem IP zur Übertragung. Zusätzlich überprüft TCP die Korrektheit der Übertragung über eine Prüfsumme, nachdem die Daten beim Empfänger wieder endgültig zusammengesetzt worden sind.
- Das **IP** (**Internet Protocol**) ist dafür zuständig, dass die Daten über verschiedene Schaltstellen und Router gelenkt werden und trotzdem am gewünschten Ziel ankommen.

Struktur eines HTML-Dokumentes

In einem HTML-Dokument ist die Information, die auf der Webseite erscheinen soll, mit sogenannten „tags“ versehen. Jeder „tag“ muss geöffnet und wieder geschlossen werden. So bedeutet <html> ... </html>, dass alle Informationen, die zwischen den „tags“ <html> und </html> kodiert sind, über das Internetprotokoll http lesbar sind. Dieses Dokument muss in dem Format *.html zur Verfügung stehen, um vom Web-Browser dargestellt werden zu können. Die Webseite besteht aus einem Kopf (head) und aus einem Körper (body). Der Kopf wird im Gegensatz zum Körper nicht auf der Webseite dargestellt. Weitere Kodierungen zur Gliederung des Inhaltes sind z.B. die Angabe des Titels (<title> ... </title>), eines Zeilenumbruchs (
 = break) oder „tags“ zur Darstellung von Tabellen (<table>), Listen (<ul> = unordered list, <ol> = ordered list) und Paragrafen (<p>). Freie und käufliche HTML-Editoren helfen bei der Erstellung einer Webseite. Auch das MS-Office Programm WORD bietet die Option, ein Dokument in das HTML-Format umzuwandeln und zu speichern („Datei“ – „Als Webseite speichern“). Ein einfacher Webeditor ist „Composer“ von Netscape. Das Programm kann von der Netscape-Menüleiste „Fenster“ aus gestartet werden. Anregungen zur Kodierung einer Webseite bietet ein Seitenquelltext, der für jede angezeigte Seite aufgerufen

werden kann („Ansicht“ – „Seitenquelltext“). Weitere Tipps zur Herstellung von Internetseiten gibt es z.B. unter http://www.self-html.de.

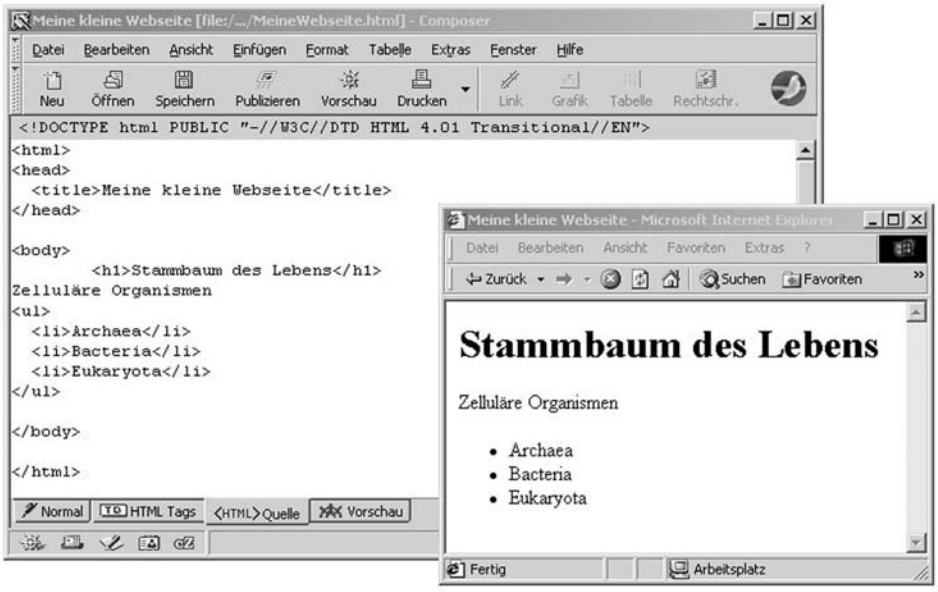

Abbildung 1.1:
Diese Webseite wurde in einem Texteditor-Programm geschrieben, als „Meine_kleine_Webseite.html“ gespeichert und im Internet-Browser Mozilla-Firefox aufgerufen.

Ein Internet-Browser (Web-Browser) bietet die grafische Oberfläche zur Darstellung von Webseiten. Unter Angabe des „Unique Resource Locator“ (URL) bzw. unter der Webadresse werden die Webseiten aufgerufen. Die drei am häufigsten verwendeten Web-Browser sind der Internet-Explorer (IE), Mozilla-Produkte wie Firefox oder SeaMonkey und Netscape. Zur effektiven Benutzung des Internets lohnt es sich, sich mit der Menüleiste seines bevorzugten Browsers intensiv vertraut zu machen. So sind viele Anwender mit der Verwaltung und Organisation von Lesezeichen noch vertraut, in der Verwaltung von Kennwörtern, Cookies und Mail-Einstellungen jedoch weniger geübt. Unter dem Menüpunkt „Hilfe“ kann der Anwender sowohl online als auch offline mehr Informationen zur Benutzung eines Web-Browsers aufrufen.

Beispiele für freie Web-Browser sind

- Microsoft Internet Explorer 7 (http://www.microsoft.de/)
- Mozilla Produkte wie Firefox oder SeaMonkey (http://www.mozilla.org/products/)
- Netscape 7.1 http://www.netscape.de/

Suchen und finden

Eine Informationssuche, sei es im Internet oder in einer Datenbank, liefert fast immer Ergebnisse. Oft führt die Recherche sogar zu einer sehr hohen Anzahl von Treffern. Der Anwender erachtet jeden Treffer als „richtig" und relevant und fängt an, sich von Treffer zu Treffer weiterzuhangeln, ohne auch nur die Möglichkeit einer fokussierten Suchanfrage in Erwägung zu ziehen. Die hier vorgestellten Strategien zur Informationssuche unterscheiden sich vom sogenannten „Browsen" oder „Stöbern" im Internet dadurch, dass es sich um eine zielgerichtete Informationssuche handelt. Das Internet zeichnet sich jedoch durch Besonderheiten aus, die einer starken Kontrolle der gefundenen Information sowie einer genauen Dokumentation über die Auffindungsparameter (Ort/Zeit) bedürfen. Diese Besonderheiten des Internets sind:

(1) **Fehlende Organisation** – Niemand koordiniert oder kontrolliert, wer was wo und wie veröffentlicht.
(2) **Fehlende Strukturierung** – Eine Veröffentlichung im Internet unterliegt keinerlei Standards. Es kann sich um ein Buch, eine Datenbank oder nur eine kurze Notiz handeln. Niemand muss Inhaltsangaben, Sachregister oder Stichwortkataloge erstellen.
(3) **Beliebigkeit** – Nur auf Initiative von Einzelpersonen oder einer Institution kommen Informationen in das Netz. Es gibt keine Pflichtexemplar-Regelung.
(4) **Dynamik** – Täglich kommen neue Einträge hinzu, andere verschwinden und wieder andere werden verlegt oder geändert.

Suchhilfen im Internet

Längst ist es unmöglich geworden, sich die für den Eigenbedarf nützlichen URLs (Internetadressen) zu merken, abzuspeichern oder aus gedruckter Literatur herauszusuchen. Zur Informationsbeschaffung bietet das Internet Suchhilfen an, die je nach Anbieter unterschiedliche Aspekte und Webinhalte berücksichtigen und eigene Suchtreffersortierungen vornehmen.

(1) **Internet-Suchmaschinen** wie z.B. Google oder AltaVista sind roboterbasierte Programme (sog. Spider oder Robots). Sie suchen nach Webseiten, um sie zu indexieren. Dabei werden nahezu alle Wörter auf einer Webseite in den Suchindex eingetragen. Eine Suchmaschine zu benutzen, ist günstig, wenn man konkret weiß, was man sucht (bestimmte Firmen, Namen, Projekte, Programme etc.).
(2) **Internet – Thematische Verzeichnisse** wie z.B. Web.de oder Yahoo!, sind intellektuell bearbeitete Register von Webseiten. Die Webseiten sind thematisch und oft hierarchisch sortiert. Thematische Verzeichnisse dienen als Einstieg in eine Internetsuche, wenn man noch keinen speziellen Suchbegriff hat, oder sich einen Überblick über die gesuchte Thematik verschaffen will. Ein Verzeichnis dient auch dem Einstieg ins „Deep Web". Durch ein Verzeichnis kann man durchklicken (browsen).

(3) **Hybride Suchhilfen** wie z.B. Web.de oder Yahoo! versuchen die Vorzüge der Suchmaschinen und der thematischen Verzeichnisse miteinander zu vereinen.
(4) **Metasuchmaschinen** wie z.B. MetaGer oder MetaCrawler ermöglichen eine Internetrecherche unter gleichzeitiger Verwendung mehrerer Suchmaschinen.
(5) **Suchhilfen auf Servern mit Datenbankanbindungen** wie z.B. Bestandskataloge von Bibliotheken (Web-OPACS) dienen dem Einstieg für eine Recherche in einer dieser Datenbanken oder in anderen Verzeichnissen.

Vorbereitung der Suche

Im Vordergrund der Informationssuche steht die Frage nach dem Suchort. Nicht immer ist eine Internetrecherche für die Lösung eines Problems geeignet. Eventuell befindet sich die gesuchte Information in einem (Fach-)Buch, in öffentlichen Registern oder Listen oder in einer Datenbank. Anhand folgender Fragen soll dargestellt werden, wie wichtig der richtige Suchort für das Rechercheergebnis ist. Wo z.B. könnte man suchen nach:

- Literatur zu einem medizinischen Thema?
- Literatur zu einer wissenschaftlichen Untersuchung?
- Nachrichten aus der Rubrik „Wissen“ aus einer Tageszeitung von vor zwei Wochen?
- Information zu einer Proteinsequenz?
- Firmeninformationen
  - *Produktinformationen*?
  - *Wirtschaftsinformationen (Portfolio, Startkapital, Kapitalgeber etc.)?*
- Patentinformationen?
- Anleitungen für wissenschaftliche Experimente?
- Elektronenmikroskopische Aufnahmen von Viren?
- Informationen über Medikamente und ihre Nebenwirkungen?

Wo ist die Information, die ich suche?

Wenn der Anwender eine Informationssuche im Internet durchführen will, muss er überlegen, wo und wie die Information untergebracht sein könnte.

- Im Internet
- In einer **Datenbank**

In einer Datenbank werden Informationen und Fakten gesammelt und zusammengestellt, die aus der Sicht des Datenbankherstellers zusammengehören, z.B. personenbezogene Daten in einer Personaldatenbank oder Sequenzdaten in einer Sequenzdatenbank. Es gibt hierarchische, relationale, multidimensionale und objektorientierte Datenbanken. Der Zugang zu einer Datenbank kann, muss aber nicht, über das Internet erfolgen. Für eine Datenbank gibt es Suchmasken, über die die Suchanfrage an die Datenbank gestellt werden muss.

Suchmaschinen können bisher nur die Startseiten von Datenbanken finden, nicht aber eine Suche in der Datenbank selbst ersetzen. Neue Entwicklungen zeigen jedoch, dass auch Datenbankinhalte über eine Internet-Suchmaschine erschlossen werden können. Ein Beispiel hierfür ist die Suchmaschine von Google für wissenschaftliche Literatur aus kostenlos zugänglichen Literaturdatenbanken (Google-Scholar).

Im **Deep Web**

Das Deep Web (auch Hidden Web oder Invisible Web) bezeichnet den Teil des Internets, der bei einer Internetrecherche nicht über normale Suchmaschinen auffindbar ist. Im Gegensatz zum Deep Web werden die über Suchmaschinen zugänglichen Webseiten „Visible Web" (Sichtbares Web) oder „Surface Web" (Oberflächenweb) genannt. Die Inhalte im Deep Web können grob unterteilt werden in „Inhalte, die nicht frei zugänglich sind" und „Inhalte, die nicht von Suchmaschinen indexiert werden". Die Größe des Deep Web kann nur geschätzt werden – es wird davon ausgegangen, dass es ein Vielfaches des direkt zugänglichen Webs umfasst.

Zum Deep Web gehören die von Suchmaschinen absichtlich vernachlässigten Daten, Webseiten, die indexiert werden könnten, aber auf Grund von Zugangsbeschränkungen des Webmasters nicht indexiert werden (z.B. Seiten des Intranets), Webseiten, die indexiert werden könnten, die jedoch nur nach Anerkennung einer Nutzungsbedingung zugänglich sind (kostenlos oder kostenpflichtig, z.B. webbasierte Fachdatenbanken), und ganz und gar unsichtbare Webseiten wie z.B. dynamisch erstellte Webseiten, Seiten mit Dateiformaten, die nicht erfasst werden können (z.B. Flash), komprimierte Daten, Webseiten mit einer Benutzernavigation, die Grafiken oder Scripte benutzen oder Inhalte auf einem FTP-Server.

Recall vs. Precision

Zur Vorbereitung der Suchanfrage stehen neben den Überlegungen zum Suchort auch Überlegungen zur Suchgenauigkeit an: Soll die Recherche alles zum Thema hervorbringen (Vollständigkeit) oder die am meisten relevanten Dokumente (Genauigkeit; s. Abb. 1.2)?

Ermittlung und Sammlung von Wortmaterial zum Problem

Die Schwierigkeit bei der Suche im Internet ist das Fehlen eines kontrollierten Vokabulars. Auch wenn Webseiten zum selben Thema angeboten werden, ist nicht gewährleistet, dass die Webseitenanbieter dieselben Wörter zur Beschreibung einer Problematik verwendet haben. Um eine Suchanfrage so genau wie möglich zu stellen, kann der Anwender mehrere Suchworte zur Thematik logisch miteinander verknüpfen. Hierdurch kann eine Suche sowohl erweitert als auch eingegrenzt werden.

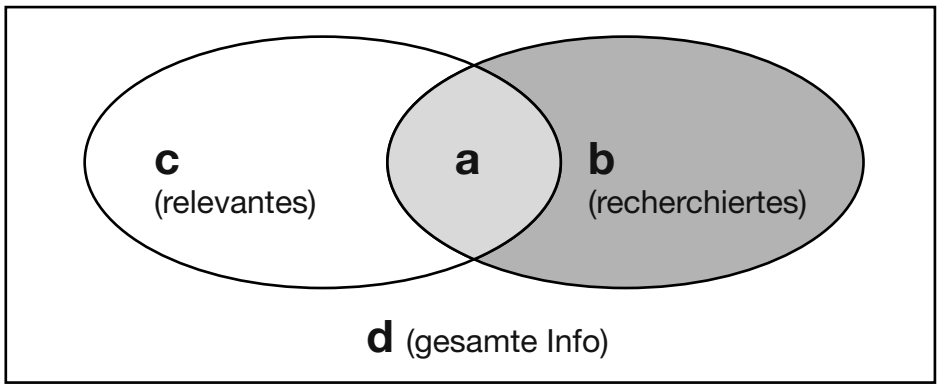

Abbildung 1.2: **Recall vs. Precision**. Das cab-d Modell veranschaulicht den Zusammenhang zwischen der Wiederauffindungsrate und der Vollständigkeitsrate (Recall) sowie zwischen der Trefferquote und der Relevanzquote (Precision). In der Menge der gesamten Informationen (d) sollten alle relevanten Einträge gefunden werden. Die Suchanfrage muss so formuliert werden, dass alles Recherchierte auch das Gesuchte (a) darstellt. In der Grafik wären dann die Schnittmengen c und b deckungsgleich.

Formulierung einer Suchanfrage für Informationen zur Bluterkrankheit (Hämophilie) in Königsfamilien.

(1) Wortsammlung mit Hilfe von Thesauri und Wörterbüchern (alternative Namen, Synonyme etc.) und Gliederung der Thematik
(2) Logische Verknüpfung der Suchworte zu einer Suchwortkette mit Hilfe von Boole'schen Operatoren (AND, OR, NOT; s. u.)

Lösungsansatz:

Es können zwei Themenkomplexe erarbeitet werden.

a) Hämophilie, Bluterkrankheit, Bluter, Haemophilia
b) Königsfamilie, Adel, Adelsfamilie, König, Kaiser

Die Themen müssen in Klammern organisiert und logisch miteinander verknüpft werden. Folgende Suchanfrage kann jetzt formuliert werden:

(Hämophilie **OR** Bluterkrankheit **OR** Bluter **OR** Haemophilia)
AND
(Königsfamilie **OR** Adel **OR** Adelsfamilie **OR** König **OR** Kaiser)

Boole'sche Operatoren AND, OR oder NOT

Boole'sche Operatoren (genannt nach George Boole; engl. Mathematiker um 1850) dienen der logischen Verknüpfung von Suchbegriffen. Die Operatoren werden immer in Großbuchstaben geschrieben. In vielen Suchmaschinen reicht das Einfügen der algebraischen Zeichen „+" für „AND" und „–" für „NOT". Das „OR" wird von einer Suchmaschine entweder in Deutsch („ODER") oder in Englisch („OR") akzeptiert (siehe Hilfsdokumentation der Suchmaschinen). Ein

weiterer Operator ist „NEAR“ für alle Wörter, die in unmittelbarer Nachbarschaft im Text vorhanden sind. Im Folgenden sind die gebräuchlichsten Boole'schen Operatoren erläutert.

AND (+)	Findet Dokumente mit allen angegebenen Wörtern oder Phrasen. Beispiel: **+Hämophilie +Königsfamilie** findet Dokumente, die sowohl das Wort Hämophilie als auch Königsfamilie enthalten.
OR (ODER)	Findet Dokumente, die mindestens eines der gesuchten Wörter oder Phrasen enthalten. Beispiel: **Hämophilie ODER Königsfamilie** findet Dokumente, die entweder das Wort Hämophilie oder Königsfamilie enthalten. Die gefundenen Dokumente können auch beide Begriffe enthalten, müssen es aber nicht.
NOT (–)	Schließt Dokumente aus, die das angegebene Wort oder die Phrase enthalten. Beispiel: **Hämophilie –Königsfamilie** findet alle Dokumente, die das Wort Hämophilie enthalten, nicht aber den Begriff Königsfamilie.

Die Internetrecherche

Suchmaschinen unterscheiden sich nicht nur, wie oben erwähnt, in ihren indexierten Inhalten, sondern auch in der Interpretation der Suchanfrage und der Beurteilung der Treffer. Bei der Benutzung einer Suchmaschine ist es daher wichtig zu wissen, wie die Suchanfrage von der Maschine übersetzt wird, um die gesuchten Informationen zu finden. So wurden in den Anfangszeiten von AltaVista bei einer Suchanfrage mit zwei oder mehr Suchwörtern diese Suchwörter mit ODER verknüpft. Hierdurch wurden bei einer Suche mit AltaVista sehr viel mehr Treffer erzielt, als z.B. mit Google. Erst später wurde den Entwicklern der Suchmaschine bewusst, dass der Anwender bei einer Eingabe von mehreren Wörtern ein AND in die Suchanfrage impliziert. Die Suchmaschine hat sich durch die Änderung der Suchinterpretation zu AND dem allgemeinen Verhalten eines Anwenders angepasst. Der Druck der Anwender auf die Suchmaschinen führt dazu, dass sich die Suchfunktionen immer weiter aneinander angleichen. Viele Suchmaschinen bieten sowohl eine einfache als auch eine erweiterte Suchoberfläche an; oft bleibt es dem Nutzer jedoch verborgen, wie die Suchanfrage an das System gestellt wurde. Hier sollen ein paar Tipps und Beispiele einen Beitrag zu den Vorüberlegungen einer Internetsuche leisten.

Die einfache Suchoberfläche (Simple Search/Anfänger-Suche)

- **Großschreibung/Kleinschreibung**: Was wird gesucht, wenn das Suchwort Großbuchstaben enthält bzw. nur in Kleinbuchstaben geschrieben ist?
- **Umlaute**: Wie geht die Suchmaschine mit den deutschen Umlauten um?
- **Trunkierung**: Kennt die Suchmaschine eine Verkürzung des Wortstamms? Wenn ja, welches Zeichen muss dafür benutzt werden?

- **Singular/Plural**: Wird von der Suchmaschine automatisch nach dem Plural gesucht, wenn nur der Singular angegeben ist? (Im Zweifelsfall beide Formen, z.B. „Elefant“ und „Elefanten“ benutzen)
- **Stoppwörter**: Welche Wörter werden von der Suchmaschine ignoriert (oft werden die Artikel, Präpositionen, „http“ und ähnliche Wörter übergangen)? Über das „+“-Zeichen können diese Wörter oft in die Suche mit einbezogen werden.
- **Suche nach Wortgruppen**: Bestimmte Zeichen dienen als Verbindung von Wortgruppen. Hierzu gehören Bindestriche, Schrägstriche, Anführungszeichen, Gleichheitszeichen und das Apostroph, z.B. Der-alte-Mann-und-das-Meer.
- **Stichwörter**: Sollten sorgfältig gewählt werden, möglichst präzise Angaben machen, z.B. „Dackel“ anstelle von „Hund“.

Erweiterte Suchoberflächen (Advanced (extended) Search / Experten-Suche)
Bei den erweiterten Suchoberflächen werden die Suchwörter logisch miteinander verknüpft, ohne dass eine komplexe Suchanfrage in die Suchmaske eingegeben werden muss. Die Suchmaschine wird die Suchanfrage unter Anwendung von Boole'schen Operatoren in eine logisch verknüpfte Suchwortkette übersetzen. In diesen Oberflächen können oft weitere Eingrenzungen z.B. zum Aktualisierungszeitpunkt der Webseite vorgenommen werden. Eine komplexe Suchanfrage mit einer Suchwortkette, wie sie im Lösungsansatz am Beispiel der Recherche zu Hämophilie in Königsfamilien in diesem Kapitel dargestellt ist, ist oft nicht in den erweiterten Suchoberflächen möglich. Diese Suchanfrage muss in das Suchfeld, einschließlich der gesetzten Klammern, eingegeben werden.

Die Trefferanzeige

Suchmaschinen sortieren die Treffer nach unterschiedlichen Gesichtspunkten. Kriterien für die Relevanzberechnung der Treffer sind u.a.:
- die Anzahl der gefundenen Suchwörter auf der Webseite
- die Position der Wörter auf der Webseite
- die Anzahl der Suchwörter bezogen auf die Länge der Webseite
- nur die Länge der Webseite
- die Häufigkeit des Abrufens von einzelnen Webseiten
- die Position der Datei im Verzeichnisbaum des Servers
- die Anzahl der Links, die auf eine Seite gesetzt wurden (z.B. bei Google)

Der zuerst aufgeführte Treffer einer Suchmaschine ist demnach nicht immer der beste Treffer für den Benutzer, auch wenn er seine Suchanfrage korrekt und vollständig gestellt hat. Für die richtige Interpretation eines Suchergebnisses ist es demnach sehr wichtig zu wissen, wie die jeweils benutzte Suchmaschine die Anfrage verarbeitet bzw. welche Kriterien in das Ranking der Treffer mit einfließen.

Trefferbearbeitung

Die Treffer einer Suchmaschine können in unterschiedlichen Formaten vorliegen. So können von Google die Dateiformate HTML, PDF und PPT schon sehr effizient indexiert werden. Zusätzlich kann jeder Treffer über folgende Optionen bearbeitet werden:

- „Ähnliche Seiten" – initiiert eine Suche nach ähnlichen Webseiten (Google)
- „Archiv-Seiten" – ruft die Seiten aus dem „Cache" auf (Google, Yahoo)
- „Weitere Seiten dieser Webseite" – sucht die Seiten in derselben Domäne (Yahoo)
- „Diese Seite übersetzen" – übersetzt die Seite in eine gewünschte Sprache (Google)

Die Beurteilung von Internet-Seiten

Wenn das Internet als zuverlässige Informationsquelle verwendet werden soll, muss eine Auswertung vorgenommen werden, die die Suchanfrage kritisch widerspiegelt. In der folgenden Liste sind die wichtigsten Auswertungskriterien zusammengestellt.

Autoren-schaft	Wer ist der Autor? Was sind seine/ihre Referenzen? Ist er/sie einem Institut zugehörig? Hat das Institut ein Renommee? Wird die Seite von einer kommerziellen Einrichtung angeboten?
Objektivität und Richtigkeit	Wer fördert den Internetauftritt? Welches Ziel verfolgt die Einrichtung mit der Seite? Vertritt der Autor die Meinung einer Gruppe/seiner Einrichtung? Gibt es eine politische Perspektive? Gibt es eine kulturelle oder religiöse Perspektive? Gibt es Werbeanzeigen auf der Seite? Ist die Seite gut und fehlerfrei geschrieben? Ist die Seite durch andere überprüft und redigiert (peer reviewed)? Werden Quellen zitiert? Wie wurden Statistiken oder Daten gesammelt und dargestellt?
Aktualität	Sind die Informationen aktuell? Wie häufig wird die Seite aktualisiert? Welche Zeitspanne wird dargestellt?
Darstellung	Ist die Seite leicht zu navigieren? Ist die Information übersichtlich dargestellt? Sind die Formate und die Geschwindigkeit annehmbar? Gibt es einen Index oder ein Inhaltsverzeichnis?
Zweck	Wer ist das beabsichtigte Publikum (user)? Ist der Zweck zu informieren oder zu überzeugen? Sind die Informationen förderlich? Sind die Informationen urheberrechtlich gesichert?

Im Vergleich zu anderen Quellen	Sind andere Quellen besser (Bücher, Zeitschriften, usw.)? Gibt es Kosten für den Service? Sind die Informationen für mich nützlich?

Die Zukunft der Internet-Recherche

Suchmaschinen werden sich in ihrer Bedienung und bei der Suchanwendung immer ähnlicher (Zusammenlegungen, Druck von Seiten der Nutzer). Trotzdem gibt es noch unberücksichtigte Aspekte, die eine Weiterentwicklung vorantreiben. Zu diesen Entwicklungen gehören Suchmaschinen mit grafischer Darstellung der Ergebnisse (z.B. http://www.kartoo.com), Richtlinien zur Homogenisierung von Webseiten (etwa durch Anwendung des „Dublin Core", ein Metadaten-Schema zur Beschreibung von Dokumenten und anderen Objekten im Internet; Urheber dieses Schemas ist die „Dublin Core Metadata Initiative" (DCMI), s. Wikipedia.de), die Spam-Indexierung, die Entwicklung neuer Konzepte, wie z.B. die mobile, regionale oder die semantische Suche, sowie die Erschließung von Nicht-Text-Informationen wie z.B. Gesichter und Fotos.

Webadressen

Freie Internet-Browser (Software)
- Microsoft Internet Explorer 7 (http://www.microsoft.de/)
- Mozilla-Produkte, z.B. Firefox oder SeaMonkey (http://www.mozilla.org/products/)
- Netscape 7.1 http://www.netscape.de/

HTML selbst beigebracht
- Self-HTML (http://de.selfhtml.org/)

Suchmaschinen (Beispiele)
- Google (http://www.google.de)
- Alta Vista (http://www.altavista.com)
- Web.de (http://www.web.de)
- Yahoo! (http://www.yahoo.de)
- Metager (http://www.metager.de/)
- Metacrawler (http://www.metacrawler.de/)
- Kartoo (http://www.kartoo.de)

Informationen zur Informationssuche
- Die Suchfibel: Wie findet man Informationen im Internet? (http://www.suchfibel.de)
- Suchmaschinen (http://www.suchfibel.de/3allgem/index.htm)